В. Н. ВАПНИК

ВОССТАНОВЛЕНИЕ ЗАВИСИМОСТЕЙ ПО ЭМПИРИЧЕСКИМ ДАННЫМ

МОСКВА «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
1979

32.81
В 17
УДК 62-52

Восстановление зависимостей по эмпирическим данным. Вапник В. Н. Главная редакция физико-математической литературы издательства «Наука», М., 1979, 448 стр.

Монография посвящена проблеме восстановления зависимостей по эмпирическим данным. В ней исследуется метод минимизации риска на выборках ограниченного объема, согласно которому при восстановлении функциональной зависимости следует выбирать такую функцию, которая удовлетворяет определенному компромиссу между величиной, характеризующей ее «сложность», и величиной, характеризующей степень ее приближения к совокупности эмпирических данных.

Рассмотрено применение этого метода к трем основным задачам восстановления зависимостей: задаче обучения распознаванию образов, восстановления регрессии, интерпретации результатов косвенных экспериментов. Показано, что учет ограниченности объема эмпирических данных позволяет решать задачи распознавания образов при большой размерности пространства признаков, восстанавливать регрессионные зависимости при отсутствии модели восстанавливаемой функции, получать устойчивые решения некорректных задач интерпретации результатов косвенных экспериментов. Приведены соответствующие алгоритмы восстановления зависимостей.

Илл. 22, табл. 5, библ. 105.

В $\frac{30501-041}{053(02)-79}$ 169-79. 1502000000

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Задача восстановления зависимостей по эмпирическим данным была и, вероятно, всегда будет центральной в прикладном анализе. Эта задача является математической интерпретацией одной из основных проблем естествознания: как найти существующую закономерность по разрозненным фактам.

В наиболее простой постановке, той, которой и посвящена книга, проблема состоит в восстановлении функции по ее значениям в некоторых точках. Необходимо сформулировать общие принципы восстановления функциональных зависимостей, а затем в соответствии с ними построить алгоритмы восстановления.

Обычно, когда ищется общий принцип, предназначенный для решения широкого класса задач, выделяется наиболее простая, базовая задача. Эта задача подвергается тщательному теоретическому анализу, а полученная для нее схема решения распространяется на все задачи класса.

При изучении проблемы восстановления функциональных зависимостей по существу принята следующая базовая задача — восстановить функцию, принимающую лишь одно значение (восстановить константу). Считается, что константа измеряется с ошибками. Требуется, имея ряд измерений, определить ее.

Существуют различные варианты конкретизации постановки этой задачи. Они основаны на разных моделях «измерения с ошибками». Однако каковы бы ни были эти модели, изучение базовой задачи приводит к утверждению следующего классического принципа восстановления функциональных зависимостей по эмпирическим данным:

— Следует из допустимого множества функций выбирать такую функцию, которая наилучшим образом приближается к совокупности имеющихся эмпирических данных.

Этот принцип является достаточно общим. Он оставляет свободу в толковании того, что является мерой качества приближения функции к совокупности эмпирических данных. Возможны различные определения меры, такие, например, как величина среднеквадратичного уклонения значений функции, величина среднего уклонения, величина наибольшего уклонения и т. д. Каждое определение меры порождает свой метод восстановления зависимостей (метод наименьших квадратов, наименьших модулей и т. д.). Однако во всех случаях принцип отыскания решения — поиск функции, наилучшим образом приближающейся к эмпирическим данным, — остается неизменным.

Основное содержание книги связано с исследованием другого, неклассического принципа восстановления зависимостей:

— Следует из допустимого множества функций выбирать такую, которая удовлетворяет определенному соотношению между величиной, характеризующей качество приближения функции к заданной совокупности эмпирических данных, и величиной, характеризующей «сложность» приближающей функции.

Этот принцип нуждается в пояснении. Дело в том, что с увеличением «сложности» приближающей функции удается получать все лучшие и лучшие приближения к имеющимся эмпирическим данным и даже, может быть, построить функцию, проходящую через заданные точки.

Сформулированный принцип, в отличие от классического, утверждает, что не следует добиваться приближения к эмпирическим данным любыми средствами (т. е. за счет выбора чрезмерно «сложной» приближающей функции). Для каждого объема эмпирических данных существует свое соотношение между «сложностью» приближающей функции и достигнутым качеством приближения, при соблюдении которого восстановленная зависимость наиболее точно характеризует истинную. Дальнейшее приближение к эмпирическим данным за счет «усложнения» приближающей функции может привести к тому, что восстановленная функция будет лучше приближать эти конкретные эмпирические данные, но хуже — искомую функцию.

Неклассический принцип восстановления отражает попытку учесть то обстоятельство, что зависимость вос-

станавливается в условиях ограниченного объема эмпирических данных.

Мысль о том, что в условиях ограниченного объема эмпирических данных выбранная функция должна не просто приближать эмпирические данные, но и обладать некоторыми экстремальными свойствами, существовала давно. Однако впервые она получила теоретическое обоснование при исследовании задачи обучения распознаванию образов. Дело в том, что математическая постановка задачи обучения распознаванию образов приводит к необходимости восстанавливать функцию, которая принимает не одно, как в базовой задаче, а два значения. Такое усложнение по сравнению с базовой задачей неожиданно оказалось принципиальным. Множество функций, принимающих два значения, намного «разнообразнее» множества констант (т. е. функций, принимающих лишь одно значение).

Важным здесь оказалось то, что «структура» функций-констант «простая и однородная», в то время как «структура» множества функций, принимающих два значения, достаточно богатая и допускает упорядочение по степени «сложности». Учет упорядоченности функций и оказался существенным при восстановлении зависимостей в условиях ограниченного объема эмпирических данных.

Таким образом, исследование задачи обучения распознаванию образов показало, что классическая базовая задача не содержит всех проблем восстановления зависимостей — класс функций, в котором ведется восстановление константы, настолько беден, что вопрос о его расслоении просто не возникает.

В книге в качестве базовой принята задача обучения распознаванию образов. Для ее решения используются разные методы: как те, которые основаны на классических идеях статистического анализа, так и те, которые связаны с неклассическим принципом восстановления.

Все эти методы перенесены на две другие задачи восстановления: задачу восстановления регрессии и задачу интерпретации результатов косвенных экспериментов.

Для новой базовой задачи оказалось возможным различать две постановки: восстановление функции и восстановление значений функции в заданных точках. (При восстановлении констант эти постановки совпадают.)

Различать эти две постановки целесообразно потому, что в условиях ограниченного объема эмпирических данных имеющейся информации может не хватить, чтобы удовлетворительно восстановить функцию в целом, но в то же время оказаться достаточно, чтобы восстановить k чисел — значений функции в заданных точках.

Итак, книга посвящена проблемам восстановления зависимостей в условиях ограниченного объема эмпирических данных. Основная ее мысль состоит в следующем: попытка учесть ограниченность объема эмпирических данных приводит к утверждению неклассического принципа восстановления зависимостей. Использование же этого принципа позволяет решать «тонкие» задачи восстановления, такие, как определение экстремального набора признаков при распознавании образов, определение структуры приближающей функции при восстановлении регрессии, построение регуляризующего функционала при решении некорректных задач интерпретации косвенных экспериментов, т. е. задачи, которые возникают вследствие ограниченности объема эмпирических данных и которые не могут быть решены в рамках классических схем.

Книга содержит двенадцать глав. Первые две главы носят вводный характер. В них разные задачи восстановления зависимостей рассматриваются с единых позиций минимизации среднего риска по эмпирическим данным и обсуждаются различные возможные пути минимизации риска.

Следующие три главы (III, IV, V) посвящены исследованию классических идей минимизации риска: восстановлению функции плотности распределения вероятностей с помощью параметрических методов и использованию восстановленной плотности для минимизации риска. В главе III эти идеи реализованы на задаче обучения распознаванию образов, а в главах IV и V — на задаче восстановления регрессии.

Начиная с главы VI, в книге исследуются неклассические пути минимизации риска. В главах VI и VII устанавливаются условия применимости метода минимизации эмпирического риска для решения задачи минимизации среднего риска на выборках ограниченного объема, а в главах VIII — X на базе полученных условий конструируется новый метод минимизации риска — метод упо-

рядоченной минимизации, который существенно учитывает ограниченность объема эмпирических данных. Этот метод и реализует неклассический принцип восстановления зависимостей. В главе VIII рассмотрено применение метода упорядоченной минимизации риска к задачам распознавания образов и восстановления регрессии, а в главе IX — к решению некорректных задач интерпретации результатов косвенных экспериментов. В главе X на основе метода упорядоченной минимизации исследована задача восстановления значений функции в заданных точках.

Наконец, в главах XI и XII приведено описание алгоритмов упорядоченной минимизации риска.

Книга рассчитана на широкий круг читателей: студентов старших курсов, аспирантов, инженеров, научных работников. В ней изложение ведется так, чтобы доказательства не заслоняли основного хода рассуждений и вместе с тем, чтобы все принципиальные утверждения были доказаны полностью.

Автор старался избегать возможно и важных, но малосодержательных с точки зрения развития основных идей книги обобщений. Поэтому всюду рассматриваются наиболее простые случаи: квадратичная функция потерь, равноточность измерений, независимость помех и т. д. Как правило, соответствующие обобщения осуществимы и могут быть получены по стандартным схемам.

Чтение основной части книги не требует знания специальных разделов математики. Однако при разборе доказательств от читателя потребуется определенный навык в обращении с математическими понятиями.

Книга не является обзором принятой теории, она во многом тенденциозна. Тем не менее автор надеется, что она будет интересна и полезна читателю.

В. Вапник

Москва, 1978

ГЛАВА I

ЗАДАЧА ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ ПО ЭМПИРИЧЕСКИМ ДАННЫМ

§ 1. Проблема минимизации среднего риска по эмпирическим данным

Всякий раз, когда возникает проблема выбора функциональной зависимости, рассматривается одна и та же схема: среди множества возможных зависимостей необходимо найти такую, которая наилучшим образом удовлетворяет заданному критерию качества.

Формально это означает, что на векторном пространстве Z задан класс функций $\{g(z)\}$, $z \in Z$ (класс возможных зависимостей) и определен функционал — критерий качества выбираемой зависимости

$$I(g) = \int \Phi(z,\ g(z))\, dz. \tag{1.1}$$

Требуется среди функций $\{g(z)\}$ найти такую $g^*(z)$, которая доставляет минимум функционалу (1.1) (будем полагать, что наилучшему качеству соответствует минимум функционала и что минимум (1.1) в $\{g(z)\}$ существует).

В том случае, когда класс функций $\{g(z)\}$ и функционал $I(g)$ заданы явно, отыскание функции $g^*(z)$, минимизирующей $I(g)$, является предметом исследования вариационного исчисления.

В этой книге рассмотрен иной случай, когда на Z определена плотность распределения вероятностей $P(z)$, а функционал задан как математическое ожидание

$$I(g) = M\Phi(z,\ g(z)) = \int \Phi(z,\ g(z))\, P(z)\, dz. \tag{1.2}$$

Проблема же состоит в том, чтобы минимизировать функционал (1.2), если плотность $P(z)$ неизвестна, но зато дана выборка

$$z_1, \ldots, z_l, \tag{1.3}$$

полученная в результате случайных независимых испытаний согласно $P(z)$.

Ниже в §§ 2, 3, 4 мы убедимся в том, что к минимизации функционала (1.2) по эмпирическим данным (1.3) сводятся все основные задачи восстановления функциональных зависимостей, а пока заметим, что проблемы, возникающие при минимизации функционала (1.1) и минимизации функционала (1.2) по эмпирическим данным (1.3), существенно различаются.

При минимизации функционала (1.1) проблема состоит в организации поиска в классе $\{g(z)\}$ функции $g^*(z)$, минимизирующей (1.1). При минимизации же функционала (1.2) по эмпирическим данным (1.3) основная проблема заключается не в организации поиска функции в $\{g(z)\}$, а в формулировке конструктивного критерия, согласно которому должен проводиться выбор функции. (Сам функционал (1.2) не может служить критерием выбора, так как в нем плотность $P(z)$ неизвестна.)

Таким образом, в первом случае ищется ответ на вопрос: «Как найти минимум функционала в заданном классе функций?», в то время как во втором: «Что следует минимизировать для того, чтобы выбрать в $\{g(z)\}$ функцию, гарантирующую «малую» величину функционала (1.2)?»

Минимизация функционала (1.2) по эмпирическим данным (1.3) является задачей математической статистики. Назовем ее *задачей минимизации среднего риска по эмпирическим данным*.

При постановке задачи минимизации среднего риска будем класс функций $\{g(z)\}$ задавать в параметрическом виде $\{g(z, \alpha)\}$ [1]). Здесь α — параметр, принадлежащий множеству Λ, конкретное значение которого $\alpha = \alpha^*$ определяет конкретную функцию $g(z, \alpha^*)$ класса $g(z, \alpha)$. Найти нужную функцию в этом случае, значит установить нужное значение параметра α. Изучение лишь параметрического класса функций не является сколько-нибудь серьезным ограничением в постановке задачи, так как множество Λ, которому принадлежит параметр α, произвольно: оно может быть множеством скалярных величин, множеством векторов или множеством абстрактных элементов.

---

[1]) В дальнейшем всюду при записи класса функций будем опускать фигурные скобки. Функцию и класс функций будем различать в зависимости от того, зафиксирован ли параметр α.

В новых обозначениях функционал (1.2) переписывается в виде

$$I(\alpha)=\int Q(z,\ \alpha)\,P(z)\,dz,\quad \alpha\in\Lambda, \tag{1.4}$$

где обозначено

$$Q(z,\ \alpha)=\Phi(z,\ g(z,\ \alpha)).$$

Функция $Q(z,\ \alpha)$ двух групп переменных z и α носит название *функции потерь*.

Задача минимизации среднего риска имеет простую интерпретацию: считается, что каждая функция $Q(z,\ \alpha^*)$, $\alpha^*\in\Lambda$ (переменного z при фиксированном $\alpha=\alpha^*$), определяет величину потери при появлении вектора z. Средняя по z величина потерь для функции $Q(z,\ \alpha^*)$ определяется интегралом

$$I(\alpha^*)=\int Q(z,\ \alpha^*)\,P(z)\,dz.$$

Суть задачи состоит в том, чтобы для неизвестного закона появления z по наблюдениям за случайными независимыми реализациями $z_1,\ \ldots,\ z_l$ выбрать в $Q(z,\ \alpha)$ такую функцию $Q(z,\ \alpha^*)$, которая минимизирует среднюю величину потерь.

Задача минимизации среднего риска по эмпирическим данным является достаточно общей. Будем выделять в ней частную постановку. Особенность этой постановки заключается в том, что вектор z состоит из $n+1$ координаты — координаты y и n координат $x^1,\ \ldots,\ x^n$, образующих вектор x. Функция потерь $Q(z,\ \alpha)$ задана в виде

$$Q(z,\ \alpha)=\Phi(y-F(x,\ \alpha)),$$

где $F(x,\ \alpha)$ — параметрический класс функций. Необходимо минимизировать функционал

$$I(\alpha)=\int\Phi(y-F(x,\ \alpha))\,P(x,\ y)\,dx\,dy, \tag{1.5}$$

если плотность $P(x,\ y)$ неизвестна, но зато дана случайная независимая выборка пар

$$x_1,\ y_1;\ \ldots;\ x_l,\ y_l \tag{1.6}$$

(*обучающая последовательность*).

Задачу минимизации функционала (1.5) по эмпирическим данным (1.6) будем называть задачей *восстановления*

функциональной зависимости. Исследованию этой задачи и посвящена книга [1]). В ней будут рассмотрены три основные задачи восстановления функциональных зависимостей:

— задача обучения распознаванию образов,

— задача восстановления регрессии,

— задача интерпретации результатов косвенных экспериментов.

В следующих параграфах мы убедимся, что все они сводятся к минимизации функционала (1.5) по эмпирическим данным (1.6).

§ 2. Задача обучения распознаванию образов

Задача обучения распознаванию образов была сформулирована в конце 50-х годов.

Содержательная ее постановка состоит в следующем. Некто (часто говорят — учитель) наблюдает возникающие ситуации и определяет, к какому из k классов каждая из них относится. Требуется построить такое устройство, которое после наблюдения за работой учителя проводило бы классификацию примерно так же, как и учитель.

Такая постановка на формальном языке имеет простое выражение. В некоторой среде, которая характеризуется плотностью распределения вероятностей $P(x)$, случайно и независимо появляются ситуации x. Учитель относит эти ситуации к одному из k классов. (В дальнейшем для простоты будем считать, что $k=2$. Такое предположение не снижает общности постановки, так как последовательным разделением ситуаций на два класса можно получить разделение и на k классов.) Предположим, что учитель осуществляет эту классификацию с помощью функции условного распределения вероятностей $P(\omega \mid x)$, где $\omega = \{0; 1\}$ ($\omega = 0$ означает, что учитель отнес ситуацию x к первому классу, а $\omega = 1$ означает, что ситуация x отнесена ко второму классу).

Ни свойства среды $P(x)$, ни решающее правило $P(\omega \mid x)$ не известны. Однако известно, что обе функции сущест-

---

[1]) В книге мы ограничимся рассмотрением задачи восстановления зависимости для квадратичной функции потерь $\Phi(y - F(x, \alpha)) = (y - F(x, \alpha))^2$. Однако полученные результаты могут быть перенесены и на общий случай.

вуют. Пусть теперь задано параметрическое множество функциональных зависимостей $F(x, \alpha)$ (класс решающих правил). Все функции класса $F(x, \alpha)$ — характеристические, т. е. они могут принимать только два значения — нуль и единица.

Требуется, наблюдая l пар

$$x_1, \omega_1; \ldots; x_l, \omega_l$$

(ситуация — x, реакция на нее учителя — ω), выбрать в классе характеристических функций $F(x, \alpha)$ такую функцию, для которой вероятность классификации, отличной от классификации учителя, была бы минимальной. Иначе говоря, достигался бы минимум функционала

$$I(\alpha) = \sum_{\omega=0,1} \int (\omega - F(x, \alpha))^2 P(\omega \mid x) P(x)\, dx.$$

Функционал $I(\alpha)$ будем записывать в виде

$$I(\alpha) = \int_{X,\ \omega} (\omega - F(x, \alpha))^2 P(x, \omega)\, dx\, d\omega,$$

а функцию $P(x, \omega) = P(\omega \mid x) P(x)$ будем называть совместной плотностью пар x, ω, заданной на пространстве $X\omega$.

Задача обучения распознаванию образов свелась, таким образом, к задаче минимизации среднего риска по эмпирическим данным. Особенность ее заключается в том, что класс функций $Q(z, \alpha)$ не обладает таким произволом, как в общей постановке. На него наложены ограничения:

1) Вектор z состоит из $n+1$ координаты: координаты ω, которая может принимать только два значения (нуль и единица), и n координат $x^1, \ldots, x^n$, образующих вектор x.

2) Класс функций $Q(z, \alpha)$ задан в виде

$$Q(z, \alpha) = (\omega - F(x, \alpha))^2,$$

где $F(x, \alpha)$ принимают также только два значения — нуль и единица.

Таким образом, в задаче обучения распознаванию образов значение функции потерь равно либо нулю, либо единице. Эта особенность задачи минимизации риска и определяет специфику обучения распознаванию образов [1]).

[1]) При постановке задачи можно учесть различные цены ошибок первого и второго рода. Однако это принципиально не меняет существа дела: важно, что функция потерь будет принимать лишь конечное число (три) значений.

§ 3. Задача восстановления регрессии

Два множества элементов X и Y связаны функциональной зависимостью, если каждому элементу $x \in X$ может быть поставлен в однозначное соответствие элемент $y \in Y$. Эта зависимость называется функцией, если множество X — векторы, а множество Y — скаляры. Однако существуют и такие зависимости, где каждому вектору x ставится в соответствие число y, полученное с помощью случайного испытания, согласно условной плотности $P(y \mid x)$. Иначе говоря, каждому x ставится в соответствие закон $P(y \mid x)$, согласно которому в случайном испытании реализуется выбор y.

Существование такого рода зависимостей отражает наличие стохастических связей между вектором x и скаляром y. Полное знание стохастических связей требует восстановления условной плотности $P(y \mid x)$. Задача же восстановления условной плотности чрезвычайно трудна. Однако часто на практике (например, в задачах обработки результатов измерения) нужно знать не плотность $P(y \mid x)$, а лишь одну из ее характеристик: функцию условного математического ожидания, т. е. функцию, которая каждому x ставит в соответствие число $y(x)$, равное математическому ожиданию скаляра y

$$y(x) = \int yP(y \mid x)\, dy.$$

Функция $y(x)$ называется *регрессией*, а задача восстановления функции условного математического ожидания — *задачей восстановления регрессии*.

Рассмотрим постановку этой задачи. В некоторой среде, которая характеризуется плотностью распределения вероятностей $P(x)$, случайно и независимо появляются ситуации x. В этой среде работает преобразователь, который каждому вектору x ставит в соответствие число y, полученное в результате реализации случайного испытания согласно закону $P(y \mid x)$. Ни свойства среды $P(x)$, ни закон $P(y \mid x)$, вообще говоря, неизвестны. Однако известно, что существует регрессия

$$\bar{y} = y(x).$$

Требуется по случайной независимой выборке пар

$$x_1, y_1; \ldots; x_l, y_l$$

восстановить регрессию, т. е. в классе функций $F(x, \alpha)$ отыскать функцию $F(x, \alpha^*)$, наиболее близкую к регрессии $y(x)$.

Задача восстановления регрессии является одной из основных задач прикладной статистики. К ней приводится проблема *интерпретации результатов прямых экспериментов*. Пусть интересующая нас закономерность связывает функциональной зависимостью величину $\bar{y}$ с вектором x

$$\bar{y} = y(x).$$

Пусть нашей целью является определение функциональной зависимости $\bar{y} = y(x)$ в ситуации, когда в любой точке x^* может быть проведен прямой эксперимент по определению этой зависимости, т. е. проведены прямые измерения величины $\bar{y}^* = y(x^*)$. Однако вследствие несовершенства эксперимента результат измерения определит истинную величину с некоторой случайной ошибкой. Иначе говоря, в каждой точке x удается определить не величину $y(x)$, а величину $y = y_x$, где $y - y(x) = \xi$ — ошибка эксперимента, $M\xi^2 < \infty$.

Считается (эта гипотеза и определяет возможность интерпретации экспериментов), что ни в одной точке x условия эксперимента не допускают систематической ошибки, т. е. математическое ожидание измерения y_x функции в каждой фиксированной точке x равно значению функции $y(x)$ в этой точке

$$My_x = y(x). \tag{1.7}$$

Кроме того, будем считать, что случайные величины y_{x_i} и y_{x_j} $(i \neq j)$ независимы. В этих условиях необходимо по конечному числу прямых экспериментов восстановить функцию $\bar{y} = y(x)$. Таким образом, интересующая нас зависимость есть регрессия (1.7), а суть проблемы состоит в отыскании регрессии по последовательности пар

$$x_1, y_1; \ldots; x_l, y_l.$$

В задачах интерпретации результатов прямых экспериментов принято различать два типа экспериментов: *закрытый* и *открытый*. Закрытый эксперимент предполагает, что закон $P(x)$, по которому определяется выбор экспериментальных точек, исследователю не известен. Откры-

тым экспериментом считается такой эксперимент, в котором закон $P(x)$ выбора точек измерения x известен исследователю (его часто задает сам исследователь). Итак, задача восстановления регрессии содержит проблему интерпретации результатов прямых экспериментов.

В свою очередь задача восстановления регрессии сводится к задаче восстановления зависимостей.

В самом деле, рассмотрим функционал

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy, \tag{1.8}$$

где обозначено $P(x, y) = P(y \mid x) P(x)$.

Покажем, что если регрессия $\bar{y} = y(x)$ принадлежит классу $F(x, \alpha)$ ($y(x) \equiv F(x, \alpha_0)$), то она минимизирует функционал (1.8), если же регрессия не принадлежит $F(x, \alpha)$, то минимум достигается на ближайшей к регрессии функции $F(x, \alpha^*)$. Близость функций $f_1(x)$ и $f_2(x)$ понимается в следующем смысле (в метрике L_P^2):

$$\rho_L(f_1(x), f_2(x)) = \left(\int (f_1(x) - f_2(x))^2 P(x)\, dx\right)^{1/2}.$$

Действительно, обозначим

$$\Delta F(x, \alpha) = F(x, \alpha) - y(x). \tag{1.9}$$

Тогда функционал (1.8) может быть записан в виде

$$I(\alpha) = \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (\Delta F(x, \alpha))^2 P(x)\, dx - 2\int \Delta F(x, \alpha)(y - y(x)) P(x, y)\, dx\, dy.$$

В этом выражении третье слагаемое равно нулю, так как в силу (1.7)

$$\int \Delta F(x, \alpha)(y - y(x)) P(x, y)\, dx\, dy = \int \Delta F(x, \alpha) P(x) \left[\int (y - y(x)) P(y \mid x)\, dy\right] dx = 0.$$

Таким образом, мы установили, что

$$I(\alpha) = \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (F(x, \alpha) - y(x))^2 P(x)\, dx.$$

Так как первое слагаемое не зависит от α, то точка минимума $I(\alpha)$ совпадает с точкой минимума второго слагаемого, и, следовательно, минимум $I(\alpha)$ достигается на регрессии,

если $y(x) \in F(x, \alpha)$, или на ближайшей к ней функции, если $y(x) \notin F(x, \alpha)$.

Итак, задача восстановления регрессии также сводится к схеме минимизации среднего риска. Особенность ее состоит в том, что на класс функций $Q(z, \alpha)$ наложены следующие ограничения:

1) Вектор z состоит из $n+1$ координат: координаты y и n координат $x^1, \ldots, x^n$, образующих вектор x. Однако, в отличие от задачи обучения распознаванию образов, здесь и координата y и функции $F(x, \alpha)$ могут принимать любые значения из интервала $(-\infty, \infty)$.

2) Класс функций $Q(z, \alpha)$ задан в виде

$$Q(z, \alpha) = (y - F(x, \alpha))^2.$$

Функции $Q(z, \alpha)$ принимают любые значения из интервала $(0, \infty)$.

§ 4. Задача интерпретации результатов косвенных экспериментов

В предыдущем параграфе мы рассмотрели задачу восстановления регрессии. Было показано, что к этой задаче сводится проблема интерпретации результатов прямых экспериментов, т. е. таких экспериментов, с помощью которых в фиксированных точках измеряется интересующая нас зависимость. Однако часто бывает так, что искомую функцию $f(t)$ нельзя измерить ни в одной точке t. В то же время может оказаться доступной измерению другая функция $F(x)$, которая связана с $f(t)$ операторным уравнением

$$Af(t) = F(x). \tag{1.10}$$

Требуется по результатам измерений $y_1, \ldots, y_l$ функции $F(x)$ в точках $x_1, \ldots, x_l$ найти в классе $f(t, \alpha)$ решение уравнения (1.10). Такую задачу будем называть задачей *интерпретации результатов косвенных экспериментов.*

Постановка задачи состоит в следующем: задан непрерывный оператор A, взаимно однозначно отображающий элементы $f(t, \alpha)$ метрического пространства E_1 в элементы $F(x, \alpha)$ метрического пространства E_2. Требуется найти решение операторного уравнения (1.10) в классе функций $f(t, \alpha)$, если функция $F(x)$ неизвестна, но зато даны измерения $y_1, \ldots, y_l$ функции $F(x)$ в точках $x_1, \ldots, x_l$.

Так же, как и при интерпретации прямых измерений, здесь эксперимент по измерению функции $F(x)$ не содержит систематических погрешностей, т. е. $My_i = F(x_i)$, а случайные величины y_i и y_j $(i \neq j)$ независимы. Кроме того, для простоты будем считать, что области задания функций $f(t)$ и $F(x)$ — отрезки $[0, 1]$. Эксперимент является открытым: точки x, в которых проводятся измерения функции $F(x)$, задаются на отрезке $[0, 1]$ случайно и независимо, согласно равномерной плотности распределения вероятностей [1]).

Задача интерпретации результатов косвенных экспериментов также сводится к проблеме минимизации среднего риска по эмпирическим данным.

В самом деле, рассмотрим функционал

$$I(\alpha) = \int (y - Af(t, \alpha))^2 P(y \mid x)\, dy\, dx \equiv$$
$$\equiv \int (y - F(x, \alpha))^2 P(y \mid x)\, dy\, dx.$$

Совершенно аналогично преобразованиям, проведенным в § 3, получаем

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(y \mid x)\, dy\, dx =$$
$$= \int (y - F(x))^2 P(y \mid x)\, dy\, dx + \int (\Delta F(x, \alpha))^2 dx -$$
$$- 2 \int \Delta F(x, \alpha) \left[\int (y - F(x)) P(y \mid x)\, dy \right] dx,$$

где обозначено

$$\Delta F(x, \alpha) = F(x, \alpha) - F(x).$$

Здесь, так же как и в аналогичном случае предыдущего параграфа, третье слагаемое суммы равно нулю, откуда заключаем, что минимум функционала

$$I(\alpha) = \int (y - Af(t, \alpha))^2 P(y \mid x)\, dy\, dx \qquad (1.11)$$

достигается на решении $f(t)$ операторного уравнения (1.10).

Таким образом, мы опять пришли к схеме минимизации среднего риска (1.4) по эмпирическим данным. В этой задаче функция потерь $Q(z, \alpha)$ такова, что:

1) вектор z состоит из двух координат y и x, каждая из которых может принимать значения из интервала $(-\infty, \infty)$;

---

[1]) Точки x могут задаваться любой не обращающейся в нуль на $[0, 1]$ плотностью.

2) функция потерь задана в виде

$$Q(z, \alpha) = (y - Af(t, \alpha))^2.$$

Особенность же интерпретации результатов косвенных экспериментов состоит в том, что ищется функция $f(t, \alpha^*)$, минимизирующая функционал (1.11) в условиях, когда задача решения операторного уравнения

$$Af(t) = F(x), \quad f(t) \in f(t, \alpha)$$

может быть некорректно поставленной.

§ 5. Некорректно поставленные задачи

Говорят, что решение операторного уравнения

$$Af(t) = F(x)$$

устойчиво, если малая вариация правой части $F(x) \in \in F(x, \alpha)$ приводит к малому изменению решения, т. е. если окажется, что для всякого ε найдется такое $\delta(\varepsilon)$, что, как только выполнится неравенство

$$\rho_{E_2}(F(x, \alpha_1), F(x)) \leqslant \delta(\varepsilon),$$

окажется выполненным и неравенство

$$\rho_{E_1}(f(t, \alpha_1), f(t)) \leqslant \varepsilon.$$

Здесь индексы E_2 и E_1 означают, что расстояние определяется соответственно в метриках пространства E_2 и E_1 (операторное уравнение (1.10) осуществляет отображение из пространства E_1 в пространство E_2).

Говорят также, что задача решения операторного уравнения *поставлена корректно по Адамару*, если решение уравнения:

1) *существует*, 2) *единственно*, 3) *устойчиво*.

Задача решения операторного уравнения считается *поставленной некорректно*, если решение уравнения не удовлетворяет хотя бы одному из перечисленных требований.

Ниже, в основном тексте книги, мы ограничимся решением некорректных задач интерпретации результатов косвенных экспериментов, заданных интегральными уравнениями Фредгольма I рода

$$\int_a^b K(t, x) f(t)\, dt = F(x).$$

Однако все полученные результаты будут справедливы и для уравнений, заданных любыми другими линейными непрерывными операторами.

Необходимые факты теории решения некорректно поставленных задач приведены в приложении к главе.

Итак, рассмотрим интегральное уравнение Фредгольма I рода

$$\int_0^1 K(x,\ t) f(t)\, dt = F(x), \tag{1.12}$$

заданное непрерывным почти всюду на $0 \leqslant t \leqslant 1$, $0 \leqslant x \leqslant 1$ ядром $K(x,\ t)$ и отображающее множество непрерывных на отрезке $[0,\ 1]$ функций $f(t)$ на множество непрерывных на отрезке $[0,\ 1]$ функций $F(x)$.

Покажем, что задача решения уравнения (1.12) является некорректно поставленной.

Для этого заметим, что непрерывная функция $G_\nu(x)$, образованная с помощью ядра $K(x,\ t)$,

$$G_\nu(x) = \int_0^1 K(x,\ t) \sin \nu t\, dt,$$

обладает свойством:

$$\sup_x G_\nu(x) \underset{\nu \to \infty}{\longrightarrow} 0.$$

Рассмотрим интегральное уравнение

$$\int_0^1 K(x,\ t) \hat{f}(t)\, dt = F(x) + G_\nu(x). \tag{1.13}$$

В силу линейности уравнения Фредгольма решение уравнения (1.13) имеет вид

$$\hat{f}(t) = f(t) + \sin \nu t,$$

где $f(t)$ есть решение уравнения (1.12).

При достаточно больших ν правые части уравнений (1.12) и (1.13) различаются мало (на $G_\nu(x)$), в то время как их решения разнятся на величину $\sin \nu t$.

Интегральное уравнение Фредгольма I рода является одним из основных уравнений в задаче интерпретации результатов косвенных экспериментов. Вот примеры задач, которые связаны с решением этого уравнения.

1. Обратная задача спектроскопии. Пусть с помощью некоторого реального спектроскопа наблюдается спектр $F(x)$. Так как прибор реальный, т. е. имеет конечную разрешающую способность, то наблюдаемый спектр, вообще говоря, отличается от того, который зафиксировал бы идеальный спектроскоп (т. е. спектроскоп с бесконечно высокой разрешающей способностью).

Требуется редуцировать спектр, полученный на приборе с конечным разрешением, к истинному спектру.

Часто такая задача может быть решена. Известно, например, что характеристика «сглаживания» (аппаратная функция) некоторых реальных спектроскопов имеет вид

$$K(x, t)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\}.$$

Наблюдаемый спектр $F(x)$ связан с истинным спектром $f(t)$ соотношением

$$\frac{1}{\sqrt{2\pi}\,\sigma}\int_0^\infty \exp\left[-\frac{(x-t)^2}{2\sigma^2}\right] f(t)\,dt = F(x).$$

Чем лучше прибор (меньше σ), тем менее искажается спектральная картина.

При $\sigma \to 0$ характеристика прибора стремится к идеальной:

$$\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\} \to \delta(t-x).$$

И, следовательно,

$$F(x) \to f(x).$$

Однако, как бы ни был плох реальный спектроскоп, в принципе по полученному спектру можно найти истинный, но для этого надо решить обратную задачу спектроскопии, т. е. решить интегральное уравнение

$$\frac{1}{\sqrt{2\pi}\,\sigma}\int_0^\infty \exp\left\{-\frac{(x-t)^2}{2\sigma^2}\right\} f(t)\,dt = F(x),$$

используя вместо функции $F(x)$ эмпирические данные $y_1 \ldots, y_l$,

2. Задача идентификации линейных объектов. Известно, что динамические свойства линейных однородных объектов с одним выходом полностью описываются импульсной переходной (весовой) функцией $f(\tau)$. Функция $f(\tau)$ представляет собой реакцию объекта на единичный импульс, подаваемый на систему в момент $\tau=0$.

Зная эту функцию, можно вычислить реакцию объекта на любое возмущение $x(t)$ по формуле

$$y(t)=\int_0^t x(t-\tau)\,f(\tau)\,d\tau.$$

Таким образом, определение динамических характеристик объекта сводится к отысканию весовой функции $f(\tau)$.

Известно, что для линейного однородного объекта справедливо уравнение Винера — Хопфа

$$\int_0^\infty R_{xx}(t-\tau) f(\tau)\, d\tau = R_{yx}(t). \tag{1.14}$$

Уравнение (1.14) связывает автокорреляционную функцию $R_{xx}(t)$ стационарного случайного процесса на входе объекта, весовую функцию объекта $f(\tau)$ и взаимную корреляционную функцию входного и выходного сигналов $R_{yx}(t)$.

Задача идентификации линейного объекта, таким образом, заключается в определении весовой функции объекта по известной автокорреляционной функции входного сигнала и измеренной взаимной корреляционной функции входного и выходного сигналов, т. е. в решении интегрального уравнения (1.14) по эмпирическим данным.

3. Задача восстановления производных. Пусть даны измерения функции $F(x)$ в l точках отрезка $[0, 1]$. Точки, на которых проведены измерения, заданы случайно и независимо согласно равномерному закону распределения вероятностей.

Требуется восстановить на $[0, 1]$ производную $f(x)$ функции $F(x)$. Легко видеть, что задача сводится к решению интегрального уравнения Вольтерра I рода

$$\int_0^x f(t)\, dt = F(x) - F(0)$$

при условии, что известно l замеров $y_1, \ldots, y_l$ функции $F(x)$, произведенных в точках $x_1, \ldots, x_l$, или, что то же самое, к решению при тех же условиях уравнения Фредгольма I рода

$$\int_0^1 \theta(x-t) f(t)\, dt = F(x) - F(0),$$

где

$$\theta(z) = \begin{cases} 1, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0. \end{cases}$$

В более общем случае, когда надо восстановить k-ю производную, необходимо решить интегральное уравнение

$$\int_0^1 \frac{(x-t)^{k-1}}{(k-1)!}\, \theta(x-t) f(t)\, dt = F(x) - \sum_{j=0}^{k-1} \frac{F^{(j)}(0)}{j!},$$

используя вместо функции $F(x)$ эмпирические данные $y_1, \ldots, y_l$.

Здесь $F^{(j)}(0)$ — значение j-й производной в точке нуль.

§ 6. О точности и надежности минимизации риска по эмпирическим данным

Итак, мы рассмотрели три основные задачи восстановления зависимостей по эмпирическим данным: задачу распознавания образов, восстановления регрессии, интерпре-

тации результатов косвенных экспериментов. В основе каждой из них лежала одна и та же общая схема — схема минимизации среднего риска по эмпирическим данным: необходимо найти α^*, минимизирующее функционал

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz,$$

если плотность распределения вероятностей $P(z)$ неизвестна, но зато дана случайная независимая выборка $z_1, \ldots, z_l$ длины l.

Более того, для всех трех задач была выбрана идентичная структура функции потерь

$$Q(z, \alpha) = (y - F(x, \alpha))^2.$$

Таким образом, во всех случаях надо было найти функцию $F(x, \alpha^*)$, доставляющую минимум функционалу

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy, \qquad (1.15)$$

если плотность $P(x, y)$ неизвестна, но зато дана выборка $x_1, y_1; \ldots; x_l, y_l$, полученная в результате случайных независимых испытаний согласно этой плотности. Правда, мы различали два варианта постановки задачи восстановления регрессии: когда плотность $P(x)$ неизвестна (случай закрытого эксперимента) и когда плотность $P(x)$ известна (случай открытого эксперимента). Но эти варианты постановок отличаются не принципиально, важно, что в обоих случаях совместная плотность $P(x, y)$ неизвестна.

Мы установили, что разные задачи восстановления зависимостей различаются так, как различаются функции потерь при минимизации риска, и что в каждой задаче параметр α, доставляющий точный минимум соответствующему функционалу, определяет искомую функциональную зависимость.

Однако найти точный минимум функционала (1.15) по выборке фиксированного объема — задача, вообще говоря, невозможная — ведь любая выборка является только реализацией закона распределения вероятностей и никак не эквивалентна ему. Поэтому может стоять задача отыскания по выборке фиксированного объема не функции, доставляющей точный минимум функционалу (1.15), а функции, доставляющей функционалу величину, «близкую» к минимальной.

Более того, получение величины, «близкой» к минимальной, можно гарантировать не безусловно, а лишь с некоторой вероятностью (ведь при любой плотности не исключена вероятность того, что обучающая последовательность, полученная в случайных испытаниях, будет состоять из одной и той же пары элементов x, y, повторенной l раз).

Таким образом, заданная *точность* минимизации среднего риска (1.15) по выборке фиксированного объема может быть достигнута лишь с некоторой *надежностью*.

Будем говорить, что значение функционала $I(\alpha^*)$ $\varkappa$-близко к минимальному ($\min\limits_{\alpha} I(\alpha)$), если выполняется неравенство

$$I(\alpha^*) - \min_{\alpha} I(\alpha) \leqslant \varkappa.$$

Пусть теперь зафиксирован некоторый алгоритм A, который по выборке объема l определяет значение параметра α^*. Так как выборка случайная, то этот алгоритм будет определять случайную величину параметра α^*, которой соответствует случайное число $I(\alpha^*)$.

Будем говорить, что алгоритм A с надежностью $1-\eta$ доставляет функционалу $I(\alpha)$ значение $\varkappa$-близкое к минимальному, если для заданного числа $0<\eta<1$ справедливо неравенство

$$P\left\{I(\alpha^*) - \min_{\alpha} I(\alpha) > \varkappa\right\} < \eta.$$

При решении задачи минимизации среднего риска нашей целью является нахождение алгоритмов, которые на выборках фиксированного объема с заданной надежностью отыскивали бы функцию, доставляющую функционалу $I(\alpha)$ значение, наиболее близкое к минимальному.

§ 7. О точности восстановления зависимостей по эмпирическим данным

В конце предыдущего параграфа была сформулирована цель исследования: найти алгоритмы, которые гарантировали бы достижение риска, близкого к минимальному. Построению и обоснованию таких алгоритмов посвящена эта книга. Однако при формулировке цели исследования была по существу подменена задача. В самом деле, нашей

исходной целью было восстановление функциональных зависимостей. В §§ 2, 3, 4 было показано, что функция, доставляющая точный минимум соответствующему функционалу среднего риска, определяет искомую зависимость. С другой стороны, найти точный минимум по выборке фиксированного объема — задача малореальная. Поэтому-то и предлагалось искать функцию, доставляющую среднему риску значение, близкое к минимальному.

Однако ниоткуда не следует, что близким значениям функционала будут соответствовать близкие функции. Отыскание значения функционала, близкого к минимальному, и функции, близкой к искомой, вообще говоря, — задачи разные. Поэтому, прежде чем решать задачу восстановления функциональной зависимости по эмпирическим данным методом минимизации среднего риска, необходимо выяснить, приведет ли такая подмена задач к успеху, т. е. гарантирует ли близость функционалов близость функций.

Для того чтобы начать исследование в этом направлении, надо прежде всего договориться о том, как мы будем понимать близость функций. В отличие от близости функционалов, которая определяется естественным образом как расстояние между двумя точками числовой оси (значениями этих функционалов), близость функций должна быть определена как расстояние между двумя элементами функционального пространства.

В функциональном анализе приняты различные способы метризации (введения понятия расстояния). Мы же будем использовать два таких понятия (две метрики): среднеквадратичное уклонение с весом и равномерное уклонение. Расстояние между двумя функциями $f_1(x)$ и $f_2(x)$ в смысле среднеквадратичного уклонения с весом $P(x)$ (метрика L_P^2) определяется функционалом

$$\rho_L(f_1(x),\ f_2(x)) = \left(\int (f_1(x) - f_2(x))^2 P(x)\, dx\right)^{1/2},$$

где $P(x)$ — неотрицательная функция, такая, что $\int P(x)\, dx = 1$. Расстояние же в смысле равномерного уклонения (метрика C) определяется функционалом

$$\rho_C(f_1(x),\ f_2(x)) = \sup_x |f_1(x) - f_2(x)|.$$

Таким образом, две функции близки в метрике L_P^2, если

$$\left(\int (f_1(x) - f_2(x))^2 P(x)\, dx\right)^{1/2} \leqslant \varkappa, \tag{1.16}$$

и близки в метрике C, если

$$\sup_x |f_1(x) - f_2(x)| \leqslant \varkappa. \tag{1.17}$$

Заметим, что требование равномерной близости (1.17) является более сильным, чем среднеквадратичной. Из выполнения неравенства (1.17) следует выполнение неравенства (1.16). Обратное утверждение, вообще говоря, неверно.

Итак, будем использовать понятия близости в следующих смыслах:

1) близость качества функций (значений функционалов),
2) близость функций в метрике L_P^2,
3) близость функций в метрике C.

Выбор понятия близости определяется не формальной, а содержательной постановкой задачи.

Как же задается близость в различных задачах восстановления зависимостей?

В задаче распознавания образов требуется в заданном классе характеристических функций найти такую, которая минимизирует вероятность ошибочной классификации (т. е. по постановке требуется минимизировать функционал). Поэтому здесь естественно считать две функции близкими, если их качества близки; здесь близость определяется близостью функционалов.

При восстановлении регрессии проблема состоит не в том, чтобы минимизировать функционал, а в том, чтобы найти функцию, близкую к регрессии. В этой задаче близость определяется с помощью метрики L_P^2 или метрики C в зависимости от того, как в дальнейшем предполагается использовать восстановленную функцию.

Пусть, например, решается задача восстановления регрессии $\bar{y} = y(x)$ в схеме интерпретации прямых экспериментов. Восстановленную зависимость $\bar{y} = F(x, \alpha^*)$ при этом предполагается использовать для целей прогноза величины $\bar{y}$ в зависимости от ситуации x. Точность прогноза для всякой фиксированной ситуации x естественно измерять величиной

$$(y(x) - F(x, \alpha^*))^2.$$

Точность прогноза в целом с помощью восстановленной функции часто измеряют как среднюю точность по мере множества x, т. е. величиной

$$\rho_L(y(x),\ F(x,\ \alpha)) = \left(\int (y(x) - F(x,\ \alpha))^2 P(x)\, dx\right)^{1/2},$$

или, иначе говоря, в этом случае близость определяется метрикой L_P^2.

Однако существуют задачи, где близость в метрике L_P^2 недостаточна. Пусть, например, некоторое количество $\bar{y}$

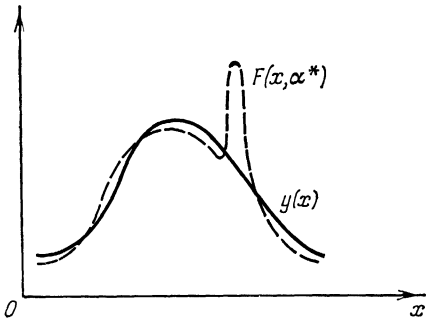

Рис. 1.

связано с технологическими параметрами x функциональной зависимостью. Требуется найти такой вектор параметров x^*, который обеспечит максимум количества $\bar{y}$. Эту задачу решают по следующей схеме: восстанавливают функциональную зависимость $\bar{y} = y(x)$, а затем ищут значение x^*, доставляющее максимум восстановленной функции. Однако если в этом случае в качестве восстановленной функции взять функцию $F(x, \alpha^*)$, близкую к истинной в метрике L_P^2, то возможна такая ситуация, которая изображена на рис. 1. Восстановленная функция достаточно хорошо приближает истинную почти всюду, за исключением множества x малой меры, где имеется большой выброс. Максимум же восстановленной функции отражает не точку, доставляющую максимум количеству $\bar{y}$, а точку «выброса» восстановленной функции.

Для того чтобы исключить такую ситуацию, необходимо чтобы восстановленная функция приближалась

к истинной равномерно во всей области задания функции, т. е. в метрике C

$$\rho_C(y(x),\ F(x,\ \alpha)=\sup_x|y(x)-F(x,\ \alpha)|.$$

Таким образом, в задаче восстановления регрессии используется понятие близость как в метрике L_P^2, так и в метрике C.

В задаче интерпретации данных косвенного эксперимента также используются два понятия близости: близость в метрике L^2 (в L_P^2 с весом $P(x)\equiv 1$):

$$\rho_L(f(t,\ \alpha_1),\ f(t,\ \alpha_2))=\left(\int(f(t,\ \alpha_1)-f(t,\ \alpha_2))^2\,dt\right)^{1/2},$$

и близость в метрике C:

$$\rho_C(f(t,\ \alpha_1),\ f(t,\ \alpha_2))=\sup_t|f(t,\ \alpha_1)-f(t,\ \alpha_2)|.$$

Как и при восстановлении регрессии, здесь выбор метрики определяется тем, как в дальнейшем предполагается использовать восстановленную функцию.

§ 8. Особенности задач восстановления зависимостей

Итак, выше мы установили, что все три задачи восстановления зависимостей сводятся к одной и той же схеме — схеме минимизации среднего риска, и что возможно лишь приближенное решение задачи минимизации среднего риска по эмпирическим данным. Спрашивается, обеспечит ли приближенное решение задачи нужную близость найденной зависимости к истинной?

Ответ на этот вопрос различен для разных задач восстановления зависимостей. Для задачи обучения распознавания образов ответ безусловный — да, обеспечит просто по определению (ведь согласно постановке требуется найти функцию, доставляющую функционалу величину, близкую к минимальной).

В задаче восстановления регрессии ответ не столь определенный. Легко можно показать, что если близость функций понимать в смысле метрики L_P^2, то из близости функционала к минимальному следует близость найденной функции к регрессии.

Доказательство этого утверждения немедленно вытекает из тождества

$$\int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy =$$
$$= \int (y - y(x))^2 P(x, y)\, dx\, dy + \int (y(x) - F(x, \alpha))^2 P(x)\, dx,$$

где $\bar{y} = y(x)$ — регрессия, $F(x, \alpha)$ — любая функция из заданного класса.

Однако из того, что значение функционала близко к минимальному, никак не следует близость в смысле метрики C соответствующей функции к регрессии. Чтобы гарантировать такую близость, просто минимизации функционала уже недостаточно. Необходимо, кроме того, удовлетворить еще и некоторым специальным требованиям.

Наконец, в задаче интерпретации результатов косвенных экспериментов близость функционала к минимальному не гарантирует близость восстанавливаемой функции к истинной ни в метрике L_P^2, ни в метрике C. Основная трудность при решении этой задачи как раз и состоит в том, что решение соответствующего операторного уравнения есть, возможно, задача некорректно поставленная, а в этом случае функции, доставляющие функционалу значения, близкие к минимальному, могут сколь угодно сильно отличаться от искомого решения. Поэтому главная проблема здесь состоит в том, чтобы установить, каким дополнительным условиям должно удовлетворять выбираемое решение, чтобы из того, что оно доставляет функционалу значение, близкое к минимальному, следовала бы близость решения к искомой функции.

Таким образом, несмотря на то, что для всех задач восстановления зависимостей функции, доставляющие точный минимум функционалу, определяют решение, приближенная минимизация не всегда приводит к цели. Поэтому прежде, чем применить конкретный метод минимизации среднего риска по эмпирическим данным, необходимо убедиться, что этот метод минимизации обеспечит приближение к искомому решению.

В последующих главах будут рассмотрены различные методы минимизации среднего риска по эмпирическим данным. Все они будут изучены применительно к каждой конкретной задаче восстановления зависимостей.

Основные утверждения главы I

1. Одной из центральных задач прикладной статистики является задача минимизации среднего риска по эмпирическим данным: требуется минимизировать функционал

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz,$$

если плотность распределения вероятностей $P(z)$ неизвестна, но задана функция потерь $Q(z, \alpha)$ и случайная независимая выборка

$$z_1, \ldots, z_l,$$

полученная согласно плотности $P(z)$.

2. Постановка задач обучения распознаванию образов, восстановления регрессии, интерпретации результатов косвенных экспериментов сводится к схеме минимизации среднего риска по эмпирическим данным: требуется минимизировать функционал

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy,$$

используя выборку

$$x_1, y_1; \ldots; x_l, y_l.$$

В задаче обучения распознаванию образов $y = \omega$ — число, которое может принимать только два значения — нуль и единица, $F(x, \alpha)$ — некоторое параметрическое множество характеристических функций.

В задаче восстановления регрессии y — действительное число, $F(x, \alpha)$ — некоторое параметрическое множество интегрируемых с квадратом функций.

В задаче интерпретации результатов косвенных экспериментов y — число, $F(x, \alpha)$ — некоторое параметрическое множество непрерывных функций, которые связаны со своими прообразами $f(t, \alpha)$ соотношением

$$Af(t, \alpha) = F(x, \alpha).$$

Задача решения операторного уравнения

$$Af(t) = F(x)$$

в $f(t, \alpha)$ *может быть некорректно поставленной.*

3. Отыскать точный минимум функционала $I(\alpha)$ *по эмпирическим данным, вообще говоря, невозможно. Можно лишь найти функцию, доставляющую функционалу* $I(\alpha)$ *значение, близкое к минимальному. Такое решение может быть найдено не наверное, а лишь с определенной вероятностью.*

В соответствии с содержательными постановками задач рассматриваются различные понятия близости функций.

— В задаче обучения распознаванию образов две функции считаются близкими, если близки соответствующие величины функционалов.

— В задачах восстановления регрессии и интерпретации результатов косвенных экспериментов приняты два определения близости функций: близость в метрике L_P^2 *и в метрике* C.

4. Приближенные методы минимизации функционала обеспечивают отыскание решений, близких к искомому, в задаче обучения распознаванию образов. В задаче восстановления регрессии близость к минимальному значению функционала гарантирует близость функции к искомой лишь в смысле метрики L_P^2. *В метрике* C *функции могут не быть близкими. Наконец, в задачах интерпретации результатов косвенных экспериментов прообраз* $f(t, \alpha^*)$ *функции* $F(x, \alpha^*)$, *доставляющей функционалу* $I(\alpha)$ *значение, близкое к минимальному, вообще говоря, не является близким к решению операторного уравнения ни в метрике* L_P^2, *ни в метрике* C.

5. Предлагается решать задачу восстановления зависимостей методом минимизации функционала $I(\alpha)$, *а в тех случаях, когда близость функционала к минимуму не гарантирует близости функции к искомой, найти дополнительные условия, при которых возможна такая гарантия.*

ПРИЛОЖЕНИЕ К ГЛАВЕ I

МЕТОДЫ РЕШЕНИЯ НЕКОРРЕКТНО ПОСТАВЛЕННЫХ ЗАДАЧ

§ П1. Задача решения операторного уравнения

Говорят, что два множества элементов, множество $\mathcal{M}$ и множество $\mathcal{N}$, связаны функциональной зависимостью, если каждому элементу $f \in \mathcal{M}$ может быть поставлен в однозначное соответствие элемент $F \in \mathcal{N}$.

Эта функциональная зависимость называется функцией, если $\mathcal{M}$ и $\mathcal{N}$ — множества чисел, функционалом, если $\mathcal{M}$ — множество функций, а $\mathcal{N}$ — множество чисел и оператором, если $\mathcal{M}$ — множество функций и $\mathcal{N}$ — также множество функций.

Каждый оператор A однозначно отображает элементы множества $\mathcal{M}$ в элементы множества $\mathcal{N}$. Этот факт будем записывать равенством

$$A\mathcal{M} = \mathcal{N}.$$

Среди множества операторов выделим такие, которые осуществляют взаимно однозначное отображение $\mathcal{M}$ в $\mathcal{N}$. Для них можно рассматривать задачу решения операторного уравнения

$$Af(t) = F(x), \qquad (\text{П}.1)$$

как задачу отыскания в $\mathcal{M}$ такого элемента $f(t)$, которому в $\mathcal{N}$ соответствует элемент $F(x)$.

Для операторов, осуществляющих взаимно однозначное отображение элементов $\mathcal{M}$ в $\mathcal{N}$, и функции $F(x) \in \mathcal{N}$ существует единственное решение операторного уравнения (П.1).

Однако найти способ решения операторного уравнения столь общей природы — задача безнадежная. Поэтому иссле-

дование операторных уравнений будем проводить лишь для случая непрерывных операторов.

Пусть элементы $f \in \mathcal{M}$ принадлежат метрическому пространству E_1 с метрикой $\rho_1(\cdot)$, а элементы $F \in \mathcal{N}$ — принадлежат, вообще говоря, другому метрическому пространству E_2 с метрикой $\rho_2(\cdot)$. Оператор A называется *непрерывным*, если близкие (в метрике ρ_1) элементы в E_1 он отображает в близкие (в метрике ρ_2) элементы в E_2. Формально это означает, что для всякого ε найдется такое $\delta(\varepsilon)$, что если только в E_1 выполнится неравенство

$$\rho_1(f_1, f_2) \leqslant \delta(\varepsilon),$$

то в E_2 окажется выполненным и неравенство

$$\rho_2(Af_1, Af_2) \leqslant \varepsilon.$$

Будем рассматривать операторное уравнение, заданное непрерывным оператором, осуществляющим взаимно однозначное отображение из $\mathcal{M}$ в $\mathcal{N}$. Решение такого операторного уравнения существует и единственно, т. е. существует обратный оператор A^{-1} из $\mathcal{N}$ в $\mathcal{M}$:

$$\mathcal{M} = A^{-1}\mathcal{N}.$$

Принципиальным оказывается вопрос, является ли обратный оператор непрерывным?

Если оператор A [1] непрерывный, то близким функциям из $\mathcal{N}$ соответствуют близкие их прообразы, т. е. решение операторного уравнения устойчиво. Если же обратный оператор не является непрерывным, то решение операторного уравнения, вообще говоря, не является устойчивым.

В последнем случае, согласно определению Адамара (см. § 5 гл. I), задача решения операторного уравнения считается некорректно поставленной.

Оказывается, что во многих важных случаях, например для вполне непрерывных операторов A, обратный оператор A^{-1} не является непрерывным, и, следовательно, задача решения соответствующего операторного уравнения является некорректно поставленной.

Определение. *Говорят, что линейный оператор A, определенный в линейном нормированном пространстве E_1 с областью значений в линейном нормированном пространстве E_2, является вполне непрерывным, если он отображает всякое ограниченное множество пространства E_1 в компакт-*

ное множество пространства E_2, т. е. если он всякую ограниченную бесконечную в E_1 последовательность

$$f_1, f_2, \ldots, f_i \ldots, \qquad \|f_j\| \leqslant c \tag{П.2}$$

($\|f_j\|$ — *норма в* E_1) *отображает в* E_2 *в такую последовательность*

$$Af_1, \ldots, Af_i, \ldots, \tag{П.3}$$

из которой может быть извлечена сходящаяся подпоследовательность

$$Af_{i_1}, \ldots, Af_{i_k}, \ldots \tag{П.4}$$

Покажем, что если пространство E_1 содержит ограниченные некомпактные множества, то оператор A^{-1}, обратный вполне непрерывному оператору A, не может быть непрерывным.

В самом деле, рассмотрим в E_1 некоторое ограниченное некомпактное множество. В этом множестве выделим бесконечную последовательность (П.2), никакая подпоследовательность которой не сходится. В E_2 ей соответствует бесконечная последовательность (П.3), из которой может быть извлечена сходящаяся подпоследовательность (П.4). Если бы оператор A^{-1} был непрерывным, то последовательности (П.4) в E_1 соответствовала бы сходящаяся последовательность

$$f_{i_1}, \ldots, f_{i_k} \ldots, \tag{П.5}$$

которая является подпоследовательностью (П.2) в противоречии с тем, как была выбрана последовательность.

Итак, задача решения операторного уравнения, заданного вполне непрерывным оператором, является некорректно поставленной. В основном тексте книги мы ограничимся рассмотрением линейных интегральных операторов

$$Af = \int_a^b K(t, x) f(t)\, dt \tag{П.6}$$

с непрерывными ядрами $K(t, x)$ в области $a \leqslant t \leqslant b$, $a \leqslant x \leqslant b$.

Операторы (П.6) являются вполне непрерывными при отображении непрерывных на $[a, b]$ функций. Доказательство этого факта содержится во всех руководствах по функциональному анализу (см., например, [28]).

§ П2. Задачи, корректные по Тихонову

Задачу решения операторного уравнения

$$Af = F$$

называют *корректной по Тихонову* на множестве $\mathcal{M}' \subset \mathcal{M}$, а само множество $\mathcal{M}'$ — ее *множеством* (*классом*) *корректности*, если:

а) точное решение задачи существует и принадлежит $\mathcal{M}'$;

б) принадлежащее $\mathcal{M}'$ решение единственно для любого $F \in A\mathcal{M}' = \mathcal{N}'$;

в) принадлежащие множеству $\mathcal{M}'$ решения устойчивы относительно $F \in \mathcal{N}'$.

При $\mathcal{M}' = \mathcal{M}$ и $\mathcal{N}' = \mathcal{N}$ понятие корректности по Тихонову совпадает с понятием корректности по Адамару. Смысл определения корректности по Тихонову заключается в том, что корректность может быть достигнута за счет сужения рассматриваемого множества решений $\mathcal{M}$ до класса корректности $\mathcal{M}'$.

Следующая лемма устанавливает, что если множество решений $\mathcal{M}$ сужено до компакта $\mathcal{M}'$, то оно образует класс корректности.

Лемма. *Если на компакте* $\mathcal{M}' \subset \mathcal{M}$ *задан непрерывный взаимно однозначный оператор* A, *то обратный оператор* A^{-1} *непрерывен на множестве* $\mathcal{N}' = A\mathcal{M}'$.

Доказательство. Выберем произвольный элемент $F_0 \in \mathcal{N}'$ и произвольную сходящуюся к нему последовательность

$$\{F_n\} \subset \mathcal{N}', \quad F_n \xrightarrow[n\to\infty]{} F_0.$$

Требуется доказать сходимость

$$f_n = A^{-1}F_n \xrightarrow[n\to\infty]{} A^{-1}F_0 = f_0.$$

Так как $\{f_n\} \subset \mathcal{M}'$, а $\mathcal{M}'$ — компакт, то последовательность $\{f_n\}$ имеет предельные точки, принадлежащие $\mathcal{M}'$. Пусть f_0 — такая предельная точка. Поскольку f_0 — предельная точка, существует сходящаяся к ней последовательность $\{f_{n_k}\}$, которой соответствует последовательность $\{F_{n_k}\}$, сходящаяся к F_0. Поэтому, переходя к пределу в равенстве

$$Af_{n_k} = F_{n_k}$$

и пользуясь непрерывностью оператора A, получим

$$Af_0 = F_0.$$

В силу однозначности оператора A^{-1} имеем $A^{-1}F_0 = f_0$, откуда следует единственность предельной точки последовательности $\{f_{n_k}\}$. Остается проверить, что к f_0 сходится вся последовательность $\{f_{n_k}\}$. Действительно, если бы к f_0 сходилась не вся последовательность, то нашлась бы окрестность точки f_0, вне которой имеется бесконечное число членов последовательности $\{f_{n_k}\}$. Ввиду компактности множества $\mathcal{M}'$ у этой последовательности есть предельная точка f_0', которая по доказанному выше обязана совпадать с f_0, а это противоречит допущению о том, что выбранная последовательность лежит вне окрестности точки f_0. Лемма доказана.

Таким образом, корректность по Тихонову на компакте $\mathcal{M}'$ вытекает из одних условий существования и единственности решения операторного уравнения: третье условие — устойчивость решения операторного уравнения — выполняется автоматически.

Этот факт, по существу, и лежит в основе всех конструктивных идей решения некорректных операторных уравнений. Рассмотрим некоторые из них.

§ П3. Метод регуляризации

Метод регуляризации был предложен А. Н. Тихоновым в 1963 г.

Пусть необходимо решить операторное уравнение

$$Af = F, \tag{П.7}$$

заданное непрерывным взаимно однозначным из $\mathcal{M}$ в $\mathcal{N}$ оператором A. И пусть решение (П.7) существует.

Введем в рассмотрение непрерывный функционал $\Omega(f)$, который назовем *стабилизатором* и который обладает следующими тремя свойствами:

1) решение операторного уравнения принадлежит области определения $D(\Omega)$ функционала $\Omega(f)$;

2) на области определения функционал $\Omega(f)$ принимает вещественные неотрицательные значения;

3) все множества

$$\mathcal{M}_c = \{f : \Omega(f) \leqslant c\}, \quad c \geqslant 0,$$

являются компактами.

Идея метода регуляризации состоит в том, чтобы найти решение (П.7) как элемент, минимизирующий некоторый функционал, но не функционал

$$\rho = \rho_2(Af, F)$$

(такая задача была бы эквивалентна решению уравнения (П.7) и потому тоже некорректна), а «исправленный» функционал

$$R^\gamma(f, F) = \rho_2^2(Af, F) + \gamma\Omega(f), \quad f \in D(\Omega) \qquad (\text{П.8})$$

с *параметром регуляризации* $\gamma > 0$. Задача минимизации функционала (П.8) устойчива, т. е. близким функциям F и F_δ ($\rho_2(F, F_\delta) \leqslant \delta$) соответствуют близкие элементы f^γ и f_δ^γ, минимизирующие функционалы $R^\gamma(f, F)$ и $R^\gamma(f, F_\delta)$.

Проблема состоит в том, чтобы установить, в каком соотношении должны находиться величины δ и γ, чтобы последовательность решений f_δ^γ регуляризованных задач $R^\gamma(f; F_\delta)$ сходилась при $\delta \to 0$ к решению операторного уравнения (П.7). Эти соотношения устанавливает следующая теорема.

Теорема П.1. *Пусть E_1 и E_2 — метрические пространства, и пусть для $F \in \mathcal{N}$ существует решение уравнения* (П.7) $f \in D(\Omega)$. *Тогда, если вместо точной правой части F уравнения* (П.7) *известны приближения*[1]) $F_\delta \in E_2$ *такие, что* $\rho_2(F, F_\delta) \leqslant \delta$, *а значения параметра γ выбираются так, что*

$$\gamma(\delta) \to 0 \quad \text{при} \quad \delta \to 0, \quad \lim_{\delta \to 0} \frac{\delta^2}{\gamma(\delta)} \leqslant r < \infty, \qquad (\text{П.9})$$

то элементы $f_\delta^{\gamma(\delta)}$, минимизирующие функционалы $R^{\gamma(\delta)}(f, F_\delta)$ на $D(\Omega)$, сходятся к точному решению f при $\delta \to 0$.

Доказательство теоремы использует следующий факт: для всякого фиксированного $\gamma > 0$ и любого $F \in \mathcal{N}$ существует элемент $f^\gamma \in D(\Omega)$, минимизирующий на $D(\Omega)$ функционал $R^\gamma(f, F)$.

---

[1]) Элементы F_δ могут не принадлежать множеству $\mathcal{N}$.

Пусть γ и δ удовлетворяют соотношению (П.9). Рассмотрим последовательность элементов $f_\delta^{\gamma(\delta)}$, минимизирующих $R^{\gamma(\delta)}(f, F_\delta)$, и покажем, что имеет место сходимость

$$f_\delta^{\gamma(\delta)} \underset{\delta \to 0}{\to} f.$$

По определению имеем

$$R^{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) \leqslant R^{\gamma(\delta)}(f, F_\delta) = \rho_2^2(Af, F_\delta) + \gamma(\delta)\,\Omega(f) \leqslant$$
$$\leqslant \delta^2 + \gamma(\delta)\,\Omega(f) = \gamma(\delta)\left(\Omega(f) + \frac{\delta^2}{\gamma(\delta)}\right).$$

Учитывая, что

$$R^{\gamma(\delta)}(f_\delta^{\gamma(\delta)}, F_\delta) = \rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) + \gamma(\delta)\,\Omega(f_\delta^{\gamma(\delta)}),$$

заключаем

$$\Omega(f_\delta^{\gamma(\delta)}) \leqslant \Omega(f) + \frac{\delta^2}{\gamma(\delta)},$$
$$\rho_2^2(Af_\delta^{\gamma(\delta)}, F_\delta) \leqslant \gamma(\delta)\left(\Omega(f) + \frac{\delta^2}{\gamma(\delta)}\right).$$

Так как выполнены условия (П.9), то все элементы ряда $f_\delta^{\gamma(\delta)}$ для достаточно малых $\delta > 0$ принадлежат компакту $\mathcal{M}_{c^*}$, где $c^* = \Omega(f) + r + \varepsilon$; $\varepsilon > 0$, а их образы $F_\delta^{\gamma(\delta)} = Af_\delta^{\gamma(\delta)}$ сходятся:

$$\rho_2(F_\delta^{\gamma(\delta)}, F) \leqslant \rho_2(F_\delta^{\gamma(\delta)}, F_\delta) + \delta \leqslant$$
$$\leqslant \delta + \sqrt{\delta^2 + \gamma(\delta)\,\Omega(f)} \underset{\delta \to 0}{\to} 0.$$

Отсюда на основании леммы заключаем, что сходятся и прообразы

$$f_\delta^{\gamma(\delta)} \to f \quad \text{при} \quad \delta \to 0,$$

что и требовалось доказать.

В гильбертовом пространстве для линейного оператора A функционал $\Omega(f)$ может быть взят равным $\|f\|^2$. И хотя множества $\mathcal{M}_c$ при этом оказываются слабо компактными, сходимость регуляризованных решений в силу свойств гильбертова пространства, как будет показано ниже, оказывается сильной. Такой выбор регуляризующего функционала удобен еще и тем, что область его определения $D(\Omega)$ совпадает со всем пространством E_1. Однако в этом случае условия на параметр γ более жесткие, чем в теореме П.1: γ должно стремиться к нулю медленнее, чем δ^2.

Итак, справедлива теорема.

Теорема П.2. *Пусть* E_1 — *гильбертово пространство и* $\Omega(f) = \|f\|^2$. *Тогда при* $\gamma(\delta)$, *удовлетворяющем соотношениям* (П.9) *с* $r = 0$, *регуляризованные элементы* $f_\delta^{\gamma(\delta)}$ *сходятся при* $\delta \to 0$ *к точному решению* f *в метрике пространства* E_1.

Доказательство. Из геометрии гильбертовых пространств известно, что сфера $\|f\|^2 \leqslant c$ является слабым компактом и что из свойств слабой сходимости элементов f_i к элементу f и сходимости норм $\|f_i\|$ к $\|f\|$ вытекает сильная сходимость

$$\|f_i - f\| \xrightarrow[i \to \infty]{} 0.$$

Кроме того, из слабой сходимости $f_i \to f$ вытекает

$$\|f\| \leqslant \varliminf_{i \to \infty} \|f_i\|. \tag{П.10}$$

Используя эти свойства гильбертова пространства, докажем теорему. Для этого заметим, что для слабой сходимости в пространстве E_1 справедлива предыдущая теорема: $f_\delta^{\gamma(\delta)}$ слабо сходятся к f при $\delta \to 0$.

Поэтому, согласно (П.10), справедливо неравенство

$$\|f\| \leqslant \varliminf_{\delta \to 0} \|f_\delta^{\gamma(\delta)}\|.$$

С другой стороны, учитывая то, что $\Omega(f) = \|f\|^2$, и то, что $r = 0$, получаем

$$\varliminf_{\delta \to 0} \|f_\delta^{\gamma(\delta)}\|^2 \leqslant \varlimsup_{\delta \to 0} \left(\|f\|^2 + \frac{\delta^2}{\gamma(\delta)} \right) = \|f\|^2.$$

Следовательно, имеет место сходимость норм

$$\|f_\delta^{\gamma(\delta)}\| \xrightarrow[\delta \to 0]{} \|f\|,$$

а это вместе с фактом существования слабой сходимости влечет в силу свойств гильбертова пространства сильную сходимость

$$\|f_\delta^{\gamma(\delta)} - f\| \xrightarrow[\delta \to 0]{} 0,$$

что и требовалось доказать.

Приведенные теоремы являются центральными в теории регуляризации. С их помощью устанавливается прин-

ципиальная возможность решения некорректных задач. Однако при решении практических задач вопросы сходимости последовательности регуляризованных решений не являются наиболее актуальными. Обычно правая часть операторного уравнения задана с конечной точностью δ, и проблема заключается в том, чтобы определить величину константы регуляризации $\gamma(\delta)$, при которой будет обеспечено наилучшее приближение к искомому решению. Для этой ситуации утверждения теорем П.1, П.2, где величина γ определяется с точностью до константы r (да и то при достаточно малых δ), являются явно недостаточными.

В настоящее время нет сколько-нибудь надежных методов выбора константы регуляризации. Однако имеются многочисленные примеры того, что при надлежащем выборе константы γ могут быть получены достаточно хорошие приближения к решению некорректных задач.

§ П4. Метод квазирешений

Определение. *Говорят, что уравнение* (П.7) *имеет на множестве* $\mathcal{M}' \subset \mathcal{M}$ *квазирешение* $f' \in \mathcal{M}'$, *если для элемента* f' *справедливо равенство*

$$\rho_2(Af', F) = \inf_{f \in \mathcal{M}'} \rho_2(Af, F).$$

Иными словами, квазирешение f' — это такая точка множества $\mathcal{M}'$, образ которой Af' является ближайшим к F элементом на множестве $A\mathcal{M}'$.

Понятие квазирешения было введено В. К. Ивановым. Оно обобщает понятие решения: для существования квазирешения не требуется, чтобы решение операторного уравнения принадлежало области $\mathcal{M}'$.

Справедлива

Теорема П.3. *Пусть* E_1 *и* E_2 — *банаховы пространства, причем пространство* E_2 *строго выпукло,* A — *взаимно однозначный непрерывный линейный оператор. Тогда задача отыскания квазирешения операторного уравнения* (П.7) *на выпуклом компакте* $\mathcal{M}'$ *поставлена корректно по Адамару.*

Доказательство этой теоремы аналогично доказательству теоремы П.1.

Пусть f' — квазирешение уравнения (П.7) и $F' = Af'$. Очевидно, F' есть проекция элемента F на множество $\mathcal{N}' = A\mathcal{M}'$. Так как проекция определяется однозначно, то в силу взаимной однозначности отображения множества $\mathcal{M}$ на множество $\mathcal{N}$ следует единственность квазирешения f'.

Очевидно, что $f' = A^{-1}F' = A^{-1}\mathrm{Pr}F$ (Pr — оператор проектирования в $A\mathcal{M}'$). Согласно лемме о непрерывности обратного отображения на компакт, оператор A^{-1} непрерывен на $\mathcal{N}'$. Оператор проектирования непрерывен на $\mathcal{N}$. Поэтому $A^{-1}\mathrm{Pr}$ — непрерывный на $\mathcal{N}$ оператор, и, следовательно, квазирешение f' непрерывно зависит от F. Теорема доказана.

Рассмотрим теперь расширяющуюся систему выпуклых компактов в пространстве E_1:

$$\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_n \subset \mathcal{M}_0, \qquad (\text{П.11})$$

и рассмотрим для них последовательность квазирешений $f^i = f(F, \mathcal{M}_i)$ операторного уравнения

$$Af = F.$$

Справедлива

Теорема П.4. *Если в условиях теоремы* П.3 *выбрать систему выпуклых компактов такую, что*

$$\overline{U\mathcal{M}_i} = \mathcal{M}_0$$

(черта означает замыкание множества), то квазирешения $f^{(i)} = f(F, \mathcal{M}_i)$ *будут сходиться к квазирешению* $f = f(F, \mathcal{M}_0)$ *при* $n \to \infty$.

Доказательство теоремы основано на том факте, что из непрерывности оператора A и плотности множества $\bigcup\limits_{n \geqslant 1} \mathcal{M}_n$ в $\mathcal{M}_0$ следует, что множество $\bigcup\limits_{n \geqslant 1} A\mathcal{M}_i$ также плотно в $A\mathcal{M}_0$. Поэтому точки F_n ($F_n \in A\mathcal{M}_n$), реализующие расстояние $\rho(F, A\mathcal{M}_n)$ при $n \to \infty$, стремятся к точке $F \in A\mathcal{M}$, реализующей расстояние $\rho(F, A\mathcal{M}_0)$. Остальное доказывается ссылкой на лемму о непрерывности оператора A^{-1} на множестве $A\mathcal{M}_0$.

Применим метод квазирешений для отыскания решения операторного уравнения

$$Af = F.$$

Пусть

$$\varphi_1(t), \ldots, \varphi_n(t), \ldots$$

— полная в L^2 ортонормированная система функций.

Будем искать решение на компакте $\mathcal{M}$:

$$\mathcal{M}=\left\{f: f(t)=\sum_{k=1}^{\infty}\alpha_k\varphi_k(t), \quad |\alpha_k|\leqslant\frac{c}{k}\right\}, \qquad (\text{П.12})$$

где c — константа. Наряду с компактом $\mathcal{M}$ рассмотрим компакт $\mathcal{M}_n$:

$$\mathcal{M}_n=\left\{f: f(t)=\sum_{k=1}^{n}\alpha_k\varphi_k(t), \quad |\alpha_k|\leqslant\frac{c}{k}\right\}.$$

Согласно теореме П.3 квазирешение $f(F, \mathcal{M}_n)$ при больших n аппроксимирует решение $f(F, \mathcal{M})$ и потому может быть принято за решение исходного уравнения.

На практике при решении некорректно поставленных задач с неточно заданной правой частью проблема состоит в том, чтобы определить компакт (размерность подпространства), в котором следует искать квазирешение.

Выбор такого компакта (определение числа членов разложения) — проблема, эквивалентная определению константы регуляризации в методе регуляризации.

В основной части книги будут рассмотрены методы определения подходящего числа членов разложения для отыскания решения некорректных задач измерений.

Подробно теория некорректных задач рассмотрена в монографии [56].

ГЛАВА II

МЕТОДЫ МИНИМИЗАЦИИ СРЕДНЕГО РИСКА

§ 1. Два пути минимизации среднего риска

Существуют два пути решения задачи минимизации среднего риска

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz \tag{2.1}$$

по эмпирическим данным

$$z_1, \ldots, z_l. \tag{2.2}$$

Первый путь связан с идеей конструирования по выборке (2.2) и функции $Q(z, \alpha)$ *эмпирического функционала*

$$I_э(\alpha) = \Phi(Q(z_1, \alpha), \ldots, Q(z_l, \alpha); z_1, \ldots, z_l), \tag{2.3}$$

т. е. такого функционала, который не зависит от неизвестной плотности распределения вероятностей $P(z)$. В отличие от (2.1), функционал (2.3) можно минимизировать. Точку минимума функционала (2.3) примем за точку минимума исходного функционала (2.1). Такой метод минимизации среднего риска называется *методом минимизации эмпирического функционала*.

Основная проблема, которая возникает при изучении метода минимизации эмпирического функционала — определить для каждого типа аппроксимации (2.3) величину ошибки и указать такую аппроксимацию функционала (2.1) эмпирическим функционалом (2.3), при которой гарантируется отыскание функции, доставляющей функционалу (2.1) значение, близкое к минимальному.

Второй путь связывает нахождение минимума функционала (2.1) с использованием итеративной процедуры:

$$\alpha(i) = \alpha(i-1) + \gamma(i) S(i, z_i). \tag{2.4}$$

Согласно процедуре (2.4) уточнение вектора параметров α на i-м шаге определяется величиной i-го шага $\gamma(i)$ и направлением $S(i, z_i)$.

Оказывается, что если направление движения $S(i, z_i)$ выбирать так, чтобы на каждом шаге выполнялось неравенство [1])

$$(\nabla_\alpha I(\alpha(i-1)))^\tau MS(i, z) \geqslant \delta > 0, \qquad (2.5)$$

где $\nabla_\alpha I(\alpha)$ — градиент по α функционала (2.1), $MS(i, z)$ — математическое ожидание направления i-го шага, то при некоторых дополнительных условиях, ограничивающих рост вектора $S(i, z)$ (например, функцией $|z|$) и величину шага $\gamma(i)$ (требуется, чтобы $\sum_{i=1}^{\infty} \gamma^2(i) < \infty$ и в то же время $\sum_{i=1}^{\infty} \gamma(i) = \infty$), процедура (2.4) и случайная выборка $z_1, \ldots, z_l, \ldots$ индуцируют последовательность $\alpha(i)$, сходящуюся к вектору параметров α_0, доставляющему минимум функционалу (2.1) [45].

Итеративная процедура (2.4) является развитием градиентных методов поиска минимума. В самом деле, если бы плотность распределения вероятностей $P(z)$ была известна, то при определенных условиях можно было бы вычислить градиент

$$\nabla_\alpha I(\alpha) = \int \nabla_\alpha Q(z, \alpha) P(z)\, dz. \qquad (2.6)$$

Тогда процедура спуска представляла бы собой следующее правило:

$$\alpha(i) = \alpha(i-1) - \gamma(i) \nabla_\alpha I(\alpha(i-1)). \qquad (2.7)$$

Процедура (2.4) отличается от (2.7) тем, что на каждом шаге в качестве направления движения выбирается не направление градиента, а такое направление, движение по которому происходит «в среднем примерно в ту же сторону, что и по градиенту». Слова «в среднем примерно в ту же сторону» формализуются неравенством (2.5).

Таким образом, основной результат теории итеративных методов состоит в том, что даже при достаточно общих условиях, определяющих выбор направления движения и величину шага, итеративные процедуры (2.4) приводят к цели. Однако именно в силу универсальности итеративной процедуры, отыскание значения функционала, близкого к минимальному, можно гарантировать лишь асимптотически. Для решения задач минимизации среднего риска по выборке фиксированного объема итератив-

[1]) Здесь и далее вектор определен как вектор-столбец, т — знак транспонирования.

ные методы оказываются малоприспособленными. Поэтому в дальнейшем мы не будем их рассматривать. Решение задачи минимизации функционала (2.1) по эмпирическим данным (2.2) будем связывать с конструированием эмпирического функционала (2.3) и последующей его минимизацией.

§ 2. Проблема больших выбросов

Нашей целью является построение метода, гарантирующего с заданной вероятностью отыскание функции, доставляющей функционалу

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

значение, близкое к минимальному, если плотность $P(z)$ неизвестна, но дана выборка $z_1, \ldots, z_l$.

Однако решить эту задачу без привлечения априорной информации невозможно. В самом деле, рассмотрим одну из самых простых задач восстановления зависимостей по эмпирическим данным. Требуется минимизировать функционал

$$I'(\alpha) = \int (t-\alpha)^2 P(t)\, dt, \tag{2.8}$$

если плотность $P(t)$ неизвестна, но дана случайная независимая выборка $t_1, \ldots, t_l$.

Минимум функционала (2.8) достигается при

$$\alpha = \int tP(t)\, dt. \tag{2.9}$$

Таким образом, проблема состоит в том, чтобы для неизвестной плотности $P(t)$ найти способ, гарантирующий с заданной вероятностью достаточно точную оценку среднего по выборке фиксированного объема l.

Оказывается, что, не имея априорных сведений о плотности $P(t)$, получить гарантированную оценку среднего нельзя.

Действительно, пусть случайная величина t принимает два значения — нуль и K, причем значение нуль она принимает с вероятностью $1-\varepsilon$ и K — с вероятностью ε ($P(t=0) = 1-\varepsilon$; $P(t=K) = \varepsilon$). Пусть теперь ε — настолько малая величина, что с большой вероятностью $(1-\delta)$ случайная независимая выборка $t_1, \ldots, t_l$ состоит из одних

нулей, и, следовательно, величина эмпирического среднего

$$\alpha_э = \frac{1}{l} \sum_{i=1}^{l} t_i.$$

равна нулю. (Вероятность этого события $(1-\varepsilon)^l = 1-\delta$.) С другой стороны, математическое ожидание случайной величины t равно

$$Mt = 0\,(1-\varepsilon) + K\varepsilon = K\varepsilon,$$

и в зависимости от величины K может принимать любые значения, в том числе и достаточно большие (например, когда $K = 1/\varepsilon^2$).

Итак, в нашем примере, несмотря на то, что почти любая величина эмпирического среднего, образованная по выборке длины l, равнялась нулю, никаких надежных заключений о величине математического ожидания сделать было нельзя.

Это произошло потому, что даже при малом ε произведение $K\varepsilon$ могло быть большой величиной. Иначе говоря, распределение случайной величины t было таким, что на «малой мере» ε оказалась сосредоточена большая величина K. О таких ситуациях в статистике говорят, что случайная величина допускает «большой выброс».

В каких же случаях по величине эмпирического среднего можно надежно судить о математическом ожидании?

Ответ на этот вопрос следует из неравенства Чебышева. Согласно этому неравенству вероятность уклонения случайной величины t от своего математического ожидания Mt может быть оценена так:

$$P\{|t - Mt| \geqslant \sigma\varkappa\} \leqslant \frac{1}{\varkappa^2},$$

где σ^2 — дисперсия случайной величины t.

Рассмотрим теперь случайную величину

$$\xi = \frac{1}{l} \sum_{i=1}^{l} t_i,$$

где $t_1, \ldots, t_l$ — случайная независимая выборка длины l. Заметим, что

$$M\xi = Mt, \quad \sigma_\xi = \frac{\sigma}{\sqrt{l}}.$$

Неравенство Чебышева для этой величины имеет вид

$$P\left\{\left|\frac{1}{l}\sum_{i=1}^{l} t_i - Mt\right| \geqslant \frac{\sigma\varkappa}{\sqrt{l}}\right\} \leqslant \frac{1}{\varkappa^2}. \qquad (2.10)$$

Запишем неравенство (2.10) в иной форме. Для этого приравняем правую часть неравенства (2.10) величине η:

$$\frac{1}{\varkappa^2} = \eta$$

и разрешим равенство относительно $\varkappa$: $\varkappa = 1/\sqrt{\eta}$.

Теперь утверждение, что с вероятностью $1-\eta$ имеют место неравенства

$$\frac{1}{l}\sum_{i=1}^{l} t_i - \frac{\sigma}{\sqrt{l\eta}} < Mt < \frac{1}{l}\sum_{i=1}^{l} t_i + \frac{\sigma}{\sqrt{l\eta}}, \qquad (2.11)$$

полностью эквивалентно утверждению (2.10).

Если бы быаа известна дисперсия σ^2 случайной величины t, то неравенства (2.11) определяли бы величину доверительного интервала для математического ожидания Mt и, тем самым, *гарантированную оценку среднего*, т. е. такую оценку, которая выполняется с заданной вероятностью. Поэтому, для того чтобы получить гарантированную оценку среднего по величине эмпирического среднего, достаточно знать либо *абсолютную оценку* $\tau^2_{\text{абс}}$ *дисперсии*

$$\sigma^2 \leqslant \tau^2_{\text{абс}}, \qquad (2.12)$$

либо при условии, что искомое среднее есть величина положительная, оценку *относительной величины дисперсии*

$$\left(\frac{\sigma}{Mt}\right)^2 \leqslant \tau^2_{\text{отн}}. \qquad (2.13)$$

Действительно, из (2.11) и (2.12) следует, что знание абсолютной оценки дисперсии немедленно приводит к построению гарантированной оценки вида

$$\frac{1}{l}\sum_{i=1}^{l} t_i - \frac{\tau_{\text{абс}}}{\sqrt{\eta l}} < Mt < \frac{1}{l}\sum_{i=1}^{l} t_i + \frac{\tau_{\text{абс}}}{\sqrt{\eta l}}. \qquad (2.14)$$

А из (2.11) и (2.13) следует, что знание оценки относительной величины дисперсии приводит к построению

гарантированной оценки вида

$$\frac{\frac{1}{l}\sum_{i=1}^{l} t_i}{1+\frac{\tau_{\text{отн}}}{\sqrt{\eta l}}} < Mt < \frac{\frac{1}{l}\sum_{i=1}^{l} t_i}{1-\frac{\tau_{\text{отн}}}{\sqrt{\eta l}}} \qquad \left(\frac{\tau_{\text{отн}}}{\sqrt{\eta l}} < 1\right). \tag{2.15}$$

Пусть теперь случайная величина t неотрицательна (именно этот случай и рассматривается в книге, ведь $t_\alpha = Q(z, \alpha) = (y - F(x, \alpha))^2$). Тогда заведомо $Mt > 0$, и, следовательно, можно воспользоваться информацией об оценке относительной дисперсии.

При получении доверительных интервалов (2.14) и (2.15) мы использовали неравенство Чебышева. Это неравенство справедливо для любых распределений и потому для некоторых типов распределений может оказаться грубым. В частности, если распределение таково, что величина t положительна и не превосходит τ (в этом случае $\sigma \leqslant \tau/2$), то, как будет показано в главе VII, имеет место более сильная, чем в неравенстве Чебышева, оценка

$$P\left\{\left|\frac{1}{l}\sum_{i=1}^{l} t_i - Mt\right| \geqslant \varkappa\right\} \leqslant le^{-\frac{\varkappa^2 l}{4\tau^2}}. \tag{2.16}$$

С помощью (2.16) можно получить более точную гарантированную оценку величины математического ожидания.

Чтобы иметь возможность использовать неравенство (2.16), будем требовать вместо априорного знания абсолютной оценки дисперсии положительной случайной величины знания абсолютной оценки τ самой случайной величины t (конечно, в том случае, когда эта оценка существует).

Итак, для того чтобы иметь возможность оценить среднее по величине эмпирического среднего, достаточно знать либо абсолютную оценку τ случайной величины t, либо оценку $\tau_{\text{отн}}$ относительной величины дисперсии случайной величины t.

В этой книге мы будем изучать распределение не одной случайной величины t, а целого множества случайных величин

$$t_\alpha = Q(z, \alpha) = (y - F(x, \alpha))^2,$$

зависящих от параметра α. Для получения равномерных гарантированных оценок средних этих величин нам понадобятся равномерные для этих величин характеристики больших выбросов.

Возможный выброс на множестве $t_\alpha = Q(z, \alpha)$ будем характеризовать абсолютной оценкой величины потерь

$$\tau_{\text{абс}} = \sup_{\alpha, z} Q(z, \alpha) \tag{2.17}$$

или оценкой относительной величины дисперсии

$$\tau_{\text{отн}} = \sup_{\alpha} \left[\frac{D\{Q(z, \alpha)\}}{(MQ(z, \alpha))^2}\right]^{1/2} = \sup_{\alpha} \sqrt{\frac{MQ^2(z, \alpha)}{(MQ(z, \alpha))^2} - 1}. \tag{2.18}$$

Ниже будет показано, что если известна хотя бы одна из характеристик выброса (абсолютная или относительная оценка), то по случайной выборке фиксированного объема l может быть дана гарантированная оценка величины среднего риска, а при некоторых дополнительных ограничениях решена задача минимизации среднего риска.

§ 3. Априорная информация в задачах восстановления зависимостей по эмпирическим данным

Итак, для того чтобы получить гарантированное решение задачи минимизации среднего риска по ограниченному объему эмпирических данных, необходимо использовать априорную информацию о возможных выбросах случайных величин $t_\alpha = Q(z, \alpha)$. Величины возможных выбросов могут быть охарактеризованы либо абсолютной оценкой величины потерь (2.17), либо оценкой относительной величины дисперсии (2.18).

Насколько же обременительно получение априорной информации об абсолютной или относительной оценках для рассматриваемых в этой книге задач восстановления зависимостей: обучения распознаванию образов, восстановления регрессии, интерпретации результатов косвенных экспериментов?

Замечательная особенность задачи обучения распознаванию образов состоит в том, что для нее абсолютная величина потерь всегда не превосходит единицу. Действительно, согласно постановке задачи распознавания,

функция потерь

$$Q(z, \alpha) = (\omega - F(x, \alpha))^2$$

равна либо нулю, либо единице.

Таким образом, существование априорной абсолютной оценки величины потерь в задаче обучения распознаванию образов является тривиальным фактом.

В отличие от задачи распознавания образов, в задачах восстановления регрессии или интерпретации результатов косвенных экспериментов существование абсолютной оценки величины потерь — факт далеко не тривиальный. Чаще оказывается, что абсолютной оценки не существует. Такая ситуация возникает уже при восстановлении линейной регрессии. В самом деле, функция потерь в этом случае равна

$$Q(z, \alpha) = (y - F(x, \alpha))^2,$$

и если на значения параметров α не наложено никаких специальных ограничений, то в классе линейных функций $F(x, \alpha)$ найдется такая функция, что величина потерь может стать достаточно большой, даже если величины y и x ограничены.

Поэтому при решении задач восстановления регрессии и интерпретации результатов косвенных экспериментов будем использовать информацию не об абсолютной оценке возможных потерь, а об оценке относительной величины дисперсии потерь.

В каком же соотношении находится эта априорная информация с обычно используемой априорной информацией в задаче восстановления зависимостей?

Зафискируем в $Q(z, \alpha) = (y - F(x, \alpha))^2$ функцию $F(x, \alpha^*)$. Тогда плотность распределения вероятностей $P(x, y)$ индуцирует случайную величину

$$t'_{\alpha^*} = y - F(x, \alpha^*),$$

и, следовательно, оценка относительной величины дисперсии есть априорная информация о плотности распределения вероятностей случайных величин $(t'_{\alpha^*})^2$.

Если бы величина t'_{α^*} для любого α^* была распределена по нормальному закону, то оценка относительной

величины дисперсии потерь была бы равна

$$\tau_{\text{отн}} = \sup_{\alpha} \sqrt{\frac{M(t'_\alpha)^4}{(M(t'_\alpha)^2)^2} - 1} =$$

$$= \sup_{\mu,\,\sigma} \sqrt{\frac{\frac{1}{\sqrt{2\pi}\,\sigma}\int\limits_{-\infty}^{\infty}(t'_\alpha)^4 \exp\left\{-\frac{(t'_\alpha-\mu)^2}{2\sigma^2}\right\} dt'_\alpha}{\left(\frac{1}{\sqrt{2\pi}\,\sigma}\int\limits_{-\infty}^{\infty}(t'_\alpha)^2 \exp\left\{-\frac{(t'_\alpha-\mu)^2}{2\sigma^2}\right\} dt'_\alpha\right)^2} - 1} = \sqrt{2}$$

независимо от параметров закона.

Если бы величина t'_{α^*} для любого α^* была распределена по равномерному закону, то оценка равнялась бы

$$\tau_{\text{отн}} = \sup_{a,\,b} \sqrt{\frac{\frac{1}{b-a}\int\limits_{a}^{b}(t'_\alpha)^4\,dt'_\alpha}{\left(\frac{1}{b-a}\int\limits_{a}^{b}(t'_\alpha)^2\,dt'_\alpha\right)^2} - 1} = \sqrt{\frac{4}{5}} = \sqrt{0{,}8}.$$

Наконец, если бы величина t_{α^*} для любого α была распределена по закону Лапласа, то оценка равнялась бы

$$\tau_{\text{отн}} = \sup_{\mu,\,\Delta} \sqrt{\frac{\frac{1}{2\Delta}\int\limits_{-\infty}^{\infty}(t'_\alpha)^4 \exp\left\{-\left|\frac{t'_\alpha-\mu}{\Delta}\right|\right\} dt'_\alpha}{\left(\frac{1}{2\Delta}\int\limits_{-\infty}^{\infty}(t'_\alpha)^2 \exp\left\{-\left|\frac{t'_\alpha-\mu}{\Delta}\right|\right\} dt'_\alpha\right)^2} - 1} = \sqrt{5}.$$

Эти оценки также не зависят от параметров закона.

Априорная информация о распределении в терминах оценки относительной величины дисперсии потерь является минимальной априорной информацией, которая будет использоваться в книге.

Другим видом априорной информации, которая обычно используется при восстановлении функциональной зависимости (см. гл. IV и V), является тип плотности распределения вероятностей случайной величины $t'_{\alpha^*} = y - F(x, \alpha^*)$ (например, задается нормальный закон, или закон Лапласа). Необходимость задания такой априорной

информации является значительно более сильным требованием, чем задание оценки относительной величины дисперсии потерь.

В самом деле, гипотеза о том, что $\tau_{\text{отн}} < 2{,}5$, допускает возможность нормального закона, равномерного закона, закона Лапласа и многих других, в то время как гипотеза о конкретном виде распределения позволяет получать гарантированные результаты только при этом фиксированном типе распределения.

§ 4. Два механизма минимизации среднего риска

В этом параграфе мы будем полагать, что нам известна абсолютная оценка величины возможных потерь

$$\sup_{z,\,\alpha} Q(z,\ \alpha) = \tau_{\text{абс}}.$$

Наша цель состоит в том, чтобы по случайной независимой выборке

$$z_1,\ \ldots,\ z_l \tag{2.19}$$

сконструировать такой эмпирический функционал

$$I_{\text{э}}(\alpha) = \Phi(Q(z_1,\ \alpha),\ \ldots,\ Q(z_l,\ \alpha);\ z_1,\ \ldots,\ z_l),$$

точка минимума которого $\alpha = \alpha^*$ с заданной вероятностью $1-\eta$ доставляет функционалу среднего риска

$$I(\alpha) = \int Q(z,\ \alpha)\, P(z)\, dz \tag{2.20}$$

значение, близкое к минимальному.

Существует «естественный» способ построения такого эмпирического функционала. Надо восстановить по выборке (2.19) плотность распределения вероятностей $\hat{P}(z)$, а затем подставить в (2.20) восстановленную плотность $\hat{P}(z$ вместо $P(z)$.

Полученный таким образом функционал не зависит от неизвестной плотности и принципиально может быть минимизирован.

Казалось бы, проблема минимизации среднего риска по эмпирическим данным сводится к восстановлению плотности распределения вероятностей. Задача же восстановления по случайной независимой выборке плотности распределения вероятностей является центральной в матема-

тической статистике, и, таким образом, решение одной из частных проблем статистики — минимизация среднего риска по эмпирическим данным — ставится в зависимость от решения ее центральной проблемы.

В следующем параграфе мы подробно рассмотрим постановку задачи о восстановлении плотности распределения вероятностей, цель же этого параграфа — установить, что существуют два различных механизма, позволяющих решать задачу минимизации среднего риска по эмпирическим данным. Один из этих механизмов действительно опирается на то, что восстанавливаемая плотность $\hat{P}(z)$ приближается к истинной, в то время как другой механизм имеет совершенно иную теоретическую основу.

Итак, пусть $0 \leqslant Q(z, \alpha) \leqslant \tau$. Рассмотрим два типа эмпирических функционалов: эмпирический функционал типа

$$I'_{\text{э}}(\alpha) = \int Q(z, \alpha)\, \hat{P}(z)\, dz, \tag{2.21}$$

где $\hat{P}(z)$ — эмпирическая плотность, восстановленная по выборке $z_1, \ldots, z_l$, и эмпирический функционал

$$I_{\text{э}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha). \tag{2.22}$$

Эмпирический функционал (2.22) принято называть *функционалом эмпирического риска*.

Формально функционал эмпирического риска является частным видом эмпирического функционала (2.21). В самом деле, если в качестве аппроксимирующей плотности в (2.21) использовать плотность

$$\hat{\hat{P}}_{\varepsilon}(z) = \frac{1}{l} \sum_{i=1}^{l} \pi_{\varepsilon}(z - z_i), \tag{2.23}$$

где

$$\pi_{\varepsilon}(z) = \frac{1}{(\sqrt{2\pi}\,\varepsilon)^n} \exp\left\{-\frac{z^{\mathrm{T}} z}{2\varepsilon^2}\right\}$$

(n — размерность вектора z), то при $\varepsilon \to 0$ окажется, что $I'_{\text{э}}(\alpha) \to I_{\text{э}}(\alpha)$. (Здесь использован факт $\lim\limits_{\varepsilon \to 0} \pi_{\varepsilon}(z) = \delta(z)$).

Однако имеет смысл выделять функционал (2.22), так как успех в минимизации среднего риска путем миними-

зации функционалов (2.21) и (2.22) может определяться разными причинами. В первом случае успех может быть обеспечен за счет близости восстановленной плотности к истинной, тогда как во втором случае плотность $\hat{\hat{P}}_\varepsilon(z)$ при малых ε не приближается к $P(z)$, и тем не менее возможны условия, когда точка минимума функционала эмпирического риска доставляет функционалу (2.20) значение, близкое к минимальному.

Действительно, пусть плотность $\hat{P}(z)$ близка к $P(z)$, т. е.

$$\int |P(z) - \hat{P}(z)|\, dz \leqslant \varepsilon,$$

и пусть минимум эмпирического функционала достигается при $\alpha = \alpha_э$, а минимум среднего риска при $\alpha = \alpha_0$. Тогда справедлива цепочка неравенств

$$I(\alpha_э) - I(\alpha_0) \leqslant I(\alpha_э) - I'_э(\alpha_э) + I'_э(\alpha_0) - I(\alpha_0) \leqslant$$
$$\leqslant \int Q(z, \alpha_э)\,|P(z) - \hat{P}(z)|\, dz + \int Q(z, \alpha_0)\,|P(z) - \hat{P}(z)|\, dz \leqslant$$
$$\leqslant 2\tau\varepsilon,$$

откуда следует близость минимумов функционалов (2.20) и (2.21).

Покажем теперь, что аппроксимирующая плотность (2.23) при $\varepsilon \to 0$ не приближается к истинной. Пусть $P(z)$ — ограниченная функция. Разобьем множество Z на два подмножества: множество $\bar{Z}$ малой меры, содержащее все элементы выборки, и множество $Z/\bar{Z}$.

Нетрудно проверить, что для достаточно малого ε может быть выбрано такое множество $\bar{Z}$, что

$$\int_Z |P(z) - \hat{\hat{P}}_\varepsilon(z)|\, dz \approx \int_{Z/\bar{Z}} P(z)\, dz + \int_{\bar{Z}} \hat{\hat{P}}_\varepsilon(z)\, dz \simeq 2.$$

Таким образом, успех минимизации среднего риска (2.20) методом минимизации функционала эмпирического риска (2.22) определяется не близостью плотностей, как в первом случае, а иным механизмом. Ниже в § 6 мы покажем, что этот механизм опирается на свойство равномерной сходимости эмпирических средних к математическим ожиданиям по некоторому множеству событий.

§ 5. Задача восстановления плотности распределения вероятностей

Задачи, которые решают теория вероятностей и математическая статистика соотносятся между собой как прямые и обратные.

Задачи теории вероятностей можно было бы описать следующей схемой: известен состав генеральной совокупности и закон распределения вероятностей. Требуется для заданной схемы эксперимента оценить вероятность исходов эксперимента.

Математическая статистика решает обратные задачи: по результату эксперимента определяет свойства закона распределения. Исчерпывающей характеристикой закона распределения является плотность распределения вероятностей.

Таким образом, задача восстановления плотности распределения вероятностей по выборке является центральной проблемой математической статистики. В этом параграфе мы убедимся, что задача восстановления плотности является, вообще говоря, некорректно поставленной.

Пусть задана выборка $t_1, \ldots, t_l$ и достаточно широко определен класс функций, которому принадлежит плотность распределения вероятностей $P(t)$ (например, известно лишь, что $P(t)$ принадлежит непрерывным функциям). Требуется восстановить плотность распределения вероятностей.

Рассмотрим сначала одномерный случай. Согласно определению плотность распределения вероятностей $P(t)$ связана с функцией распределения вероятностей $F(z) = P\{t \leqslant z\}$ интегральным соотношением

$$\int_{-\infty}^{z} P(t)\,dt = F(z).$$

Или, что то же самое, соотношением

$$\int_{-\infty}^{\infty} \theta(z-t)\,P(t)\,dt = F(z), \tag{2.24}$$

где обозначено

$$\theta(x) = \begin{cases} 1, & \text{если } x \geqslant 0, \\ 0, & \text{если } x < 0. \end{cases}$$

Для непрерывных плотностей существует единственное решение уравнения (2.24).

Определим теперь эмпирическую функцию распределения вероятностей: $F_l(z) = k/l$, если величина z превосходит k элементов выборки $z_1, \ldots, z_l$.

Центральная теорема математической статистики — теорема Гливенко — Кантелли — утверждает, что с ростом объема выборки l эмпирическая функция распределения равномерно приближается к истинной.

Теорема (Гливенко — Кантелли). *Пусть $F(z)$ — функция распределения случайной величины z, $F_l(z)$ — эмпирическая функция распределения. Тогда при $l \to \infty$ справедливо*

$$P\left\{\sup_z |F(z) - F_l(z)| \xrightarrow[l\to\infty]{} 0\right\} = 1.$$

Мы не будем приводить здесь доказательство этой теоремы. В главе VI будет доказана теорема о равномерной сходимости частот появления событий к их вероятностям, из которой теорема Гливенко — Кантелли следует как частный случай.

Вернемся к интегральному уравнению (2.24), решение которого определяет плотность распределения вероятностей. Будем искать приближенное решение этого уравнения в ситуации, когда вместо функции распределения случайной величины $F(z)$ известна эмпирическая функция $F_l(z)$, найденная по конечной выборке. В главе IX, используя оценку скорости равномерной сходимости $F_l(z)$ к $F(z)$, мы покажем, что существует такая процедура получения приближенных решений уравнения (2.24), при которой с ростом l последовательность решений стремится к искомой плотности вероятностей.

Таким образом, существует принципиальная возможность восстанавливать непрерывную плотность распределения вероятностей. Однако восстановление плотности связано с решением некорректно поставленной задачи численного дифференцирования (2.24) в условиях, когда правая часть уравнения задана неточно.

Правда, при восстановлении плотности распределения вероятностей заранее известно, что решением интегрального уравнения (2.24) окажется не любая непрерывная функция, а функция $P(t)$, принимающая лишь

неотрицательные значения и удовлетворяющая условию

$$\int_{-\infty}^{\infty} P(t)\,dt = 1.$$

Однако этой априорной информации недостаточно, чтобы задача решения интегрального уравнения (2.24) перестала быть некорректно поставленной.

Подобно одномерному случаю может быть поставлена задача восстановления многомерной плотности распределения вероятностей. Для этого также выпишем интегральное уравнение, связывающее многомерную плотность с многомерной функцией распределения вероятностей:

$$\int_{-\infty}^{z^1}\ldots\int_{-\infty}^{z^n} P(t^1, \ldots, t^n)\,dt^1, \ldots, dt^n = P(t^1 \leqslant z^1; \ldots; t^n \leqslant z^n), \quad (2.25)$$

и определим многомерную эмпирическую функцию распределения

$$F_l(z^1, \ldots, z^n) = \frac{k}{l}, \quad (2.26)$$

где k — число элементов выборки $z_1, \ldots, z_l$, попадающих в область $t^1 \leqslant z^1, \ldots, t^n \leqslant z^n$.

Оказывается, что справедлив многомерный аналог теоремы Гливенко — Кантелли: с ростом объема выборки эмпирическая функция распределения равномерно сходится к функции распределения вероятностей. Справедливость обобщенной теоремы Гливенко — Кантелли также будет следовать из общей теории равномерной сходимости частот к вероятностям, рассмотренной в главе VI.

С помощью этой теоремы аналогично одномерному случаю устанавливается принципиальная возможность восстановления многомерных плотностей по эмпирическим данным.

Таким образом, задача восстановления плотности распределения вероятностей в классе непрерывных функций сводится к некорректной задаче численного дифференцирования функции распределения вероятностей [1]).

Заметим, что приведенная здесь постановка задачи численного дифференцирования отличается от задачи численного дифференцирования, рассмотренной в примере 3

[1]) Существуют непараметрические методы восстановления плотности (например, метод Парзена), которые, казалось бы, позволяют избежать необходимости решать некорректно поставленные задачи. Однако, как будет показано в главе IX, проблемы, которые возникают при реализации этих методов, оказываются эквивалентными проблемам решения некорректной задачи численного дифференцирования.

главы I. В главе I рассматривались некорректные задачи измерения, т. е. такие постановки некорректных задач, у которых ошибки являлись результатом измерения — значения правой части интегрального уравнения (2.24) определялись в l точках статистически независимо. В нашем же случае разность между точным значением правой части и функцией, полученной в результате измерения, является случайной функцией.

Таким образом, задача восстановления плотности распределения вероятностей является задачей более общей, чем интерпретация результатов косвенных экспериментов. И, следовательно, решать задачу минимизации среднего риска по эмпирическим данным путем восстановления плотности распределения вероятностей, вообще говоря, нерационально. (Наоборот, в главе IX мы рассмотрим задачу восстановления плотности как проблему минимизации среднего риска по эмпирическим данным).

Однако возможны вырожденные случаи, когда имеется настолько большая априорная информация об искомой плотности распределения вероятностей, что задача перестает быть некорректно поставленной.

Так, задача восстановления плотности может оказаться корректно поставленной, если плотность известна с точностью до конечного числа параметров (здесь важно, что число параметров конечно и заранее известно).

Методы восстановления плотности распределения вероятностей, заданной с точностью до конечного числа параметров, получили название *методов параметрической статистики*. Они образуют специальный класс методов восстановления плотности, который существенно отличается от общих методов восстановления плотности распределения вероятностей (иногда их называют *методами непараметрической статистики*).

§ 6. Равномерная близость эмпирических средних к математическим ожиданиям

Выше мы установили, что существуют два механизма минимизации среднего риска по эмпирическим данным.

Первый из них связан с минимизацией эмпирического функционала, построенного по восстановленной плотности. Однако промежуточная задача — восстановление плот-

ности — является, вообще говоря, более сложной, чем задача минимизации риска по эмпирическим данным.

Поэтому решать задачу минимизации среднего риска путем восстановления плотности, вообще говоря, нерационально.

В этом параграфе мы рассмотрим второй механизм минимизации среднего риска по эмпирическим данным. Будем решать задачу минимизации среднего риска

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

по эмпирическим данным

$$z_1, \ldots, z_l$$

путем минимизации функционала эмпирического риска

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha).$$

Для каждого фиксированного $\alpha = \alpha^*$ функционал $I(\alpha^*)$ определяет математическое ожидание случайной величины $t_{\alpha^*} = Q(z, \alpha^*)$, в то время как функционал $I_э(\alpha^*)$ — эмпирическое среднее этой случайной величины.

Согласно классическим теоремам теории вероятностей в достаточно общих случаях эмпирическое среднее случайной величины t_{α^*} с ростом l сходится к математическому ожиданию этой случайной величины.

Однако из этих теорем никак не следует, что значение параметра $\alpha_э$, доставляющего минимум эмпирическому риску $I_э(\alpha)$, будет доставлять среднему риску $I(\alpha)$ величину, близкую к минимальной. Это утверждение является важным, и потому разберем его подробнее.

Предположим для наглядности, что параметр α есть скалярная величина, лежащая в интервале $[0, 1]$. Каждому α ставится в соответствие величина $I(\alpha)$. Рассмотрим функцию $I(\alpha)$. Наряду с этой функцией рассмотрим функцию $I_э(\alpha)$, которая для каждого α определяет эмпирическое среднее, найденное по выборке длины l (рис. 2).

Метод минимизации эмпирического риска предлагает по минимуму функции $I_э(\alpha)$ судить о минимуме функции $I(\alpha)$. Для того же, чтобы по точке минимума и минимальному значению функции $I_э(\alpha)$ можно было судить о мини-

мальном значении функции $I(\alpha)$, достаточно, чтобы кривая $I_э(\alpha)$ целиком находилась внутри $\varkappa$-трубки кривой $I(\alpha)$. Выброс хотя бы в одной точке (как на рис. 2) может привести к тому, что в качестве точки, минимизирующей $I(\alpha)$, будет выбрана точка выброса $I_э(\alpha)$. В этом случае минимум $I_э(\alpha)$ никак не характеризует минимум $I(\alpha)$. Если же функция $I_э(\alpha)$ приближает $I(\alpha)$ равномерно по α с точностью $\varkappa$, то минимум $I_э(\alpha)$ отстоит от минимума $I(\alpha)$ на величину, не превосходящую $2\varkappa$.

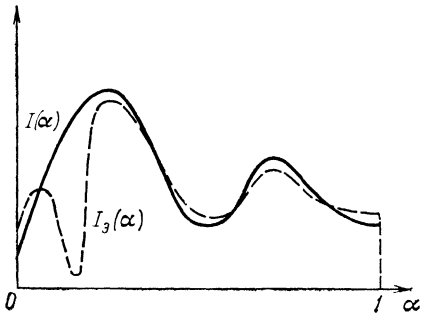

Рис. 2.

Формально это означает, что нас интересуют не классические условия, когда для любого α и $\varkappa$ справедливо

$$P\{|I(\alpha)-I_э(\alpha)|>\varkappa\}\underset{l\to\infty}{\to}0, \tag{2.27}$$

а более сильные условия, когда для любого $\varkappa$ справедливо

$$P\left\{\sup_{\alpha}|I(\alpha)-I_э(\alpha)|>\varkappa\right\}\underset{l\to\infty}{\to}0. \tag{2.28}$$

В случае, когда выполняется (2.28), будем говорить, что имеет место *равномерная по параметру α сходимость эмпирических средних к их математическим ожиданиям.*

Итак, второй механизм минимизации риска связан с равномерной по параметру α сходимостью эмпирических средних к математическим ожиданиям. Однако для наших целей — минимизации среднего риска на выборках фиксированного объема — просто факта равномерной сходимости недостаточно. Для того чтобы с заданной вероятно-

стью можно было гарантировать отыскание решения, доставляющего функционалу значение, близкое к минимальному, надо, чтобы была известна оценка скорости равномерной сходимости. Действительно, выполнение неравенства

$$P\left\{\sup_{\alpha} |I(\alpha) - I_э(\alpha)| \geqslant \varkappa\right\} < \eta(l, \varkappa),$$

$$\lim_{l\to\infty} \eta(l, \varkappa) = 0$$

эквивалентно утверждению: с вероятностью $1-\eta(l, \varkappa)$ одновременно для всех α справедлива оценка

$$I_э(\alpha) - \varkappa \leqslant I(\alpha) \leqslant I_э(\alpha) + \varkappa. \tag{2.29}$$

И если $\eta(l, \varkappa)$ — убывающая по l и $\varkappa$ функция, то для заданного уровня надежности $1-\eta$,

$$\eta(l, \varkappa) = \eta, \tag{2.30}$$

величина доверительного интервала $\varkappa = \varkappa(l, \eta)$, полученная как решение уравнения (2.30), уменьшается с ростом l. Следовательно, для больших l точка $\alpha_э$ минимума эмпирического риска доставит величине среднего риска значение, близкое к минимальному. При любом же фиксированном l можно утверждать, что с вероятностью $1-\eta$ точка $\alpha_э$ доставит величине среднего риска значение из интервала

$$I_э(\alpha_э) - \varkappa \leqslant I(\alpha_э) \leqslant I_э(\alpha_э) + \varkappa.$$

§ 7. Обобщение теоремы Гливенко — Кантелли и задача распознавания образов

В этом параграфе мы рассмотрим частный случай: функция потерь $Q(z, \alpha)$ функционала

$$I(\alpha) = \int Q(z, \alpha)\, P(z)\, dz \tag{2.31}$$

принимает лишь два значения — нуль и единица. Как уже отмечалось, к этому случаю приводится задача обучения распознаванию образов.

Обозначим через $S(\alpha^*)$ множество векторов z, для которых данная функция потерь $Q(z, \alpha^*)$ принимает значение единица. Иными словами, $S(\alpha^*)$ есть событие $S(\alpha^*) = \{z\colon Q(z, \alpha^*) = 1\}$. Для фиксированного $\alpha = \alpha^*$ функционал (2.31) определяет вероятность того, что вектор z принадлежит множеству $S(\alpha^*)$, т. е. вероятность события $S(\alpha^*)$.

Соответственно для каждого фиксированного $\alpha = \alpha^*$ функционал эмпирического риска

$$I_э(\alpha^*) = \frac{1}{l} \sum_{i=1}^{l} Q(z_l, \alpha^*) \tag{2.32}$$

определяет частоту события $S(\alpha^*)$, найденную по выборке $z_1, \ldots, z_l$ длины l. Для того чтобы выделить этот важный частный случай, будем обозначать функционал (2.31) через $P(\alpha)$, а функционал (2.32) — через $\nu(\alpha)$. В этих обозначениях условие (2.28) перепишется в виде

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} \xrightarrow[l\to\infty]{} 0.$$

Оно означает равномерную сходимость частот появления событий к их вероятностям по классу событий $S(\alpha)$. В этих терминах утверждение теоремы Гливенко — Кантелли о том, что эмпирическая функция распределения вероятностей равномерно сходится к истинной функции, есть утверждение о существовании равномерной сходимости частот появления событий к их вероятностям для одной специальной системы событий.

В самом деле, рассмотрим прямую z и множество лучей $z \leqslant \alpha$. Это множество лучей задает систему событий $S^1(\alpha)$ (событие $S^1(\alpha^*)$ заключается в том, что точка z принадлежит лучу $z \leqslant \alpha^*$). В этих терминах утверждение теоремы Гливенко — Кантелли состоит в следующем: «имеет место равномерная сходимость частот появления событий к их вероятностям по классу событий $S^1(\alpha)$».

Рассмотрим теперь следующий класс событий $S^n(\alpha)$: вектор $z = (z^1, \ldots, z^n)^T$ принадлежит событию $S^n(\alpha^*)$ (здесь $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T$), если одновременно для всех n координат выполнятся неравенства $z^1 \leqslant \alpha_1^*, \ldots, z^n \leqslant \alpha_n^*$. Множество всех событий $S^n(\alpha^*)$ и есть $S^n(\alpha)$. В этих терминах многомерным аналогом теоремы Гливенко — Кантелли является утверждение о существовании равномерной сходимости частот появления событий к их вероятностям по классу событий $S^{(n)}(\alpha)$.

Таким образом, требование равномерной сходимости частот появления событий к их вероятностям для различных систем событий, которое возникает при исследовании задачи обучения распознаванию образов, приводит к необходимости обобщения теоремы Гливенко — Кантелли.

§ 8. Замечания о двух механизмах минимизации среднего риска по эмпирическим данным

Итак, существуют два механизма минимизации среднего риска по эмпирическим данным. Один из них связан с возможностью восстановить плотность распределения вероятностей, другой — с возможностью обеспечить равномерную сходимость эмпирических средних к математическим ожиданиям.

Восстанавливать плотность распределения вероятностей рационально лишь в вырожденных случаях, а именно тогда, когда о плотности имеется достаточно большая априорная информация. Если же априорная информация ограничена, решение промежуточной задачи — восстановление плотности оказывается никак не проще проблемы минимизации среднего риска. В этом случае возможность восстановления плотности распределения вероятностей опирается на теорему Гливенко — Кантелли, т. е. на существование равномерной сходимости частот появления событий к их вероятностям для специального класса событий.

Второй механизм минимизации риска непосредственно опирается на существование равномерной сходимости эмпирических средних к их математическим ожиданиям.

Таким образом, вопрос о равномерной сходимости эмпирических средних к математическим ожиданиям оказался центральным в теории минимизации среднего риска.

Ниже в главах VI — VII будет показано, что достаточные условия существования равномерной сходимости средних к математическим ожиданиям определяются особенностью функций потерь. Для задач восстановления зависимостей эта особенность выразится в том, что класс функции, в котором ведется восстановление, должен быть достаточно узким.

Существование двух механизмов минимизации среднего риска отражает наличие условий двух типов, при которых в принципе возможна минимизация среднего риска по эмпирическим данным.

Условия первого типа связывают возможность минимизации риска с информацией о классе плотностей, которому принадлежит восстанавливаемая плотность распределения вероятностей. В том случае, когда удается ее восстановить, оказывается, что независимо от того, какова функция потерь (лишь бы она не допускала больших выбросов), можно добиться успеха в минимизации среднего риска.

Условия второго типа накладывают определенные ограничения на свойства функций потерь, и тогда независимо от того, какова плотность $P(z)$, можно добиться успеха в минимизации среднего риска.

При решении задачи восстановления зависимостей по эмпирическим данным в условиях, когда функция потерь

не допускает больших выбросов, разница в этих двух подходах скажется на схемах возможных утверждений.

Схема утверждений первого типа. *Если природа задач угадана хорошо (найден «узкий» класс плотностей $\{P(z)\}$, которому принадлежит искомая плотность), то независимо от особенностей класса функций, в котором ведется восстановление, минимум эмпирического функционала будет близок к минимуму среднего риска.*

Схема утверждений второго типа. *Если восстановление ведется в достаточно «узком» классе функций $F(x, \alpha)$, то независимо от того, какова природа задач (какова плотность $P(z)$), минимум эмпирического риска будет близок к минимуму среднего риска.*

Следует заметить, что с формальной точки зрения существует определенное предпочтение в использовании алгоритмов, относительно которых возможны утверждения второго типа. В самом деле, в утверждении первого типа содержится два требования. Надо, чтобы:

1) класс плотностей, в котором ведется восстановление, был достаточно узким;

2) искомая плотность принадлежала этому классу.

Во втором утверждении содержится лишь одно требование. Надо, чтобы класс функций, в котором ведется восстановление, был достаточно узким. На практике широту класса плотностей, так же, как и широту класса функций, в котором ведется восстановление, нетрудно регулировать.

Вопрос о том, принадлежит ли восстанавливаемая плотность заданному классу, всегда остается открытым.

Основное содержание этой книги составляет отыскание условий равномерной сходимости и использование их для восстановления зависимостей по выборкам ограниченного объема. Используя оценки скорости равномерной сходимости средних к математическим ожиданиям, удается не только обосновать метод минимизации эмпирического риска, но и построить новый метод минимизации риска (метод упорядоченной минимизации), позволяющий в условиях ограниченного объема эмпирических данных находить решение, которое доставляет среднему риску наименьшее гарантированное значение.

Рассмотрению методов минимизации риска, основанных на использовании механизма равномерной сходимости, посвящены главы VI — X. Однако, прежде чем перейти к систематическому изучению этого механизма, мы рассмотрим классические методы минимизации риска, основанные на идее минимизации функционала, построенного с помощью восстановленной плотности. Как уже указывалось выше, в том исключительном случае, когда плотность известна с точностью до параметров, задача восстановления может оказаться устойчивой, и ее решение, а вместе с ней и решение задачи восстановления зависимостей по эмпирическим данным, может быть успешно проведено методами параметрической статистики. В главе III мы рассмотрим применение методов параметрической статистики к решению задачи обучения распознаванию образов, а в главах IV и V — к задаче восстановления регрессии.

Основные утверждения главы II

1. Решение задачи минимизации среднего риска

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

по ограниченному множеству эмпирических данных

$$z_1, \ldots, z_l$$

связано с аппроксимацией функционала $I(\alpha)$ *эмпирическим функционалом*

$$I_э(\alpha) = \Phi(Q(z_1, \alpha), \ldots, Q(z_l, \alpha); z_1, \ldots, z_l)$$

и последующей его минимизацией.

2. Однако получение гарантий того, что функция, минимизирующая $I_э(\alpha)$*, будет доставлять величине среднего риска* $I(\alpha)$ *значение, близкое к минимальному, возможно лишь при наличии определенной априорной информации.*

В качестве такой априорной информации может быть использована либо абсолютная оценка величины потерь

$$\sup_{\alpha, z} Q(z, \alpha) = \tau_{абс},$$

либо оценка относительной величины дисперсии потерь

$$\sup_{\alpha} \sqrt{\frac{MQ^2(z, \alpha)}{(MQ(z, \alpha)^2} - 1} = \tau_{\text{отн}}.$$

В задаче распознавания образов $\tau_{\text{абс}} = 1$. *В задачах же восстановления регрессии и интерпретации результатов косвенных экспериментов используются сведения об относительной оценке* $\tau_{\text{отн}}$. *Эта информация значительно меньше, чем обычно принятая (о виде распределения случайных величин* $t_\alpha = Q(z, \alpha)$, *индуцированных плотностью распределения вероятностей* $P(z)$).

3. Конструирование эмпирического функционала $I_э(\alpha)$ *основано на аппроксимации плотности* $P(z)$, *входящей в функционал среднего риска* $I(\alpha)$ *эмпирической плотностью* $\hat{P}(z)$ *и использовании* $\hat{P}(z)$ *вместо* $P(z)$:

$$I'_э(\alpha) = \int Q(z, \alpha) \hat{P}(z)\, dz.$$

Существуют две различные причины, в силу которых минимум $I'_э(\alpha)$ *оказывается близким к минимуму* $I(\alpha)$.

Первая причина состоит в том, что плотность $\hat{P}(z)$ *близка к* $P(z)$. *Вторая — в том, что имеет место равномерная сходимость эмпирических средних к математическим ожиданиям:*

$$P\left\{\sup_{\alpha} |I(\alpha) - I'_э(\alpha)| > \varkappa\right\} \xrightarrow[l \to \infty]{} 0.$$

4. Восстановление плотности распределения вероятностей, вообще говоря, является некорректно поставленной задачей. Поэтому гарантировать успех восстановления плотности по выборке можно лишь, когда имеется достаточно большая априорная информация об искомой плотности. Например, когда плотность распределения вероятностей известна с точностью до параметров. Тогда можно, восстановив плотность, использовать функционал $I'_э(\alpha)$ *для отыскания минимума среднего риска. В иных случаях задача минимизации среднего риска связана с проблемой равномерной сходимости эмпирических средних к математическим ожиданиям.*

ГЛАВА III

МЕТОДЫ ПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ В ЗАДАЧЕ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ

§ 1. Параметрические методы в задаче распознавания образов

Пусть требуется минимизировать функционал

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \qquad (3.1)$$

в условиях, когда плотность распределения вероятностей $P(x, y)$ неизвестна, но зато дана выборка

$$x_1, y_1; \ldots; x_l, y_l, \qquad (3.2)$$

полученная в случайных независимых испытаниях согласно $P(x, y)$.

Будем решать эту задачу по следующей схеме:

1) восстановим по выборке (3.2) плотность $\hat{P}(x, y)$;

2) сконструируем с помощью восстановленной плотности функционал

$$I_э(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy; \qquad (3.3)$$

3) найдем минимум этого функционала и объявим функцию $F(x, \alpha_э)$, доставляющую минимум (3.3), решением исходной задачи минимизации (3.1).

Как указывалось в главе II, реализация этой схемы может привести к успеху лишь тогда, когда имеется значительная априорная информация о плотности $P(x, y)$, а именно, когда плотность распределения вероятностей известна с точностью до параметров.

Иначе говоря, успех возможен тогда, когда заранее известна «модель» восстанавливаемой плотности. «Модель» же искомых плотностей оказывается существенно разной для различных задач восстановления зависимостей.

В этой главе мы рассмотрим задачу обучения распознаванию образов. Для нее характерно, что неизвестная

плотность распределения вероятностей $P(x, \omega)$ [1]) может быть представлена как объединение двух плотностей $P(x \mid \omega = 0)$ и $P(x \mid \omega = 1)$, заданных на разных подпространствах X, 0 и X, 1:

$$P(x, \omega) = P(x \mid \omega = 0)\, P(\omega = 0)\,(1 - \omega) + \\ + P(x \mid \omega = 1)\, P(\omega = 1)\,\omega. \qquad (3.4)$$

Множество пар x, ω состоит из двух непересекающихся подпространств размерности n, а именно:

$$X \subset E_n, \quad \omega = 0 \quad \text{и} \quad X \subset E_n, \quad \omega = 1.$$

Формула (3.4) утверждает, что на первом подпространстве плотность равна $P(x \mid \omega = 0) \cdot P(\omega = 0)$, а на втором — $P(x \mid \omega = 1)\, P(\omega = 1)$. В формуле $P(x \mid \omega = 0)$ и $P(x \mid \omega = 1)$ — состав объединения; $P(\omega = 0)$, $P(\omega = 1) = 1 - P(\omega = 0)$ — пропорция объединения.

Пусть плотность $P(x, \omega)$ известна с точностью до конечного числа $m_1 + m_2 + 1$ параметров:

$$P(x, \omega) = P_\beta(x \mid \omega = 0)\, P(\omega = 0)\,(1 - \omega) + \\ + P_\gamma(x \mid \omega = 1)\, P(\omega = 1)\,\omega, \qquad (3.5)$$

где β — неизвестный m_1-мерный вектор параметров плотности $P_\beta(x \mid \omega = 0)$; γ — неизвестный m_2-мерный вектор параметров плотности $P_\gamma(x \mid \omega = 1)$; $P(\omega = 0)$ — скалярный параметр.

Теперь, для того чтобы реализовать нашу схему, необходимо уметь решать две задачи:

1) находить для заданной плотности распределения вероятностей $P(x, \omega)$ минимум функционала (3.3);

2) восстанавливать по выборке (3.2) плотность распределения вероятностей $P(x, \omega)$.

Первая задача называется в статистике *задачей дискриминантного анализа*, вторая задача — *задачей восстановления плотности распределения вероятностей в параметрическом классе функции*. Рассмотрим обе эти задачи.

---

[1]) Чтобы подчеркнуть, что y принимает только два значения — нуль и единица, здесь используется обозначение $y = \omega$.

§ 2. Задача дискриминантного анализа

Итак, пусть требуется найти минимум функционала (3.3) для заданной плотности распределения вероятностей (заданных состава объединения $P(x \mid \omega = 0)$, $P(x \mid \omega = 1)$ и пропорции объединения $P(\omega = 0)$, $P(\omega = 1) = 1 - P(\omega = 0)$).

Рассмотрим сначала простой случай: класс возможных решающих правил $F(x, \alpha)$ никак не ограничен. В этой ситуации легко может быть построено решающее правило, минимизирующее функционал (3.3).

В самом деле, согласно формуле Байеса вероятность того, что вектор x принадлежит первому (второму) классу, определяется так:

$$\left(P(\omega = 0 \mid x) = \frac{P(x \mid \omega = 0)\, P(\omega = 0)}{P(x \mid \omega = 0)\, P(\omega = 0) + P(x \mid \omega = 1)\, P(\omega = 1)},\right.$$
$$\left.P(\omega = 1 \mid x) = \frac{P(x \mid \omega = 1)\,(1 - P(\omega = 0))}{P(x \mid \omega = 0)\, P(\omega = 0) + P(x \mid \omega = 1)\, P(\omega = 1)}\right). \tag{3.6}$$

Минимальные потери (минимум вероятности ошибки) будут получены при такой классификации, при которой вектор x относят к первому классу, если более вероятной оказывается его принадлежность к первому классу, чем ко второму, т. е. если

$$P(\omega = 0 \mid x) > P(\omega = 1 \mid x).$$

В противном случае вектор x относят ко второму классу.

Иначе говоря, учитывая (3.6), вектор x должен быть отнесен к первому классу, если выполнится неравенство

$$\frac{P(x \mid \omega = 1)}{P(x \mid \omega = 0)} < \frac{P(\omega = 0)}{1 - P(\omega = 0)},$$

или, что то же самое, оптимальная классификация векторов осуществляется с помощью характеристической функции

$$F(x) = \theta\left[\ln P(x \mid \omega = 1) - \ln P(x \mid \omega = 0) + \ln \frac{1 - P(\omega = 0)}{P(\omega = 0)}\right], \tag{3.7}$$

где

$$\theta(z) = \begin{cases} 1, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0. \end{cases}$$

Таким образом, знание плотности распределения вероятностей (состава и пропорции объединения (3.5)) позво-

ляет немедленно построить оптимальное решающее правило.

Однако задача отыскания оптимального решающего правила значительно усложняется, если класс возможных решающих правил $F(x, \alpha)$ ограничен. В частности, трудной оказывается задача отыскания оптимального линейного решающего правила, т. е. правила вида

$$F(x, \alpha) = \theta[\alpha^{\mathrm{T}} x + \alpha_0]. \tag{3.8}$$

Вектор $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ определяет направление линейной дискриминантной функции, а параметр α_0 — пороговое значение. Задача отыскания минимума (3.3) в классе (3.8) получила название задачи *линейного дискриминантного анализа*.

В 30-х годах Р. Фишер предложил в качестве направления линейной дискриминантной функции выбирать направление, на котором достигается максимум величины относительного расстояния между математическими ожиданиями проекций векторов различных классов, т. е. направление α, на котором достигается максимум величины

$$T(\alpha) = \frac{(m_1(\alpha) - m_2(\alpha))^2}{\sigma_1^2(\alpha) + \sigma_2^2(\alpha)}, \tag{3.9}$$

где

$$m_1(\alpha) = \int \alpha^{\mathrm{T}} x P_\beta(x \mid \omega = 0)\, dx,$$

$$m_2(\alpha) = \int \alpha^{\mathrm{T}} x P_\gamma(x \mid \omega = 1)\, dx,$$

$$\sigma_1^2(\alpha) = \int (\alpha^{\mathrm{T}} x - m_1(\alpha))^2 P_\beta(x \mid \omega = 0)\, dx,$$

$$\sigma_2^2(\alpha) = \int (\alpha^{\mathrm{T}} x - m_2(\alpha))^2 P_\gamma(x \mid \omega = 1)\, dx,$$

$$\alpha^{\mathrm{T}} \alpha = 1.$$

Отыскание максимума (3.9) для произвольных плотностей — задача чрезвычайно трудная. Поэтому основные исследования в области линейного дискриминантного анализа были направлены на то, чтобы установить для определенных типов плотностей, что, во-первых, линейная дискриминантная функция Фишера действительно определяет решение задачи линейного дискриминантного анализа, а во-вторых, найти алгоритмы вычисления дискриминантной функции. Основной результат здесь заключается в том, что для объединения двух нормальных законов

$$P(x \mid \omega = 0) = N(\mu_1, \Delta_1), \qquad P(x \mid \omega = 1) = N(\mu_2, \Delta_2)$$

(μ_1 — вектор средних, Δ_1 — матрица ковариации для первого многомерного нормального закона; μ_2, Δ_2 — аналогичные элементы для второго закона), взятых в пропорции $P(\omega = 0)$ и $(1 - P(\omega = 0))$, оптимальная линейная дискриминантная функция задается вектором

направления

$$\alpha_{t^*}=(\mu_1-\mu_2)^{\mathrm{T}}\left(t^*\Delta_1+(1-t^*)\,\Delta_2\right)^{-1}, \tag{3.10}$$

где $0\leqslant t^*\leqslant 1$. Значение t^* определяется из условия обращения в нуль так называемой *резольвентной функции*

$$f(t)=t\sigma_1^2(\alpha_t)+(1-t)\,\sigma_2^2(\alpha_t)-\ln\left(\frac{P(\omega=0)}{1-P(\omega=0)}\cdot\frac{\sigma_2^2(\alpha_t)}{\sigma_1^2(\alpha_t)}\right). \tag{3.11}$$

При $P(\omega=0)=1/2$ направление (3.10) линейной дискриминантной функции максимизирует функционал

$$I(\alpha)=\frac{(m_1(\alpha)-m_2(\alpha))^2}{t^*\sigma_1^2(\alpha)+(1-t^*)\,\sigma_2^2(\alpha)}.$$

Вычисление нулей резольвентного уравнения (3.8) — задача достаточно трудная. Поэтому на практике при построении линейной дискриминантной функции полагают $t^*=1/2$. Тем самым в качестве решения задачи принимается линейная дискриминантная функция Фишера. (Подробнее смотри [71].)

Таким образом, проблемы, которые возникают в дискриминантном анализе, связаны с тем, что класс возможных решающих правил, на котором ищется минимум функционала (3.3), ограничен. Поэтому может показаться, что задача дискриминантного анализа надумана. В самом деле, если уж удается восстановить плотность распределения вероятностей, то для чего отыскивать решающее правило, доставляющее функционалу условный минимум, когда легко можно найти решающее правило (см. (3.7)), доставляющее функционалу (3.3) абсолютный минимум?

Суть, однако, заключается в том, что если плотность восстанавливается неточно, то величина гарантированного уклонения минимума эмпирического функционала от минимума функционала среднего риска будет большей для функции, выбранной из более широкого класса. Поэтому может оказаться, что меньшее гарантированное значение среднего риска будет достигнуто не на функции, доставляющей абсолютный минимум эмпирическому функционалу, а на функции, принадлежащей более узкому классу и доставляющей условный минимум.

Такой результат связан с эффектом второго механизма минимизации среднего риска (см. § 4 гл. II). Идеи сужения класса решающих правил для получения меньшей гарантированной величины среднего риска будут реализованы ниже в главах VIII — IX. В этой же главе мы рассмотрим параметрические методы восстановления плот-

ностей. Согласно (3.7) знание плотности распределения вероятностей немедленно приводит к построению решающего правила, доставляющего абсолютный минимум (3.3).

§ 3. Решающие правила в задаче распознавания образов

Алгоритмы обучения распознаванию образов, основанные на восстановлении плотности распределения вероятностей (состава объединения $P(x \mid \omega = 0)$, $P(x \mid \omega = 1)$ и его пропорции $P(\omega = 0)$), традиционно связаны с двумя классами распределений.

Первый класс распределений. Распределение вероятностей $P_\omega(x) = P(x \mid \omega)$ таково, что координаты вектора $x = (x^1, \ldots, x^n)^T$ статистически независимы, т. е.

$$P_\omega(x) = P_\omega(x^1) \ldots P_\omega(x^n), \quad \omega = 0,1, \tag{3.12}$$

и, кроме того, каждая координата x^i вектора x может принимать лишь фиксированное число значений. Будем считать, что каждая координата x^i принимает τ_i значений $c^i(1), \ldots, c^i(\tau_i)$.

Таким образом, рассматривается случай, когда законы распределения случайных величин $P_{\omega=0}(x)$ и $P_{\omega=1}(x)$ заданы выражением (3.12), где функция $P_\omega(x^i)$ может быть записана так:

$$P_\omega(x^i) = \begin{cases} p^i_\omega(1), & \text{если } x^i = c^i(1), \\ \cdots\cdots\cdots\cdots\cdots \\ p^i_\omega(\tau_i), & \text{если } x^i = c^i(\tau_i), \end{cases} \tag{3.13}$$

$$\sum_{j=1}^{\tau_i} p^i_\omega(j) = 1.$$

Здесь $p^i_\omega(j)$ — вероятность того, что для вектора из класса $\omega = \{0, 1\}$ значение координаты x^i будет равно $c^i(j)$. Восстановить распределение вероятностей для такого объединения — значит найти значения

$$r = 2 \sum_{i=1}^{n} \tau_i + 1$$

параметров ($\sum_{i=1}^{n} \tau_i$ параметров для восстановления каждого закона $P_\omega(x)$ и один параметр — пропорцию объединения).

Согласно (3.7) оптимальным решающим правилом для объединений, образованных двумя законами (3.12), будет следующая линейная дискриминантная функция:

$$F(x) = \theta\left(\sum_{i=1}^{n} \ln \frac{P_{\omega=1}(x^i)}{P_{\omega=0}(x^i)} - \ln \frac{p}{1-p}\right),$$

где p, $(1-p)$ — пропорция объединения.

Второй класс распределений. Здесь в каждом классе $\omega = \{0, 1\}$ векторы x распределены согласно многомерному нормальному закону

$$P_\omega(x) = \frac{1}{(2\pi)^{n/2} |\Delta_\omega|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_\omega)^T \Delta_\omega^{-1}(x-\mu_\omega)\right\},$$

где μ_ω — вектор средних, Δ_ω — ковариационная матрица.

Из (3.7) следует, что оптимальным решающим правилом в этом случае оказывается квадратичная дискриминантная функция

$$F(x) = \theta\left[\frac{1}{2}(x-\mu_0)^T\Delta_0^{-1}(x-\mu_0) - \frac{1}{2}(x-\mu_1)^T\Delta_1^{-1}(x-\mu_1) + \right.$$
$$\left. + \ln\frac{|\Delta_0|}{|\Delta_1|} - \ln\frac{p}{1-p}\right], \qquad (3.14)$$

где μ_0, Δ_0; μ_1, Δ_1 — параметры нормальных законов, образующих объединение (3.5); p, $(1-p)$ — пропорция объединения.

В частном случае, когда $\Delta_0 = \Delta_1 = \Delta$ — квадратичная дискриминантная функция (3.14) вырождается в линейную

$$F(x) = \theta\left[(\mu_1-\mu_0)^T\Delta^{-1}x + \frac{1}{2}(\mu_0^T\Delta^{-1}\mu_0 - \mu_1^T\Delta^{-1}\mu_1) - \ln\frac{p}{1-p}\right].$$

§ 4. Об оценке качества алгоритмов восстановления плотности вероятностей

Итак, построение дискриминантной функции по эмпирическим данным сводится к восстановлению законов распределения $P(x \mid \omega = 0)$ и $P(x \mid \omega = 1)$ и оценке параметра p.

Параметр p определяет долю пар x, ω с $\omega = 0$ и может быть оценен величиной $\tilde{p} = m/l$, где m — число пар в выборке с $\omega = 0$, l — объем выборки [1]).

Какие же алгоритмы следует использовать для восстановления плотности распределения вероятностей $P(x \mid \omega = 0)$, $P(x \mid \omega = 1)$?

Для того чтобы ответить на этот вопрос, надо, прежде всего, договориться о том, как следует оценивать качество алгоритмов восстановления плотности на выборках ограниченного объема.

Качество фиксированного алгоритма A, восстанавливающего по выборке $x_1, \ldots, x_l$ плотность $P(x, \alpha_0)$, естественно определить как расстояние от этой плотности до восстановленной функции $P_A(x \mid x_1, \ldots, x_l)$, т. е. величиной

$$\rho(P(x, \alpha_0), P_A(x \mid x_1, \ldots, x_l)) = \rho_{\alpha_0, A}(x_1, \ldots, x_l).$$

Будем определять близость плотностей метрикой L^2, т. е.

$$\rho_{\alpha_0, A}(x_1, \ldots, x_l) = \left(\int (P(x, \alpha_0) - P_A(x \mid x_1, \ldots, x_l))^2 \, dx\right)^{1/2}. \qquad (3.15)$$

Поскольку выбор плотности $P_A(x \mid x_1, \ldots, x_l)$ зависит от состава выборки $x_1, \ldots, x_l$, величина $\rho_{\alpha_0, A}(x_1, \ldots, x_l)$ является случайной. Будем характеризовать качество алгоритма A математическим ожиданием величины $\rho^2_{\alpha_0, A}(x_1, \ldots, x_l)$:

$$R(\alpha_0, A) = \int \rho^2_{\alpha_0, A}(x_1, \ldots, x_l) P(x_1) \ldots P(x_l) \, dx_1, \ldots, dx_l.$$

Для восстановления плотности $P(x, \alpha_0)$ на выборках длины l тот алгоритм лучше, для которого величина $R(\alpha_0, A)$ меньше.

Итак, определено, как должно измеряться качество алгоритма A, предназначенного для восстановления фиксированной плотности $P(x, \alpha_0)$. Теперь следует договориться о том, как измерять качество алгоритма, предназначенного для восстановления любой плотности, принадлежащей заданному классу $P(x, \alpha)$ (в нашем слу-

---

[1]) В § 6 будет показано, что более точной оценкой является $p = \frac{m+1}{l+2}$.

чае класс плотностей задан с точностью до значения вектора параметров α).

В теории статистических решений в таких ситуациях используются два принципа:

— *принцип Байеса*,

— *принцип минимакса*.

Принцип Байеса состоит в том, чтобы оценивать качество алгоритма как среднее качество по множеству восстанавливаемых плотностей. Для того чтобы оценить среднее качество алгоритма, надо знать, как часто придется восстанавливать тот или иной закон из $P(x, \alpha)$, т. е., в нашем случае, знать плотность распределения вероятностей $P(\alpha)$ вектора параметров α. Тогда качество алгоритма определится так:

$$R_{\text{Б}}(A) = \int R(\alpha, A) P(\alpha)\, d\alpha. \tag{3.16}$$

Тот алгоритм A считается лучшим, для которого величина $R_{\text{Б}}(A)$ меньше.

Принцип минимакса состоит в том, чтобы оценивать качество алгоритма по наиболее неблагоприятному для данного алгоритма распределению $P(x, \alpha^*)$.

Здесь, напротив, совершенно не принимается во внимание то, какие плотности придется восстанавливать на практике. Поэтому может оказаться, что качество алгоритма определяет случай, который никогда не встретится.

Качество алгоритма, согласно принципу минимакса, определяется так:

$$R_{\text{м}}(A) = \sup_{\alpha} R(\alpha, A). \tag{3.17}$$

Тот алгоритм считается лучшим, для которого величина $R_{\text{м}}(A)$ меньше.

§ 5. Байесов алгоритм восстановления плотности

Определим структуру алгоритмов, обеспечивающих решение байесовой задачи восстановления плотности, т. е. минимизирующих функционал

$$R_{\text{Б}}(A) = \int R(\alpha, A) P(\alpha)\, d\alpha.$$

Пусть по выборке $x_1, \ldots, x_l$ восстанавливается плотность, принадлежащая классу $P(x, \alpha)$, и пусть дана априорная плотность вероятностей $P(\alpha)$.

Найдем с помощью формулы Байеса

$$P(\alpha \mid x_1, \ldots, x_l) = \frac{P(x_1, \ldots, x_l \mid \alpha)\, P(\alpha)}{P(x_1, \ldots, x_l)}$$

плотность распределения апостериорных вероятностей $P(\alpha \mid x_1, \ldots, x_l)$, характеризующую возможность реализации значений параметров α после того, как к априорной информации $P(\alpha)$ добавлена информация о выборке $x_1, \ldots, x_l$. Здесь $P(x_1, \ldots, x_l \mid \alpha)$ — условная, а $P(x_1, \ldots, x_l)$ — безусловная плотность вероятностей появления выборки $x_1, \ldots, x_l$:

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l \mid \alpha)\, P(\alpha)\, d\alpha.$$

Ниже мы покажем, что решением байесовой задачи является апостериорное среднее, т. е. функция

$$P_{\text{Б}}(x \mid x_1, \ldots, x_l) = \int P(x, \alpha)\, P(\alpha \mid x_1, \ldots, x_l)\, d\alpha. \quad (3.18)$$

Вообще говоря, полученная в результате усреднения функций $P(x, \alpha)$ по мере $P(\alpha \mid x_1, \ldots, x_l)$ плотность $P_{\text{Б}}(x \mid x_1, \ldots, x_l)$ вовсе не обязана принадлежать рассматриваемому параметрическому семейству $P(x, \alpha)$. Поэтому, строго говоря, метод построения апостериорного среднего (3.18) нельзя называть восстановлением функции в классе $P(x, \alpha)$.

Итак, найдем функцию $\pi(x; x_1, \ldots, x_l)$, которая минимизирует функционал

$$R_{\text{Б}}(\pi) = \int (P(x \mid \alpha) - \pi(x; x_1, \ldots, x_l))^2 \times$$
$$\times P(x_1, \ldots, x_l \mid \alpha)\, P(\alpha)\, d\alpha\, dx\, dx_1, \ldots, dx_l. \quad (3.19)$$

Обозначим

$$r(x; x_1, \ldots, x_l) =$$
$$= \int (P(x \mid \alpha) - \pi(x; x_1, \ldots, x_l))^2\, P(x_1, \ldots, x_l \mid \alpha)\, P(\alpha)\, d\alpha.$$

Изменим порядок интегрирования в (3.19), после чего окажется

$$R_{\text{Б}}(\pi) = \int r(x; x_1, \ldots, x_l)\, dx\, dx_1 \ldots dx_l. \quad (3.20)$$

Преобразуем теперь функцию

$$r(x; x_1, \ldots, x_l) = \int P^2(x \mid \alpha) P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha - $$
$$- 2\pi(x; x_1, \ldots, x_l) \int P(x \mid \alpha) P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha + $$
$$+ \pi^2(x; x_1, \ldots, x_l) \int P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha. \qquad (3.21)$$

Обозначим

$$\hat{P}(x \mid x_1, \ldots, x_l) = \frac{\int P(x \mid \alpha) P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha}{P(x_1, \ldots, x_l)},$$

где

$$P(x_1, \ldots, x_l) = \int P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha,$$

и перепишем равенство (3.21) в виде

$$r(x; x_1, \ldots, x_l) = \int P^2(x \mid \alpha) P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha - $$
$$- \hat{P}^2(x \mid x_1, \ldots, x_l) P(x_1, \ldots, x_l) + $$
$$+ [\hat{P}(x \mid x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2 P(x_1, \ldots, x_l).$$

Подставим выражение $r(x; x_1, \ldots, x_l)$ в (3.20). В результате получим функционал, который может быть представлен в виде двух слагаемых:

$$R_{\text{Б}}(\pi) = R_1 + R_2(\pi),$$

где

$$R_1 = \int [P^2(x \mid \alpha) P(x_1, \ldots, x_l \mid \alpha) P(\alpha)\, d\alpha - $$
$$- P(x_1, \ldots, x_l) \hat{P}^2(x \mid x_1, \ldots, x_l)]\, dx\, dx_1 \ldots dx_l,$$

$$R_2(\pi) =$$
$$= \int [\hat{P}(x \mid x_1, \ldots, x_l) - \pi(x; x_1, \ldots, x_l)]^2\, dx\, dx_1 \ldots dx_l.$$

Первое слагаемое не зависит от $\pi(x; x_1, \ldots, x_l)$. Поэтому минимизация $R_{\text{Б}}(\pi)$ эквивалентна минимизации второго слагаемого $R_2(\pi)$.

Минимум этого слагаемого равен нулю и достигается тогда, когда

$$\pi(x; x_1, \ldots, x_l) = \hat{P}(x \mid x_1, \ldots, x_l) \equiv P_{\text{Б}}(x \mid x_1, \ldots, x_l).$$

В следующих параграфах для некоторых априорных законов $P(\alpha)$ будут найдены байесовы приближения

плотностей. Построение байесова приближения для фиксированного априорного закона $P(\alpha)$ зависит от того, удается ли провести аналитическое интегрирование выражения (3.18).

§ 6. Байесова оценка распределения вероятностей дискретных независимых признаков

В § 3 была введена функция распределения вероятностей дискретных независимых признаков (3.12), (3.13). Здесь мы покажем, что при минимальных априорных сведениях относительно значений параметров $p^i(j)$: для каждого i параметры $p^i(1), \ldots, p^i(\tau_i)$ равномерно распределены на симплексе

$$C_i = \left\{ p: \sum_{j=1}^{\tau_i} p^i(j) = 1, \quad p^i(j) \geqslant 0 \right\},$$

байесова оценка распределения вероятностей дискретных независимых признаков равна

$$P_{\text{Б}}(x) = \prod_{i=1}^{n} P_{\text{Б}}(x^i),$$

где

$$P_{\text{Б}}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)+1}{l+\tau_i}, \\ \ldots\ldots\ldots \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)+1}{l+\tau_i}, \end{cases}$$

$m_i(j)$ — число векторов выборки, у которых i-я координата принимает j-е значение, τ_i — число значений, которые принимает i-я координата, l — объем выборки.

Получим байесову оценку распределения вероятностей дискретных независимых признаков. Для этого вычислим функцию

$$P_{\text{Б}}(x^i) = \frac{\int P(x^i \mid p)\, P\left(x_1^i, \ldots, x_l^i \mid p\right) P(p)\, dp}{\int P\left(x_1^i, \ldots, x_l^i \mid p\right) P(p)\, dp}. \tag{3.22}$$

В нашем случае

$$P(x^i \mid p) = \begin{cases} p^i(1), & \text{если } x^i = c^i(1), \\ \ldots\ldots\ldots\ldots\ldots \\ p^{(i)}(\tau_i), & \text{если } x^i = c^i(\tau_i). \end{cases}$$

Вычислим сначала знаменатель (3.22). Учитывая, что выборка $x_1, \ldots$ $\ldots x_l$ — случайная и независимая, получаем

$$\int P\left(x_1^i, \ldots x_l^i \mid p\right) P(p)\, dp =$$

$$= \frac{1}{v} \int_{C_i} \prod_{j=1}^{\tau_l} [p^i(j)]^{m_i(j)}\, dp^i(1), \ldots, dp^i(\tau_i), \tag{3.23}$$

где v — объем симплекса C_i.

Известно [52], что определенный интеграл (3.23) может быть вычислен аналитически:

$$P\left(x_1^i, \ldots, x_l^i\right) = \frac{1}{v} \frac{\Gamma(m_i(1)+1) \cdot \ldots \cdot \Gamma(m_i(\tau_i)+1)}{\Gamma(m_i(1)+\ldots+m_i(\tau_i)+\tau_i)}, \tag{3.24}$$

где $\Gamma(n)$ — гамма-функция. Для целых n она вычисляется так:

$$\Gamma(n) = (n-1)!.$$

Найдем теперь числитель выражения (3.22) для случая $x^i = c^i(k)$:

$$I_k^i = \int_{C_i} P\left(x^i = c^i(k) \mid p\right) P\left(x_1^i, \ldots, x_l^i \mid p\right) P(p)\, dp =$$

$$= \frac{1}{v} \int_{C_i} p^i(k) \prod_{j=1}^{\tau_i} [p^i(j)]^{m_i(j)}\, dp^i(1), \ldots, dp^i(\tau_i).$$

Определенный интеграл I_k^i равен [52]

$$I_k^i = \frac{1}{v} \frac{\Gamma(m_i(1)+1) \ldots \Gamma(m_i(\tau_i)+1)\, \Gamma(m_i(k)+2)}{\Gamma(m_i(1)+\ldots+m_i(\tau_i)+\tau_i+1)\, \Gamma(m_i(k)+1)}. \tag{3.25}$$

Разделив (3.25) на (3.24), получим

$$P_{\text{Б}}(x^i = c^i(k)) = \frac{\Gamma(m_i(k)+2)\, \Gamma(l+\tau_i)}{\Gamma(m_i(k)+1)\, \Gamma(l+\tau_i+1)} = \frac{m_i(k)+1}{l+\tau_i}.$$

Таким образом,

$$P_{\text{Б}}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)+1}{l+\tau_i}, & \text{если } x^i = c^i(1), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)+1}{l+\tau_i}, & \text{если } x^i = c^i(\tau_i). \end{cases}$$

В силу независимости признаков $P_{\text{Б}}(x) = \prod_{i=1}^{n} P_{\text{Б}}(x^i)$.

§ 7. Байесовы приближения плотности нормального закона

Найдем байесовы приближения плотности нормального закона для некоторых специальных случаев априорного распределения параметров. Сначала мы найдем байесово приближение для одномерного нормального закона, построенное в предположении, что параметры нормального закона $N(\mu, \sigma)$ распределены равномерно в прямоугольнике $0 \leqslant \sigma \leqslant \Pi$, $-T \leqslant \mu \leqslant T$.

Окажется, что если величины Π и T достаточно большие, то байесово приближение равно

$$P_{\text{Б}}(x) = \frac{E(l)}{\sigma_{\text{э}}}\left[1 + \frac{(x - x_{\text{э}})^2}{(l+1)\,\sigma_{\text{э}}^2}\right]^{-\left(\frac{l-1}{2}\right)}, \tag{3.26}$$

где

$$E(l) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{\sqrt{(l+1)\cdot\pi}\,\Gamma\left(\frac{l}{2}-1\right)},$$

$$x_{\text{э}} = \frac{1}{l}\sum_{i=1}^{l} x_i; \quad \sigma_{\text{э}}^2 = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{э}})^2.$$

Затем мы найдем байесово приближение n-мерного нормального закона для специального априорного закона распределения параметров μ и Δ (μ — n-мерный вектор средних и Δ — матрица ковариации $n \times n$).

Окажется, что в этом случае байесово приближение равно

$$P_{\text{Б}}(x) = \frac{E(l)}{|S|^{1/2}}\left[1 + \frac{(x - x_{\text{э}})^{\text{т}} S^{-1}(x - x_{\text{э}})}{l+1}\right]^{-\frac{l+n}{2}}, \tag{3.27}$$

где

$$E(l) = \frac{\Gamma\left(\frac{l+n}{2}\right)}{((l+1)\,\pi)^{n/2}\,\Gamma(l/2)},$$

вектор $x_{\text{э}}$ — оценка вектора средних:

$$x_{\text{э}} = \frac{1}{l}\sum_{i=1}^{l} x_i,$$

S — эмпирическая матрица ковариаций:

$$S=\frac{1}{l}\sum_{i=1}^{l}(x_i-x_э)(x_i-x_э)^{\mathrm{т}}.$$

Заметим, что оба приближения нормальных законов (3.26) и (3.27) не принадлежат классу нормальных. Однако легко можно убедиться, что в обоих случаях при $l\to\infty$

$$P_{\mathrm{Б}}(x)\underset{l\to\infty}{\to}N(\mu,\ \Delta).$$

И еще одно замечание. Для того чтобы вычислить байесово приближение многомерного нормального закона (см. ниже п. 2) пришлось рассмотреть специальный закон априорного распределения параметров, отличный от равновероятного, принятого при выводе одномерного случая (см. п. 1) Однако байесово приближение для одномерного закона, полученное из (3.27) при $n=1$, оказалось близким к байесову приближению, полученному в предположении равномерного распределения параметров (3.26).

1. Байесово приближение одномерного нормального закона. Пусть величина x распределена по нормальному закону

$$P(x;\ \mu,\ \sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Кроме того, пусть априорное распределение параметров μ и σ подчиняется равномерному закону в прямоугольнике $0\leqslant\sigma\leqslant\Pi$; $-T\leqslant\mu\leqslant T$. Так как выборка $x_1,\ \ldots,\ x_l$ случайная и независимая, то

$$P(x_1,\ \ldots,\ x_l;\ \mu,\ \sigma)=\frac{1}{(2\pi)^{l/2}\sigma^l}\exp\left\{-\frac{\sum_{i=1}^{l}(x_i-\mu)^2}{2\sigma^2}\right\}.$$

Байесова оценка плотности распределения вероятностей, согласно (3.18), равна

$$P_{\mathrm{Б}}(x)=$$

$$=\left(\frac{1}{2T\Pi}\,\frac{1}{(2\pi)^{\frac{l+1}{2}}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l+1}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\}d\mu\,d\sigma\right)\times$$

$$\times\left(\frac{1}{2T\Pi}\,\frac{1}{(2\pi)^{\frac{l}{2}}}\int_{-T}^{T}\int_{0}^{\Pi}\frac{1}{\sigma^{l}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(x_i-\mu)^2\right\}d\mu\,d\sigma\right)^{-1}. \qquad (3.28)$$

Будем считать, что интервалы $[-T, T]$ и $[0, \Pi]$ столь велики, что пределы интегрирования в (3.28) могут быть расширены до $(-\infty, \infty)$ и $(0, \infty)$. Во всяком случае, это можно сделать, если $l \geqslant 2$. В этом случае интегралы в (3.28) сходятся. Вычислим числитель выражения (3.28):

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{1}{\sigma^{l+1}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2\right)\right\} d\mu\, d\sigma. \quad (3.29)$$

Для этого обозначим

$$T(\mu) = \sum_{i=1}^{l}(x_i-\mu)^2+(x-\mu)^2, \quad y = \frac{\sqrt{T(\mu)}}{\sigma}.$$

Тогда интеграл (3.29) перепишется в виде

$$I(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{y^{l-1}}{T^{l/2}(\mu)} \exp\left\{-\frac{1}{2}y^2\right\} dy\, d\mu =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)} \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy.$$

Обозначим

$$c(l) = \int_{0}^{\infty} y^{l-1} \exp\left\{-\frac{y^2}{2}\right\} dy,$$

где $c(l)$ не зависит ни от μ, ни от σ. Интеграл (3.29) может быть переписан в виде

$$I(x) = \frac{c(l)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d\mu}{T^{l/2}(\mu)}.$$

Преобразуем теперь выражение $T(\mu)$. Для этого заметим, что

$$\sum_{i=1}^{l}(x_i-\mu)^2 = l\sigma_э^2 + l(\mu - x_э)^2,$$

где обозначено

$$x_э = \frac{1}{l}\sum_{i=1}^{l} x_i, \quad \sigma_э^2 = \frac{1}{l}\sum_{i=1}^{l}(x_i-x_э)^2.$$

Соответственно преобразуется и выражение

$$T(\mu) = l\sigma_э^2 + l(\mu - x_э)^2 + (x-\mu)^2.$$

Положим теперь

$$\bar{x}=\frac{x_э l+x}{l+1}$$

и перепишем $T(\mu)$:

$$T(\mu)=l\sigma_э^2+\frac{l}{l+1}(x-x_э)^2+(\bar{x}-\mu)^2(l+1).$$

Запишем теперь интеграл $I(x)$ в виде

$$I(x)=\frac{c(l)}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{d\mu}{\left[l\sigma_э^2+\frac{l}{l+1}(x-x_э)^2+(\bar{x}-\mu)^2(l+1)\right]^{l/2}}=$$

$$=\frac{c(l)}{\sqrt{2\pi(l+1)}}\left(l\sigma_э^2+\frac{l(x-x_э)^2}{(l+1)}\right)^{-\frac{l-1}{2}}\int_{-\infty}^{\infty}\frac{dz}{(1+z^2)^{l/2}}.$$

Заметим, что подынтегральное выражение не зависит от параметров. Таким образом, оказывается, что

$$I(x)=c'(l,\ \sigma_э)\left(1+\frac{(x-x_э)^2}{(l+1)\sigma_э^2}\right)^{-\frac{l-1}{2}}. \tag{3.30}$$

Для получения байесовой оценки нам остается нормировать к единице выражение (3.30):

$$P_Б(x)=\frac{I(x)}{\int_{-\infty}^{\infty}I(x)\,dx}.$$

Известно [52], что интеграл в знаменателе равен следующему выражению:

$$\int_{-\infty}^{+\infty}I(x)\,dx=c''(l,\ \sigma_э)\int_{-\infty}^{+\infty}\frac{dx}{\left(1+\frac{(x-x_э)^2}{(l+1)\sigma_э^2}\right)^{\frac{l-1}{2}}}=$$

$$=\frac{c''(l,\ \sigma_э)\,\sigma_э\sqrt{l+1}\cdot\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{l}{2}-1\right)}{\Gamma\left(\frac{l-1}{2}\right)}.$$

Обозначим

$$E(l)=\frac{\Gamma\left(\frac{l-1}{2}\right)}{\sqrt{l+1}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{l}{2}-1\right)}=\frac{\Gamma\left(\frac{l-1}{2}\right)}{\sqrt{\pi(l+1)}\,\Gamma\left(\frac{l}{2}-1\right)}.$$

Таким образом,

$$P_{\text{Б}}(x)=\frac{E(l)}{\sigma_{\text{э}}}\left(1+\frac{(x-x_{\text{э}}')^2}{(l+1)\,\sigma_{\text{э}}^2}\right)^{-\frac{l-1}{2}}.$$

2. Байесово приближение n-мерного нормального закона. При получении байесова приближения n-мерного нормального закона используются следующие два факта теории многомерных нормальных законов.

1. Свертка двух многомерных нормальных законов $N(0, \Delta)$ и $N(\mu, \gamma\Delta)$, где γ — положительное число, есть нормальный закон $N(\mu, (1+\gamma)\Delta)$. Иначе говоря, справедливо равенство [4]

$$\int\limits_{E_n} N(\mu-t, \gamma\Delta)\cdot N(t, \Delta)\,dt = N(\mu, (1+\gamma)\Delta).$$

2. Распределение эмпирических оценок S ковариационной матрицы Δ, вычисляемых по формуле

$$S=\frac{1}{l}\sum_{i=1}^{l}(x_i-x_{\text{э}})(x_i-x_{\text{э}})^{\text{т}},\quad x_{\text{э}}=\frac{1}{l}\sum_{i=1}^{l}x_i,$$

задается законом Уишарта [5]:

$$W_{l,n}(S;\Delta)=$$
$$=\begin{cases} C_{n,l}\,|\Delta|^{-\frac{l-1}{2}}\,|S|^{\frac{l-n-2}{2}}\exp\left\{-\frac{l}{2}\operatorname{Sp}[\Delta^{-1}S]\right\}, & \text{если } |S|>0,\\ 0, & \text{если } |S|\leqslant 0 \end{cases}$$

где предполагается, что $l>n+1$, $\operatorname{Sp}\|a_{ij}\|=\sum\limits_{i=1}^{n}a_{ii}$. Величина $C_{n,l}$ есть константа, равная

$$C_{n,l}=\left(\left(\frac{l}{2}\right)^{-\frac{(l-1)n}{2}}\cdot\pi^{\frac{n(n-1)}{4}}\cdot\prod_{i=1}^{n}\Gamma\left(\frac{l-i}{2}\right)\right)^{-1}. \qquad (3.31)$$

При получении байесова приближения будет использован факт нормированности к единице распределения Уишарта:

$$\int\limits_{|S|>0}|S|^{\frac{l-n-2}{2}}\exp\left\{-\frac{l}{2}\operatorname{Sp}[\Delta^{-1}S]\right\}dS=\frac{1}{C_{n,l}}\,|\Delta|^{\frac{l-1}{2}}. \qquad (3.32)$$

Обозначим матрицу $\Delta^{-1}=\mathscr{D}$. Очевидно, $|\Delta|=1/|\mathscr{D}|$. Пусть априорное распределение параметров μ и Δ n-мерного нормального закона $N(\mu, \Delta)$ задано в виде

$$P_{a,A}(\mu, \mathscr{D})=P_a(\mu\,|\,\mathscr{D})\cdot P_A(\mathscr{D}),$$

где вектор параметров μ распределен по нормальному закону:

$$P_a(\mu \mid \mathscr{D}) = c_1 \mid \mathscr{D} \mid^{1/2} \exp\left\{-\frac{\omega}{2}(\mu-a)^{\mathrm{T}}\mathscr{D}(\mu-a)\right\};$$

здесь c_1 — константа, $\omega > 0$ — число, a — вектор, $\mathscr{D}$ — матрица, распределенная по закону Уишарта:

$$P_A(\mathscr{D}) =$$

$$= \begin{cases} C_{n,\nu} \mid \omega A \mid^{\frac{\nu-1}{2}} \mid \mathscr{D} \mid^{\frac{\nu-n-2}{2}} \exp\left\{-\frac{\nu\omega}{2}\operatorname{Sp}[A\mathscr{D}]\right\}, & \text{если } \mid \mathscr{D} \mid > 0, \\ 0, & \text{если } \mid \mathscr{D} \mid \leqslant 0. \end{cases}$$

Здесь $\nu > n+2$ — константа, A — матрица. Заметим, что

$$\operatorname{Sp}[\mathscr{D}xx^{\mathrm{T}}] = \operatorname{Sp}[xx^{\mathrm{T}}\mathscr{D}] = x^{\mathrm{T}}\mathscr{D}x, \tag{3.33}$$

где $\mathscr{D}$ — симметричная матрица, x — вектор-столбец. Выпишем совместную плотность $P(x_1, \ldots, x_l \mid \mu, \mathscr{D})$ для случайной независимой выборки $x_1, \ldots, x_l$:

$$P(x_1, \ldots, x_l \mid \mu, \mathscr{D}) = c_2 \mid \mathscr{D} \mid^{l/2} \exp\left\{\frac{-\sum_{i=1}^{l}(x_l-\mu)^{\mathrm{T}}\mathscr{D}(x_l-\mu)}{2}\right\} =$$

$$= c_2 \mid \mathscr{D} \mid^{l/2} \exp\left\{-\frac{1}{2}\operatorname{Sp}\left[l\mathscr{D}S + l\mathscr{D}(x_э-\mu)(x_э-\mu)^{\mathrm{T}}\right]\right\}.$$

Здесь и далее c_0, c_1, c_2, c_3 — константы, которые определяются условиями нормировки. Согласно формуле Байеса апостериорная плотность $P(\mu, \mathscr{D} \mid x_1, \ldots, x_l)$ равна

$$P(\mu, \mathscr{D} \mid x_1, \ldots, x_l) = c_0 P(x_1, \ldots, x_l \mid \mu, \mathscr{D})\, P_a(\mu \mid \mathscr{D})\, P_A(\mathscr{D}). \tag{3.34}$$

Вычислим правую часть выражения (3.34)

$$P(\mu, \mathscr{D} \mid x_1, \ldots, x_l) =$$

$$= c_0 \mid \mathscr{D} \mid^{l/2} \exp\left\{-\frac{1}{2}\operatorname{Sp}\left[l\mathscr{D}S + l\mathscr{D}(x_э-\mu)(x_э-\mu)^{\mathrm{T}}\right]\right\} \times$$

$$\times c_1 \mid \mathscr{D} \mid^{l/2} \exp\left\{-\frac{1}{2}\operatorname{Sp}\left[\mathscr{D}\omega(\mu-a)(\mu-a)^{\mathrm{T}}\right]\right\} \times$$

$$\times c_2 \cdot C_{n,\nu} \mid \mathscr{D} \mid^{\frac{\nu-n-2}{2}} \mid \omega A \mid^{\frac{\nu-1}{2}} \exp\left\{-\frac{1}{2}\operatorname{Sp}[\nu\mathscr{D}A\omega]\right\} =$$

$$= c_3 \mid \mathscr{D} \mid^{\frac{l+\nu-n-1}{2}} \exp\left\{-\frac{1}{2}\operatorname{Sp}\left[l\mathscr{D}S + l\mathscr{D}(x_э-\mu)(x_э-\mu)^{\mathrm{T}} + \right.\right.$$

$$\left.\left. + \omega\mathscr{D}(\mu-a)(\mu-a)^{\mathrm{T}} + \nu\omega\mathscr{D}A\right]\right\}. \tag{3.35}$$

Преобразуя выражение в показателе экспоненты (3.35), получим

$$\mathscr{D}\left(lS + l(x_э-\mu)(x_э-\mu)^{\mathrm{T}} + \omega(\mu-a)(\mu-a)^{\mathrm{T}} + \nu\omega A\right) =$$

$$= \mathscr{D}\left[(l+\omega)(\mu-b)(\mu-b)^{\mathrm{T}} + (l+\nu)B\right],$$

где обозначено

$$b=\frac{lx_{\text{э}}+a\omega}{l+\omega},\quad B=\frac{\left(lS+\omega\nu A+\dfrac{l\omega}{l+\omega}(x_{\text{э}}-a)(x_{\text{э}}-a)^{\text{т}}\right)}{l+\nu}. \tag{3.36}$$

В этих обозначениях перепишем (3.35):

$$P(\mu,\ \mathscr{D}\mid x_1,\ \ldots,\ x_l)=$$
$$=c_{\text{з}}\mid\mathscr{D}\mid^{\frac{l+\nu-n-1}{2}}\exp\left\{-\frac{1}{2}\operatorname{Sp}\left[\mathscr{D}\left((l+\omega)(\mu-b)(\mu-b)^{\text{т}}+\right.\right.$$
$$\left.\left.+(l+\nu)B\right)\right]\right\}. \tag{3.37}$$

Теперь из условия нормировки можно восстановить константу $C_{\text{з}}$

$$c_{\text{з}}^{-1}=\int\mid\mathscr{D}\mid^{\frac{l+\nu-n-2}{2}}\exp\left\{-\frac{l+\nu}{2}\operatorname{Sp}[\mathscr{D}B]\right\}d\mathscr{D}\int\mid\mathscr{D}\mid^{1/2}\times$$
$$\times\exp\left\{-\frac{l+\omega}{2}\operatorname{Sp}\left[\mathscr{D}(\mu-b)(\mu-b)^{\text{т}}\right]\right\}d\mu=$$
$$=\left(\frac{2\pi}{l+\omega}\right)^{n/2}\left(C_{n,\,l+\nu}\mid(l+\nu)B\mid^{\frac{l+\nu-1}{2}}\right)^{-1}.$$

При вычислении внешнего интеграла было использовано равенство (3.32). Наконец, найдем байесову оценку

$$P_{\text{Б}}(x)=\int P(x\mid\mu,\ \mathscr{D})\,P(\mu,\ \mathscr{D}\mid x_1,\ \ldots,\ x_l)\,d\mu\,d\mathscr{D}=$$
$$=\int(2\pi)^{-n/2}\mid\mathscr{D}\mid^{1/2}\exp\left\{-\frac{1}{2}(x-\mu)^{\text{т}}\mathscr{D}(x-\mu)\right\}c_{\text{з}}\mid\mathscr{D}\mid^{\frac{l+\nu-n-1}{2}}\times$$
$$\times\exp\left\{-\frac{l+\omega}{2}(\mu-b)^{\text{т}}\mathscr{D}(\mu-b)\right\}\exp\left\{-\frac{l+\nu}{2}\operatorname{Sp}[\mathscr{D}B]\right\}d\mu\,d\mathscr{D}=$$
$$=\left(\frac{2\pi}{l+\omega}\right)^{n/2}\int c_{\text{з}}\mid\mathscr{D}\mid^{\frac{l+\nu-n-2}{2}}\exp\left\{-\frac{l+\nu}{2}\operatorname{Sp}[\mathscr{D}B]\right\}d\mathscr{D}\times$$
$$\times\int(2\pi)^{-n/2}\mid\mathscr{D}\mid^{1/2}(2\pi)^{-n/2}\mid(l+\omega)\mathscr{D}\mid^{1/2}\times$$
$$\times\exp\left\{-\frac{1}{2}(x-\mu)^{\text{т}}\mathscr{D}(x-\mu)\right\}\exp\left\{-\frac{l+\omega}{2}(\mu-b)^{\text{т}}\mathscr{D}(\mu-b)\right\}d\mu.$$

Заметим, что внутренний интеграл по μ есть свертка двух нормальных законов, поэтому получим

$$P_{\text{Б}}(x)=c_{\text{з}}\int(l+\omega+1)^{-n/2}\mid\mathscr{D}\mid^{\frac{l+\nu-n-1}{2}}\times$$
$$\times\exp\left\{-\frac{1}{2}\operatorname{Sp}\left[\mathscr{D}\left(B(l+\nu)+\frac{l+\omega}{l+\omega+1}(x-b)(x-b)^{\text{т}}\right)\right]\right\}d\mathscr{D}. \tag{3.38}$$

Согласно (3.32) получаем

$$P_{\text{Б}}(x) = \frac{c_3 (l+\omega+1)^{-n/2}}{C_{n,\ l+\nu+1}} \left| (l+\nu) B + \right.$$

$$\left. + \frac{l+\omega}{l+\omega+1} (x-b)(x-b)^{\mathrm{T}} \right|^{-\frac{l+\nu}{2}} = \left(\frac{l+\omega}{l+\omega+1} \frac{1}{2\pi}\right)^{n/2} \frac{C_{n,\ l+\nu}}{C_{n,\ l+\nu+1}} \times$$

$$\times \frac{|(l+\nu) B|^{\frac{l+\nu-1}{2}}}{\left|(l+\nu) B + \frac{l+\omega}{l+\omega+1} (x-b)(x-b)^{\mathrm{T}}\right|^{\frac{l+\nu}{2}}}. \qquad (3.39)$$

Преобразуем выражение (3.39):

$$P_{\text{Б}}(x) = \left(\frac{1}{\pi} \frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+\nu}{2}\right)}{\Gamma\left(\frac{l+\nu-n}{2}\right)} \times$$

$$\times \frac{|(l+\nu) \cdot B|^{-1/2}}{\left|I + \frac{l+\omega}{l+\omega+1} \frac{1}{l+\nu} (x-b)(x-b)^{\mathrm{T}} B^{-1}\right|^{\frac{l+\nu}{2}}}. \qquad (3.40)$$

В знаменателе выражения (3.40) I — единичная матрица. Заметим, что матрица $(x-b)(x-b)^{\mathrm{T}}$, а следовательно, и матрица $(x-b) \times (x-b)^{\mathrm{T}} B^{-1}$ имеют ранг, равный единице. Поэтому только одно ее собственное число отлично от нуля, откуда следует, что знаменатель выражения (3.40) равен

$$\left|I + \frac{l+\omega}{l+\omega+1} \frac{1}{l+\nu} (x-b)(x-b)^{\mathrm{T}} B^{-1}\right|^{\frac{l+\nu}{2}} =$$

$$= \left(1 + \frac{l+\omega}{l+\omega+1} \cdot \frac{1}{l+\nu} (x-b)^{\mathrm{T}} B^{-1} (x-b)\right)^{\frac{l+\nu}{2}}.$$

Таким образом, окончательно получим

$$P_{\text{Б}}(x) = \left(\frac{1}{\pi} \frac{l+\omega}{l+\omega+1}\right)^{n/2} \frac{\Gamma\left(\frac{l+\nu}{2}\right)}{\Gamma\left(\frac{l+\nu-n}{2}\right)} \times$$

$$\times \frac{|(l+\nu) \cdot B|^{-1/2}}{\left(1 + \frac{l+\omega}{l+\omega+1} \cdot \frac{1}{l+\nu} (x-b)^{\mathrm{T}} B^{-1} (x-b)\right)^{\frac{l+\nu}{2}}}$$

Зададим теперь конкретные величины ν и ω для того, чтобы в условиях рассматриваемой схемы получить наиболее неопределенные априорные условия:

1) $\nu = n + \varepsilon$ $(\varepsilon > 0)$ — условие, необходимое для интегрирования распределения Уишарта;

2) $\omega \to 0$, $\varepsilon \to 0$ — условие, обеспечивающее стремление каждого элемента матрицы A к нулю.

Тогда, согласно (3.36), получим $b \to x_э$, $(l+\nu)B \to lS$, откуда заключаем, что

$$P_Б(x) = \left(\frac{1}{(l+1)\pi}\right)^{n/2} \frac{\Gamma\left(\frac{l+n}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{|S|^{-1/2}}{\left(1+\frac{1}{l+1}(x-x_э)^T S^{-1}(x-x_э)\right)^{\frac{l+n}{2}}}.$$

Наконец, для одномерного случая (полагая $n=1$) получаем

$$P_Б(x) = \sqrt{\frac{1}{\pi(l+1)}} \frac{1}{\sigma_э} \frac{\Gamma\left(\frac{l+1}{2}\right)}{\Gamma\left(\frac{l}{2}\right)} \frac{1}{\left(1+\frac{1}{l+1}\frac{(x-x_э)^2}{\sigma_э^2}\right)^{\frac{l+1}{2}}}.$$

§ 8. Несмещенные оценки

В предыдущих параграфах были получены байесовы оценки плотности распределения вероятностей для специальных априорных законов распределения параметров.

Однако при решении практических задач априорный закон распределения параметров, как правило, неизвестен. Минимаксная же схема восстановления плотности может привести к слишком грубым результатам. Поэтому хотелось бы найти достаточно хороший метод восстановления плотности распределения вероятностей, который не был бы связан с решением байесовой задачи. Как это можно сделать?

Предположим, что существует такой метод восстановления плотности, который является лучшим не только в среднем (что соответствует байесовому критерию), но и при восстановлении каждой конкретной плотности. Если бы такой равномерно лучший метод восстановления плотности существовал, то он не зависел бы от априорного закона задания плотностей.

К сожалению, равномерно лучшего метода оценивания в классе всех возможных методов оценивания нет. Действительно, существует тривиальный алгоритм восстановления плотности, который независимо от особенностей выборки восстанавливает плотность с одними и теми же фиксированными значениями параметров. Такой алгоритм восстанавливает идеально одну-единственную плотность и плохо — другие. Он и будет лучшим для своей плотности.

Но если нет равномерно наилучшего метода в классе всех возможных методов оценивания, то может быть существует такой метод в более узком классе методов? Поэтому возникает идея ограничить класс возможных методов восстановления плотности и попытаться найти в нем равномерно лучший. Оказывается, что если ограничить класс оценок так называемыми *несмещенными оценками плотности распределения вероятностей,* то задача отыскания равномерно лучшей в этом классе оценки имеет решение.

Определение. *Говорят, что функция* $\pi(x; x_1, \ldots, x_l)$ *является несмещенной оценкой плотности* $P(x, \alpha^*)$ *из класса* $P(x, \alpha)$, *построенной по выборке* $x_1, \ldots, x_l$ *длины* l, *полученной согласно* $P(x, \alpha^*)$, *если математическое ожидание оценки* $\pi(x; x_1, \ldots, x_l)$ *равно плотности* $P(x, \alpha^*)$, *т. е. если для любого* $P(x, \alpha^*)$ *из* $P(x, \alpha)$ *справедливо*

$$M_{\alpha^*}\pi(x; x_1, \ldots, x_l) = P(x, \alpha^*).$$

Заметим, что само по себе свойство несмещенности оценки не имеет никакой самостоятельной ценности и вводится исключительно для того, чтобы сузить класс возможных оценок. И если в статистике широко используется класс несмещенных оценок, то только потому, что он доступен для анализа.

В чем же заключается эта доступность? Выпишем еще раз определение несмещенной оценки:

$$\int \pi(x; x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha)\, dx_1 \ldots dx_l = P(x, \alpha). \tag{3.41}$$

Выражение (3.41) не только определяет несмещенные оценки плотности, но и указывает способ их построения: множество несмещенных оценок есть множество решений уравнения Фредгольма I рода. Однако нахождение решения уравнения (3.41) является, вообще говоря, трудной задачей. В главе I было показано, что даже в том случае, когда решение уравнения Фредгольма единственно, численное его решение является некорректно поставленной задачей. Поэтому получение несмещенных оценок плотности $P(x, \alpha)$ возможно лишь тогда, когда удается решить уравнение (3.41) аналитически.

В § 10 мы найдем наилучшую несмещенную оценку плотности многомерного нормального закона. Но прежде чем приступить к построению такой оценки, заметим, что в гл. II более общая задача — восстановление плотности в классе непрерывных функций — была также сведена к решению уравнения Фредгольма I рода. Здесь же к решению уравнения Фредгольма сводится частная постановка задачи — получение несмещенной оценки плотности, известной с точностью до параметров.

Существенная разница, однако, состоит в том, что в общем случае, рассмотренном в главе II, правая часть уравнения Фредгольма I рода была известна с точностью до помех, здесь же она задана точно.

§ 9. Достаточные статистики

Построение наилучшей несмещенной оценки оказывается возможным в терминах так называемых *достаточных статистик.* До сих пор при исследовании оценок мы исходили из того, что оценка плотности имеет вид $\pi(x; x_1, \ldots, x_l)$, т. е. оценка есть функция от $(l+1)$-го вектора, а именно: вектора x и l векторных переменных $x_1, \ldots, x_l$. Фиксируя последние l переменных, мы получали конкретный вид восстанавливаемой плотности.

Однако такой способ задания оценки плотности является не совсем удобным. Так, очевидно, что $\pi(x; x_1, \ldots, x_l)$ не должна зависеть от порядка следования векторов выборки $x_1, \ldots, x_l$. Кроме того, для другого объема выборки, например $l+1$, приходится задавать свою функцию (размерности $l+2$).

Поэтому хотелось бы найти такие k характеристик выборки

$$t_i = f_i(x_1, \ldots, x_l), \quad i = 1, \ldots, k,$$

чтобы, во-первых, вся информация о плотности, находящаяся в выборке $x_1, \ldots, x_l$, содержалась и в этих k числах, а во-вторых, количество необходимых характеристик k зависело бы не от объема выборки, а от особенности класса восстанавливаемых плотностей. В терминах этих характеристик выборки и хотелось бы получить несмещенную оценку $\pi^*(x; t_1, \ldots, t_k)$. Такими характеристиками выборки и являются достаточные статистики (см. [58]).

Определение. *Говорят, что функции* $t_i = f_i(x_1, \ldots, x_l)$ *являются достаточными статистиками для плотности* $P(x, \alpha)$, *если совместная плотность* $P(x_1, \ldots, x_l; \alpha)$ *выборки* $x_1, \ldots, x_l$ *может быть представлена в виде*

$$P(x_1, \ldots, x_l; \alpha) = P_1(t_1, \ldots, t_k; \alpha) P_2(x_1, \ldots, x_l).$$

Иначе говоря, совместная плотность $P(x_1, \ldots, x_l; \alpha)$ распадается на произведение двух сомножителей, один из которых $P_2(\cdot)$ не зависит от параметра α плотности распределения вероятностей $P(x, \alpha)$, а другой сомножитель, содержащий α, зависит лишь от значений $t_1, \ldots, t_k$ (а не от самой выборки $x_1, \ldots, x_l$).

Легко проверить, что для n-мерного нормального закона достаточными статистиками будут следующие $\frac{n(n+3)}{2}$ величины:

$$t = \frac{1}{l} \sum_{j=1}^{l} x_j, \quad t = (t_1, \ldots, t_n)^{\mathrm{T}} \text{ (всего } n \text{ величин);}$$

$$\| t_{ij} \| = \sum_{r=1}^{l} (x_r - t)(x_r - t)^{\mathrm{T}} \qquad \left(\text{всего } \frac{n(n+1)}{2} \text{ величин} \right).$$

В самом деле, для n-мерного нормального закона справедливо

$$P(x_1, \ldots, x_l; \mu, \Delta) =$$

$$= \frac{1}{(2\pi)^{nl/2} |\Delta|^{l/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{l} (x_i - \mu)^{\mathrm{T}} \Delta^{-1} (x_i - \mu) \right\} =$$

$$= (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\left\{ -\frac{1}{2} \mathrm{Sp}\left[\Delta^{-1} \sum_{i=1}^{l} (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \right] \right\} =$$

$$= (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\left\{ -\frac{1}{2} \mathrm{Sp}\left[\Delta^{-1} \left(\sum_{i=1}^{l} (x_i - t)(x_i - t)^{\mathrm{T}} + \right. \right. \right.$$

$$\left. \left. \left. + l(t - \mu)(t - \mu)^{\mathrm{T}} \right] \right\} =$$

$$= (2\pi)^{-nl/2} |\Delta|^{-l/2} \exp\left\{ -\frac{1}{2} \mathrm{Sp}\left[\Delta^{-1} (\| t_{ij} \| + l(t - \mu)(t - \mu)^{\mathrm{T}}) \right] \right\}.$$

При выводе использовалось равенство $z^{\mathrm{T}} B z = \mathrm{Sp}[z z^{\mathrm{T}} B]$.

Итак, будем искать оценку плотности как функцию достаточных статистик.

Замечательная особенность несмещенных оценок $\pi^*(x; t_1, \ldots, t_k)$ заключается в том, что они, во всяком

случае, «не хуже» оценок $\pi(x;\ x_1,\ \ldots,\ x_l)$. И вот в каком смысле [35, 58].

Теорема. *Для всякой оценки* $\pi(x;\ x_1,\ \ldots,\ x_l)$ *найдется такая оценка* $\pi^*(x;\ t_1,\ \ldots,\ t_k)$, *что для любой плотности из* $P(x,\ \alpha)$ *математическое ожидание оценок совпадает*:

$$M\pi^*(x;\ t_1,\ \ldots,\ t_k) = M\pi(x;\ x_1,\ \ldots,\ x_l) = \pi(x),$$

а дисперсия оценки $\pi^*(x;\ t_1,\ \ldots,\ t_k)$ *не больше дисперсии оценки* $\pi^*(x;\ x_1,\ \ldots,\ x_l)$, *т. е.*

$$M(\pi(x) - \pi^*(x;\ t_1,\ \ldots,\ t_k))^2 \leqslant M(\pi(x) - \pi(x;\ x_1,\ \ldots,\ x_l))^2.$$

Из этой теоремы следует, что несмещенные оценки, выраженные через достаточные статистики, содержат наилучшую.

§ 10. Вычисление наилучшей несмещенной оценки

Построим наилучшую несмещенную оценку плотности многомерного нормального закона. При построении оценки существенно будет использован тот факт, что для распределения экспоненциального типа существует единственная несмещенная оценка, выраженная через достаточные статистики [26, 35]. Иначе говоря, существует единственное решение уравнения Фредгольма I рода

$$\int \pi^*(x;\ t_1,\ \ldots,\ t_k)\,P(t_1,\ \ldots,\ t_k;\ \alpha)\,dt_1,\ \ldots,\ dt_k = P(x,\ \alpha), \tag{3.42}$$

где $P(x,\ \alpha)$ — нормальный закон, а $P(t_1,\ \ldots,\ t_k;\ \alpha)$ — плотность распределения вероятностей его достаточных статистик.

Согласно теореме, приведенной в предыдущем параграфе, решением уравнения (3.42) в силу единственности является наилучшая несмещенная оценка плотности многомерного нормального закона.

Покажем, что несмещенная оценка плотности n-мерного нормального закона распределения вероятностей равна

$$P_{\text{н}}(x) = \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\,\Gamma\left(\frac{l-n-1}{2}\right)\,|S|^{1/2}}\left[1 - \frac{(x-x_{\text{э}})^{\mathrm{T}} S^{-1}(x-x_{\text{э}})}{l-1}\right]_{+}^{\frac{l-n-3}{2}}.$$

Здесь $x_э = \frac{1}{l}\sum_{i=1}^{l} x_i$ — вектор средних, $S = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_э)\times$ $\times (x_i - x_э)^т$ — эмпирическая оценка ковариационной матрицы Δ, выражение $[z]_+$ означает

$$[z]_+ = \begin{cases} z, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0. \end{cases}$$

При выводе наилучшей несмещенной оценки плотности n-мерного нормального закона воспользуемся формулой Байеса

$$\varphi(x_l \mid t) = \frac{q(x_l, t; \alpha)}{P(t; \alpha)}, \tag{3.43}$$

где $t = (t_1, \ldots, t_k)^т$, $x_l = (x_l^1, \ldots, x_l^n)^т$, плотность $q(x_l, t; \alpha)$ задает распределение статистик x_l, t, а плотность $P(t, \alpha)$ — распределение статистики t, $\varphi(x_l \mid t)$ — условная плотность. Покажем, что условная плотность (3.43) есть несмещенная оценка плотности $P(x, \alpha)$. В самом деле,

$$\int \varphi(x_l \mid t)\, P(t; \alpha)\, dt = \int q(x, t; \alpha)\, dt = P(x, \alpha).$$

А так как несмещенная оценка, выраженная через достаточные статистики, единственна, то $\varphi(x \mid t)$ и есть наилучшая несмещенная оценка. Вычислим $\varphi(x \mid t)$. Найдем сначала $q(x, t; \alpha)$. Для нормального закона появления вектора x

$$q(x, t; \alpha) = q(x, x_э, S; \mu, \Delta),$$

где

$$x = x_l, \quad x_э = \frac{1}{l}\sum_{i=1}^{l} x_l, \quad S = \frac{1}{l}\sum_{i=1}^{l}(x_l - x_э)(x_i - x_э)^т.$$

Пусть векторы $x_1, \ldots, x_l$, из которых образуются тройки x, $x_э$, S, появляются случайно и независимо согласно плотности $N(\mu, \Delta)$.

Рассмотрим векторы $y_1, \ldots, y_l$, полученные из $x_1 - \mu, \ldots, x_l - \mu$ ортогональным преобразованием

$$\mathscr{L} = \begin{vmatrix} c_{11} & \ldots & c_{1\,l-1} & 0 \\ \ldots & \ldots & \ldots & \ldots \\ c_{l-2\,1} & \ldots & c_{l-2\,l-1} & 0 \\ \frac{1}{\sqrt{l-1}} & \ldots & \frac{1}{\sqrt{l-1}} & 0 \\ 0 & \ldots & 0 & 1 \end{vmatrix}.$$

Векторы $y_1, \ldots, y_l$ распределены независимо по закону $N(0, \Delta)$. Справедливо

$$x_l = y_l + \mu, \quad x_э = \frac{\sqrt{l-1}}{l}\, y_{l-1} + \frac{y_l}{l} + \mu.$$

Выразим матрицу S через векторы $y_1, \ldots, y_l$. Для этого воспользуемся представлением

$$S = \frac{1}{l}\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} + \\ + \frac{(x_l-\mu)(x_l-\mu)^{\mathrm{T}}}{l} - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} - \\ - \frac{l-1}{l}\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right](x_l-\mu)^{\mathrm{T}} - \frac{l-1}{l}(x_l-\mu)\left[\sum_{i=1}^{l-1}\frac{x_i-\mu}{\sqrt{l-1}}\right]^{\mathrm{T}} - \\ - \frac{1}{l^2}(x_l-\mu)(x_l-\mu)^{\mathrm{T}}$$

и тем, что для преобразования $\mathscr{L}$ справедливо

$$\sum_{i=1}^{l-1}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} = \sum_{i=1}^{l-1} y_i y_i^{\mathrm{T}}.$$

В результате получим

$$S = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}} + \left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)\cdot\left(\frac{y_{l-1}-\sqrt{l-1}\,y_l}{l}\right)^{\mathrm{T}}.$$

Обозначим

$$\mathscr{D} = \frac{1}{l}\sum_{i=1}^{l-2} y_i y_i^{\mathrm{T}}.$$

Заметим, что векторы $y_1, \ldots, y_l$ распределены по нормальному закону $N(0, \Delta)$. Кроме того, элементы y_{l-1}, y_l, $\mathscr{D}$ независимы. Так как y_{l-1}, y_l распределены по нормальному закону, а $\mathscr{D}$ — по закону Уишарта, то совместное распределение $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ равно

$$P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta) = P(y_{l-1}; 0, \Delta)\, P(y_l; 0, \Delta)\, W_{l-1}(\mathscr{D}; \Delta), \quad (3.44)$$

где $W_{l-1}(\mathscr{D}, \Delta)$ — распределение Уишарта:

$$W_{l-1}(\mathscr{D}, \Delta) = \\ = \begin{cases} C_{n,\,l-1}\dfrac{|\mathscr{D}|^{\frac{l-n-3}{2}}\exp\left\{-\frac{1}{2}\operatorname{Sp}[\Delta^{-1}\mathscr{D}]\right\}}{|\Delta|^{\frac{l-2}{2}}}, & \text{если } |\mathscr{D}|>0, \\ 0, & \text{если } |\mathscr{D}|\leqslant 0, \end{cases}$$

$C_{n,\,l}$ — константа, определенная в (3.31).

Выразим теперь $P(y_{l-1}, y_l, \mathscr{D}; 0, \Delta)$ через элементы x_l, $x_э$, S. Для этого заметим, что

$$y_l = x_l - \mu, \quad y_{l-1} = \frac{l}{\sqrt{l-1}}(x_э - \mu) - \frac{(x_l - \mu)}{\sqrt{l-1}}, \tag{3.45}$$

$$\mathscr{D} = lS - \frac{l}{l-1}(x_l - x_э)(x_l - x_э)^{\mathrm{T}}.$$

Учитывая, что якобиан преобразования (3.45) равен $\frac{l^{\frac{n(n+3)}{2}}}{(l-1)^{n/2}}$, и подставляя (3.45) в (3.44), получим

$$q(x_l, x_э, S; \mu, \Delta) =$$
$$= \frac{l^{\frac{n(n+3)}{2}}}{(l-1)^{n/2}} P\left(\frac{l}{\sqrt{l-1}}(x_э - \mu) - \frac{(x_l - \mu)}{\sqrt{l-1}}; 0, \Delta\right) \times$$
$$\times P(x_l - \mu; 0, \Delta) W_{l-1}\left(lS - \frac{l}{l-1}(x_l - x_э)(x_l - x_э)^{\mathrm{T}}; \Delta\right),$$

откуда находим

$$q(x_l, x_э, S; \mu, \Delta) =$$
$$= \begin{cases} \dfrac{l^{\frac{n(n+3)}{2}} C_{n,\,l-1} \left| lS - \dfrac{l(x_l - x_э)(x_l - x_э)^{\mathrm{T}}}{l-1} \right|^{\frac{l-n-3}{2}} |\mathscr{D}|^{\frac{l}{2}}}{(2\pi)^n (l-1)^{\frac{n(l-1)}{2}} |\Delta|^{\frac{l}{2}} \exp\left\{\dfrac{l}{2} \operatorname{Sp}\left[\Delta^{-1}(S + (x_э - \mu)(x_э - \mu)^{\mathrm{T}})\right]\right\}}, \\ \qquad \text{если } \left| S - \dfrac{(x_l - x_э)(x_l - x_э)^{\mathrm{T}}}{l-1} \right| > 0, \\ 0, \quad \text{если } \left| S - \dfrac{(x_l - x_э)(x_l - x_э)^{\mathrm{T}}}{l-1} \right| = 0. \end{cases} \tag{3.46}$$

Найдем теперь знаменатель выражения (3.43).

Для нормального распределения векторов x статистики $x_э$ и lS распределены независимо:

$$P(x_э, S; \mu, \Delta) = P(x_э; \mu, \Delta) P(S; \Delta), \tag{3.47}$$

где $x_э$ распределено по нормальному закону $N\left(\mu, \frac{1}{l}\Delta\right)$, а lS — по закону Уишарта $W_l(S; \Delta)$, откуда следует, что

$$P(x_э, S; \mu, \Delta) =$$
$$= \frac{C_{n,l}}{(2\pi)^{n/2}} \frac{l^{n/2} |S|^{\frac{l-n-2}{2}}}{|\Delta|^{l/2} \exp\left\{\dfrac{l}{2} \operatorname{Sp}[\Delta^{-1}(S + (x_э - \mu)(x_э - \mu)^{\mathrm{T}})]\right\}}, \tag{3.48}$$

если $|S| \geqslant 0$ и равно нулю в противном случае, $C_{n,\,l}$ — константа, определенная в (3.31).

Подставляя (3.46) и (3.48) в (3.43), получаем

$$\varphi(x \mid t) = \frac{\Gamma\left(\frac{l-1}{2}\right)[(l-1)\pi]^{-n/2}}{\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}} \left(\frac{\left|S - \frac{(x-x_э)(x-x_э)^T}{l-1}\right|}{|S|}\right)^{\frac{l-n-3}{2}}$$

в случае, когда $|S| > 0$ и $\left|S - \frac{(x-x_э)(x-x_э)^T}{l-1}\right| \geqslant 0$. Заметим, что

$$\frac{\left|S - \frac{(x-x_э)(x-x_э)^T}{l-1}\right|}{|S|} = \left(1 - \frac{(x-x_э)^T S^{-1}(x-x_э)}{l-1}\right).$$

Откуда окончательно получаем

$$\varphi(x \mid x_э,\ S) =$$

$$= \frac{\Gamma\left(\frac{l-1}{2}\right)}{[(l-1)\pi]^{n/2}\,\Gamma\left(\frac{l-n-1}{2}\right)|S|^{1/2}} \left[1 - \frac{(x-x_э)^T S^{-1}(x-x_э)}{l-1}\right]_+^{\frac{l-n-3}{2}},$$

где обозначено

$$[z]_+ = \begin{cases} z, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0. \end{cases}$$

§ 11. Задача оценивания параметров плотности

Итак, казалось бы, нам удалось достичь своей цели — построить байесову оценку плотности, вычислить наилучшую несмещенную оценку. Однако методы, с помощью которых были получены эти оценки, существенно используют особые свойства восстанавливаемой плотности. Поэтому рассмотренные методы не являются регулярными для восстановления плотностей различных типов.

Вот почему представляют интерес методы, которые быть может не позволяют получать столь точные приближения, как рассмотренные, но зато являются регулярными, т. е. могут быть применены для восстановления плотностей из различных параметрических классов.

Чтобы получить такие методы, подменим задачу. Будем считать, что нашей целью является не восстановление плотности, а оценка параметров плотности.

При этом мы полагаем, что если удастся решить промежуточную задачу — найти хорошие оценки параметров плотности, то мы сможем удовлетворительно восстановить и саму плотность, приняв в качестве приближения плотности функцию $P(x, \alpha^*)$, где α^* — значения восстановленных параметров.

Заметим, что при восстановлении нормального закона ни байесово приближение, ни несмещенная оценка плотности не принадлежали классу нормальных законов. В случае же восстановления плотности путем оценивания ее параметров полученное приближение будет принадлежать классу нормальных. (Сам по себе этот факт не имеет никакого значения. Он лишь косвенно указывает, насколько далеким может оказаться полученное решение, например, от байесова.)

Итак, будем оценивать параметры α_0 плотности $P(x, \alpha_0)$. Определим качество оценки $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ вектора параметров $\alpha = \alpha_0$ по выборке $x_1, \ldots, x_l$ величиной

$$d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) = (\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l))^2,$$

качество оценки вектора параметров $\alpha = \alpha_0$ на выборках длины l — математическим ожиданием величины $d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l)$, т. е.

$$d(\alpha_0, \hat{\alpha}, l) = \int d(\alpha_0, \hat{\alpha}; x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \ldots dx_l,$$

где $P(x_1, \ldots, x_l; \alpha_0)$ — плотность распределения вероятностей выборки $x_1, \ldots, x_l$.

Наконец, качество оценки, предназначенной для восстановления параметра α при априорном распределении $P(\alpha)$, — величиной

$$R_{\text{Б}}(\hat{\alpha}, l) = \int d(\alpha, \hat{\alpha}, l) P(\alpha)\, d\alpha. \qquad (3.49)$$

Оценка $\hat{\alpha}$, доставляющая минимум функционалу (3.49), называется *байесовой оценкой параметров.*

Так же как и при восстановлении плотности, априорное распределение $P(\alpha)$ параметров α обычно неизвестно, поэтому, как и раньше, имеет смысл минимаксный критерий

$$R_{\text{м}}(\hat{\alpha}, l) = \sup_{\alpha} d(\alpha, \hat{\alpha}, l).$$

Вектор $\hat{\alpha}$, доставляющий минимум $R_{\text{м}}(\hat{\alpha}, l)$, образует *минимаксную оценку параметров.* Однако построение регулярного способа оценки параметров плотности связано не с байесовым и не с минимаксным приближением, а с идеей наилучшего несмещенного оценивания.

Определение. *Будем говорить, что оценка* $\hat{\alpha} = \hat{\alpha}(x_1, \ldots, x_l)$ *является несмещенной оценкой вектора параметров* α_0, *если*

$$\int \hat{\alpha}(x_1, \ldots, x_l) P(x_1, \ldots, x_l; \alpha_0) dx_1 \ldots dx_l = \alpha_0.$$

Рассмотрим сначала случай, когда плотность распределения вероятностей $P(x, \alpha_0)$ зависит лишь от скалярного параметра α_0. Тогда для класса несмещенных оценок справедливо замечательное *неравенство Рао — Крамера*:

$$\int (\alpha_0 - \hat{\alpha}(x_1, \ldots, x_l))^2 P(x_1, \ldots, x_l; \alpha_0) dx_1 \ldots dx_l \geqslant \geqslant \frac{1}{I_\Phi}, \qquad (3.50)$$

где

$$I_\Phi = -\int \frac{d^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{d\alpha^2} P(x_1, \ldots, x_l; \alpha_0) dx_1 \ldots dx_l.$$

Величина I_Φ получила название *информационного количества Фишера.* Для независимой выборки $x_1, \ldots, x_l$ она равна

$$I_\Phi = -l \int \frac{d^2 \ln P(x, \alpha_0)}{d\alpha^2} P(x, \alpha_0) dx.$$

Вывод неравенства Рао — Крамера есть во всех современных учебниках по статистике [35, 49, 58].

Смысл этого неравенства заключается в том, что дисперсия несмещенной оценки параметра (а для несмещенных оценок величина дисперсии определяет точность оценивания) не может быть меньше обратной величины информационного количества Фишера.

Таким образом, правая часть неравенства (3.50) определяет предельную точность несмещенного оценивания параметра. Оценка, при которой неравенство (3.50) переходит в равенство, называется *эффективной.* Проблема же состоит в том, чтобы найти регулярный способ пост-

роения эффективных оценок параметров для различных параметрических классов плотностей.

Неравенство, аналогичное (3.50), может быть получено и для одновременного несмещенного оценивания нескольких параметров. В этом случае аналогом информационного количества служит *информационная матрица Фишера* I, элементы которой I_{ij} есть величины

$$I_{ij} = \\ = -\int \frac{\partial^2 \ln P(x_1, \ldots, x_l; \alpha_0)}{\partial\alpha_i\, \partial\alpha_j} P(x_1, \ldots, x_l; \alpha_0)\, dx_1 \ldots dx_l,$$

$$i, j = 1, 2, \ldots, n.$$

Для независимой выборки $x_1, \ldots, x_l$ элементы I_{ij} равны

$$I_{ij} = -l \int \frac{\partial^2 \ln P(x, \alpha)}{\partial\alpha_i\, \partial\alpha_j}\, dx.$$

Пусть информационная матрица Фишера I не особенная, и пусть несмещенными оценками параметров $\alpha_1^0, \ldots, \alpha_n^0$ будут оценки $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$. Рассмотрим для этих оценок ковариационную матрицу B, т. е. матрицу с элементами

$$b_{ij} = M\left(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l)\right)\left(\alpha_j^0 - \hat{\alpha}_j(x_1, \ldots, x_l)\right).$$

Тогда аналогом неравенства Рао — Крамера в многомерном случае будет утверждение: для любого вектора z и любых несмещенных оценок $\hat{\alpha}_1(x_1, \ldots, x_l), \ldots, \hat{\alpha}_n(x_1, \ldots, x_l)$ справедливо неравенство

$$z^{\mathrm{T}} B z \geqslant z^{\mathrm{T}} I^{-1} z. \tag{3.51}$$

Смысл этого неравенства заключается в следующем: пусть качество совместной оценки n параметров $\alpha_1^0, \ldots, \alpha_n^0$ определяется квадратом взвешенной (с весами $z = (z_1, \ldots, z_n)^{\mathrm{T}}$, $z_i \geqslant 0$) суммы уклонений по всем оцениваемым параметрам:

$$T(x_1, \ldots, x_l) = \left(\sum_{i=1}^{n} z_i \left(\alpha_i^0 - \hat{\alpha}_i(x_1, \ldots, x_l)\right)\right)^2.$$

Тогда математическое ожидание $T(x_1, \ldots, x_l)$ ограничено снизу величиной $z^{\mathrm{T}} I^{-1} z$. Иначе говоря, в каком бы смысле (при каких конкретных весах z_i) не измерялось качество

совместного несмещенного оценивания n параметров, имеет место оценка

$$MT(x_1, \ldots, x_l) \geqslant z^{\mathrm{T}} I^{-1} z.$$

В частности, из неравенства (3.51) следует, что дисперсия оценки по каждому параметру в отдельности удовлетворяет неравенству (3.50). Действительно, неравенство (3.50) получается из (3.51) при конкретном векторе $z = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$. Методы оценивания, для которых при всех z неравенство (3.51) переходит в равенство, называются *совместно эффективными*.

При несмещенном оценивании нескольких параметров наша цель состоит в получении совместно эффективных оценок.

§ 12. Метод максимума правдоподобия

К сожалению, нет регулярного метода получения эффективных оценок параметров плотности по выборкам фиксированного объема.

Существует лишь метод, позволяющий строить асимптотически эффективные оценки. Этим методом является разработанный Р. Фишером *метод максимума правдоподобия* [58]. Однако, прежде чем рассмотреть метод максимума правдоподобия, введем некоторые понятия, необходимые для классификации оценок, получаемых по выборкам большого объема.

В предыдущем параграфе для характеристики оценок параметров распределения, найденных по выборкам фиксированного объема, была введена классификация, представленная на рис. 3.

На рисунке, кроме того, показана мера эффективности несмещенной оценки параметров α_0, которая в случае оценки одного параметра определяется величиной e_l, равной

$$e_l = \frac{1}{M(\alpha_0 - \hat{\alpha}(x_1, \ldots, x_\alpha))^2 I_\Phi}. \tag{3.52}$$

В случае же одновременного оценивания нескольких параметров мера эффективности определяется величиной

$$e_l = \frac{v(B, l)}{v(I, l)}, \tag{3.53}$$

равной отношению объема $v(B, l)$ эллипсоида

$$z^{\mathrm{T}}Bz = 1$$

к объему $v(I, l)$ эллипсоида

$$z^{\mathrm{T}}I^{-1}z = 1.$$

Для выборок большого объема предлагается несколько иная классификация, в которую введены понятия асимптотически несмещенных, состоятельных и асимптотически эффективных оценок.

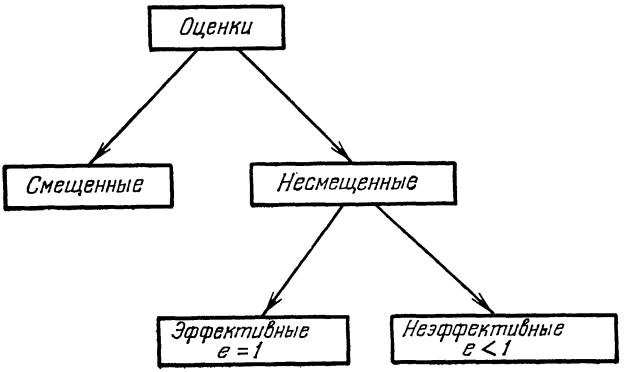

Рис. 3.

Асимптотически несмещенными называются оценки, для которых

$$M_{\alpha_0}\hat{\alpha}(x_1, \ldots, x_l) \underset{l\to\infty}{\longrightarrow} \alpha_0.$$

Состоятельными называются оценки, для которых

$$P_{\alpha_0}\{|\hat{\alpha}(x_1, \ldots, x_l) - \alpha_0| > \varepsilon\} \underset{l\to\infty}{\longrightarrow} 0$$

для всех $\varepsilon > 0$.

Асимптотически эффективными называются такие асимптотически несмещенные оценки, у которых величина

$$e_l \underset{l\to\infty}{\longrightarrow} 1,$$

где e_l в случае оценки одного параметра α определяется величиной (3.52), а при совместной оценке нескольких параметров — величиной (3.53). Такая классификация представлена на рис. 4.

Метод максимума правдоподобия связан с исследованием функции правдоподобия $P(x_1, \ldots, x_l; \alpha)$. В нашем случае, когда выборка $x_1, \ldots, x_l$ получена в результате случайных независимых испытаний согласно плотности

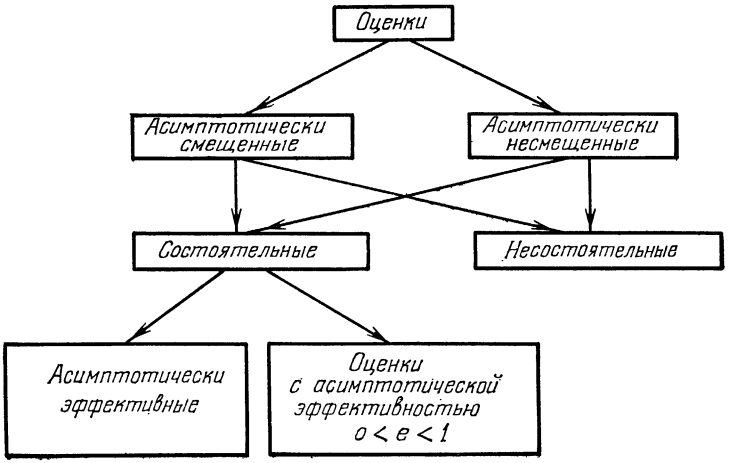

Рис. 4.

$P(x, \alpha)$, функция правдоподобия может быть представлена так:

$$P(x_1, \ldots, x_l; \alpha) = \prod_{i=1}^{l} P(x_i, \alpha). \tag{3.54}$$

Метод максимума правдоподобия состоит в том, чтобы в качестве оценки выбрать такое α, которое доставляет максимум функции (3.54). Наряду с функцией правдоподобия (3.54) принято рассматривать функцию

$$\ln P(x_1, \ldots, x_l; \alpha) = \sum_{i=1}^{l} \ln P(x_i, \alpha). \tag{3.55}$$

Максимумы функций (3.54) и (3.55) совпадают, и, следовательно, поиск оценок параметров плотности распределения вероятностей оказывается связанным с решениями системы уравнений

$$\frac{\partial P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \quad i = 1, 2, \ldots, n, \tag{3.56}$$

или системы уравнений

$$\frac{\partial \ln P(x_1, \ldots, x_l; \alpha)}{\partial \alpha_i} = 0, \quad i = 1, 2, \ldots, n. \qquad (3.57)$$

Теория метода максимума правдоподобия призвана определить, насколько оправдан предлагаемый метод оценивания параметров. Эта теория достаточно хорошо развита.

Основное ее содержание заключается в том, что для определенных классов $P(x, \alpha)$ (которым принадлежат все классы плотностей, рассматриваемые в этой книге) метод максимума правдоподобия обеспечивает асимптотическую эффективность оценки [24, 58].

Прежде чем приступить к оценке параметров плотности методом максимума правдоподобия, заметим, что получение оценок здесь свелось к более простой задаче, чем многократное интегрирование (как при реализации байесова метода) или решение уравнения Фредгольма I рода (при построении несмещенной оценки).

При реализации метода максимума правдоподобия необходимо решить систему уравнений (3.56) или (3.57). Хотя система не всегда линейная, численное ее решение обычно не составляет трудностей, тем более, что для широкого класса функций существует единственное решение.

§ 13. Оценивание параметров плотности вероятностей методом максимума правдоподобия

В этом параграфе, используя метод максимума правдоподобия, мы найдем оценки параметров распределения вероятностей для закона

$$P(x^i) = \begin{cases} p^i(1), & \text{если } x^i = c(1), \\ \cdots\cdots\cdots\cdots & \\ p^i(\tau_i), & \text{если } x^i = c(\tau_i), \end{cases} \qquad i = 1, 2, \ldots, n,$$

$$\sum_{j=1}^{\tau_i} p^i(j) = 1, \quad i = 1, 2, \ldots, n,$$

и параметров плотности нормального закона

$$N(\mu, \Delta) = \frac{1}{(2\pi)^{n/2} |\Delta|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Delta^{-1} (x-\mu)\right\}.$$

При этом окажется, что для закона $P(x^i)$ параметры равны

$$\hat{P}(x^i) = \begin{cases} \hat{p}^i(1) = \dfrac{m_i(1)}{l}, & \text{если} \quad x^i = c^i(1), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \hat{p}^i(\tau_i) = \dfrac{m_i(\tau_i)}{l}, & \text{если} \quad x^i = c^i(\tau_i), \end{cases} \tag{3.58}$$

где $m_i(j)$ — число векторов выборки, у которых i-я координата принимает значение $x^i = c^i(j)$.

Оценки параметров нормального закона, вычисленные методом максимума правдоподобия, равны

$$x_э = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

$$S = \frac{1}{l} \sum_{i=1}^{l} (x_i - x_э)(x_i - x_э)^т.$$

Таким образом, получаем оценку нормального закона

$$\hat{P}(x) = \frac{1}{(2\pi)^{n/2} \mid S \mid^{1/2}} \exp\left\{ -\frac{1}{2} (x - x_э)^т S (x - x_э) \right\}. \tag{3.59}$$

1°. Оценим параметры закона $P(x^i)$. Для этого составим функцию правдоподобия:

$$P(x_1, \ldots, x_l; p) = \prod_{j=1}^{l} \prod_{i=1}^{n} P\left(x_j^i, p^i\right),$$

где x_j^i — значение i-й координаты j-го вектора выборки.

Изменив порядок сомножителей, получим

$$P(x_1, \ldots, x_l; p) = \prod_{i=1}^{n} \prod_{j=1}^{l} P\left(x_j^i, p^i\right).$$

Перейдем теперь к функции

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{j=1}^{l} \ln P\left(x_j^i, p^i\right).$$

Рассмотрим величину

$$\sum_{j=1}^{l} \ln P\left(x_j^i, p^i\right).$$

Она может быть представлена в виде

$$\sum_{j=1}^{l} \ln P\left(x_j^i, p^i\right) = \sum_{r=1}^{\tau_i} m_i(r) \ln p^i(r),$$

где $m_l(r)$ — число векторов выборки, у которых координата x^l принимает значение $x^l = c^l(r)$.

Таким образом,

$$\ln P(x_1, \ldots, x_l; p) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_l} m_l(r) \ln p^i(r). \tag{3.60}$$

Найдем теперь максимум по $p^i(r)$ функции (3.60) при ограничениях $\sum_{r=1}^{\tau_l} p^i(r) = 1$, $i = 1, 2, \ldots, n$.

Для этого воспользуемся методом множителей Лагранжа. Составим функцию Лагранжа:

$$L(p, \lambda) = \sum_{i=1}^{n} \sum_{r=1}^{\tau_l} (m_l(r) \ln p^i(r) - \lambda_l p^i(r)), \tag{3.61}$$

где λ_i — множители Лагранжа.

Вектор p^i, доставляющий максимум функции $L(p, \lambda)$, определяется из системы уравнений

$$\frac{\partial L(p^i, \lambda)}{\partial p^i(r)} = \frac{m_l(r)}{p^i(r)} - \lambda_i = 0, \quad i = 1, \ldots, n. \tag{3.62}$$

Из (3.62), учитывая условия нормировки

$$\sum_{r=1}^{\tau_l} p^i(r) = 1,$$

получим

$$\hat{p}^i(r) = \frac{m_l(r)}{l}.$$

Заметим, что здесь оценка максимума правдоподобия оказалась несмещенной.

2°. Оценим теперь параметры μ, Δ нормального закона:

$$P(x; \mu, \Delta) = \frac{1}{(2\pi)^{n/2} |\Delta|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}} \Delta^{-1} (x-\mu)\right\}.$$

Составим функцию правдоподобия:

$$P(x_1, \ldots, x_l; \mu, \mathscr{D}) = \frac{|\mathscr{D}|^{l/2}}{(2\pi)^{ln/2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^{l} (x_l - \mu)^{\mathrm{T}} \mathscr{D} (x_l - \mu)\right\},$$

где обозначено $\Delta^{-1} = \mathscr{D}$.

Найдем ее логарифм:

$$\ln P(x_1, \ldots, x_l; \mu, \mathscr{D}) =$$

$$= -\frac{nl}{2} \ln 2\pi + \frac{l}{2} \ln |\mathscr{D}| - \frac{1}{2} \sum_{i=1}^{l} (x_l - \mu)^{\mathrm{T}} \mathscr{D} (x_l - \mu).$$

Запишем

$$\frac{\partial P(x_1, \ldots, x_l; \mu, \mathscr{D})}{\partial \mu} = \mathscr{D}\left(\sum_{i=1}^{l} x_i - l\mu\right) = 0, \tag{3.63}$$

$$\frac{\partial P(x_1, \ldots, x_l; \mu, \mathscr{D})}{\partial \mathscr{D}} = \frac{l}{2}\mathscr{D}^{-1} - \frac{1}{2}\sum_{i=1}^{l}(x_i - \mu)(x_i - \mu)^{\mathrm{T}} = 0. \tag{3.64}$$

Здесь использовано соотношение

$$\frac{d \ln |A|}{dA} = A^{-1}.$$

Из уравнений (3.63) и (3.64) находим

$$x_{\text{э}} = \frac{1}{l}\sum_{i=1}^{l} x_i, \quad S = \mathscr{D}^{-1} = \frac{1}{l}\sum_{i=1}^{l}(x_i - x_{\text{э}})(x_i - x_{\text{э}})^{\mathrm{T}}.$$

Оценка параметров ковариационной матрицы является смещенной.

§ 14. Замечания о различных методах приближения плотности

В этой главе мы рассмотрели три типа приближения плотностей, заданных с точностью до параметров: байесовы приближения, наилучшие несмещенные приближения и приближения, задаваемые параметрами, найденными методом максимума правдоподобия. Для наших частных задач восстановления плотностей двух конкретных классов (3.58) и (3.59) удалось получить все эти приближения. Какое же из приближений лучше использовать на практике, какое из них следует подставить в выражение (3.7) для получения решающих правил в задаче обучения распознаванию образов?

С теоретической точки зрения, безусловно, байесово. Это приближение экстремизирует функционал, который разумно определяет качество, предъявляемое к приближению. Однако для того, чтобы получить байесово приближение, необходимо знать априорное распределение параметров плотности, т. е. знать закон, определяющий, как часто на практике придется восстанавливать ту или иную конкретную плотность. Обычно этот закон неизвестен.

В §§ 6, 7 были получены байесовы приближения для таких априорных законов, которые, с одной стороны, содержат достаточно неопределенную информацию, а с дру-

гой стороны, способствуют максимальному упрощению вычислений. Насколько же можно доверять байесовому приближению, полученному по одному априорному закону, если на практике будет реализован другой закон? На этот вопрос существует лишь качественный ответ. С ростом объема выборки влияние априорной информации на байесово приближение падает (теорема Бернштейна).

Таким образом, выбор байесова приближения определяется верой в то, что на практике несоответствие в задании априорного закона скажется мало.

При конструировании наилучшей несмещенной оценки плотности нет необходимости учитывать априорную информацию. В этом классе оценок существует наилучшая оценка, которая не зависит от того, какие плотности из заданного класса придется восстанавливать. Казалось бы, в такой ситуации выбор наилучшей несмещенной оценки не связан ни с каким риском. На самом деле это не так. Ни откуда не следует, что в классе несмещенных оценок плотности имеются достаточно хорошие оценки. Ведь, как уже отмечалось, само свойство несмещенности оценки не имеет никакой самостоятельной ценности и вводится исключительно в целях ограничения класса оценок. Класс же несмещенных оценок узок (так, несмещенная оценка нормального закона, выраженная через достаточные статистики, единственна).

Не исключено, что сравнительно узкий класс несмещенных оценок состоит лишь из достаточно «плохих» оценок, и тогда выбор в нем наилучшей не гарантирует того, что оценка будет хорошей.

Подтверждением того, что такая ситуация вполне реальна, служит пример, приведенный Стейном: при оценке вектора μ средних n-мерного ($n > 2$) нормального закона с единичной ковариационной матрицей I равномерно лучшей оценкой, чем среднее арифметическое (наилучшая несмещенная оценка)

$$x_э = \frac{1}{l} \sum_{i=1}^{l} x_i,$$

является смещенная оценка

$$\hat{x}_э = \left(1 - \frac{n-2}{l x_э^T x_э}\right) x_э.$$

(Подробнее оценки стейновского типа будут рассмотрены в главе V.)

Пример Стейна замечателен тем, что он построен для самых простых задач оценивания параметров, и уже здесь существуют равномерно лучшие смещенные оценки.

Таким образом, выбор наилучшей несмещенной оценки определяется верой в то, что класс несмещенных оценок содержит достаточно хорошую.

Наконец, теория оценок максимального правдоподобия не дает никакого ответа на вопрос о том, каковы свойства оценок на конечных выборках. Теория лишь гарантирует приближение к эффективным оценкам с ростом объема выборки, т. е. что качество оценки максимального правдоподобия с ростом объема выборки приблизится к качеству наилучшей несмещенной оценки параметров.

Получилось так, что благодаря счастливому стечению обстоятельств, в этой главе нам удалось найти байесовы оценки — провести аналитическое интегрирование многократного интеграла (численное интегрирование в силу высокой кратности интеграла затруднительно), получить наилучшую несмещенную оценку плотности — найти аналитическое решение уравнения Фредгольма I рода (численное решение является некорректно поставленной задачей).

Однако такой результат связан с особенностью рассмотренного параметрического класса плотностей.

В общем случае вряд ли можно рассчитывать на получение подобных приближений. В этом отношении метод максимума правдоподобия имеет преимущество — он может быть применен для различных классов плотностей. Регулярность метода максимума правдоподобия связана с тем, что он сводится к решению алгебраических уравнений, т. е. к задаче, которая может эффективно решаться на вычислительных машинах.

И еще одно замечание. Рассмотренные в этой главе методы восстановления плотностей имеют смысл лишь при условии, что искомая плотность принадлежит заданному параметрическому семейству плотностей.

На практике же мы никогда не располагаем такой априорной информацией, которая позволяет выделить параметрическое семейство функций, заведомо содержащее искомую. Таким образом, оказывается, что не только вы-

бор того или иного метода приближения плотности, но и выбор самой постановки задачи восстановления зависимости по эмпирическим данным как параметрической во многом является вопросом веры.

Основные утверждения главы III

1. Решение задачи обучения распознаванию образов методами параметрической статистики связано с двумя проблемами:

— восстановлением по выборке

$$x_1,\ \omega_1;\ \ldots;\ x_l,\ \omega_l$$

плотности $\hat{P}(x, \omega)$, *известной с точностью до конечного числа параметров;*

— отысканием затем в классе решающих правил $F(x, \alpha)$ *правила, минимизирующего функционал*

$$I_э(\alpha) = \int (\omega - F(x, \alpha))^2 \hat{P}(x, \omega)\, dx\, d\omega.$$

2. Найти правило $F(x, \alpha_0)$, *минимизирующее функционал* $I_э(\alpha)$, *легко, когда класс решающих правил* $F(x, \alpha)$ *достаточно широк. Трудности возникают тогда, когда необходимо отыскать лучшее правило в узком классе решающих правил (например, линейном). Необходимость выбора правила из узкого (а не из широкого) класса возникает, когда плотность* $P(x, \omega)$ *восстанавливается недостаточно точно.*

3. Существуют различные понимания наилучшего алгоритма восстановления плотности: наилучший в среднем (байесов), наилучший для наиболее неблагоприятных условий (минимаксный), наилучший алгоритм из заданного множества (например, обеспечивающего несмещенность восстанавливаемых плотностей).

Каждое из этих определений порождает свой оптимальный алгоритм восстановления плотности. Практическая реализация этих алгоритмов часто оказывается сложной задачей.

Поэтому задачу восстановления плотности подменяют задачей оценивания параметров плотности. Оценивание параметров плотности проводится с помощью метода максимума правдоподобия. Экстремальные свойства этого метода проявляются лишь при больших объемах выборки.

4. Для плотностей $P(x|\omega)$, заданных нормальными законами и законами распределения вектора x с независимыми дискретными координатами, принимающими ограниченное число значений, могут быть найдены байесовы и наилучшие несмещенные приближения.

Для этих же законов методом максимума правдоподобия могут быть найдены оценки параметров плотностей.

ГЛАВА IV

МЕТОДЫ ПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ В ЗАДАЧЕ ВОССТАНОВЛЕНИЯ РЕГРЕССИИ

§ 1. Схема интерпретации результатов прямых экспериментов

В предыдущей главе методы параметрической статистики были применены для решения задачи обучения распознаванию образов: для минимизации функционала

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \qquad (4.1)$$

с неизвестной плотностью распределения вероятностей $P(x, y)$ по эмпирическим данным

$$x_1, y_1; \ldots; x_l, y_l \qquad (4.2)$$

сначала в параметрическом классе плотностей $\{P(x, y)\}$ восстанавливалась плотность $\hat{P}(x, y)$, затем с помощью плотности $\hat{P}(x, y)$ строился эмпирический функционал

$$I_э(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy, \qquad (4.3)$$

и наконец отыскивалось такое $\alpha_э$, которое доставляло минимум (4.3).

Для реализации этой схемы существенным было то, что координата y принимала лишь два значения — нуль и единица, множество $F(x, \alpha)$ было множеством характеристических функций, а плотность $P(x, y)$ была объединением двух плотностей. Все эти особенности определяют задачу обучения распознаванию образов.

В этой главе мы реализуем ту же самую схему минимизации риска, но применительно к задаче восстановления регрессии.

При решении этой задачи методами параметрической статистики принята своя модель плотности, отличная от той, которая рассматривалась в главе III. Считается, что случайная величина y и случайный вектор x связаны соотношением

$$y = F(x, \alpha_0) + \xi,$$

где $F(x, \alpha_0)$ — функция, принадлежащая классу $F(x, \alpha)$, а ξ — случайная не зависящая от x помеха, распределенная согласно плотности $P(\xi)$:

$$M\xi = 0, \quad M\xi^2 < \infty.$$

Таким образом, для всякого фиксированного x закон $P(\xi)$ индуцирует плотность условного распределения вероятностей величины y

$$P(y \mid x) = P(y - F(x, \alpha_0)). \tag{4.4}$$

Совместная же плотность $P(x, y)$ определяется законом

$$P(x, y) = P(y \mid x) P(x) = P(y - F(x, \alpha_0)) P(x), \tag{4.5}$$

где $P(x)$ — плотность распределения вероятностей вектора x.

Задачу восстановления регрессии $F(x, \alpha_0) \in F(x, \alpha)$ по случайной независимой выборке пар $x_1, y_1; \ldots; x_l, y_l$ можно интерпретировать как восстановление функциональной зависимости $F(x, \alpha_0)$ в классе $F(x, \alpha)$ по ее прямым измерениям, проводимым с аддитивной помехой в l случайно выбранных точках. В главе I такая задача была названа интерпретацией результатов прямых экспериментов.

Будем решать эту задачу методами параметрической статистики: восстановим плотность

$$\hat{P}(y \mid x) = \hat{P}(y - F(x, \alpha^*)),$$

а затем найдем точку минимума эмпирического функционала

$$I_э(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*)) P(x)\, dx\, dy. \tag{4.6}$$

Прежде всего покажем, что минимум функционала (4.6) достигается при $\alpha = \alpha^*$.

В самом деле, воспользуемся тождеством

$$\begin{aligned} I_э(\alpha) &= \int (y - F(x, \alpha))^2 \hat{P}(y - F(x, \alpha^*)) P(x)\, dx\, dy = \\ &= \int (y - F(x, \alpha^*))^2 \hat{P}(y - F(x, \alpha^*)) P(x)\, dx\, dy + \\ &\quad + \int (F(x, \alpha) - F(x, \alpha^*))^2 P(x)\, dx. \end{aligned} \tag{4.7}$$

Так как первое слагаемое правой части не зависит от α, то минимум $I_э(\alpha)$ достигается тогда, когда обращается в нуль неотрицательное второе слагаемое, т. е. при $\alpha = \alpha^*$.

Таким образом, значение вектора $\alpha = \alpha^*$, задающего условную плотность $\hat{P}(y \mid x) = \hat{P}(y - F(x, \alpha^*))$, немедленно определяет регрессию. Она равна $F(x, \alpha^*)$.

§ 2. Замечание о постановке задачи интерпретации результатов прямых экспериментов

В приведенной постановке задачи интерпретации результатов прямых экспериментов требуется, чтобы искомая функция $F(x, \alpha_0)$ принадлежала заданному параметрическому семейству $F(x, \alpha)$.

Это требование связано с тем, что плотность $P(y - F(x, \alpha))$ предполагается восстанавливать методами параметрической статистики (при восстановлении плотности параметрическими методами необходимо, чтобы искомая плотность принадлежала заданному семейству плотностей). Однако возможна и другая постановка, согласно которой неизвестная плотность $P(x, y)$ принадлежит заданному параметрическому множеству плотностей $P(x, y; \alpha)$, а искомая зависимость $F(x, \alpha_0)$ не принадлежит заданному множеству зависимостей $f(x, \beta)$. Иначе говоря, в схеме интерпретации результатов прямых экспериментов может быть поставлена задача: найти минимум функционала

$$I(\beta) = \int (y - f(x, \beta))^2 P(y - F(x, \alpha_0)) P(x)\, dy\, dx \qquad (4.8)$$

по выборке

$$x_1, y_1; \ldots; x_l, y_l,$$

если совместная плотность $P(x, y) = P(y - F(x, \alpha_0)) P(x)$ неизвестна, $F(x, \alpha_0) \in F(x, \alpha)$, а множество функций $f(x, \beta)$ не обязательно совпадает с $F(x, \alpha)$. Если $F(x, \alpha_0) \notin f(x, \beta)$, то минимум функционала (4.8) достигается на ближайшей к $F(x, \alpha_0)$ функции из $f(x, \beta)$. Близость здесь понимается в смысле L_2^P:

$$\rho_L(F, f) = \left(\int (F(x, \alpha_0) - f(x, \beta))^2 P(x)\, dx \right)^{1/2}.$$

Если же $F(x, \alpha_0) \in f(x, \beta)$, то минимум совпадает с регрессией. (Этот факт также немедленно следует из тождества (4.7).)

Таким образом, регрессия доставляет абсолютный минимум функционалу (4.8).

Для известной плотности $P(x)$ решение задачи минимизации функционала (4.8) также может быть проведено методами параметрической статистики: по выборке (4.2) восстанавливается плотность $\hat{P}(y - F(x, \alpha))$, а затем минимизируется эмпирический функционал

$$I_э(\beta) = \int (y - f(x, \beta))^2 \hat{P}(y - F(x, \alpha^*)) P(x)\, dx\, dy.$$

Заметим, что для задачи обучения распознаванию образов отыскание условного (в классе $f(x, \beta)$) минимума функционала (а не безусловного) составляло предмет исследования дискриминантного анализа.

Как указывалось в § 2 главы III, целесообразность такой постановки определялась ограниченностью объема выборки: по ограниченной выборке плотность восстанавливается неточно, и поэтому гарантированный минимум величины среднего риска может быть достигнут на функции, принадлежащей более узкому классу.

Совершенно аналогичная ситуация возникает и при интерпретации результатов прямых экспериментов по выборкам ограниченного объема: вследствие неточного определения плотности, гарантированная близость к регрессии может достигаться на функции, принадлежащей более узкому классу $f(x, \beta)$.

Методы сужения классов искомых зависимостей для получения меньшей гарантированной величины среднего риска будут рассмотрены в главе VIII.

§ 3. Ошибки измерений

Итак, для того чтобы восстановить регрессию в условиях схемы интерпретации результатов прямых экспериментов, достаточно восстановить плотность $P(y - F(x, \alpha_0))$, заданную с точностью до значения параметров α. Согласно же принятой модели параметрическое семейство плотностей $P(y - F(x, \alpha))$, содержащее искомую, определяется, во-первых, заданным параметрическим семейством функций $F(x, \alpha)$, содержащим регрессию $F(x, \alpha_0)$, а во-вторых, известной плотностью распределения вероятностей помехи $P(\xi)$.

Задание класса функций $F(x, \alpha)$, содержащего регрессию, является неформальным моментом в постановке задачи. Класс функций должен быть задан априори.

Что же касается плотности распределения вероятностей помехи, то с формальной точки зрения возможна любая плотность. Однако на практике при проведении прямых экспериментов возникают типичные ситуации, связанные с одинаковыми механизмами возникновения помехи при измерениях. Эти механизмы могут быть изучены.

При интерпретации результатов прямых экспериментов важную роль играют следующие три закона плотности распределения вероятностей помехи: равномерный закон, нормальный закон, закон Лапласа.

Равномерным законом плотности распределения вероятностей

$$P(\xi) = \begin{cases} \frac{1}{2\Delta}, & \text{если } |\xi| \leqslant \Delta, \\ 0, & \text{если } |\xi| > \Delta, \end{cases}$$

принято описывать ошибки округления. Пусть, например, значение некоторой большой величины x определяется с точностью до целых чисел. Тогда ошибка ξ, которая возникает вследствие округления до ближайшего целого числа, часто считается распределенной по закону

$$P(\xi)=\begin{cases}1, & \text{если } |\xi|\leqslant 0{,}5,\\ 0, & \text{если } |\xi|>0{,}5.\end{cases}$$

Нормальным законом плотности распределения вероятностей (*законом Гаусса*)

$$P(\xi)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}$$

принято описывать ошибки, которые возникают при физических измерениях, проводимых в одних и тех же условиях. Условия измерения определяют величину дисперсии σ^2. Так, например, ошибки при измерениях расстояния с помощью теодолита, проводимых в одних и тех же условиях (одинаковая освещенность, влажность, температура воздуха, степень запыленности атмосферы и т. п.) принято описывать нормальным законом.

Законом Лапласа

$$P(\xi)=\frac{1}{2\Delta}\exp\left\{\frac{|\xi|}{\Delta}\right\}$$

принято описывать ошибки, которые возникают при физических экспериментах, проводимых в меняющихся условиях. Например, если измерение расстояния происходит при различной степени облачности, в различное время суток, при различной запыленности атмосферы и т. п., то ошибки измерения принято описывать законом Лапласа.

Каждый из этих законов $P(\xi)$ порождает свое параметрическое множество плотностей $P(y-F(x,\alpha))$.

В этой главе для восстановления плотностей в различных параметрических множествах мы используем один и тот же регулярный метод — метод максимума правдоподобия. Выбор этого метода определяется тем, что его реализация не связана с техническими трудностями. Метод максимума правдоподобия может быть успешно реализован для всех интересующих нас параметрических множеств плотностей.

Итак, используем метод максимума правдоподобия для восстановления параметров условной плотности

$$P(y|x) = P(y - F(x, \alpha_0))$$

по случайной независимой выборке

$$x_1, y_1; \ldots; x_l, y_l,$$

распределенной согласно закону

$$P(x, y) = P(y - F(x, \alpha_0))\, P(x).$$

Для этого выпишем функцию правдоподобия

$$P(x_1, y_1, \ldots, x_l, y_l; \alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha))\, P(x_i). \quad (4.9)$$

Эта функция распадается на сомножители: сомножитель

$$P_1(\alpha) = \prod_{i=1}^{l} P(y_i - F(x_i, \alpha)), \quad (4.10)$$

который является функцией правдоподобия для условной плотности, и сомножитель

$$P_2 = \prod_{i=1}^{l} P(x_i).$$

Так как сомножитель P_2 не зависит от параметров α, то точки максимума (4.9) и (4.10) совпадают.

В дальнейшем максимизацию функции (4.10) будем также называть методом максимума правдоподобия.

Рассмотрим функцию правдоподобия $P_1(\alpha)$ для различных законов распределения помехи и найдем для них точку максимума.

Функция правдоподобия (4.10) для равномерного закона распределения помехи ξ имеет вид

$$P_1(\Delta, \alpha) = \prod_{i=1}^{l} \frac{1}{2\Delta}\, \delta_i(\alpha) = \frac{1}{(2\Delta)^l} \prod_{i=1}^{l} \delta_i(\alpha),$$

где

$$\delta_i(\alpha) = \begin{cases} 1, & \text{если } |y_i - F(x_i, \alpha)| \leqslant \Delta, \\ 0, & \text{если } |y_i - F(x_i, \alpha)| > \Delta. \end{cases}$$

Максимум функции правдоподобия определится такими α и Δ, для которых достигается минимум выражения

$$\Delta(\alpha) = \max_{x_i y_i} |y_i - F(x_i, \alpha)|, \tag{4.11}$$

т. е. α выбирается из условия минимизации наибольшего уклонения $F(x_i, \alpha)$ от y_i.

Для нормального закона плотности распределения вероятностей функция правдоподобия равна

$$P_1(\alpha, \sigma) = \frac{1}{(2\pi)^{l/2}\sigma^l} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(y_i - F(x_i, \alpha))^2\right\},$$

а метод максимального правдоподобия эквивалентен минимизации функционала

$$I_э(\alpha) = \sum_{i=1}^{l}(y_i - F(x_i, \alpha))^2. \tag{4.12}$$

Метод отыскания α путем минимизации функционала (4.12) называется *методом наименьших квадратов.*

Наконец, если считать, что помеха распределена по закону Лапласа, то функция правдоподобия равна

$$P_1(\Delta, \alpha) = \frac{1}{(2\Delta)^l} \exp\left\{-\frac{1}{\Delta}\sum_{i=1}^{l}|y_i - F(x_i, \alpha)|\right\},$$

и максимум правдоподобия достигается при таком векторе параметров α, при котором функционал

$$I_э(\alpha) = \sum_{i=1}^{l}|y_i - F(x_i\ \alpha)| \tag{4.13}$$

минимален. Метод минимизации функционала (4.13) получил название *метода минимальных модулей.*

Как указывалось в главе III, метод максимума правдоподобия является асимптотически эффективным методом оценивания параметров, поэтому все три алгоритма являются в определенном смысле оптимальными. Плохо лишь то, что каждый из них оптимален в своих условиях (в условиях равномерного, нормального или лапласовского распределения помехи), и решения, полученные с помощью этих алгоритмов, могут значительно различаться.

В самом деле, рассмотрим наиболее простую задачу восстановления зависимости — определение среднего значения случайной величины y по выборке объема l. Эта задача сводится к минимизации функционала

$$I(\alpha)=\int (y-\alpha)^2 P(y)\,dy \tag{4.14}$$

по выборке $y_1, \ldots, y_l$. Для нее метод минимизации наибольшего уклонения (4.11) приведет к следующему решению:

$$\alpha^*=\frac{y_{\min}+y_{\max}}{2}, \tag{4.15}$$

где $y_{\min}$ — наименьшее значение y в выборке, $y_{\max}$ — наибольшее значение, т. е. за оценку принимается величина, равная половине размаха выборки.

Метод наименьших квадратов (4.12) приведет к такой оценке параметра:

$$\alpha^*=\frac{1}{l}\sum_{i=1}^{l} y_i, \tag{4.16}$$

т. е. в качестве оценки среднего берется среднее арифметическое выборки.

Наконец, метод минимальных модулей (4.13) приводит к следующему решению: надо образовать вариационный ряд

$$y_{i_1}, \ldots, y_{i_l},$$

т. е. расположить элементы выборки в порядке неубывания, и затем найти значение среднего по формуле

$$\alpha^*=\begin{cases} y_{i_{k+1}}, & \text{если } l=2k+1, \\ \dfrac{y_{i_k}+y_{i_{k+1}}}{2}, & \text{если } l=2k. \end{cases}$$

§ 4. Экстремальные свойства законов Гаусса и Лапласа

В предыдущем параграфе было показано, что алгоритмы восстановления регрессии, полученные методами параметрической статистики зависят от принятой модели помехи. Поэтому необходимо уметь описывать ситуации, в которых следует применять ту или иную модель. Указывалось, что равномерной плотностью распределения вероятностей следует описывать помехи, которые возникают при округлении величин, нормальным законом плотности распределения вероятностей (законом Гаусса) — помехи, возникающие при измерениях, проводимых в одних и тех же условиях, законом Лапласа описываются помехи, возникающие

при измерениях в меняющихся условиях. Хотелось бы придать этим рекомендациям точный смысл.

В этом параграфе мы установим некоторые замечательные свойства нормального закона и закона Лапласа. Мы увидим, что нормальный закон обладает определенными экстремальными свойствами при абсолютной стабильности условий измерения, в то время как закон Лапласа обладает аналогичными экстремальными свойствами при «максимальной нестабильности» условий измерения.

Итак, покажем, что из всех непрерывных законов плотности распределения вероятностей, имеющих заданную величину дисперсии, нормальный закон обладает наибольшей энтропией. Иначе говоря, нормальный закон — это такой закон появления помехи, при котором величина замера будет зафиксирована в наибольшей степени неопределенно.

Будем оценивать степень неопределенности измерений в случае, когда ошибки определяются плотностью распределения вероятностей $P(\xi)$, величиной энтропии Шеннона:

$$H(P) = -\int_{-\infty}^{\infty} P(\xi) \ln P(\xi)\, d\xi. \tag{4.17}$$

Найдем функцию $P(\xi)$, удовлетворяющую ограничениям:

$$P(\xi) \geqslant 0, \tag{4.18}$$

$$\int_{-\infty}^{\infty} P(\xi)\, d\xi = 1, \tag{4.19}$$

$$\int_{-\infty}^{\infty} \xi P(\xi)\, d\xi = 0, \tag{4.20}$$

$$\int_{-\infty}^{\infty} \xi^2 P(\xi)\, d\xi = \sigma^2, \tag{4.21}$$

на которой достигается максимум энтропии (4.17). Здесь ограничения (4.18), (4.19) следуют из определения плотности, ограничение (4.20) отражает требование несмещенности помехи, а ограничение (4.21) фиксирует класс плотностей с заданной дисперсией.

Эта задача решается стандартным приемом — составляется функция Лагранжа, учитывающая ограничения

(4.19)—(4.21):

$$L = -(P(\xi)\ln P(\xi) + \lambda_1 P(\xi) + \lambda_2 \xi P(\xi) + \lambda_3 \xi^2 P(\xi)).$$

Выписывается уравнение Эйлера

$$\frac{\partial L}{\partial P} = -(\ln P(\xi) + 1 + \lambda_1 + \lambda_2 \xi + \lambda_3 \xi^2) = 0. \qquad (4.22)$$

Решение уравнения (4.22)

$$P(x) = \exp\{-(\lambda_1 + 1 + \xi\lambda_2 + \xi^2\lambda_3)\}$$

удовлетворяет ограничению (4.18) и, следовательно, определяет искомую плотность.

Для определения констант λ_1, λ_2, λ_3 используются условия (4.19) — (4.21), из которых легко устанавливается, что

$$P(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}, \qquad (4.23)$$

т. е. из всех непрерывных законов плотности распределения вероятностей, имеющих заданную величину дисперсии, нормальный закон обладает наибольшей энтропией (случайная величина распределена наиболее неопределенно).

Рассмотрим теперь более сложную модель образования помехи ξ. Каждый раз случайная величина ξ является реализацией нормального закона $P_N(\xi \mid \sigma^2)$, имеющего нулевое среднее и некоторую дисперсию σ^2. Однако каждый раз реализуется нормальный закон, имеющий свою величину дисперсии. Эта величина назначается случайно и независимо согласно плотности $P(\sigma^2)$. Таким образом, образуется закон

$$P_\Lambda(x) = \int P_N(\xi \mid \sigma^2) P(\sigma^2)\, d\sigma^2. \qquad (4.24)$$

Такая схема хорошо отражает часто встречающиеся в практике случаи, когда при фиксированных условиях измерения реализуется нормальный закон. Однако условия измерения меняются случайно и независимо, и, таким образом, закон плотности распределения вероятностей задается композицией двух законов. В примере с измерением расстояния множитель $P_N(x \mid \sigma^2)$ в (4.24) отражает ошибки, которые возникали бы при работе в одних и тех же атмосферных условиях. Сомножитель $P(\sigma^2)$ отражает случайный характер состояния атмосферы. Если

условия измерений не меняются (крайний случай, когда $P(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$, где $\delta(z)$ — дельта-функция), то композиция (4.24) определяет нормальный закон. Мы же рассмотрим другой крайний случай, когда условия эксперимента меняются относительно среднего в «наибольшей степени неопределенно», т. е. когда функция $P(\sigma^2)$ такова, что доставляет максимум энтропии

$$H(P) = -\int_0^\infty P(\sigma^2) \ln P(\sigma^2)\, d\sigma^2, \tag{4.25}$$

и при этом удовлетворяет ограничениям:

$$P(\sigma^2) \geqslant 0, \tag{4.26}$$

$$\int P(\sigma^2)\, d\sigma^2 = 1, \tag{4.27}$$

$$\int_0^\infty \sigma^2 P(\sigma^2)\, d\sigma^2 = 2\Delta^2. \tag{4.28}$$

Ограничения (4.26) и (4.27) вытекают из определения плотности распределения вероятностей. Ограничение же (4.28) задает средние условия проведения эксперимента.

Итак, найдем максимум энтропии (4.25) при условиях (4.26) — (4.28). Для этого выпишем функцию Лагранжа, учитывающую ограничения (4.27) и (4.28),

$$L = -(P(\sigma^2) \ln P(\sigma^2) + \lambda_1 P(\sigma^2) + \lambda_2 \sigma^2 P(\sigma^2)).$$

Получим уравнение Эйлера

$$\frac{\partial L}{\partial P} = -(\ln P(\sigma^2) + 1 + \lambda_1 + \lambda_2 \sigma^2) = 0. \tag{4.29}$$

Решение уравнения (4.29)

$$P(\sigma^2) = \exp\{-(\lambda_1 + 1 + \lambda_2 \sigma^2)\}$$

удовлетворяет ограничению (4.26) и, следовательно, определяет искомую плотность. Для определения констант λ_1 и λ_2 решение (4.29) подставим в (4.27) и (4.28), откуда получим $\lambda_1 + 1 = -\ln 2\Delta^2$ и $\lambda_2 = \frac{1}{2\Delta^2}$.

Таким образом, «наиболее неопределенные» условия проведения эксперимента задаются плотностью

$$P(\sigma^2) = \frac{1}{2\Delta^2} \exp\left\{\frac{\sigma^2}{2\Delta^2}\right\}. \tag{4.30}$$

Покажем теперь, что плотность распределения вероятностей $P_\Delta(\xi)$, заданная композицией плотностей (4.23) и (4.30), есть закон Лапласа, т. е.

$$P_\Delta(\xi) = \frac{1}{\sqrt{2\pi}\,2\Delta^2}\int_0^\infty \frac{1}{\sigma}\exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}\exp\left\{-\frac{\sigma^2}{2\Delta}\right\}d\sigma^2 = \\ = \frac{1}{2\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}. \quad (4.31)$$

Для того чтобы вычислить интеграл (4.31), воспользуемся следующим фактом, справедливым для интегрируемой на $(-\infty, \infty)$ функции

$$\int_0^\infty f\left[\left(\frac{x}{a}-\frac{b}{x}\right)^2\right]dx = a\int_0^\infty f(y^2)\,dy \qquad (a,\ b>0). \quad (4.32)$$

Для доказательства тождества (4.32) положим $y=\frac{x}{a}-\frac{b}{x}$. Тогда

$$\int_{-\infty}^\infty f(y^2)\,dy = \int_0^\infty f\left[\left(\frac{x}{a}-\frac{b}{x}\right)^2\right]\left(\frac{1}{a}+\frac{b}{x^2}\right)dx = \\ = \frac{1}{a}\int_0^\infty f\left[\left(\frac{x}{a}-\frac{b}{x}\right)^2\right]dx + b\int_0^\infty f\left[\left(\frac{x}{a}-\frac{b}{x}\right)^2\right]\frac{dx}{x^2}.$$

Но последний интеграл подстановкой $x=-ab/t$ приводится к виду

$$\frac{1}{a}\int_{-\infty}^0 f\left[\left(\frac{t}{a}-\frac{b}{t}\right)^2\right]dt.$$

Таким образом,

$$\int_{-\infty}^\infty f(y^2)\,dy = \frac{1}{a}\int_{-\infty}^\infty f\left[\left(\frac{x}{a}-\frac{b}{x}\right)^2\right]dx.$$

Отсюда (в силу четности подынтегральной функции) и получаем тождество (4.32).

Преобразуем теперь левую часть выражения (4.31)

$$P_{\Delta}(\xi)=\frac{1}{2\sqrt{2\pi}\,\Delta^2}\int_0^{\infty}\frac{1}{\sigma}\exp\left\{-\left(\frac{\sigma^2}{2\Delta^2}+\frac{\xi^2}{2\sigma^2}\right)\right\}d\sigma^2=$$

$$=\frac{1}{\sqrt{2\pi}\,\Delta^2}\exp\left\{-\frac{|\xi|}{\Delta}\right\}\int_0^{\infty}\exp\left\{-\frac{1}{2}\left(\frac{\sigma}{\Delta}-\frac{|\xi|}{\sigma}\right)^2\right\}d\sigma. \qquad (4.33)$$

Из (4.33) в силу тождества (4.32) находим

$$P_{\Delta}(\xi)=\frac{1}{\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}\int_0^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{y^2}{2}\right\}dy=$$

$$=\frac{1}{2\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}. \qquad (4.34)$$

То есть композиция (4.31) нормального закона и закона (4.30) задает закон плотности распределения вероятностей Лапласа (4.34).

Итак, мы установили, что при фиксированных условиях проведения эксперимента наиболее неопределенный результат измерения будет получен, если помеха распределена по нормальному закону, если же условия эксперимента меняются относительно некоторого среднего наиболее неблагоприятно, то самый неопределенный результат измерений будет получен, когда помеха возникает в соответствии с законом Лапласа.

Таким образом, выбор закона Гаусса или закона Лапласа зависит от того, являются ли условия эксперимента абсолютно стабильными или наиболее нестабильными.

На практике же редко реализуются крайние случаи. Поэтому ни закон Гаусса, ни закон Лапласа, как правило, не выполняются. Принято считать, что имеют место «промежуточные» случаи.

Таким образом, возникает ситуация, когда мы оцениваем регрессию в предположении, что справедлив некоторый гипотетический закон распределения помехи, например, Гаусса или Лапласа, в то время как на самом деле реализуется другой закон («промежуточный»). В какой мере окажутся полезными в этой ситуации рассмотренные алгоритмы (4.11) — (4.13)? Иначе говоря, вопрос заключается в том, в какой мере устойчивыми к изменению закона распределения помехи являются рассмотренные

алгоритмы восстановления зависимостей и как следует конструировать устойчивые алгоритмы? Ответ на этот вопрос составляет содержание следующих параграфов главы.

§ 5. Об устойчивых методах оценивания параметра сдвига

Пусть плотность распределения вероятностей помехи неизвестна. Известно только, что она принадлежит некоторому заданному множеству плотностей $\{P(\xi)\}$. Ниже мы уточним характер этих множеств, а пока примем, что множества выпуклы и что функции плотности распределения вероятностей имеют две непрерывные производные и являются симметричными относительно нуля. (Требование симметричности плотности является принципиальным для всей рассматриваемой ниже теории.) Изучению подлежит следующий вопрос: как в заданном классе $\{P(\xi)\}$ следует выбирать гипотетическую плотность распределения вероятностей помехи, чтобы возможная ошибка как можно меньше сказалась на оценке параметров регрессии, если известно, что истинная плотность принадлежит $\{P(\xi)\}$.

Рассмотрим сначала простой случай: требуется оценить математическое ожидание m случайной величины x по выборке $x_1, \ldots x_l$.

Такая задача эквивалентна задаче оценивания параметра сдвига m плотности $P(x-m)$ (здесь использован факт, что помеха ξ связана с замером x равенством $\xi = x-m$). При известной плотности $P(\xi)$ оценку $\hat{m}$ параметра сдвига m будем проводить методом максимума правдоподобия, т. е. максимизируя выражение

$$R(m) = \sum_{i=1}^{l} \ln P(x_i - m). \tag{4.35}$$

В этом случае оценка $\hat{m}$ является состоятельной и асимптотически эффективной.

Однако если функция $P(\cdot)$ в выражении (4.35) не совпадает с функцией плотности вероятностей помехи $P(\xi)$, оценка $\hat{m}$, доставляющая максимум выражению (4.35), вообще говоря, не является ни состоятельной, ни асимптотически эффективной.

Обозначим значение $\hat{m}$, максимизирующее (4.35) в предположении, что $P(\xi) = P_\Gamma(\xi)$, через $\hat{m} = \hat{m}(x_1, \ldots, x_l; P_\Gamma(\xi))$. Условимся, как будем измерять точность оценивания параметра, если мы считаем, что помеха распределена по закону $P_\Gamma(\xi) \in \{P(\xi)\}$, в то время как на самом деле имеет место закон $P_0(\xi) \in \{P(\xi)\}$.

Естественно за точность оценки параметра $\hat{m}$, найденной по выборке $x_1, \ldots, x_l$ в предположении, что помеха распределена по закону $P_\Gamma(\xi)$, принять величину

$$d(P_\Gamma(\xi); x_1, \ldots, x_l) = (\hat{m}(x_1, \ldots, x_l; P_\Gamma(\xi)) - m)^2,$$

равную квадрату уклонения найденного значения параметра сдвига от истинного значения. Точность оценивания параметра сдвига на выборках длины l естественно характеризовать математическим ожиданием величины $d(P_\Gamma(\xi); x_1, \ldots, x_l)$:

$$D(P_0, P_\Gamma) = Md(P_\Gamma(\xi); x_1, \ldots, x_l) = \int (\hat{m}(x_1, \ldots, x_l; P_\Gamma(\xi)) - m)^2 P_0(x_1 - m) \ldots P_0(x_l - m) \cdot dx_1 \ldots dx_l. \quad (4.36)$$

Величина $D(P_0, P_\Gamma)$ зависит от двух законов плотности распределения вероятностей, принадлежащих одному и тому же классу плотностей $\{P(\xi)\}$: от гипотетического закона $P_\Gamma(\xi)$ (согласно этому закону строилась оценка параметра сдвига $\hat{m}$) и истинного закона $P_0(\xi)$ (согласно этому закону вычислялась средняя величина квадрата уклонения).

В дальнейшем мы используем представление функции $D(P_0, P_\Gamma)$ в виде

$$D(P_0, P_\Gamma) = \frac{1}{l} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\, d\xi\right)^2} = \frac{1}{l} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \frac{P'_\Gamma(\xi) P'_0(\xi)}{P_\Gamma(\xi)}\, d\xi\right)^2}. \quad (4.37)$$

Получим это представление, проведя, возможно, недостаточно строгие, но зато наглядные рассуждения. Строгая теория устойчивого оценивания изложена в [88].

Не ограничивая общности, будем считать, что искомое значение параметра сдвига m равно нулю. Обозначим

$$f(\xi) = \frac{P'_{\Gamma}(\xi)}{P_{\Gamma}(\xi)} = (\ln P_{\Gamma}(\xi))'.$$

Тогда, согласно методу максимума правдоподобия, оценка $\hat{m}$ параметра $m=0$ находится из условия

$$\left(\sum_{i=1}^{l} \ln P_{\Gamma}(x_i - \hat{m})\right)' = \sum_{i=1}^{l} f(x_i - \hat{m}) = 0.$$

Воспользуемся приближением, справедливым для больших l и рассматриваемых симметричных плотностей:

$$\sum_{i=1}^{l} f(x_i - \hat{m}) \approx \sum_{i=1}^{l} f(x_i) - \hat{m} \sum_{i=1}^{l} f'(x_i) = 0, \quad (4.38)$$

откуда получим

$$\hat{m} = \frac{\sum_{i=1}^{l} f(x_i)}{\sum_{i=1}^{l} f'(x_i)}.$$

Пусть l настолько велико, что

$$\hat{m} \approx \frac{\frac{1}{l}\sum_{i=1}^{l} f(x_i)}{\int f'(x)\, P_0(x)\, dx}.$$

(При выводе этого соотношения мы допустили, что интеграл в знаменателе существует. Для этого достаточно, чтобы функции $f'(x)$ были ограничены. В дальнейшем будем рассматривать только такие плотности, для которых $|(\ln P(\varepsilon))''| < \text{const.}$)

Вычислим теперь $D(P_0, P_\Gamma) = M\hat{m}^2$:

$$D(P_0, P_\Gamma) = \int \hat{m}^2 P_0(x_1), \ldots, P_0(x_l)\, dx_1, \ldots, dx_l =$$

$$= \frac{1}{l^2} \frac{1}{\left[\int f'(x) P_0(x)\, dx\right]^2} \int \sum_{i,j}^{l} f(x_i) f(x_j) P_0(x_1) \ldots$$

$$\ldots P_0(x_l)\, dx_1 \ldots dx_l.$$

Так как плотности $P_0(x)$, $P_\Gamma(x)$ — функции симметричные, то

$$\int f(x_i) f(x_j) P_0(x_1) \ldots P_0(x_l)\, dx_1 \ldots dx_l = 0, \quad i \neq j.$$

Таким образом, получаем для больших l

$$D(P_0, P_\Gamma) = \frac{1}{l^2} \frac{\int \sum_{i=1}^{l} f^2(x_i) P(x_l)\, dx_l}{\left(\int f'(x) P_0(x)\, dx\right)^2} = \frac{1}{l} \frac{\int f^2(x) P_0(x)\, dx}{\left(\int f'(x) P_0(x)\, dx\right)^2}.$$

Наконец, возвращаясь к старым обозначениям, получаем представление (4.37).

Итак, мы определили критерий качества оценивания параметра сдвига в условиях, когда истинная плотность равна $P_0(\xi)$, а гипотетическая — $P_\Gamma(\xi)$. Наша цель состоит в том, чтобы выбрать такую плотность $P_\Gamma(\xi)$, которая минимизирует $D(P_0, P_\Gamma)$. Если бы плотность $P_0(\xi)$ была известна, то, как легко показывается (см. ниже), минимум $D(P_0, P_\Gamma)$ достигался бы при $P_\Gamma(\xi) = P_0(\xi)$.

Проблема заключается в том, чтобы выбрать функцию $P_\Gamma(\xi)$, если известно лишь, что $P_0(\xi)$ принадлежит классу $\{P(\xi)\}$.

Как обычно в подобных ситуациях, принимается одна из двух постановок: байесова или минимаксная.

В первом случае считается, что априори известна вероятность того, что истинной плотностью будет та или иная плотность из $\{P(\xi)\}$, качество же оценки параметра сдвига определяется как среднее (по мере $\mu(P)$) качество, т. е.

$$D_\text{Б}(P_\Gamma) = \int D(P_0, P_\Gamma)\, d\mu(P_0).$$

Минимаксная постановка предлагает принять за оценку качества величину $D(P_0, P_\Gamma)$, вычисленную для наиболее

неблагоприятной плотности $P_0(\xi) \in \{P(\xi)\}$, т. е. вычислять качество из условия

$$D_{\text{м}}(P_{\text{г}}) = \max_{P_0} D(P_0,\ P_{\text{г}}).$$

Конструирование оптимального по Байесу решения наталкивается здесь на значительные трудности, поэтому ниже мы будем изучать минимаксные решения.

Итак, будем определять качество оценки параметра сдвига, полученного с помощью гипотетической плотности $P_{\text{г}}(\xi)$, величиной

$$D_{\text{м}}(P_{\text{г}}) = \max_{P_0} D(P_0,\ P_{\text{г}}) = \max_{P_0} \frac{\int \left(\frac{P'_{\text{г}}(\xi)}{P_{\text{г}}(\xi)}\right)^2 P_0(\xi)\, d\xi}{I\left(\int \frac{P'_{\text{г}}(\xi)\, P'_0(\xi)}{P_{\text{г}}(\xi)}\, d\xi\right)^2}, \tag{4.39}$$

и попытаемся найти такую гипотетическую плотность $P_{\text{г}}(\xi)$, которая минимизирует (4.39).

Такая постановка задачи имеет игровую интерпретацию. Существуют два игрока — природа и статистик, имеющие один и тот же набор стратегий (функций $\{P(\xi)\}$) и противоположные цели. Первый игрок (природа) стремится выбрать такую стратегию (назначить истинную плотность $P_0(\xi)$), чтобы максимизировать потери второго игрока, второй игрок выбирает такую стратегию (гипотетическую плотность $P_{\text{г}}(\xi)$), чтобы минимизировать свои потери. Величина же потерь определяется функционалом (4.39).

Требуется найти оптимальную стратегию второго игрока, т. е. уметь для заданного класса плотностей выбирать такую гипотетическую плотность, использование которой гарантирует минимальные потери при самой неблагоприятной истинной плотности. Найденную плотность будем называть *устойчивой в классе* $\{P(\xi)\}$, а метод оценивания параметра сдвига, полученный применением метода максимума правдоподобия к найденной плотности, — *методом устойчивого оценивания параметра сдвига.*

Важным фактом теории устойчивого оценивания параметра сдвига является то, что на выпуклом множестве $\{P(\xi)\}$ игра с функцией потерь (4.39) имеет седловую

точку, т. е.

$$\max_{P_0 \in \{P(\xi)\}} \min_{P_\Gamma \in \{P(\xi)\}} D(P_0, P_\Gamma) = \min_{P_\Gamma \{P(\xi)\}} \max_{P_0 \in \{P(\xi)\}} D(P_0, P_\Gamma).$$

Используя этот факт, можно найти оптимальную стратегию статистика против природы.

Для этого воспользуемся неравенством Шварца

$$\left(\int a(x)\, b(x)\, d\mu(x)\right)^2 \leqslant \int a^2(x)\, d\mu(x) \int b^2(x)\, d\mu(x). \quad (4.40)$$

Преобразуем с помощью этого неравенства знаменатель (4.37):

$$D(P_0, P_\Gamma) = \frac{1}{I} \frac{\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P'_\Gamma(\xi)}{P_\Gamma(\xi)} \frac{P'_0(\xi)}{P_0(\xi)}\right) P_0(\xi)\, d\xi\right)^2} \geqslant \frac{1}{I \int \left(\frac{P'_0(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi}. \quad (4.41)$$

Заметим, что при $P_\Gamma(\xi) = P_0(\xi)$ справедливо

$$D(P_0, P_0) = \frac{1}{I \int \left(\frac{P'_0(\xi)}{P_0(\xi)}\right)^2 P_0(\xi)\, d\xi}. \quad (4.42)$$

Из (4.41) и (4.42) следует, что минимум (4.39) достигается тогда, когда $P_\Gamma(\xi) = P_0(\xi)$, т. е. оптимальные стратегии природы и статистика реализуются на одной и той же плотности. Для того чтобы ее найти, необходимо в классе $\{P(\xi)\}$ максимизировать выражение (4.42) или, что то же самое, найти в классе $\{P(\xi)\}$ такую плотность, которая минимизирует функционал

$$I_\Phi(P) = I \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi.$$

Заметим, что функционал $I_\Phi(P)$ есть информационное количество Фишера (см. § 11 гл. III).

В §§ 7, 8 для различных классов плотностей мы найдем плотности распределения вероятностей, минимизирующие информационное количество Фишера, и тем самым получим устойчивые (в этих классах) оценки параметра сдвига, а пока распространим полученный здесь результат на случай оценивания параметров регрессии.

§ 6. Устойчивое оценивание параметров регрессии

Пусть теперь надо восстановить регрессию. Будем полагать, что класс функций, в котором ведется восстановление и которому принадлежит регрессия, представим в виде

$$F(x, \alpha) = \sum_{r=1}^{n} \alpha_r \varphi_r(x),$$

где $\varphi_r(x)$ — система линейно независимых функций.

Как и раньше, истинная и гипотетическая плотности помехи $P_0(\xi)$ и $P_\Gamma(\xi)$ принадлежат классу $\{P(\xi)\}$.

Для восстановления параметров регрессии используем метод максимума правдоподобия, т. е. найдем вектор α, доставляющий максимум выражению

$$\ln P_\Gamma(x_1, y_1; \ldots; x_l, y_l; \alpha) = \sum_{i=1}^{l} \ln P_\Gamma\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right). \tag{4.43}$$

Пусть этот вектор есть $\alpha = \alpha^*$. Рассмотрим вектор уклонений найденных значений параметров регрессии α^* от истинных α_0:

$$\bar{\alpha} = (\alpha_0 - \alpha^*).$$

Образуем ковариационную матрицу B:

$$B = M\bar{\alpha} \cdot \bar{\alpha}^{\mathrm{T}},$$

которая определяет качество оценивания вектора параметров α (см. § 11 гл. III).

Ниже, совершенно аналогично (4.37), мы получим, что для достаточно больших l справедливо равенство

$$B = \frac{1}{l} \frac{\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\, d\xi}{\left(\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\, d\xi\right)^2} \| k_{ij} \|^{-1}, \tag{4.44}$$

где

$$k_{ij} = \frac{1}{l} \sum_{t=1}^{l} \varphi_i(x_t)\, \varphi_j(x_t) \cong \int \varphi_i(x)\, \varphi_j(x)\, P_0(x)\, dx.$$

Таким образом, элементы матрицы B пропорциональны величине

$$D(P_0,\ P_\Gamma)=\frac{1}{l}\,\frac{\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 P_0(\xi)\,d\xi}{\left(\int\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' P_0(\xi)\,d\xi\right)^2}.$$

В представлении (4.44) важно, что от плотностей $P_0(\xi)$, $P_\Gamma(\xi)$ зависит лишь коэффициент пропорциональности $D(P_0,\ P_\Gamma)$ (а не матрица $\|k_{ij}\|$).

Поэтому двум различным гипотетическим плотностям $P_\Gamma(\xi)$ и $\hat{P}_\Gamma(\xi)$ соответствуют две квадратичные формы z^TB_1z и z^TB_2z с равными матрицами $\|k_{ij}\|$, но различными величинами $D(P_0,\ P_\Gamma)$. Эти формы находятся в одном из отношений: в отношении

$$z^TB_1z \geqslant z^TB_2z \quad \text{для любого } z$$

или отношении

$$z^TB_1z < z^TB_2z \quad \text{для любого } z,$$

в зависимости от того, какой из коэффициентов больше: $D(P_0,\ P_\Gamma)$ или $D(P_0,\ \hat{P}_\Gamma)$. В § 11 гл. III было показано, что минимум формы z^TBz определяет совместно эффективные оценки параметров.

Таким образом, величина коэффициента $D(P_0,\ P_\Gamma)$ определяет качество оценивания параметров линейной регрессии: чем меньше $D(P_0,\ P_\Gamma)$, тем выше качество.

Это означает, что и в случае оценивания параметров регрессии задача выбора устойчивой плотности приводится к рассмотренной игре природы и статистика. В предыдущем параграфе было показано, что в этой игре оптимальная стратегия статистика состоит в том, чтобы в классе плотностей $\{P(\xi)\}$ выбрать плотность, доставляющую минимум величине информационного количества Фишера

$$I_\Phi(P)=l\int\left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\,d\xi. \tag{4.45}$$

Таким образом, для того чтобы найти наилучшую гипотетическую модель помехи в классе $\{P(\xi)\}$, надо в этом классе найти функцию, доставляющую минимум (4.45). Эту плотность и будем использовать для определения

параметров регрессии с помощью метода максимума правдоподобия.

Нам осталось получить соотношение (4.44). Получим его аналогично выводу (4.37). Обозначим $f(\xi)=\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}$. Тогда максимум функции правдоподобия (4.43) достигается на тех значениях α, которые удовлетворяют уравнениям

$$\sum_{i=1}^{l} f\left(\xi_i-\sum_{r=1}^{n}\bar{\alpha}_r\varphi_r(x_i)\right)\varphi_k(x_i)=0,\quad k=1,\ 2,\ \ldots,\ n.$$

Воспользуемся приближением (4.38):

$$\sum_{i=1}^{l} f\left(\xi_i-\sum_{r=1}^{n}\bar{\alpha}_r\varphi_r(x_i)\right)\varphi_k(x_i)\approx$$

$$\approx\sum_{i=1}^{l}\left[f(\xi_i)-f'(\xi_i)\sum_{r=1}^{n}\bar{\alpha}_r\varphi_r(x_i)\right]\varphi_k(x_i)=0.$$

Отсюда для достаточно больших l в силу независимости ξ_i и x_i получаем

$$\frac{1}{l}\sum_{i=1}^{l} f(\xi_i)\varphi_k(x_i)-$$

$$-\int f'(\xi)P_0(\xi)\,d\xi\sum_{i=1}^{l}\left(\sum_{r=1}^{n}\bar{\alpha}_r\varphi_r(x_i)\right)\varphi_k(x_i)=0,\ k=1,2,\ldots,n.$$

Или в векторной форме

$$\|k_{ij}\|\bar{\alpha}\approx\frac{1}{l}\frac{H}{\int f'(\xi)P_0(\xi)\,d\xi},\qquad(4.46)$$

где H — вектор-столбец с координатами $h_r=\sum_{i=1}^{l}\varphi_r(x_i)f(\xi_i)$.

Из (4.46) получим

$$\bar{\alpha}=\frac{1}{l}\frac{1}{\int f'(\xi)P_0(\xi)\,d\xi}\|k_{ij}\|^{-1}H.$$

Найдем теперь ковариационную матрицу:

$$B=M\bar{\alpha}\bar{\alpha}^{\mathrm{T}}\doteq\frac{1}{l}\frac{\int f^2(\xi_i)P_0(\xi)\,d\xi}{\left(\int f'(\xi)P_0(\xi)\,d\xi\right)^2}\|k_{ij}\|^{-1}.$$

Здесь мы полагаем, что матрица $\|k_{ij}\|$ не вырождена. Возвращаясь к исходным обозначениям, получим (4.44).

§ 7. Устойчивость законов Гаусса и Лапласа

Покажем, что законы Гаусса и Лапласа являются, каждый в своем классе, устойчивыми. Как указывалось в предыдущем параграфе, для этого достаточно показать, что в соответствующих классах плотностей $\{P(\xi)\}$ законы Гаусса и Лапласа доставляют минимум величине информационного количества Фишера (4.45).

Для тех конкретных классов $\{P(\xi)\}$, которые будут рассмотрены ниже, эта задача оказывается задачей неклассического вариационного исчисления (класс $\{P(\xi)\}$ задается ограничениями типа неравенств). Поэтому здесь мы не будем получать гипотетические плотности регулярным способом, т. е. решать неклассические вариационные задачи, а укажем решения и затем установим, что они действительно определят седловую точку функции

$$D(P, P_{\Gamma}) = \frac{1}{l} \frac{\int \left(\frac{P_{\Gamma}'(\xi)}{P_{\Gamma}(\xi)}\right)^2 P(\xi)\, d\xi}{\left(\int \left(\frac{P_{\Gamma}'(\xi)}{P_{\Gamma}(\xi)}\right)' P(\xi)\, d\xi\right)^2}.$$

Иначе говоря, надо будет проверить, что для указанной плотности $P_{\Gamma}(\xi)$ выполняются неравенства

$$D(P, P_{\Gamma}) \leqslant D(P_{\Gamma}, P_{\Gamma}) \leqslant D(P_{\Gamma}, P).$$

Заметим, что, согласно (4.41), всегда имеет место одно из неравенств, а именно:

$$D(P_{\Gamma}, P_{\Gamma}) \leqslant D(P_{\Gamma}, P).$$

Таким образом, для доказательства оптимальности выбранной стратегии достаточно установить справедливость неравенства

$$D(P, P_{\Gamma}) \leqslant D(P_{\Gamma}, P_{\Gamma}). \tag{4.47}$$

Рассмотрим следующие классы плотностей.

1. Класс плотностей с ограниченной дисперсией. Соответствующая вариационная задача состоит в том, чтобы минимизировать функционал (4.45) в классе

функций, удовлетворяющих условиям:

$$\begin{aligned} &1) && P(\xi) > 0, \\ &2) && \int P(\xi)\, d\xi = 1, \\ &3) && \int \xi P(\xi)\, d\xi = 0, \\ &4) && \int \xi^2 P(\xi)\, d\xi \leqslant \sigma^2. \end{aligned} \tag{4.48}$$

Условия 1), 2), 3) определяют плотность помехи, условие 4) — ограниченность дисперсии. Решением этой некласcической (вследствие 1) и 4)) вариационной задачи будет плотность

$$P(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}.$$

Действительно, подставим

$$P_{\Gamma}(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}$$

в неравенство (4.47). Получим

$$\frac{\int \frac{\xi^2}{\sigma^4} P(\xi)\, d\xi}{\left(\frac{1}{\sigma^2}\int P(\xi)\, d\xi\right)^2} = \int \xi^2 P(\xi)\, d\xi \leqslant \sigma^2. \tag{4.49}$$

Неравенство (4.49) справедливо для любой плотности из (4.48), так как класс (4.48) состоит из плотностей, для которых величина дисперсии не превосходит σ^2. Таким образом, нормальный закон плотности распределения вероятностей с нулевым средним и дисперсией σ^2 является устойчивым законом в классе всех плотностей с дисперсией, ограниченной величиной σ^2.

2. Рассмотрим теперь класс невырожденных плотностей. Этому классу принадлежат такие плотности, для которых $P(0) \geqslant 1/2\Delta$. Покажем, что устойчивым законом в этом классе плотностей будет закон Лапласа.

Для этого подставим $P_{\Gamma}(\xi) = \frac{1}{2\Delta} \exp\left\{-\frac{|\xi|}{\Delta}\right\}$ в (4.47). Получим

$$\frac{\int \left(\frac{\operatorname{sign}\xi}{\Delta}\right)^2 P(\xi)\, d\xi}{\frac{4}{\Delta^2} P^2(0)} = \frac{1}{4P^2(0)} \leqslant \Delta^2,$$

или, что то же самое,

$$P(0) \geqslant \frac{1}{2\Delta}.$$

А так как в класс $\{P(\xi)\}$ входят плотности, у которых $P(0) \geqslant 1/2\Delta$, то неравенство (4.47) выполнится для любых функций класса.

Таким образом, закон Лапласа, является устойчивым законом в классе плотностей, для которых $P(0) \geqslant \frac{1}{2\Delta}$.

Устойчивость законов Гаусса и Лапласа (каждого в своем классе) является фактом не менее замечательным, чем экстремальные свойства этих законов, установленные в § 4.

Итак, законы Гаусса и Лапласа устойчивы. Однако классы плотностей, в которых они устойчивы, часто оказываются слишком широкими. И тогда более содержательная статистическая модель должна быть построена на основе других более узких классов плотностей.

Ниже в §§ 8, 9 мы рассмотрим некоторые конкретные классы плотностей и получим для них устойчивые плотности.

§ 8. Класс плотностей, образованных смесью плотностей

Рассмотрим *класс H плотностей, образованных смесью*

$$P(\xi) = g(\xi)(1-\varepsilon) + \varepsilon h(\varepsilon) \tag{4.50}$$

некоторой фиксированной симметричной относительно начала координат плотности $g(\xi)$ и любой симметричной относительно начала координат плотности $h(\xi)$. Смесь составляется в пропорции $1-\varepsilon$ и ε. Для таких классов плотностей справедлива следующая теорема.

Теорема 4.1 (Хубер). *Пусть* — $\ln g(\xi)$ — *дважды непрерывно дифференцируемая выпуклая функция. Тогда в классе H существует устойчивая плотность*

$$P_{\Gamma}(\xi) = \begin{cases} (1-\varepsilon)\, g(\xi_0) \exp\{k(\xi-\xi_0)\}, & \text{если } \xi < \xi_0, \\ (1-\varepsilon)\, g(\xi), & \text{если } \xi_0 \leqslant \xi < \xi_1, \\ (1-\varepsilon)\, g(\varepsilon_1) \exp\{-k(\xi-\xi_1)\}, & \text{если } \xi \geqslant \xi_1, \end{cases} \tag{4.51}$$

где ξ_0, ξ_1 — *концы интервала* $[\xi_0, \xi_1]$, *на котором монотонная* (*вследствие выпуклости* $-\ln g(\xi)$) *функция* $\frac{g'(\xi)}{g(\xi)}$ *ограничена по модулю константой* k, *определяемой из условия нормировки плотности*

$$1 = (1-\varepsilon)\int_{\xi_0}^{\xi_1} g(\xi)\,d\xi + \frac{g(\xi_0)+g(\xi_1)}{k}(1-\varepsilon).$$

Доказательство. Для доказательства этой теоремы, так же как и при установлении устойчивости законов Гаусса и Лапласа, надо показать, что для функций класса (4.50) справедливо

$$D(P,\ P_\Gamma) \leqslant D(P_\Gamma,\ P_\Gamma) \leqslant D(P_\Gamma,\ P).$$

Справедливость оценки

$$D(P_\Gamma,\ P_\Gamma) \leqslant D(P_\Gamma,\ P),$$

как уже указывалось, следует из неравенства Шварца (4.40). Поэтому для доказательства теоремы достаточно показать справедливость оценки

$$D(P,\ P_\Gamma) \leqslant D(P_\Gamma,\ P_\Gamma)$$

для любой функции $P(\xi) \in H$.

Представим плотность $P_\Gamma(\xi)$ в виде смеси фиксированной плотности $g(\xi)$ и плотности $\hat{h}(\xi) = \frac{P_\Gamma(\xi) - (1-\varepsilon)g(\xi)}{\varepsilon}$. Выпишем плотность $\hat{h}(\xi)$, учитывая (4.51):

$$\hat{h}(\xi) = \begin{cases} \frac{1-\varepsilon}{\varepsilon}\left(g(\xi_0)\exp\{k(\xi-\xi_0)\} - g(\xi)\right), & \text{если } \xi < \xi_0, \\ 0, & \text{если } \xi_0 \leqslant \xi < \xi_1, \\ \frac{1-\varepsilon}{\varepsilon}\left(g(\xi_1)\exp\{-k(\xi-\xi_1)\} - g(\xi)\right), & \text{если } \xi \geqslant \xi_1. \end{cases} \tag{4.52}$$

Нетрудно убедиться, что $\hat{h}(\xi)$ есть плотность. Действительно, $\int \hat{h}(\xi)\,d\xi = 1$, а $\hat{h}(\xi) \geqslant 0$, так как по условию теоремы $-\ln g(\xi)$ — выпуклая функция и, следовательно, целиком лежит над касательной

$$-\ln g(\xi) \geqslant -\ln g(\xi_i) - (-1)^i k(\xi - \xi_i), \quad i = 0,\ 1. \tag{4.53}$$

Неравенство же (4.53) эквивалентно утверждению

$$g(\xi) \leqslant g(\xi_i) \exp\{(-1)^i k(\xi - \xi_i)\}, \quad i = 0, 1.$$

Рассмотрим неравенство

$$\frac{\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 [(1-\varepsilon) g(\xi) + \varepsilon h(\xi)]\, d\xi}{\left(\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' [(1-\varepsilon) g(\xi) + \varepsilon h(\xi)]\, d\xi\right)^2} \leqslant$$

$$\leqslant \frac{(1-\varepsilon) \int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)^2 g(\xi)\, d\xi + \varepsilon k^2}{(1-\varepsilon)^2 \left(\int \left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' g(\xi)\, d\xi\right)^2}. \qquad (4.54)$$

Убедимся, что правая часть этого неравенства есть точная верхняя грань выражения, стоящего слева, для произвольных симметричных плотностей $h(\xi)$. Для этого заметим, что функция $\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}$ равна

$$\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)} = \begin{cases} k, & \text{если } \xi < \xi_0, \\ \frac{g'(\xi)}{g(\xi)}, & \text{если } \xi_0 \leqslant \xi < \xi_1, \\ -k, & \text{если } \xi \geqslant \xi_1, \end{cases}$$

где, согласно условию теоремы $\left|\frac{g'(\xi)}{g(\xi)}\right| \leqslant k$, а функция $\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)'$ равна

$$\left(\frac{P_\Gamma'(\xi)}{P_\Gamma(\xi)}\right)' = \begin{cases} 0, & \text{если } \xi < \xi_0, \\ \left(\frac{g'(\xi)}{g(\xi)}\right)', & \text{если } \xi_0 \leqslant \xi < \xi_1, \\ 0, & \text{если } \xi \geqslant \xi_1. \end{cases}$$

Таким образом, для того чтобы максимизировать левую часть неравенства, необходимо выбрать такую плотность $h(\xi)$, которая сосредоточена на отрезках $(-\infty, \xi_0)$ и (ξ_1, ∞). Такая плотность одновременно обеспечивает максимум числителю и минимум знаменателю выражения, стоящего в левой части неравенства. Значение же выражения, стоящего слева, при этом будет в точности равно значению правой части неравенства. Плотность (4.52) как

раз и принадлежит плотностям, сосредоточенным на отрезках $(-\infty, \xi_0)$, (ξ_1, ∞). Теорема доказана.

Эта теорема замечательна тем, что позволяет конструировать различные устойчивые плотности. В частности, если выбрать в качестве $g(\xi)$ плотность нормального закона

$$g(\xi) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\},$$

и рассмотреть класс плотностей

$$P(\xi) = \frac{(1-\varepsilon)}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} + \varepsilon h(\xi),$$

то, согласно теореме, устойчивой в классе будет плотность

$$P_{\Gamma}(\xi) = \begin{cases} \dfrac{(1-\varepsilon)}{\sqrt{2\pi}\,\sigma} \exp\left\{\dfrac{k^2}{2} - \dfrac{k}{\sigma}|\xi|\right\}, & \text{если } |\xi| \geqslant k\sigma, \\ \dfrac{(1-\varepsilon)}{\sqrt{2\pi}\,\sigma} \exp\left\{-\dfrac{\xi^2}{2\sigma^2}\right\}, & \text{если } |\xi| < k\sigma, \end{cases}$$

где величина k определяется из условия нормировки

$$1 = \frac{(1-\varepsilon)}{\sqrt{2\pi}\,\sigma}\left[\int_{-k\sigma}^{k\sigma} \exp\left\{-\frac{\xi^2}{2\sigma^2}\right\} d\xi + \frac{2\exp\left\{-\dfrac{k^2}{2}\right\}}{k}\right].$$

Этот закон является «промежуточным» между законом Гаусса и законом Лапласа. На отрезке $|\xi| < k\sigma$ он с точностью до нормировочного множителя совпадает с законом Гаусса, а на отрезках $|\xi| \geqslant k\sigma$ — с законом Лапласа.

§ 9. Плотности, сосредоточенные на отрезке

Рассмотрим еще один важный класс плотностей и найдем устойчивую в нем плотность распределения вероятностей.

Рассмотрим *класс K_p плотностей, сосредоточенных в основном на отрезке* $[-A, A]$, т. е. класс плотностей $P(\xi)$, для элементов которого выполняется условие

$$\int_{-A}^{A} P(\xi)\, d\xi \geqslant p,$$

где p — известный параметр, задающий класс K_p.

Покажем, что устойчивой в этом классе будет плотность

$$P_{\Gamma}(\xi) = \begin{cases} \frac{1}{A}\left(\frac{b}{1+b}\cos^2\frac{a\xi}{A}\right), & \text{если } \left|\frac{\xi}{A}\right| < 1, \\ \frac{1}{A}\left(\frac{b}{1+b}\cos^2 a\right)\exp\left\{-2b\left(\left|\frac{\xi}{A}\right|-1\right)\right\}, & \text{если } \left|\frac{\xi}{A}\right| \geqslant 1, \end{cases} \tag{4.55}$$

где параметры a, b связаны с константой p, задающей класс K_p, соотношениями

$$\begin{aligned} p &= 1 - \frac{\cos^2 a}{1+b}, \\ b &= a \cdot \operatorname{tg} a, \qquad 0 < a < \frac{\pi}{2}. \end{aligned} \tag{4.56}$$

Не ограничивая общности, будем полагать $A = 1$ (случай $A \neq 1$ приводится к случаю $A = 1$ подстановкой $z = A\xi$). Таким образом, задача состоит в том, чтобы показать, что в классе плотностей, удовлетворяющих условию

$$\int_{-1}^{1} P(\xi)\, d\xi \geqslant p,$$

устойчивой будет плотность

$$P_{\Gamma}(\xi) = \begin{cases} \frac{b}{1+b}\cos^2 \xi a, & \text{если } |\xi| < 1, \\ \frac{b}{1+b}\cos^2 a \exp\{-2b(|\xi|-1)\}, & \text{если } |\xi| \geqslant 1. \end{cases} \tag{4.57}$$

Для того чтобы показать устойчивость в K_p плотности (4.57), достаточно показать, что $P_{\Gamma}(\xi)$ минимизирует в K_p функционал Фишера

$$I_{\Phi} = I \int \left(\frac{P'(\xi)}{P(\xi)}\right)^2 P(\xi)\, d\xi. \tag{4.58}$$

Однако мы не будем непосредственно минимизировать функционал (4.58), а воспользуемся тем, что необходимым и достаточным условием того, что $P_{\Gamma}(\xi)$ есть точка мини-

мума (4.58) является неотрицательность в K_p функционала

$$R(P_\Gamma, P) = \\ = l\int (2(-\ln P_\Gamma(\xi))'' - [(\ln P_\Gamma(x))']^2)(P(\xi) - P_\Gamma(\xi))\,d\xi. \quad (4.59)$$

Функционал $R(P_\Gamma, P)$ есть производная по параметру ε выражения

$$I_\Phi((1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)),$$

вычисленная в точке $\varepsilon = 0$, т. е.

$$\left.\frac{dI_\Phi((1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi))}{d\varepsilon}\right|_{\varepsilon=0} = R(P_\Gamma, P). \quad (4.60)$$

Неотрицательность производной в точке $\varepsilon = 0$ для плотностей $(1-\varepsilon)P_\Gamma(\xi) + \varepsilon P(\xi)$ (по любому направлению в K_p) означает, что на $P_\Gamma(\xi)$ достигается минимум I_Φ.

Итак, проверим неотрицательность выражения $R(P_\Gamma, P)$. В силу четности функции $R(P_\Gamma, P)$ достаточно проверить положительность ее на луче $0 \leqslant \xi < \infty$. Для этого найдем, что

$$(-\ln P_\Gamma(\xi))' = \begin{cases} 2a\,\mathrm{tg}\,a\xi, & \text{если } |\xi| < 1, \\ 2b\,\mathrm{sign}\,\xi, & \text{если } |\xi| \geqslant 1. \end{cases} \quad (4.61)$$

Подставим (4.61) в (4.59) и произведем вычисления

$$R(P_\Gamma, P) = \\ = 4a^2 l\int_0^1 (P(\xi) - P_\Gamma(\xi))\,d\xi - 4b^2 l\int_1^\infty (P(\xi) - P_\Gamma(\xi))\,d\xi. \quad (4.62)$$

Преобразуем выражение (4.62):

$$R(P_\Gamma, P) = \\ = 4a^2 l\int_0^1 (P(\xi) - P_\Gamma(\xi))\,d\xi - 4b^2 l\int_1^\infty (P(\xi) - P_\Gamma(\xi))\,d\xi = \\ = 4(a^2 + b^2)\,l\int_0^1 (P(\xi) - P_\Gamma(\xi))\,d\xi.$$

Таким образом, выражение $R(P_\Gamma, P)$ неотрицательно для всех $P(\xi)$, для которых

$$\int_{-1}^1 P(\xi)\,d\xi \geqslant \int_{-1}^1 P_\Gamma(\xi)\,d\xi = 1 - 2\int_1^\infty P_\Gamma(\xi)\,d\xi = p,$$

т. е. для всех функций из K_p.

§ 10. Устойчивые методы восстановления регрессии

В предыдущих параграфах мы рассмотрели некоторые классы плотностей и нашли для них устойчивые плотности.

Теперь в нашей схеме интерпретации результатов прямых экспериментов можно ослабить требования к априорной информации о статистических свойствах помехи. Достаточно знать класс плотностей, которому она принадлежит. В этом случае для восстановления параметров регрессии методами параметрической статистики можно использовать устойчивую в классе плотность вместо истинной. Конечно при такой подмене ухудшается асимптотическая скорость сходимости параметров регрессии. Она становится пропорциональной не предельно достижимой для несмещенного оценивания параметра сдвига величине (см. § 11 гл. III)

$$I_{\min} = \frac{1}{I \int \left(\frac{P_0'(\xi)}{P_0(\xi)} \right)^2 P_0(\xi)\, d\xi},$$

($P_0(\xi)$ — истинная плотность помехи), а величине I, лежащей в интервале

$$I_{\min} \leqslant I \leqslant I_{\max},$$

где

$$I_{\max} = \sup_{P(\xi) \in \{P(\xi)\}} \frac{1}{I \int \left(\frac{P'(\xi)}{P(\xi)} \right)^2 P(\xi)\, d\xi}.$$

Однако если класс $\{P(\xi)\}$ не очень широкий, такая возможная потеря в скорости может быть не слишком большой.

Основным конструктивным результатом рассмотренной здесь теории устойчивого оценивания является выделение четырех классов плотностей с указанием в них устойчивой плотности [1]).

Выпишем эти классы и плотности.

1. Класс плотностей с дисперсией, ограниченной величиной σ^2. Устойчивой плотностью

---

[1]) Известны и другие классы плотностей, для которых найдены устойчивые плотности [46].

в этом классе является плотность нормального закона

$$P_{\Gamma}(\xi)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{\xi^2}{2\sigma^2}\right\}.$$

2. Класс невырожденных плотностей (для которых $P(0)>1/2\Delta$). Устойчивой в этом классе является плотность

$$P_{\Gamma}(\xi)=\frac{1}{2\Delta}\exp\left\{-\frac{|\xi|}{\Delta}\right\}.$$

3. Класс плотностей, образованный смесью известной плотности, например нормальной $P_N(\xi)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{\xi^2}{2\sigma^2}}$, с любой возможной плотностью, взятыми в пропорции $1-\varepsilon$ и ε. Устойчивой в этом классе является плотность

$$P_{\Gamma}(\xi)=\begin{cases} c\exp\left\{-\dfrac{\xi^2}{2\sigma^2}\right\}, & \text{если } |\xi|<k\sigma,\\ c\exp\left\{\dfrac{k^2}{2}-k\left|\dfrac{\xi}{\sigma}\right|\right\}, & \text{если } |\xi|\geqslant k\sigma, \end{cases}$$

где c, k — определяемые через ε и σ константы.

4. Класс плотностей, сосредоточенных в основном на отрезке $[-A, A]$ $\left(\int\limits_{-A}^{A}P(\xi)\,d\xi\geqslant p\right)$. Устойчивой в этом классе является плотность

$$P_{\Gamma}(\xi)=\begin{cases} c\cos^2\dfrac{a\xi}{A}, & \text{если } \left|\dfrac{\xi}{A}\right|<1,\\ c\cos^2 a\exp\left\{-2b\left(\left|\dfrac{\xi}{A}\right|-1\right)\right\}, & \text{если } \left|\dfrac{\xi}{A}\right|\geqslant 1, \end{cases}$$

где c, a, b — определяемые через A и p константы.

Теперь, если вместо истинной плотности помехи $P_0(\xi)$ взять устойчивую в классе $P_{\Gamma}(\xi)$, определить с ее помощью плотность условного распределения вероятностей

$$P_{\Gamma}\left(y-\sum_{r=1}^{n}\alpha_r\varphi_r(x)\right)$$

и, наконец, воспользоваться для оценки параметров методом максимума правдоподобия, то получим следующий

алгоритм восстановления регрессии по выборке:

$$x_1, y_1; \ldots; x_l, y_l.$$

Надо минимизировать функционал

$$I_э(\alpha) = \sum_{i=1}^{l} d\left(y_i - \sum_{r=1}^{n} \alpha_r \varphi_r(x_i)\right),$$

где

$$d(z) = z^2,$$

если истинная плотность помехи принадлежит классу плотностей с ограниченной дисперсией;

$$d(z) = |z|,$$

если истинная плотность помехи принадлежит классу невырожденных плотностей;

$$d(z) = \begin{cases} \frac{z^2}{2\sigma^2}, & \text{если } |z| < k\sigma, \\ -\frac{k^2}{2} + \frac{k}{\sigma}|z|, & \text{если } |z| \geqslant k\sigma, \end{cases}$$

если истинная плотность есть смесь нормального закона с любым возможным;

$$d(z) = \begin{cases} -2\ln\cos\frac{a}{A}z, & \text{если } |z| < A, \\ b\left(\left|\frac{z}{a}\right| - 1\right) - 2\ln\cos\frac{a}{A}z, & \text{если } |z| \geqslant A, \end{cases}$$

если истинная плотность сосредоточена в основном на отрезке $[-A, A]$.

Среди этих методов метод наименьших квадратов ($d(z) = z^2$) и метод минимальных модулей ($d(z) = |z|$) не содержат свободных параметров. Метод наименьших модулей является более универсальным — он определяется устойчивой плотностью в более широком классе плотностей.

Два других метода оценивания содержат параметры, которые вычисляются в зависимости от величин, задающих классы плотностей. Эти методы следует применять, когда удается достаточно точно определить классы плотностей, содержащие искомую плотность.

Итак, при восстановлении регрессии нам удалось снять требование точного знания модели помехи. Достаточно знать класс функций, содержащий регрессию и класс плотностей, которому принадлежит плотность помехи. Однако вся построенная выше теория является принципиально асимптотической (при выводе основного соотношения (4.37) существенно использовался закон больших чисел). Поэтому гарантией того, что найденные алгоритмы окажутся работоспособными на выборках ограниченного объема, может служить лишь вера в то, что асимптотика наступает рано.

Основные утверждения главы IV

1. Восстановление регрессии методами параметрической статистики проводится для схемы интерпретации измерений, согласно которой измеряемая величина y связана с вектором x соотношением

$$y = F(x, \alpha_0) + \xi,$$

где $F(x, \alpha_0)$ — искомая функция, принадлежащая заданному параметрическому семейству $F(x, \alpha)$, ξ — случайная независимая ошибка измерения, возникающая, согласно плотности $P_0(\xi)$, с нулевым математическим ожиданием.

Методы параметрической статистики при восстановлении регрессии направлены на оценку параметров плотности условного распределения вероятностей

$$P(y \mid x) = P_0(y - F(x, \alpha)),$$

2. Применение метода максимума правдоподобия для оценивания параметров условной плотности по выборке x_1, y_1; ...; x_l, y_l приводит к минимизации функционала

$$R(\alpha) = \sum_{i=1}^{l} \ln P_0(y_i - F(x_i, \alpha)),$$

зависящего от плотности помехи $P_0(\xi)$. Таким образом, различным статистическим моделям $P_0(\xi)$ помехи соответствуют разные алгоритмы восстановления регрессии.

3. На практике закон образования помехи $P_0(\xi)$, как правило, неизвестен. Считается, что может быть установлен лишь класс плотностей, которому принадлежит плотность $P_0(\xi)$.

В этом случае в качестве статистической модели помехи следует выбирать устойчивую плотность, т. е. такую, которая при самом неблагоприятном стечении обстоятельств обеспечит максимальную асимптотическую скорость оценивания параметров регрессии.

4. Устойчивые методы восстановления регрессии могут быть получены для разных классов плотностей. Они возможны лишь при использовании выборок большого объема.

ГЛАВА V

ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЕГРЕССИИ

§ 1. Задача оценивания параметров регрессии

В предыдущей главе были рассмотрены методы восстановления регрессии в условиях, когда объем выборки стремился к бесконечности.

Однако, строго говоря, полученные результаты относились не к задаче *восстановления регрессии*, а к другой задаче — *оценивания параметров регрессии*. Такая подмена задач (вместо приближения функции — оценивание ее параметров) правомерна для достаточно больших объемов выборки. С ростом объема выборки оцениваемые параметры стремятся к истинным и, следовательно, построенная с помощью найденных параметров функция стремится к регрессии. Однако для выборки ограниченного объема задача восстановления регрессии не всегда эквивалентна задаче оценивания ее параметров.

Действительно, качество оценки $\hat{\alpha}$ параметров α_0 регрессии $y(x)=F(x,\ \alpha_0)$ определяется близостью векторов α_0 и $\hat{\alpha}$:

$$\rho(\alpha_0,\ \hat{\alpha})=\|\hat{\alpha}-\alpha_0\|. \tag{5.1}$$

Качество же приближения функций $F(x,\ \hat{\alpha})$ к регрессии $F(x,\ \alpha_0)$ — близостью функций.

В гл. I мы условились рассматривать среднеквадратичную меру близости

$$\rho_L(F(x,\ \alpha_0);\ F(x,\ \hat{\alpha}))= \\ =\left(\int (F(x,\ \hat{\alpha})-F(x,\ \alpha_0))^2\,P(x)\,dx\right)^{1/2}. \tag{5.2}$$

Критерии (5.1) и (5.2) не идентичны, и поэтому решение, лучшее по одному из них, может быть худшим по другому.

Пример. Пусть в классе функций

$$F(x, \alpha) = \alpha^0 + \alpha^1 x + \alpha^2 x^2$$

на отрезке $[1, 2]$ восстанавливается регрессия

$$y = x^2.$$

Рассмотрим два решения (рис. 5): первое — полином

$$F(x, \hat{\alpha}) = 0{,}5x^2$$

и второе — полином

$$F(x, \hat{\hat{\alpha}}) = 3x - 2.$$

С точки зрения критерия оценки параметров первое решение лучше второго (как бы ни понималась норма (5.1), вектор $\hat{\alpha} = (0, 0, 0, 5)^T$ ближе к $\alpha_0 = (0, 0, 1)^T$, чем вектор $\hat{\hat{\alpha}} = (-2, 3, 0)^T$).

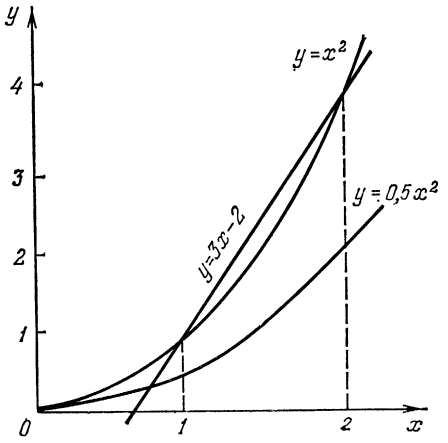

Рис. 5.

Однако с точки зрения критерия (5.2) лучшим будет второе решение $F(x, \hat{\hat{\alpha}})$. При любой мере $P(x)$ справедливо неравенство

$$\rho_L(3x - 2, x^2) < \rho_L(0{,}5x^2, x^2).$$

Когда же задача оценивания параметров регрессии по выборкам ограниченного объема эквивалентна задаче восстановления регрессии?

Допустим, что класс функций, которому принадлежит регрессия, линеен по параметрам

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x), \qquad (5.3)$$

и пусть $\varphi_1(x), \ldots, \varphi_n(x)$ — система ортонормальных с весом $P(x)$ функций, т. е. таких функций, для которых

$$\int_a^b \varphi_p(x)\,\varphi_q(x)\,P(x)\,dx = \begin{cases} 1, & \text{если } p=q, \\ 0, & \text{если } p\neq q. \end{cases} \tag{5.4}$$

В этом случае величины, характеризующие близость функций в метрике L_P^2 и близость параметров в евклидовой метрике, совпадают, и задача приближения на $[a, b]$ функции к регрессии становится эквивалентной задаче оценивания параметров. Действительно,

$$\rho_L^2\left(F(x, \hat{\alpha}),\ F(x, \alpha)\right) = \int_a^b \left(\sum_{i=1}^n \hat{\alpha}_i\varphi_i(x) - \sum_{i=1}^n \alpha_i\varphi_i(x)\right)^2 P(x)\,dx = \sum_{i=1}^n (\hat{\alpha}_i - \alpha_i)^2. \tag{5.5}$$

Условия (5.3), (5.4) достаточны для того, чтобы заменить задачу восстановления регрессии задачей оценивания ее параметров. Однако, для того чтобы построить ортогональную систему функций, надо знать плотность $P(x)$. В этой главе мы будем полагать, что плотность $P(x)$ нам известна.

§ 2. Теория нормальной регрессии

Теория оценивания параметров регрессии по выборкам ограниченного объема разработана для случая, когда, во-первых, класс функций, которому принадлежит регрессия, линеен по параметрам

$$F(x, \alpha) = \sum_{i=1}^n \alpha_i\varphi_i(x), \tag{5.6}$$

а, во-вторых, структура измерений подчиняется *схеме Гаусса — Маркова.*

Считается, что измерения функциональной зависимости

$$y(x) = \sum_{i=1}^n \alpha_i^0\varphi_i(x)$$

производятся в l фиксированных точках

$$x_1, \ldots, x_l.$$

(Эти точки не являются случайными.)

Измерения производятся с аддитивной помехой ξ, которая возникает случайно согласно плотности $P(\xi)$, имеет нулевое среднее значение $\left(\text{т. е. } \int \xi P(\xi)\, d\xi = 0\right)$ и конечную дисперсию $\left(\int \xi^2 P(\xi)\, d\xi < \infty\right)$. Помехи в точках x_i и x_j $(i \neq j)$ некоррелированы.

Результатом измерений функции $\bar{y} = y(x)$ в точках $x_1, \ldots, x_l$ является случайный вектор $Y = (y_1, \ldots, y_l)^{\mathrm{T}}$, координаты которого равны

$$y_j = \sum_{i=1}^{n} \alpha_i^0 \varphi_i(x_j) + \xi_j = \bar{y}_j + \xi_j, \quad j = 1, 2, \ldots, l.$$

Или в векторной форме

$$Y = \Phi\alpha_0 + \xi, \tag{5.7}$$

где Φ — матрица $l \times n$ с элементами $\varphi_i(x_j)$ $(j = 1, 2, \ldots, l;\ i = 1, 2, \ldots, n)$, α_0 — вектор параметров, ξ — вектор помех.

Таким образом, схему Гаусса — Маркова определяют равенства

$$MY = \Phi\alpha_0, \quad M\{(Y - MY)(Y - MY)^{\mathrm{T}}\} = \sigma^2 I, \tag{5.8}$$

где I — единичная матрица.

При построении теории оценивания параметров регрессии в схеме Гаусса — Маркова выделяется случай, когда помеха ξ задается нормальным законом распределения вероятностей.

Для нормального закона распределения помехи справедлива так называемая теория *нормальной регрессии*, в основе которой лежит следующий факт: экстремальным методом оценивания параметров нормальной регрессии является метод наименьших квадратов, согласно которому в качестве оценки параметров α следует выбирать такой вектор $\alpha_э$, который доставляет минимум функционалу

$$I_э(\alpha) = \frac{1}{l} \sum_{j=1}^{l} \left(y_j - \sum_{i=1}^{n} \alpha_i \varphi_i(x_j) \right)^2.$$

Справедлива

Теорема 5.1. *Оценки метода наименьших квадратов параметров нормальной регрессии являются совместно эффективными.*

Ниже мы докажем эту теорему, а затем построим метод восстановления нормальной регрессии лучший, чем тот, который основан на методе наименьших квадратов.

Доказательство. Запишем плотность распределения вероятностей помехи в виде

$$P(\xi_j) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(y_j - \sum_{i=1}^{n} \alpha_i^0 \varphi_i(x_j)\right)^2\right\}. \quad (5.9)$$

Здесь задача оценивания параметров регрессии эквивалентна оцениванию параметров распределения (5.9) по результатам измерения функции $\bar{y} = y(x)$ в точках $x_1, \ldots, x_l$.

Выпишем функцию правдоподобия [1]):

$$P(y_1, \ldots, y_l; \alpha) = P(\alpha) =$$
$$= \frac{1}{(2\pi)^{l/2}\,\sigma^l} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n} \alpha_i \varphi_i(x_j)\right)^2\right]\right\}. \quad (5.10)$$

Согласно неравенству Рао — Крамера (см. § 11 гл. III) информационная матрица Фишера $\|f_{ij}\|$ (матрица с элементами $f_{ij} = -M\dfrac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j}$) определяет предельную точность совместных оценок вектора параметров α в классе несмещенных методов оценивания. А именно, для любого вектора z справедливо неравенство

$$z^{\mathrm{T}} \|f_{ij}\|^{-1} z \leqslant z^{\mathrm{T}} B z,$$

где B — ковариационная матрица несмещенных оценок вектора параметров. Таким образом, предельная точность в классе несмещенных оценок достигается при таком способе оценивания, при котором

$$B = \|f_{ij}\|^{-1}. \quad (5.11)$$

Покажем, что для нормальной помехи равенство (5.11) достигается при оценивании параметров регрессии методом наименьших квадратов. Действительно, вычислим

---

[1]) Далее для сокращения записи функция правдоподобия $P(y_1, \ldots, y_l; \alpha)$ обозначена $P(\alpha)$.

элементы f_{ij} матрицы Фишера. Учитывая (5.10), получим

$$f_{ij} = -M\frac{\partial^2 \ln P(\alpha)}{\partial\alpha_i\,\partial\alpha_j} = \frac{1}{\sigma^2} M \sum_{r=1}^{l} \varphi_i(x_r)\,\varphi_j(x_r).$$

Или в матричной форме

$$\|f_{ij}\| = \frac{1}{\sigma^2} M\Phi^{\mathrm{T}}\Phi, \tag{5.12}$$

где Φ — матрица $l \times n$ с элементами $\varphi_i(x_j)$, $i = 1, \ldots, n$ $j = 1, \ldots, l$.

Вычислим теперь элементы b_{ij} ковариационной матрицы B оценок метода наименьших квадратов. Для этого найдем оценку параметров регрессии по методу наименьших квадратов, т. е. вектор $\alpha_э$, минимизирующий функционал

$$I_э(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{i=1}^{n}\alpha_i\varphi_i(x_j)\right)^2. \tag{5.13}$$

Минимизация по α функционала (5.13) эквивалентна решению следующего уравнения:

$$\Phi^{\mathrm{T}}\Phi\alpha = \Phi^{\mathrm{T}}Y. \tag{5.14}$$

Уравнение (5.14) называется *нормальным уравнением*. Решение нормального уравнения относительно вектора параметров α равно [1])

$$\alpha = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y.$$

Заметим, что оценка метода наименьших квадратов принадлежит классу несмещенных

$$M\alpha = M[(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y] = \alpha_0.$$

Выпишем вектор $(\alpha - \alpha_0)$ уклонений оценки параметров регрессии от истинных значений параметров

$$(\alpha - \alpha_0) = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\bar{\xi},$$

где $\bar{\xi}$ — вектор помех измерений.

---

[1]) Здесь принято, что матрица $(\Phi^{\mathrm{T}}\Phi)$ не вырождена. Для вырожденных матриц вместо обратной матрицы $(\Phi^{\mathrm{T}}\Phi)^{-1}$ используется псевдообратная матрица $(\Phi^{\mathrm{T}}\Phi)^{+}$.

Найдем теперь ковариационную матрицу:

$$B = M(\alpha - \alpha_0)(\alpha - \alpha_0)^{\mathrm{T}} = (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} M\xi\xi^{\mathrm{T}}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1}.$$

Учитывая, что $M\xi\xi^{\mathrm{T}} = \sigma^2 I$, получим

$$B = \sigma^2(\Phi^{\mathrm{T}}\Phi)^{-1}.$$

Таким образом, для нормального закона распределения помех ковариационная матрица вектора оценок равна обратной информационной матрице Фишера. Тем самым доказана эффективность метода наименьших квадратов в задаче восстановления параметров регрессии при структуре измерений, определяемой схемой Гаусса — Маркова.

Следует заметить, что метод наименьших квадратов является эффективным средством оценивания параметров лишь в схеме Гаусса — Маркова. В схемах с нефиксированными точками измерений x_i даже при нормальном законе появления помех метод наименьших квадратов оказывается лишь асимптотически эффективным.

Так, уже в случае оценивания одного параметра

$$y^0 = ax$$

при измерениях, проводимых с аддитивной нормальной помехой

$$y = ax + \xi$$

в точках $x_1, \ldots, x_l$, заданных случайно и независимо согласно $P(x)$, оценка параметра a не будет эффективной.

Действительно, точно так же, как и выше, можно найти величину информационного количества Фишера:

$$I_\Phi = \frac{M\sum_{i=1}^{l} x_i^2}{\sigma^2},$$

и вычислить дисперсию оценки параметра a:

$$D(a) = M\frac{\sigma^2}{\sum_{i=1}^{l} x_i^2}.$$

Заметим, что в силу выпуклости функции $1/x^2$ имеет место неравенство

$$M\frac{1}{\sum_{i=1}^{l} x_i^2} \geqslant \frac{1}{M\sum_{i=1}^{l} x_i^2}, \tag{5.15}$$

откуда следует, что в рассмотренном примере

$$D(a) \geqslant I_{\Phi}^{-1}.$$

Единственный случай, когда неравенство (5.15) переходит в равенство — фиксированность точек измерений. Этот случай и реализует схема Гаусса — Маркова.

§ 3. Методы восстановления нормальной регрессии, равномерно лучшие метода наименьших квадратов

Итак, в схеме Гаусса — Маркова метод наименьших квадратов является эффективным средством оценивания параметров нормальной регрессии. Однако в этом утверждении есть две оговорки:

1. Измерения проводятся в условиях нормальной помехи.

2. Метод наименьших квадратов является наилучшим не безусловно, а лишь среди несмещенных методов оценивания.

Возникает вопрос, существенны ли эти оговорки? Оказывается, что обе оговорки существенны. Метод наименьших квадратов сохраняет свои экстремальные свойства лишь при нормально распределенных помехах ξ. При числе измерений $l \geqslant 2n+1$ (n — размерность базиса) из эффективности метода наименьших квадратов следует, что помеха распределена по нормальному закону [23].

Не менее существенна и вторая оговорка: даже в условиях нормально распределенной помехи в классе смещенных методов оценивания существуют оценки равномерно лучшие, чем оценки метода наименьших квадратов.

Определение. *Будем говорить, что для функции потерь*

$$\| \alpha - \alpha_0 \|^2 = (\alpha - \alpha_0)^{\mathrm{T}} (\alpha - \alpha_0)$$

метод оценивания $\alpha_A(y_1, \ldots, y_l)$ *вектора параметров* α_0 *равномерно лучше метода оценивания* $\alpha_B(y_1, \ldots, y_l)$, *если для любого* α_0 *выполняются неравенства*

$$M \| \alpha_A(y_1, \ldots, y_l) - \alpha_0 \|^2 \leqslant M \| \alpha_B(y_1, \ldots, y_l) - \alpha_0 \|^2.$$

В этом параграфе мы построим алгоритмы приближения к регрессии равномерно лучшие (лучшие для любого α_0), чем те, которые вытекают из процедуры метода наименьших квадратов.

Основой этих алгоритмов служат методы оценивания вектора средних многомерного нормального закона и, в частности, следующая

Теорема 5.2 (Джеймс — Стейн). *Пусть x — n-мерный ($n \geqslant 3$) случайный вектор, распределенный по нормальному закону $N(\alpha, \sigma^2 I)$ с вектором средних α и ковариационной матрицей $\sigma^2 I$. Пусть S — случайная величина, независимая от x, распределенная согласно центральному $\sigma^2\chi^2$-распределению с q степенями свободы. Тогда оценка среднего*

$$\hat{\alpha}(x, S) = \left(1 - \frac{n-2}{q+2}\frac{S}{\|x\|^2}\right)_+ x, \qquad (z)_+ = \begin{cases} z, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0 \end{cases} \qquad (5.16)$$

равномерно лучше оценки $\hat{\hat{\alpha}}(x) = x$.

Иначе говоря, теорема 5.2 утверждает, что в качестве оценки вектора α следует брать не реализацию x, а вектор $\hat{\alpha}(x, S)$, коллинеарный вектору реализации, но отличающийся от x величиной модуля. Эта теорема является частным случаем более общего утверждения, доказанного в следующем параграфе.

Используя теорему 5.2, построим алгоритм приближения к регрессии равномерно лучший, чем тот, который вытекает из метода наименьших квадратов.

Итак, пусть в точках $x_1, \ldots, x_l$ проведены измерения $y_1, \ldots, y_l$, и пусть нашей целью является построение метода приближения нормальной регрессии лучшего, чем метод наименьших квадратов.

По-прежнему близость функций определяется метрикой L_P^2:

$$\rho_L(F(x, \hat{\alpha}), F(x, \alpha)) = \left(\int (F(x, \hat{\alpha}) - F(x, \alpha))^2 P(x)\, dx\right)^{1/2}.$$

Перейдем к новому дважды ортогональному базису

$$\psi_1(x), \ldots, \psi_n(x), \qquad (5.17)$$

т. е. базису, для которого выполняются равенства

$$\int \psi_i(x)\psi_j(x) P(x)\, dx = \begin{cases} \lambda_i, & \text{если } i = j, \\ 0, & \text{если } i \neq j, \end{cases} \qquad \sum_{r=1}^{l} \psi_i(x_r)\psi_j(x_r) = \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j, \end{cases} \qquad (5.18)$$

и будем искать регрессию в разложении по базису (5.17) [1]):

$$F(x,\ \alpha)=\sum_{i=1}^{n}\alpha_i\psi_i(x).$$

В новом базисе близость функции $F(x,\ \alpha)$ к регрессии $F(x,\ \alpha_0)$ определяется величиной

$$\rho_L^2(F(x,\ \alpha),\ F(x,\ \alpha_0))\equiv\rho_L^2(\alpha,\ \alpha_0)=$$
$$=\int\left(\sum_{i=1}^{n}(\alpha_i^0-\alpha_i)\psi_i(x)\right)^2P(x)\,dx=\sum_{i=1}^{n}\lambda_i(\alpha_i^0-\alpha_i)^2.$$

Таким образом, нашей целью является отыскание такого алгоритма $\hat{\alpha}(y_1,\ \ldots,\ y_l)$ оценивания параметра α_0, для которого величина

$$M\rho_L^2(\hat{\alpha}(y_1,\ \ldots,\ y_l),\alpha_0)=M\sum_{i=1}^{n}\lambda_i(\hat{\alpha}_i(y_1,\ldots,y_l)-\alpha_i^0)^2 \quad (5.19)$$

меньше

$$M\rho_L^2(\alpha_{\text{мнк}},\ \alpha_0)=M\sum_{i=1}^{n}\lambda_i(\alpha_{\text{мнк}}^i-\alpha_i^0)^2,$$

где $\alpha_{\text{мнк}}=(\alpha_{\text{мнк}}^1,\ \ldots,\ \alpha_{\text{мнк}}^n)^{\mathrm{T}}$ — оценка метода наименьших квадратов.

Рассмотрим оценку параметров регрессии, которую определяет метод наименьших квадратов. В базисе (5.17) эта оценка равна

$$\alpha_{\text{мнк}}=\Phi^{\mathrm{T}}Y,$$

где Φ — матрица $l\times n$ с элементами $\psi_i(x_j)$, $j=1,\ 2,\ \ldots,\ l$, $i=1,\ 2,\ \ldots,\ n$, Y — вектор измерений.

Вектор $\alpha_{\text{мнк}}$ является случайным вектором, распределенным по нормальному закону с вектором средних α_0

$$M\alpha_{\text{мнк}}=M\Phi^{\mathrm{T}}Y=\alpha_0$$

и матрицей ковариации σ^2I

$$M(\alpha_{\text{мнк}}-\alpha_0)(\alpha_{\text{мнк}}-\alpha_0)^{\mathrm{T}}=M\Phi^{\mathrm{T}}\xi\xi^{\mathrm{T}}\Phi=\sigma^2I.$$

---

[1]) Согласно теореме о приведении двух квадратичных форм к диагональному виду с помощью линейного преобразования, такой базис существует, он может быть построен методами линейной алгебры.

Таким образом, проблема оценивания параметра α_0 регрессии сводится к оцениванию вектора среднего α_0 нормального закона $N(\alpha_0, \sigma^2 I)$ по его реализации $\alpha_{\text{мнк}}$.

Если бы в (5.19) оказалось, что $\lambda_1 = \lambda_2 = \ldots = \lambda_n$, то можно было бы, воспользовавшись теоремой 5.2, сконструировать алгоритм восстановления регрессии лучший, чем метод наименьших квадратов.

Действительно, как будет показано ниже, статистика

$$S = Y^{\mathrm{T}}Y - \alpha_{\text{мнк}}^{\mathrm{T}}\alpha_{\text{мнк}} \tag{5.20}$$

не зависит от $\alpha_{\text{мнк}}$ и распределена согласно центральному $\sigma^2\chi^2$-распределению с $l-n$ степенями свободы.

Поэтому, согласно теореме 5.2, оценка

$$\hat{\alpha} = \left(1 - \frac{n-2}{l-n+2}\,\frac{Y^{\mathrm{T}}Y - \alpha_{\text{мнк}}^{\mathrm{T}}\alpha_{\text{мнк}}}{\alpha_{\text{мнк}}^{\mathrm{T}}\alpha_{\text{мнк}}}\right)_{+} \alpha_{\text{мнк}} \tag{5.21}$$

равномерно лучше, чем $\alpha_{\text{мнк}}$, т. е. доставляет критерию (5.19) (при $\lambda_1 = \ldots = \lambda_n$) меньшее значение, чем $\alpha_{\text{мнк}}$.

Однако в построенной дважды ортогональной системе (5.17) обычно не все величины λ_i равны между собой. Таким образом, получение лучшего приближения к регрессии в случае несовпадающих λ_i связано с отысканием способа оценивания параметров, доставляющего меньшую величину критерию (5.19), чем метод наименьших квадратов.

Конструирование такого алгоритма оценивания также опирается на результаты теоремы 5.2. Будем считать, что функции ψ_i перенумерованы в порядке невозрастания величин $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n \geqslant 0$. Введем обозначения: пусть $\alpha_0(p)$ — вектор размерности p, составленный из первых p координат вектора $\alpha_0 = (\alpha_1^0, \ldots, \alpha_n^0)^{\mathrm{T}}$; $\alpha_{\text{мнк}}(p)$ — вектор составленный из первых p координат вектора оценок метода наименьших квадратов $\alpha_{\text{мнк}}$.

Определим n чисел $f_1, \ldots, f_n$:

$$f_1 = 1,$$

$$f_p = \left(1 - \frac{S}{\alpha_{\text{мнк}}^{\mathrm{T}}(p)\,\alpha_{\text{мнк}}(p)}\,\frac{p-2}{l-p+2}\right)_{+}, \quad p = 2, \ldots, n.$$

С помощью чисел $f_1, \ldots, f_n$ образуем n чисел h_p по правилу

$$h_p = \frac{\sum_{i=p}^{n} (\lambda_i - \lambda_{i+1}) f_i}{\lambda_p}, \quad \text{где} \quad \lambda_{n+1} = 0, \quad p = 1, 2, \ldots, n.$$

Справедлива

Теорема 5.3 (Бхаттачария). *Для риска* (5.19) *оценка*

$$\hat{\alpha}(y_1, \ldots, y_l) = (\hat{\alpha}^1_{\text{мнк}} h_1, \ldots, \alpha^n_{\text{мнк}} h_n)^{\text{т}}, \quad n \geqslant 3, \quad (5.22)$$

равномерно лучше оценки $\alpha_{\text{мнк}} = (\alpha^1_{\text{мнк}}, \ldots, \alpha^n_{\text{мнк}})^{\text{т}}$.

Доказательство. Доказательство теоремы 5.3 опирается на теорему 5.2, согласно которой для любого p справедливо неравенство

$$M \| \alpha_{\text{мнк}}(p) f_p - \alpha_0(p) \|^2 \leqslant M \| \alpha_{\text{мнк}}(p) - \alpha_0(p) \|^2. \quad (5.23)$$

Рассмотрим рандомизированную оценку

$$g\alpha_{\text{мнк}} = (\alpha^1_{\text{мнк}} g_1, \ldots, \alpha^n_{\text{мнк}} g_n), \quad (5.24)$$

где g_k — случайные, не зависящие от S и y величины, имеющие распределение

$$P\{(q_k = f_j)\} = \frac{\lambda_j - \lambda_{j+1}}{\lambda_k}, \quad k = 1, 2, \ldots, n; \quad j = k, \ldots, n,$$

$$\lambda_{n+1} = 0.$$

Величина риска (5.19) при такой оценке равна

$$\rho^2_L(G\alpha_{\text{мнк}}, \alpha_0) = M \sum_{k=1}^{n} \lambda_k (g_k \alpha^k_{\text{мнк}} - \alpha^0_k)^2 =$$

$$= \sum_{k=1}^{n} \sum_{j=k}^{n} \frac{\lambda_j - \lambda_{j+1}}{\lambda_k} \lambda_k M (f_j \alpha^k_{\text{мнк}} - \alpha^0_k)^2.$$

Воспользуемся неравенством (5.23):

$$\rho_L^2(G\alpha_{\text{мнк}}, \alpha_0) =$$

$$= \sum_{k=1}^{n} \sum_{j=k}^{n} (\lambda_j - \lambda_{j+1}) M \left(\alpha_{\text{мнк}}^k f_j - \alpha_k^0\right)^2 =$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \sum_{k=1}^{j} \left(\alpha_{\text{мнк}}^k f_j - \alpha_k^0\right)^2 =$$

$$= \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \| \alpha_{\text{мнк}}(j) f_j - \alpha_0(j) \|^2 \leqslant$$

$$\leqslant \sum_{j=1}^{n} (\lambda_j - \lambda_{j+1}) M \| \alpha_{\text{мнк}}(j) - \alpha_0(j) \|^2 \leqslant$$

$$\leqslant \sum_{j=1}^{n} \lambda_j M \left(\alpha_{\text{мнк}}^j - \alpha_j^0\right)^2.$$

Таким образом, величина риска для рандомизированной оценки значений параметров меньше величины риска для оценки, полученной методом наименьших квадратов. С другой стороны, из свойств выпуклости функции потерь (5.19) следует, что нерандомизированная оценка (5.22) не хуже рандомизированной оценки (5.24).

Следовательно, приближение к регрессии, определяемое параметрами (5.22), равномерно лучше приближения, полученного методом наименьших квадратов. Теорема доказана.

Нам осталось показать, что статистика $S = Y^{\mathrm{T}}Y - \alpha_{\text{мнк}}^{\mathrm{T}}\alpha_{\text{мнк}}$ не зависит от $\alpha_{\text{мнк}}$ и распределена согласно центральному $\sigma^2\chi^2$-распределению с $l-n$ степенями свободы.

Для этого дополним ортонормальную на $x_1, \ldots, x_l$ систему из n векторов $\psi_1, \ldots, \psi_n$:

$$\psi_i = (\psi_i(x_1), \ldots, \psi_i(x_l))^{\mathrm{T}},$$

$$\psi_i^{\mathrm{T}}\psi_j = \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j, \end{cases} \quad i, j = 1, 2, \ldots, n,$$

до полной системы, состоящей из l ортонормальных векторов

$$\psi_1, \ldots, \psi_n, \psi_{n+1}, \ldots, \psi_l,$$

$$\psi_i^{\mathrm{T}}\psi_j = \begin{cases} 1, & \text{если } i=j, \\ 0, & \text{если } i\neq j, \end{cases} \quad i, j = 1, 2, \ldots, l.$$

Разложим Y по этой системе:

$$Y = \sum_{i=1}^{n} \gamma_i\psi_i + \sum_{j=n+1}^{l} \gamma_j\psi_j, \tag{5.25}$$

где

$$\gamma_i = Y^{\mathrm{T}}\psi_i = \alpha^i_{\text{мнк}}, \quad i = 1, 2, \ldots, n,$$
$$\gamma_j = Y^{\mathrm{T}}\psi_j, \qquad j = n+1, \ldots, l.$$

Подставляя (5.25) в (5.20), получаем

$$S = \sum_{j=n+1}^{l} \gamma_j^2, \tag{5.26}$$

и, следовательно, S не зависит от $\alpha^i_{\text{мнк}}$ (а зависит лишь от γ_j, $j = n+1, \ldots, l$). Так как по условию $Y = Y_0 + \xi$, а вектор Y_0 разложим по неполной системе (5.17)

$$Y_0 = \sum_{i=1}^{n} \alpha_i^0\psi_i,$$

то имеет место равенство

$$\gamma_j = \xi^{\mathrm{T}}\psi_j.$$

Подставляя в (5.26) значение γ_j, получим

$$S = \sum_{j=n+1}^{l} \gamma_j^2 = \sum_{j=n+1}^{l} \left(\sum_{i=1}^{l} \xi_i^{\mathrm{T}}\psi_j(x_i) \right)^2 = \sum_{j=n+1}^{l} \xi_j^2,$$

и, следовательно, статистика S распределена согласно центральному $\sigma^2\chi^2$-распределению с $l-n$ степенями свободы.

§ 4. Теорема об оценивании вектора средних многомерного нормального закона

В этом параграфе мы получим семейство оценок вектора средних, равномерно лучших, чем $\alpha(x, S) = x$. Этому семейству принадлежит оценка (5.21).

Итак, пусть x — случайный вектор, распределенный согласно $N(\alpha_0, \sigma^2 I)$, S — случайная величина, распределенная согласно центральному $\sigma^2\chi^2$-распределению с q степенями свободы. Обозначим $F = \frac{x^T x}{S}$.

Справедлива

Теорема 5.4 (Баранчик). *Оценка n-мерного ($n \geqslant 3$) вектора средних*

$$\hat{\alpha}(x, S) = \left(1 - \frac{r(F)}{F}\right) x,$$

где $r(F)$ — монотонная неубывающая функция, лежащая в пределах

$$0 \leqslant r(F) \leqslant 2\,\frac{n-2}{q+2}, \tag{5.27}$$

равномерно лучше оценки $\alpha(x, S) = x$.

Замечание. Теорема 5.2 является частным случаем теоремы 5.4, если положить

$$r(F) = \begin{cases} \dfrac{n-2}{q+2}, & \text{если } F \geqslant \dfrac{n-2}{q+2}, \\ F, & \text{если } F < \dfrac{n-2}{q+2}. \end{cases}$$

Доказательство. При доказательстве теоремы 5.4 используется следующий факт: математическое ожидание случайной величины $f(\chi^2(n, b))$, взятое по мере $\mu(\chi^2(n, b))$, где $\chi^2(n, b)$ — случайная величина, распределенная согласно нецентральному χ^2-распределению с n степенями свободы и параметром b, представимо в виде

$$Mf(\chi^2(n, b)) = Mf(\chi^2_{n+2k}),$$

χ^2_{n+2k} — случайная величина, распределенная согласно центральному χ^2-распределению с $n+2k$ степенями свободы, k — случайная величина, распределенная согласно закону Пуассона с параметром b:

$$P(k) = \exp\{-b\}\,\frac{b^k}{k!}$$

(математическое ожидание в правой части равенства вычисляется как по x, так и по k).

Таким образом,

$$Mf(\chi^2(n,\ b)) = Mf(\chi^2_{n+2k}) = \exp\{-b\}\sum_{t=0}^{\infty}\frac{b^t}{t!}Mf(\chi^2_{n+2t}). \tag{5.28}$$

Перейдем непосредственно к доказательству теоремы. Нашей целью является доказательство того, что разность

$$H = M\|\hat{\alpha}(x,\ S) - \alpha_0\|^2 - M\|x - \alpha_0\|^2 \tag{5.29}$$

— величина неположительная. Обозначим

$$g(F) = 1 - \frac{r(F)}{F}$$

и преобразуем (5.29)

$$H = M[x^{\mathrm{T}}xg^2(F)] - 2\alpha_0^{\mathrm{T}}Mg(F)x + \|\alpha_0\|^2 - n\sigma^2. \tag{5.30}$$

Следующие выражения с (5.31) по (5.34) мы получим в предположении, что величина S фиксирована. Согласно (5.28) имеем

$$M\left[x^{\mathrm{T}}xg^2\left(\frac{x^{\mathrm{T}}x}{S}\right)\right] =$$

$$= \exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}\frac{\|\alpha_0\|^{2t}}{t!\,(2\sigma^2)^t}M\left[\sigma^2\chi^2_{n+2t}g^2\left(\frac{\sigma^2\chi^2_{n+2t}}{S}\right)\right]. \tag{5.31}$$

Преобразуем теперь выражение

$$\alpha_0^{\mathrm{T}}Mg(F)x = \alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x.$$

Для этого произведем ортогональное преобразование векторов x в векторы z так, чтобы в новой системе координат вектор средних был равен $(\|\alpha_0\|, 0, \ldots, 0)$ (отлична от нуля лишь первая координата, которая равна норме вектора средних). При таком преобразовании величина S не меняется.

Получим

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x = \|\alpha_0\|\,Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1,$$

где z_1 — первая координата вектора $z = (z_1, \ldots, z_n)^{\mathrm{T}}$.

Заметим теперь, что

$$M\left[g\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1\right]=\frac{\sigma^2}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\times$$

$$\times\frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2-2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\ldots dz_n$$

Таким образом, получим

$$\|\alpha_0\|Mg\left(\frac{z^{\mathrm{T}}z}{S}\right)z_1=\frac{\sigma^2\|\alpha_0\|}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\times$$

$$\times\frac{d}{d\|\alpha_0\|}\int g\left(\frac{\sum_{i=1}^{n}z_i^2}{S}\right)\exp\left\{-\frac{\sum_{i=1}^{n}z_i^2-2\|\alpha_0\|z_1}{2\sigma^2}\right\}dz_1\ldots dz_n=$$

$$=\sigma^2\|\alpha_0\|\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\frac{d}{d\|\alpha_0\|}\exp\left\{\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}Mg\left(\frac{\sigma^2\chi_n^2+2k}{S}\right),$$

где k — случайная величина, распределенная по закону Пуассона со средним $\|\alpha_0\|^2/(2\sigma^2)$. Окончательно получим

$$\alpha_0^{\mathrm{T}}Mg\left(\frac{x^{\mathrm{T}}x}{S}\right)x=$$

$$=2\sigma^2\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}t\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t\frac{Mg(\sigma^2\chi_n^2+2t\mid S)}{t!}. \quad (5.32)$$

Наконец, учитывая, что $\|\alpha_0\|^2/(2\sigma^2)$ есть среднее случайной величины k, распределенной по закону Пуассона, выразим третье слагаемое в сумме (5.30) в виде

$$\|\alpha_0\|^2=2\sigma^2\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}t\frac{\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!}. \quad (5.33)$$

Представим выражение (5.30) в виде

$$H=\sigma^2\exp\left\{-\frac{\|\alpha_0\|^2}{2\sigma^2}\right\}\sum_{t=0}^{\infty}\frac{\left(\frac{\|\alpha_0\|^2}{2\sigma^2}\right)^t}{t!}\left[M\chi_{n+2t}^2g^2\left(\frac{\sigma^2\chi_n^2+2t}{S}\right)-\right.$$

$$\left.-4tMg\left(\frac{\sigma^2\chi_n^2+2t}{S}\right)-n+2t\right]. \quad (5.34)$$

Пусть теперь $S = \sigma^2\chi_q^2$ есть случайная величина, распределенная согласно центральному $\sigma^2\chi^2$-распределению с q степенями свободы. Тогда теорема будет доказана, если мы установим, что выражение

$$h = M\left[\chi^2_{n+2t} g^2\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - 4tg\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right) - n + 2t\right] \quad (5.35)$$

неположительно ни при каком t.

Обозначим $\frac{\chi^2_{n+2t}}{\chi^2_q} = u$ и заметим, что

$$u(1 - g(u)) = r(u). \quad (5.36)$$

Поэтому из условия теоремы (5.27) следует

$$g(u) > 1 - 2\frac{n-2}{q+2}u^{-1}. \quad (5.37)$$

Преобразуем выражение (5.35), используя обозначения (5.36) и тот факт, что $M\chi^2_{n+2t} = n + 2t$:

$$h = M\left[-2r(u)\chi^2_q + r(u)(1 - g(u))\chi^2_q + 4t\frac{r(u)}{u}\right] =$$
$$= M\left[r(u)\chi^2_q\left(-1 - g(u) + \frac{4t}{\chi^2_{n+2t}}\right)\right].$$

Учитывая (5.37), получим, что величина h не превосходит

$$\hat{h} = M(r(u)\zeta) = M\left[M\left\{r\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right)\zeta \mid \chi^2_q\right\}\right],$$

где обозначено

$$\zeta = \chi^2_q\left[-2 + \left(4t + 2\frac{n-2}{q+2}\chi^2_q\right)\frac{1}{\chi^2_{n+2t}}\right].$$

Для всякого фиксированного χ^2_q определим такую константу a, для которой

$$-2 + \left(4t + 2\frac{n-2}{q+2}\chi^2_q\right)\frac{1}{a} = 0. \quad (5.38)$$

Заметим, что для всякого $\chi^2_{n+2t} > a$ справедливо $\zeta < 0$. Поэтому, учитывая, что, согласно условию теоремы,

функция $r(u)$ не убывает, оценим величину

$$M\left\{r\left(\frac{\chi^2_{n+2t}}{\chi^2_q}\right)\zeta \mid \chi^2_q\right\} \leqslant$$

$$\leqslant r\left(\frac{a}{\chi^2_q}\right) M\{\zeta \mid \chi^2_{n+2t} \leqslant a\}\ P\{\chi^2_{n+2t} \leqslant a\} +$$

$$+ r\left(\frac{a}{\chi^2_q}\right) M\{\zeta \mid \chi^2_{n+2t} > a\} P\{\chi^2_{n+2t} > a\} = r\left(\frac{a}{\chi^2_q}\right) M\{\zeta \mid \chi^2_q\} =$$

$$= r\left(\frac{a}{\chi^2_q}\right)\chi^2_q\left[-2+\left(4t+2\frac{n-2}{q+2}\chi^2_q\right)\frac{1}{n+2t-2}\right] =$$

$$= 2\frac{n-2}{n+2t-2} r\left(\frac{a}{\chi^2_q}\right)\chi^2_q\left(-1+\frac{\chi^2_q}{q+2}\right). \quad (5.39)$$

(Здесь мы воспользовались равенством $M\frac{1}{\chi^2_m} = \frac{1}{m-2}$ $(m \geqslant 3)$.)

Подставим теперь в (5.39) значение a, удовлетворяющее (5.38), и вычислим математическое ожидание (5.39)

$$2\frac{n-2}{n+2t-2} M\left\{r\left(\frac{2t}{\chi^2_q}+\frac{n-2}{q+2}\right)\chi^2_q\left[-1+\frac{\chi^2_q}{q+2}\right]\right\}.$$

Учитывая, что $r(u)$ — неубывающая функция, получим оценку

$$M\left\{r\left(\frac{2t}{\chi^2_q}+\frac{n-2}{q+2}\right)\chi^2_q\left[-1+\frac{\chi^2_q}{q+2}\right]\right\} \leqslant$$

$$\leqslant r\left(\frac{n+2t-2}{q+2}\right) M\left\{\chi^2_q\left[-1+\frac{\chi^2_q}{q+2}\right] \mid \chi^2_q \leqslant q+2\right\} +$$

$$+ r\left(\frac{n+2t-2}{q+2}\right) M\left\{\chi^2_q\left(-1+\frac{\chi^2_q}{q+2}\right) \mid \chi^2_q > q+2\right\} =$$

$$= r\left(\frac{n+2t-2}{q+2}\right) M\left\{\chi^2_q\left[-1+\frac{\chi^2_q}{q+2}\right]\right\} = 0.$$

(Для центрального χ^2-распределения справедливо $M\chi^2_q = q$, $M(\chi^2_q)^2 = q(q+2)$.)

Таким образом, величина (5.35) не превосходит нуля. Теорема доказана.

§ 5. Теорема Гаусса — Маркова

До сих пор при восстановлении регрессии мы считали, что помехи измерения подчиняются нормальному закону. Теперь мы откажемся от этого предположения. Будем

считать, что закон, согласно которому распределены помехи, нам неизвестен. Известно лишь, что он имеет ограниченную дисперсию. Требуется в этих условиях построить наилучший алгоритм восстановления регрессии.

Выше при построении теории нормальной регрессии мы сначала установили, что в классе алгоритмов, приводящих к несмещенным оценкам параметров, метод наименьших квадратов является наилучшим, а затем в более широком классе алгоритмов нашли алгоритмы лучшие, чем метод наименьших квадратов. Сейчас мы поступим аналогичным образом. Сначала покажем, что в некотором узком классе алгоритмов оценивания параметров метод наименьших квадратов является наилучшим, а затем укажем в более широком классе алгоритмов способы оценивания лучшие, чем метод наименьших квадратов.

В условиях нормальной помехи метод наименьших квадратов является наилучшим в классе несмещенных методов оценивания. В этом параграфе мы покажем, что в более узком классе оценок, являющихся одновременно и линейными и несмещенными, метод наименьших квадратов реализует наилучший алгоритм оценивания независимо от того, по какому закону распределена помеха.

Определение. *Говорят, что оценка параметра α является линейной по измерениям $Y=(y, \ldots, y_l)^{\mathrm{T}}$, если она представима в виде*

$$\alpha = LY \qquad \left(\alpha_j = \sum_{i=1}^{l} \beta_{ij} y_i\right), \tag{5.40}$$

где L — матрица с элементами β_{ij} $(i=1, \ldots, l;\ j=1, \ldots, n)$.

Справедлива

Теорема 5.5 (Гаусс — Марков). *Среди всех линейных несмещенных оценок оценка наименьших квадратов обладает минимальными дисперсиями координат.*

Мы проведем доказательство теоремы Гаусса — Маркова в более общей формулировке — для случая линейных смещенных оценок. Обозначим через α_0 вектор параметров линейной регрессии

$$MY = \Phi\alpha_0 \qquad (Y = \Phi\alpha_0 + \xi). \tag{5.41}$$

Определим оценку $\alpha(B)$ как решение уравнения

$$(\Phi^{\mathrm{T}}\Phi + B)\,\alpha(B) = \Phi^{\mathrm{T}}Y, \tag{5.42}$$

где B — симметричная неотрицательно определенная матрица $n \times n$, которая задает вектор μ смещения оценки. Покажем, что оценка $\alpha(B)$ обладает экстремальными свойствами. А именно, справедлива теорема.

Теорема. *Среди всех линейных оценок вектора параметров α с векторами смещения, равными μ, оценка $\alpha(B)$ обладает минимальными дисперсиями координат.*

Доказательство. Из (5.42) получим

$$M\alpha(B) = M(\Phi^T\Phi + B)^{-1}\Phi^T Y = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi\alpha_0. \quad (5.43)$$

Пусть $\hat{\alpha} = LY$ — произвольная линейная оценка такая, что

$$M\hat{\alpha} = M\alpha(B) = \mu. \quad (5.44)$$

Тогда из (5.42) получим

$$MLY = L\Phi\alpha_0 = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi\alpha_0. \quad (5.45)$$

Так как равенство (5.45) справедливо для любых α_0, то

$$L\Phi = (\Phi^T\Phi + B)^{-1}\Phi^T\Phi. \quad (5.46)$$

Выпишем теперь дисперсию i-й координаты оценки $\hat{\alpha}$:

$$M(\hat{\alpha}_i - \mu_i)^2 = M(\hat{\alpha}_i - \alpha_i(B) + \alpha_i(B) - \mu_i)^2 \geqslant$$
$$\geqslant M(\alpha_i(B) - \mu_i)^2 + 2M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i), \quad (5.47)$$

где μ_i — i-я координата вектора смещения μ.

Покажем, что второе слагаемое в правой части (5.47) равно нулю. Действительно, используя (5.44), (5.46) получим

$$M(\hat{\alpha}_i - \alpha_i(B))(\alpha_i(B) - \mu_i) = M(\hat{\alpha}_i - \alpha_i(B))\alpha_i(B) =$$
$$= \sigma^2 \| (L - (\Phi^T\Phi + B)^{-1}\Phi^T)\Phi(\Phi^T\Phi + B)^{-1} \|_{ii} =$$
$$= \sigma^2 \| (L\Phi - (\Phi^T\Phi + B)^{-1}\Phi^T\Phi)(\Phi^T\Phi + B)^{-1} \|_{ii} = 0,$$

где $\| A \|_{ii}$ означает элемент A_{ii} матрицы $\| A \|$.

Таким образом,

$$M(\hat{\alpha}_i - \mu_i)^2 \geqslant M(\alpha_i(B) - \mu_i)^2.$$

Теорема доказана.

Теорема Гаусса — Маркова следует из доказанной, если в (5.42) положить $B = 0$. В этом случае $\mu = 0$.

§ 6. Наилучшие линейные оценки

Итак, среди линейных несмещенных оценок оценки метода наименьших квадратов лучшие независимо от того, каков закон распределения помехи.

В следующих параграфах мы рассмотрим более широкий класс оценок — линейные оценки, и найдем в них наилучшие. Эти оценки будут отличаться от оценок метода наименьших квадратов при наличии нетривиальной априорной информации об оцениваемых параметрах. В тех же случаях, когда нетривиальной априорной информации нет, наилучшей линейной оценкой остается оценка метода наименьших квадратов.

Пусть в схеме Гаусса — Маркова оцениваются параметры регрессии

$$\bar{y} = y(x) = \sum_{i=1}^{n} \alpha_i^0 \psi_i(x) \tag{5.48}$$

по эмпирическим данным $x_1, y_1; \ldots; x_l, y_l$. Пусть $\hat{\psi}_1(x), \ldots, \hat{\psi}_n(x)$ — дважды ортогональный базис

$$\int \hat{\psi}_i(x)\,\hat{\psi}_j(x)\,P(x)\,dx = \begin{cases} \lambda_i, & \text{если} \quad i=j, \\ 0, & \text{если} \quad i\neq j, \end{cases}$$

$$\sum_{r=1}^{l} \hat{\psi}_i(x_r)\,\hat{\psi}_j(x_r) = \begin{cases} l, & \text{если} \quad i=j, \\ 0, & \text{если} \quad i\neq j. \end{cases} \tag{5.49}$$

Рассмотрим класс линейных оценок:

$$\hat{\alpha}_p = \theta_p^{\mathrm{T}} Y + \beta_0^p, \tag{5.50}$$

где

$$\theta_p = (\theta_1^p, \ldots, \theta_l^p)^{\mathrm{T}}, \quad Y = (y_1, \ldots, y_l)^{\mathrm{T}}.$$

Введем систему ортогональных векторов:

$$\chi_1, \ldots, \chi_l; \quad \chi_i^{\mathrm{T}}\chi_j = \begin{cases} l, & \text{если} \quad i=j, \\ 0, & \text{если} \quad i\neq j, \end{cases} \tag{5.51}$$

у которой первые n векторов есть

$$\chi_i = (\hat{\psi}_i(x_1), \ldots, \hat{\psi}_i(x_l))^{\mathrm{T}}, \qquad i = 1, \ldots, n.$$

Представим теперь вектор θ_p в разложении по (5.51):

$$\theta_p = \sum_{i=1}^{l} \beta_i^p \chi_i. \tag{5.52}$$

Тогда равенство (5.50) перепишется в виде

$$\hat{\alpha}_p = \sum_{i=1}^{l} \beta_i^p \chi_i^T Y + \beta_0^p. \tag{5.53}$$

Выразим величину уклонения $M(\hat{\alpha}_p - \alpha_p^0)^2$ через параметры β. Для этого воспользуемся тождеством

$$M(\hat{\alpha}_p - \alpha_p^0)^2 = (M(\hat{\alpha}_p - \alpha_p^0))^2 + M(\hat{\alpha}_p - M\hat{\alpha}_p)^2. \tag{5.54}$$

Первое слагаемое правой части равно

$$(M(\hat{\alpha}_p - \alpha_p^0))^2 = \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2.$$

Второе слагаемое равно

$$M(\hat{\alpha}_p - M\hat{\alpha}_p)^2 = l\sigma^2 \sum_{i=1}^{l} (\beta_i^p)^2.$$

Таким образом,

$$M(\hat{\alpha}_p - \alpha_p^0)^2 =$$

$$= \sigma^2 l \sum_{i=1}^{l} (\beta_i^p)^2 + \left(l \sum_{i=1}^{n} \beta_i^p \alpha_i^0 + \beta_0^p - \alpha_p^0\right)^2 = \mathscr{D}^p(\beta \mid \alpha, \sigma). \tag{5.55}$$

Наилучшей линейной оценкой является такая оценка, которая минимизирует (5.55).

§ 7. Критерии качества оценок

Найти наилучшую линейную оценку можно, непосредственно минимизируя по $\beta_1, \ldots, \beta_l$ правую часть равенства (5.55). Минимум выражения (5.55) достигается при $\beta_1^p = \beta_2^p = \ldots = \beta_l^p = 0$ и $\beta_0^p = \alpha_p^0$ и равен нулю.

Таким образом, для каждой конкретной задачи (конкретных α_0 и σ) может быть указана тривиальная оценка, доставляющая минимум квадрату уклонения. Проблема же состоит в том, чтобы построить линейную оценку, предназначенную для решения не одной задачи, а класса задач.

Зададим множество задач $R(\alpha, \sigma)$, на которые рассчитан алгоритм, неравенствами

$$\begin{aligned} a_p \leqslant \alpha_p \leqslant b_p, \\ d \leqslant \sigma \leqslant e. \end{aligned} \tag{5.56}$$

Определим качество алгоритма оценивания параметра α_p из множества $R(\alpha, \sigma)$.

Как обычно в такой ситуации рассмотрим две идеи: байесову и минимаксную. В соответствии с этими идеями введем разные понятия качества линейной оценки.

Согласно принципу Байеса наилучшим методом оценивания считается тот, для которого среднее значение критерия по множеству задач из $R(\alpha, \sigma)$ минимально (мера на этом множестве задается распределением $P(\alpha, \sigma)$).

Определение. *Оценка*

$$\alpha_p^{(1)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

называется линейной наилучшей в среднем, если она среди всех линейных оценок доставляет минимум функционалу

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta \mid \alpha, \sigma) P(\alpha, \sigma) \, d\alpha_1 \ldots d\alpha_n \, d\sigma. \tag{5.57}$$

Ниже мы вычислим байесову оценку для случая, когда параметры α и σ распределены независимо согласно равномерному закону на соответствующих интервалах, т. е.

$$P(\alpha, \sigma) = \begin{cases} \prod_{i=1}^{n} \frac{1}{(b_i - a_i)} \cdot \frac{1}{e-d}, & \text{если } a_p \leqslant \alpha_p \leqslant b_p, \ d \leqslant \sigma \leqslant e, \\ 0 & \text{в противном случае.} \end{cases} \tag{5.58}$$

Таким образом, качество оценки определяется функционалом

$$\mathscr{D}_1^p(\beta) = \int \mathscr{D}^p(\beta \mid \alpha, \sigma) \prod_{i=1}^{n} \frac{d\alpha_i}{(b_i - a_i)} \frac{d\sigma}{(e-d)}. \tag{5.59}$$

В соответствии с принципом минимакса наилучшим методом оценивания считается такой метод, который доставляет минимум $\mathscr{D}^p(\beta \mid \alpha, \sigma)$ для самой неблагоприятной задачи (пары α, σ).

Определение. *Оценка*

$$\alpha_p^{(2)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

называется наилучшей линейной минимаксной оценкой для класса $R(\alpha, \sigma)$, *если она в классе линейных оценок доставляет минимум функционалу*

$$\mathscr{D}_2^p(\beta) = \sup_{\alpha,\ \sigma} \mathscr{D}^p(\beta \mid \alpha, \sigma). \tag{5.60}$$

Вообще говоря, в классе $R(\alpha, \sigma)$ могут существовать задачи, для которых введенные оценки $\alpha_p^{(1)}$ и $\alpha_p^{(2)}$ будут хуже оценок метода наименьших квадратов $\beta_{\text{мнк}}^p = (0, \ldots, 1/l, \ldots 0)^{\mathrm{r}}$, $\beta_0^p = 0$ (отлична от нуля лишь p-я координата вектора $\beta_{\text{мнк}}^p$). Поэтому определим третью оптимальную оценку так, чтобы она была равномерно лучше оценки метода наименьших квадратов. Для этого введем функцию потерь

$$\mathscr{D}_3^p(\beta) = \sup_{\alpha,\ \sigma} (\mathscr{D}^p(\beta \mid \alpha, \sigma) - \mathscr{D}^p(\beta_{\text{мнк}} \mid, \alpha, \sigma)) \tag{5.61}$$

и потребуем, чтобы оптимальная оценка доставляла минимум выражению (5.61).

Определение. *Оценка*

$$\alpha_p^{(3)} = \sum \beta_i^p \chi_i^{\mathrm{T}} Y + \beta_0^p$$

называется линейной равномерно лучшей, чем оценка наименьших квадратов, если она доставляет в классе линейных оценок минимум функционалу (5.61).

§ 8. Вычисление наилучших линейных оценок

Основное содержание теории наилучшего линейного оценивания заключено в следующих теоремах.

Теорема 5.6 (Кощеев). *Наилучшие линейные оценки параметра* α_p *для класса* $R(\alpha, \sigma)$ *имеют вид*

$$\alpha_p^{(i)} = \frac{\alpha_{\text{мнк}}^p + \dfrac{c_p}{l}\rho_p^{(i)}}{1 + \dfrac{1}{l}\rho_p^{(i)}}, \quad i = 1, 2, 3, \tag{5.62}$$

где $c_p = \frac{a_p + b_p}{2}$, $\alpha^p_{\text{мнк}}$ *— оценка метода наименьших квадратов,* $\alpha^{(1)}_p$ *— наилучшая в среднем оценка,*

$$\rho^{(1)}_p = 4\,\frac{d^2 + de + e^2}{(a_p - b_p)^2}, \tag{5.63}$$

$\alpha^{(2)}_p$ *— наилучшая минимаксная оценка,*

$$\rho^{(2)}_p = 4\,\frac{e^2}{(a_p - b_p)^2}, \tag{5.64}$$

$\alpha^{(3)}_p$ *— равномерно лучшая оценка,*

$$\rho^{(3)}_p = 4\,\frac{d^2}{(a_p - b_p)^2}. \tag{5.65}$$

Таким образом, оказывается, что наилучшие линейные оценки будут смещенными. Структура оценок задается выражением (5.62), где $\rho^{(i)}_p$ определяется по формулам (5.63)—(5.65) в зависимости от конкретного содержания понятия качества оценки. Существует простое соотношение, которое показывает, во сколько раз байесова или минимаксная оценка лучше оценок метода наименьших квадратов.

Теорема 5.7 (Кощеев). *Справедливо равенство*

$$\mathscr{D}^p_l(\alpha^{(i)}_p) = \frac{1}{1 + \frac{1}{l}\,\rho^{(i)}_p}\,\mathscr{D}^p_l(\alpha^p_{\text{мнк}}), \quad i = 1, 2. \tag{5.66}$$

Согласно теореме 5.7 величина $\left(1 + \frac{1}{l}\,\rho^{(i)}_p\right)$ определяет, во сколько раз оптимальные оценки $\alpha^{(i)}_p$ лучше оценки метода наименьших квадратов. Преимущество оценок $\alpha^{(i)}_p$ тем больше, чем меньше объем выборки l.

Ниже приведено доказательство теоремы 5.6. Справедливость же теоремы 5.7 будет следовать из более общей теоремы, рассмотренной в следующем параграфе.

Доказательство теоремы 5.6.

1. Вывод наилучшей в среднем линейной оценки. Выпишем функционал, минимум которого определяет в наших

условиях наилучшую оценку в среднем:

$$\mathscr{D}_1^p(\beta)=\int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}\int_d^e\left[l\sigma^2\sum_{i=1}^{l}(\beta_i^p)^2+\left(l\sum_{i=1}^{n}\beta_i^p\alpha_i+\beta_0^p-\alpha_p\right)^2\right]\times$$

$$\times\prod_{i=1}^{n}\frac{d\alpha_i}{b_i-a_i}\,\frac{d\sigma}{e-d}. \qquad (5.67)$$

Этот интеграл легко может быть вычислен

$$\mathscr{D}_1^p(\beta)=\frac{l}{3}\,\frac{e^3-d^3}{e-d}\sum_{i=1}^{l}(\beta_i^p)^2+$$

$$+\prod_{j=1}^{n}\frac{1}{(b_j-a_j)}\int_{a_1}^{b_1}\cdots\int_{a_n}^{b_n}\left(l\sum_{i=1}^{n}\beta_i^p\alpha_i+\beta_0^p-\alpha_p\right)^2 d\alpha_1\ldots d\alpha_n.$$

Обозначим $\frac{a_i+b_i}{2}=c_i$; $\frac{a_i-b_i}{2}=\mathscr{M}_i$; $t_i=\alpha_i-c_i$ и произведем замену переменных

$$\mathscr{D}_1^p(\beta)=\frac{l}{3}\,\frac{e^3-d^3}{e-d}\sum_{i=1}^{l}(\beta_i^p)^2+$$

$$+\prod_{j=1}^{n}\frac{1}{2\mathscr{M}_j}\int_{-\mathscr{M}_1}^{\mathscr{M}_1}\cdots\int_{-\mathscr{M}_n}^{\mathscr{M}_n}\left(l\sum_{i=1}^{n}\beta_i^p(t_i+c_i)+\right.$$

$$\left.+\beta_0^p-(t_p+c_p)\right)^2 dt_1\ldots dt_n. \qquad (5.68)$$

Так как интегрирование идет по симметричным интервалам $[-\mathscr{M}, \mathscr{M}]$, то линейные по t члены обращаются в нуль. Получаем

$$\mathscr{D}_1^p(\beta)=\frac{l}{3}(e^2+ed+d^2)\sum_{i=1}^{l}(\beta_i^p)^2+\left(\beta_0^p+\sum_{i=1}^{n}(l\beta_i^p-\delta_{ip})c_i\right)^2+$$

$$+\prod_{j=1}^{n}\frac{1}{2\mathscr{M}_j}\int_{-\mathscr{M}_1}^{\mathscr{M}_1}\cdots\int_{-\mathscr{M}_n}^{\mathscr{M}_n}\sum_{i=1}^{n}(l\beta_i^p-\delta_{ip})^2 t_i^2\,dt_1\ldots dt_n. \qquad (5.69)$$

В выражении (5.69) использовано обозначение

$$\delta_{ip}=\begin{cases}1, & \text{если } i=p,\\ 0, & \text{если } i\neq p.\end{cases}$$

Окончательно получаем

$$\mathscr{D}_1^p(\beta)=\frac{l}{3}(e^2+ed+d^2)\sum_{i=1}^{l}(\beta_i^p)^2+$$

$$+\Big(\beta_0^p+\sum_{i=1}^{n}(l\beta_i^p-\delta_{ip})c_i\Big)^2+\sum_{i=1}^{n}\frac{\mathscr{M}_i^2}{3}(l\beta_i^p-\delta_{ip})^2. \quad (5.70)$$

Для того чтобы найти наилучшую в среднем линейную оценку, нам осталось минимизировать по параметрам β выражение (5.70).

Приравнивая нулю частные производные выражения (5.70), получим, что

$$\begin{aligned}&\beta_i^p=0, \quad \text{если} \quad i\neq p,\\ &\beta_0^p=-c_p(l\beta_p^p-1),\\ &\beta_p^p=\frac{\dfrac{\mathscr{M}_p^2}{e^2+ed+d^2}}{1+\dfrac{\mathscr{M}_p^2 l}{e^2+ed+d^2}}.\end{aligned} \quad (5.71)$$

Подставим найденные значения (5.71) в (5.53):

$$\alpha_p^{(1)}=\frac{\dfrac{\mathscr{M}_p^2}{e^2+ed+d^2}}{1+\dfrac{\mathscr{M}_p^2 l}{e^2+ed+d^2}}\chi_p^{\mathrm{T}}Y+\frac{c_p}{1+\dfrac{\mathscr{M}_p^2 l}{e^2+ed+d^2}}.$$

Введем теперь обозначения $\rho_p^{(1)}=\dfrac{e^2+ed+d^2}{\mathscr{M}_p^2}$. В этих обозначениях

$$\alpha_p^{(1)}=\frac{\dfrac{1}{l}\chi_p^{\mathrm{T}}Y+\dfrac{c_p}{l}\rho_p^{(1)}}{1+\dfrac{1}{l}\rho_p^{(1)}}.$$

Заметим, что величина $\dfrac{1}{l}\chi_p^{\mathrm{T}}Y$ есть оценка параметра α_p^0, полученная методом наименьших квадратов. Таким образом,

$$\alpha_p^{(1)}=\frac{\alpha_{\text{мнк}}^p+\dfrac{c_p}{l}\rho_p^{(1)}}{1+\dfrac{1}{l}\rho_p^{(1)}}.$$

Первая часть теоремы доказана.

2. Вывод наилучшей минимаксной оценки. Функционал, минимум которого определяет наилучшую минимаксную

оценку, равен

$$\mathscr{D}_2^p(\beta)=\sup_{\sigma,\,\alpha}\left[\sigma^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+\left(l\sum_{i=1}^{n}\left(\beta_i^p\alpha_i+\beta_0^p-\alpha_p\right)\right)^2\right]. \qquad (5.72)$$

Используем обозначения

$$c_i=\frac{b_i+a_i}{2},\quad \mathscr{M}_i=\frac{b_i-a_i}{2},\quad t_i=\alpha_i-c_i,$$

и произведем замену переменных в (5.72):

$$\mathscr{D}_2^p(\beta)=$$

$$=e^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+\sup_{|t_i|\leqslant\mathscr{M}_i}\left[\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)\left(t_i+c_i\right)+\beta_0^p\right]^2=$$

$$=e^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+\sup_{|t_i|\leqslant\mathscr{M}_i}\left[\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)t_i+\right.$$

$$\left.+\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)c_i+\beta_0^p\right]^2=e^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+$$

$$+\left[\sum_{i=1}^{n}\left|l\beta_i^p-\delta_{ip}\right|\mathscr{M}_i+\left|\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)c_i+\beta_0^p\right|\right]^2.$$

Таким образом,

$$\mathscr{D}_2^p(\beta)=e^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+$$

$$+\left[\sum_{i=1}^{n}\left|l\beta_i^p-\delta_{ip}\right|\mathscr{M}_i+\left|\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)c_i+\beta_0^p\right|\right]^2. \qquad (5.73)$$

Найдем минимум (5.73). Выбором β_0^p второй член суммы в квадратных скобках можно обратить в нуль:

$$\beta_0^p=-\sum_{i=1}^{n}\left(l\beta_i^p-\delta_{ip}\right)c_i.$$

Поэтому достаточно минимизировать

$$\mathscr{D}_2^p(\beta)=e^2 l\sum_{i=1}^{l}\left(\beta_i^p\right)^2+\left(\sum_{i=1}^{n}\left|l\beta_i^p-\delta_{ip}\right|\mathscr{M}_i\right)^2. \qquad (5.74)$$

Минимум же (5.74) достигается при

$$\beta_i^p=0\quad\text{для}\quad i\neq p, \qquad (5.75)$$

откуда при $\beta_i^p=0$ $(i\neq p)$ функционал (5.74) равен

$$\mathscr{D}_2^p(\beta)\Big|_{\beta_i^p=0\,(i\neq p)}=le^2\left(\beta_p^p\right)^2+\left(l\beta_p^p-1\right)^2\mathscr{M}_p^2. \qquad (5.76)$$

Минимум выражения (5.76) достигается при

$$\beta_p^p = \frac{\mathcal{M}_p^2}{e^2 + \mathcal{M}_{pl}^2}. \tag{5.77}$$

Подставляя (5.75), (5.77) в (5.53), найдем наилучшую минимаксную оценку

$$\alpha_p^{(2)} = \frac{\mathcal{M}_p^2}{e^2 + \mathcal{M}_{pl}^2} \chi_p^{\mathrm{T}} Y + \left(\frac{l\mathcal{M}_p^2}{e^2 + \mathcal{M}_{pl}^2} - 1 \right) c_p = \frac{l\mathcal{M}_p^2 \alpha_{\text{мнк}}^p + c_p e^2}{e^2 + \mathcal{M}_{pl}^2}.$$

Введя обозначение $\rho_p^{(2)} = \dfrac{e^2}{\mathcal{M}_p^2}$, получаем

$$\alpha_p^{(2)} = \frac{\alpha_{\text{мнк}}^p + \dfrac{c_p}{l} \rho_p^{(2)}}{1 + \dfrac{1}{l} \rho_p^{(2)}}.$$

3. Вывод равномерно лучшей линейной оценки. Для вычисления равномерно лучшей оценки надо минимизировать функционал

$$\mathscr{D}_3^p(\beta) = \sup_{\alpha,\ \sigma} \left(\mathscr{D}^p(\beta \mid \alpha,\ \sigma) - \mathscr{D}^p(\beta_{\text{мнк}} \mid \alpha,\ \sigma) \right),$$

или в явном виде

$$\mathscr{D}_3^p(\beta) = \sup_{d \leqslant \sigma \leqslant e} \left[l\sigma^2 \left(\sum_{i=1}^{l} (\beta_i^p)^2 - 1 \right) \right] + \\ + \sup_{a_i \leqslant \alpha_i \leqslant b_i} \left[\sum_{i=1}^{n} \left(l\beta_i^p - \delta_{ip} \right) \alpha_i + \beta_0^p \right]^2. \tag{5.78}$$

Нетрудно видеть, что все вычисления в этом случае совпадают с только что проделанными, за исключением того, что если

$$\sum_{i=1}^{l} \beta_i^2 - 1 < 0, \tag{5.79}$$

то вместо $e = \sup \sigma$ следует брать $d = \inf \sigma$.

Следовательно,

$$\beta_0^p = -\sum_{i=1}^{l} \left(l\beta_i^p - \delta_{ip} \right) c_i,$$

$$\beta_i^p = \begin{cases} 0, & \text{если } i \neq p, \\ \dfrac{\mathcal{M}_i^2}{s^2 + \mathcal{M}_i^2 l}, & \text{если } i = p, \end{cases} \tag{5.80}$$

где s — либо $\inf \sigma$, либо $\sup \sigma$ в зависимости от знака выражения $\sum_{i=1}^{n} \beta_i^p - 1$.

Но при значениях (5.80) выражение (5.79) отрицательно

$$\sum_{i=1}^{l} \beta_i - 1 = \frac{\mathscr{M}_p^2}{s^2 + \mathscr{M}_p^2} - 1 = -\frac{s^2}{s^2 + \mathscr{M}_p^2} < 0.$$

Следовательно, $s = \inf \sigma = d$. Таким образом, равномерно лучшая линейная оценка равна

$$\alpha_p^{(3)} = \frac{\alpha_{\text{мнк}}^p + \frac{c_p}{l} \rho_p^{(3)}}{1 + \frac{1}{l} \rho_p^{(3)}},$$

где на этот раз

$$\rho_p^{(3)} = \frac{d^2}{\mathscr{M}_p^2}.$$

§ 9. Использование априорной информации

Итак, согласно теореме 5.6, знание априорной информации:

1) интервала $[a_i, b_i]$, которому принадлежит оцениваемый параметр α_p,

2) интервала $[d, e]$, которому принадлежит дисперсия помехи σ, позволяют строить наилучшие линейные оценки. Согласно же теореме 5.7 функционал, определяющий качество наилучшей линейной оценки, в $\left(1 + \frac{\rho_p^{(i)}}{l}\right)$ раз меньше функционала для оценки, полученной методом наименьших квадратов.

Обычно получение априорной информации при решении практических задач в схеме Гаусса — Маркова не вызывает серьезных трудностей. Как правило, заранее известны интервалы, в которых лежат измеряемые значения y:

$$\tau_i \leqslant y_i \leqslant T_i. \tag{5.81}$$

Сведения об этих интервалах отражают либо длительный опыт, либо знание законов природы. Например, при построении регрессии для прогноза температуры в заданном пункте Земли в 166-й день года заранее известно, что прогнозируемая величина t лежит в заданных пределах $+10° \leqslant t \leqslant +35°$. Знание оценок (5.81) позволяет найти

интервалы оцениваемых параметров. Из равенства $\alpha_p^0 = M\frac{1}{l}\chi_p^{\mathrm{T}}Y$ следует

$$b_p = \sup_Y \frac{1}{l}\chi_p^{\mathrm{T}}Y \leqslant \frac{1}{l}\left(\sum_{i=1}^{l}{}' T_i\hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' \tau_i\hat{\psi}_p(x_i)\right).$$

Здесь в первую сумму $\sum'$ включены положительные координаты вектора $\chi_p = (\hat{\psi}_p(x_1), \ldots, \hat{\psi}_p(x_l))^{\mathrm{T}}$, во вторую сумму $\sum''$ — отрицательные координаты. Аналогично находятся оценки

$$a_p = \inf_Y \frac{1}{l}\chi_p^{\mathrm{T}}Y \geqslant \frac{1}{l}\left(\sum_{i=1}^{l}{}' \tau_i\hat{\psi}_p(x_i) + \sum_{i=1}^{l}{}'' T_i\hat{\psi}_p(x_i)\right).$$

Для оценки интервала дисперсии также могут быть привлечены опыт и знание законов образования помех. Однако если полученный интервал дисперсии окажется слишком широким, может быть применен вероятностный подход, который заключается в том, чтобы выбирать интервал, который с большой вероятностью содержит истинное значение дисперсии.

Известно, что величина

$$\sigma_э^2 = \frac{\sum_{i=1}^{l} y_i^2 - l\sum_{p=1}^{n}(\alpha_{\text{мнк}}^p)^2}{l-n}$$

является несмещенной оценкой дисперсии помехи. Воспользуемся неравенством Чебышева

$$P\left\{\sigma_э^2 \geqslant \frac{\sigma^2}{\eta}\right\} \leqslant \eta,$$

откуда следует, что с вероятностью $1-\eta$

$$\sigma^2 \geqslant \sigma_э^2\eta. \tag{5.82}$$

Оценка (5.82) может быть уточнена, если известен характер распределения помехи.

По интервалу дисперсии $d \leqslant \sigma \leqslant e$ и интервалу принадлежности параметра α_p находятся параметры $\rho_p^{(i)}$, $c_p^{(i)}$, с помощью которых строятся оптимальные линейные оценки.

Заметим, что чем более неопределенна априорная информация (шире интервал принадлежности оценок), тем меньше величина $\rho_p^{(i)}$ и тем ближе наилучшая линейная оценка к оценке метода наименьших квадратов. Можно показать, что при тривиальной априорной информации ($-\infty < \alpha_p < \infty$; $0 < \sigma < \infty$) наилучшая линейная оценка совпадает с оценкой метода наименьших квадратов.

Для завершения теории наилучшего линейного оценивания нам остается выяснить, насколько чувствительны к точности априорной информации методы линейного оценивания. Ответ на этот вопрос дает теорема 5.8.

Теорема 5.8 (Кощеев). *Пусть* $\hat{\alpha}_p^i = \alpha_p(\hat{\rho}_p^{(i)}, \hat{c}_p)$ *— наилучшая линейная оценка, вычисленная по приближенным значениям параметров* $\hat{\rho}_p^{(i)}$, $\hat{c}_p$, $\hat{\mathscr{M}}_p$, *в то время как истинные значения параметров равны* $\rho_p^{(i)}$, c_p, $\mathscr{M}_p$. *Тогда качество полученной оценки будет равно*

$$\mathscr{D}_i^p(\hat{\alpha}_p(\hat{\rho}_p^{(i)}, \hat{c}_p)) = \frac{1 + \nu_i \frac{(\hat{\rho}_p^{(i)})^2}{\rho_p^{(i)}}}{(1+\hat{\rho}_p^{(i)})^2} \mathscr{D}_i^p(\alpha_{\text{мнк}}^p) \quad (i = 1, 2), \tag{5.83}$$

где

$$\nu_1 = 1 + 3\left(\frac{\hat{c}_p - c_p}{\hat{\mathscr{M}}_p}\right)^2, \quad \nu_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\hat{\mathscr{M}}_p}\right)^2. \tag{5.84}$$

Заметим, что теорема 5.7 является частным случаем теоремы 5.8 при $\hat{c}_p = c_p$ и $\hat{\rho}_p^{(i)} = \rho_p^{(i)}$.

Из равенства (5.83) следует, что если значение параметра $\hat{\rho}_p^{(i)}$ связано с $\rho_p^{(i)}$, ν_i соотношением

$$\rho_p^{(i)} > \frac{\hat{\rho}_p^{(i)} \nu_i}{2 + \hat{\rho}_p^{(i)}}, \tag{5.85}$$

то полученная с помощью $\hat{\rho}_p^{(i)}$, $\hat{c}_p$ оценка будет лучше оценки метода наименьших квадратов. Следовательно, выбирать $\hat{\rho}_p^{(i)}$ приходится, исходя из двух противоречивых соображений. Для того чтобы получить оценку не хуже оценки метода наименьших квадратов, необходимо занижать $\hat{\rho}_p^{(i)}$ (чтобы выполнилось (5.85)). Однако при этом падает выигрыш, приближенно равный $\dfrac{\mathscr{D}_i(\alpha_{\text{мнк}}^p)}{1+\hat{\rho}^{(i)}}$.

Доказательство теоремы 5.8. Прежде всего вычислим значение критерия (5.55) на оценке $\hat{\alpha}_p(\hat{\rho}^{(i)}c_p)$:

$$M(\alpha_p(\hat{\rho}_p^{(i)},\ \hat{c}_p)-\alpha_p^0)^2=$$

$$=M\left(\frac{\alpha_{\text{мнк}}^p+\frac{\hat{c}_p}{l}\hat{\rho}_p^{(i)}}{1+\frac{\hat{\rho}_p^{(i)}}{l}}-\frac{\alpha_p^0+\frac{\hat{c}_p}{l}\hat{\rho}_p^{(i)}}{1+\hat{\rho}_p^{(i)}\frac{1}{l}}\right)^2+\left(\frac{\alpha_p^0+\frac{\hat{c}_p}{l}\hat{\rho}_p^{(i)}}{1+\hat{\rho}_p^{(i)}\frac{1}{l}}-\alpha_p^0\right)^2=$$

$$=\frac{\frac{\sigma^2}{l}}{\left(1+\frac{\hat{\rho}_p^{(i)}}{l}\right)^2}+\frac{\left(\frac{\hat{\rho}_p^{(i)}}{l}\right)^2(\hat{c}_p-\alpha_p^0)^2}{\left(1+\frac{\hat{\rho}_p^{(i)}}{l}\right)^2}.$$

Оба соотношения теоремы подтверждаются элементарными выкладками

$$\mathscr{D}_1^p(\hat{\alpha})=\int_{c_p-\mathscr{M}_p}^{c_p+\mathscr{M}_p}\int_d^e\frac{\frac{\sigma^2}{l}+\left(\frac{\hat{\rho}_p^{(1)}}{l}\right)^2(\hat{c}_p-\alpha_p^0)}{\left(1+\frac{1}{l}\hat{\rho}_p^{(1)}\right)^2}\frac{d\sigma}{e-d}\frac{d\alpha_p^0}{2\mathscr{M}_p}=$$

$$=\frac{\frac{1}{3}\frac{e^3-d^3}{e-d}+\left(\frac{\hat{\rho}_p^{(1)}}{l}\right)^2\left(\frac{\mathscr{M}_p^3}{3}+(c_p-\hat{c}_p)^2\right)}{\left(1+\frac{1}{l}\hat{\rho}_p^{(1)}\right)^2},$$

$$\mathscr{D}_1^p(\alpha_{\text{мнк}}^p)=\int_{c_p-\mathscr{M}_p}^{c_p+\mathscr{M}_p}\frac{d\alpha}{2\mathscr{M}_p}\int_d^e\frac{\sigma^2}{l}\frac{d\sigma}{e-d}=\frac{e^2+ed+d^2}{3l},$$

откуда следует

$$\frac{\mathscr{D}_1^p(\hat{\alpha})}{\mathscr{D}_1^p(\alpha_{\text{мнк}})}=\frac{1+\frac{1}{\rho_p^1}\left(\frac{\hat{\rho}_p^{(1)}}{l}\right)^2\nu_1}{\left(1+\frac{1}{l}\hat{\rho}_p^{(1)}\right)^2},\qquad \nu_1=1+3\left(\frac{\hat{c}_p-c_p}{\mathscr{M}_p}\right)^2.$$

Вычислим

$$\mathscr{D}_2^p(\hat{\alpha})=$$

$$=\sup_{\alpha,\ \sigma}\frac{\frac{\sigma^2}{l}+\left(\frac{\hat{\rho}_p^{(2)}}{l}\right)^2(c_p-\alpha_p^0)^2}{\left(1+\frac{\hat{\rho}_p^{(2)}}{l}\right)^2}=\frac{\frac{e^2}{l}+\left(\frac{\hat{\rho}_p^{(2)}}{l}\right)^2(|\hat{c}_p-c_p|+\mathscr{M}_p)^2}{\left(1+\hat{\rho}_p^2\frac{1}{l}\right)^2}.$$

С другой стороны,

$$\mathscr{D}_2^p(\alpha_{\text{мнк}}) = \sup_{\sigma} \frac{\sigma^2}{l} = \frac{e^2}{l},$$

откуда следует

$$\frac{\mathscr{D}_2^p(\hat{\alpha}_p)}{\mathscr{D}_2^p(\alpha_{\text{мнк}})} = \frac{1 + \frac{1}{\rho_p^{(2)}} \left(\frac{\hat{\rho}_p^{(2)}}{l}\right)^2 \nu_2}{\left(1 + \frac{1}{l}\hat{\rho}_p^{(2)}\right)^2}, \qquad \nu_2 = \left(1 + \frac{|c_p - \hat{c}_p|}{\mathscr{M}_p}\right)^2.$$

Теорема доказана

Итак, мы рассмотрели теорию оценивания параметров регрессии. В основе теории лежит факт экстремальности метода наименьших квадратов в некотором узком классе методов (в классе несмещенных методов оценивания в теории нормальной регрессии и в классе линейных несмещенных методов в общей теории регрессии).

Затем оказалось, что в классе смещенных методов оценивания можно строить оценки лучшие, чем те, которые следуют из метода наименьших квадратов.

Такие нелинейные смещенные методы оценивания были найдены для оценивания параметров нормальной регрессии и линейные смещенные методы в общей схеме восстановления регрессии.

Приведенные методы оценивания удается использовать для восстановления регрессии, если известна плотность $P(x)$, а регрессия действительно является линейной по параметрам функцией.

Основные утверждения главы V

1. Существуют два пути решения задачи восстановления регрессии: оценивание параметров регрессии и приближение функции регрессии.

Эти пути равнозначны, если восстановление линейной по параметрам регрессии проводится в классе функций, заданных разложением по ортонормальной с весом $P(x)$ фундаментальной системе функций.

2. Восстановление параметров регрессии исследуется в схеме Гаусса — Маркова. В этой схеме для нормальной

регрессии применение метода наименьших квадратов гарантирует получение совместно эффективных оценок.

3. Использование нелинейных смещенных оценок вектора средних нормального закона позволяет получать приближения к нормальной регрессии лучшие, чем те, которые следуют из метода наименьших квадратов.

4. В классе оценок, являющихся одновременно линейными и несмещенными, метод наименьших квадратов оказывается наилучшим методом оценивания независимо от закона распределения помехи.

5. Более точные методы оценивания в этом случае могут быть реализованы в классе линейных смещенных оценок при наличии нетривиальной априорной информации о задаче. Такая априорная информация на практике может быть получена, и на ее основе могут быть построены оптимальные (в разных смыслах) линейные методы оценивания параметров.

ГЛАВА VI

МЕТОД МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА В ЗАДАЧЕ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ

§ 1. Метод минимизации эмпирического риска

В предыдущих трех главах восстановление зависимостей мы связывали с методами восстановления плотности вероятностей. Отыскание функции, минимизирующей средний риск

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{6.1}$$

по эмпирическим данным

$$x_1, y_1; \ldots; x_l, y_l, \tag{6.2}$$

мы сводили к восстановлению плотности $\hat{P}(x, y)$ по выборке (6.2) и минимизации функционала

$$I_э(\alpha) = \int (y - F(x, \alpha))^2 \hat{P}(x, y)\, dx\, dy.$$

Как отмечалось в главе II, этот путь минимизации риска (6.1), вообще говоря, не является рациональным — задача восстановления плотности более трудная, чем минимизация среднего риска. И лишь когда об искомой плотности $P(x, y)$ имеется настолько большая априорная информация, что функция $P(x, y)$ может быть задана с точностью до параметров, такой путь оказывается приемлемым. Разработанные для этого случая методы параметрической статистики и были использованы в предыдущих главах.

Однако при решении конкретных задач структура плотности $P(x, y)$ неизвестна. Таким образом, успех применения методов параметрической статистики оказывается основанным на вере в то, что используемая гипотетическая структура плотности соответствует истинной.

Начиная с этой главы, мы будем изучать методы восстановления зависимостей, для реализации которых не нужно восстанавливать плотность. В основе этих методов лежит принцип минимизации эмпирического риска, согласно

которому за точку минимума функционала (6.1) принимается точка минимума эмпирического функционала

$$I_э(\alpha) = \frac{1}{l}\sum_{i=1}^{l}(y_i - F(x_i, \alpha))^2, \tag{6.3}$$

построенного по случайной независимой выборке (6.2). Пусть минимум функционала (6.3) достигается на функции $F(x, \alpha_э)$. Проблема состоит в том, чтобы установить, в каких случаях найденная функция $F(x, \alpha_э)$ близка к функции $F(x, \alpha_0)$, минимизирующей (6.1) в $F(x, \alpha)$.

Ранее (§ 6 гл. II) мы связали эту проблему с проблемой существования равномерной сходимости средних к математическим ожиданиям, т. е. с ситуацией, когда для любой заданной величины уклонения $\varkappa$ может быть указано неравенство

$$P\{\sup_{\alpha} |I(\alpha) - I_э(\alpha)| > \varkappa\} < \eta. \tag{6.4}$$

Пусть выполнено (6.4). Тогда справедливо неравенство

$$P\{I(\alpha_э) - I(\alpha_0) > 2\varkappa\} < \eta. \tag{6.5}$$

Иначе говоря, в ситуации (6.4) с вероятностью $1-\eta$ уклонение наилучшего в классе $F(x, \alpha)$ решения $F(x, \alpha_0)$ от решения, доставляющего минимум эмпирическому риску $F(x, \alpha_э)$, составит величину, не превышающую $2\varkappa$.

Действительно, из условия (6.4) следует, что с вероятностью $1-\eta$ одновременно выполнятся два неравенства

$$I(\alpha_э) - I_э(\alpha_э) < \varkappa, \qquad I_э(\alpha_0) - I(\alpha_0) < \varkappa. \tag{6.6}$$

Кроме того, поскольку $\alpha_э$ и α_0 — точки минимума $I_э(\alpha)$ и $I(\alpha)$, то справедливо неравенство

$$I_э(\alpha_э) \leqslant I_э(\alpha_0). \tag{6.7}$$

Из неравенств (6.6), (6.7) вытекает, что

$$I(\alpha_э) - I(\alpha_0) < 2\varkappa. \tag{6.8}$$

А так как оба неравенства (6.6) одновременно выполняются с вероятностью $1-\eta$, то и неравенство (6.8) выполнится с вероятностью $1-\eta$. Следовательно,

$$P\{I(\alpha_э) - I(\alpha_0) > 2\varkappa\} < \eta. \tag{6.9}$$

В этой главе мы рассмотрим теорию равномерной сходимости средних к математическим ожиданиям применительно к задаче обучения распознаванию образов, т. е. для случая, когда функция потерь в функционале среднего риска принимает только два значения — нуль и единица. В гл. VII для задачи восстановления регрессии мы распространим полученные здесь результаты на общий случай, когда функция потерь принимает произвольные значения из интервала $(0, \infty)$.

§ 2. Равномерная сходимость частот появления событий к их вероятностям

Рассмотрим функционал, минимизация которого составляет суть задачи обучения распознаванию образов:

$$I(\alpha) = P(\alpha) = \int (\omega - F(x, \alpha))^2 P(x, \omega)\, dx\, d\omega. \qquad (6.10)$$

Как уже указывалось, этот функционал для каждого решающего правила определяет вероятность ошибочной классификации. Эмпирический функционал

$$I_э(\alpha) = \nu(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (\omega_i - F(x_i, \alpha))^2, \qquad (6.11)$$

вычисленный по обучающей последовательности

$$x_1, \omega_1; \ldots; x_l, \omega_l, \qquad (6.12)$$

для каждого решающего правила определяет частоту неправильной классификации.

Согласно классическим теоремам теории вероятностей частота появления любого события сходится к вероятности этого события при неограниченном увеличении числа испытаний. Формально это означает, что для любых фиксированных α и $\varkappa$ имеет место соотношение

$$\lim_{l \to \infty} P\{|P(\alpha) - \nu(\alpha)| > \varkappa\} = 0. \qquad (6.13)$$

Однако (см. гл. II, § 6) из условия (6.13) не следует, что правило, минимизирующее (6.11), будет доставлять функционалу (6.10) значение, близкое к минимальному. Для

достаточно больших l близость найденного решения к наилучшему следует из более сильного условия, когда для любого $\varkappa$ выполняется равенство

$$\lim_{l \to \infty} P\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\} = 0. \qquad (6.14)$$

В этом случае говорят, что имеет место *равномерная сходимость частот появления событий к их вероятностям по классу событий* $S(\alpha)$. Каждое событие $S(\alpha^*)$ в классе $S(\alpha)$ задается решающим правилом $F(x, \alpha^*)$ как множество пар x, ω, на которых выполняется равенство $(\omega - F(x, \alpha^*))^2 = 1$.

Ниже мы приведем условия, обеспечивающие равномерную сходимость частот появления событий к их вероятностям, и тем самым укажем область применимости метода минимизации эмпирического риска.

Однако прежде заметим, что применение метода минимизации эмпирического риска само по себе не гарантирует успеха при решении задачи восстановления зависимостей.

Вот пример алгоритма обучения распознаванию образов, который минимизирует эмпирический риск и вместе с тем не способен обучаться (т. е. для которого нельзя гарантировать, что построенное решающее правило близко к наилучшему в классе). Алгоритм заключается в следующем. Запоминаются элементы обучающей последовательности, и каждая ситуация, предъявленная для распознавания, сравнивается с примерами, хранящимися в памяти. Если предъявленная ситуация совпадает с одним из примеров, то она будет отнесена к тому классу, к которому принадлежит пример. Если же в памяти нет аналогичного примера, то ситуацию относят к первому классу. Понятно, что подобное устройство ничему научиться не может, так как обучающую последовательность обычно составляет лишь ничтожная доля ситуаций, которые могут возникнуть при контроле. А вместе с тем такое устройство классифицирует элементы обучающей последовательности безошибочно, т. е. его алгоритм минимизирует (вплоть до нуля) эмпирический риск.

В дальнейшем мы убедимся, что этот алгоритм использует множество решающих правил, образующих систему событий, по которой нет равномерной сходимости.

§ 3. Частный случай

Когда же имеет место равномерная сходимость частот к вероятностям? Рассмотрим простой случай: множество решающих правил $F(x, \alpha)$ конечно и состоит из N правил:

$$F(x, \alpha_1), \ldots, F(x, \alpha_N).$$

В соответствие каждому решающему правилу $F(x, \alpha_i)$ может быть поставлено событие A_i, состоящее из тех пар x, ω, на которых $(\omega - F(x, \alpha_i))^2 = 1$. Таким образом, определено конечное число N событий $A_1, \ldots, A_N$.

Для каждого фиксированного события справедлив закон больших чисел (частота сходится к вероятности при неограниченном увеличении числа испытаний). Одним из конкретных выражений этого закона является оценка (неравенство Бернштейна)

$$P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\} < \exp\{-\varkappa^2 l\}. \tag{6.15}$$

Нас, однако, интересует равномерная сходимость, т. е. вероятность одновременного выполнения неравенств

$$|P(\alpha_i) - \nu(\alpha_i)| \leqslant \varkappa, \quad i = 1, 2, \ldots, N.$$

Такая вероятность легко может быть оценена, коль скоро оценивается вероятность выполнения отдельно каждого неравенства (6.15), а именно:

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} \leqslant \sum_{i=1}^{N} P\{|P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}.$$

Учитывая неравенство (6.15), получаем

$$P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} < N \exp\{-\varkappa^2 l\}. \tag{6.16}$$

Из неравенства (6.16) вытекает, что для конечного числа событий всегда имеет место равномерная сходимость частот появления событий к их вероятностям, т. е. справедливо

$$\lim_{l \to \infty} P\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\} = 0.$$

Потребуем теперь, чтобы вероятность выполнения события

$$\left\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\right\}$$

не превосходила величину η, т. е. выполнялось неравенство

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\} \leqslant \eta. \tag{6.17}$$

Как следует из оценки (6.16), неравенство (6.17), во всяком случае, выполнится, если величины N, l, $\varkappa$, η будут связаны соотношением

$$N \exp\{-\varkappa^2 l\} = \eta. \tag{6.18}$$

Если разрешить равенство (6.18) относительно $\varkappa$, то для данных N, l, η получится оценка максимального уклонения частоты от соответствующей вероятности в рассматриваемом классе событий

$$\varkappa = \sqrt{\frac{\ln N - \ln \eta}{l}}. \tag{6.19}$$

Если же разрешить равенство (6.18) относительно l, то будет найдено, какова должна быть длина обучающей последовательности, чтобы с вероятностью не меньшей $1 - \eta$, можно было утверждать, что наибольшее уклонение частоты от вероятности по этому классу не превосходит $\varkappa$:

$$l = \frac{\ln N - \ln \eta}{\varkappa^2}. \tag{6.20}$$

Итак, доказана теорема.

Теорема 6.1. *Пусть множество решающих правил состоит из N элементов, и пусть для решающих правил $F(x, \alpha_i)$ частоты ошибок на обучающей последовательности длины l равны $\nu(\alpha_i)$. Тогда с вероятностью $1 - \eta$ можно утверждать, что одновременно для всех решающих правил выполнятся неравенства*

$$\nu(\alpha_i) - \sqrt{\frac{\ln N - \ln \eta}{l}} < P(\alpha_i) < \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln \eta}{l}}.$$

Замечание. Так как неравенства справедливы для всех N правил, то теорема 6.1 устанавливает доверительный интервал для качества решающего правила $F(x, \alpha_э)$, минимизирующего среди N правил эмпирический риск. Он равен

$$\nu(\alpha_э) - \sqrt{\frac{\ln N - \ln \eta}{l}} < P(\alpha_э) < \nu(\alpha_э) + \sqrt{\frac{\ln N - \ln \eta}{l}}.$$

В дальнейшем для нас важна будет верхняя грань: с вероятностью $1-\eta$ одновременно для всех решающих правил (в том числе и того, которое минимизирует эмпирический риск) справедлива оценка

$$P(\alpha_i) < \nu(\alpha_i) + \sqrt{\frac{\ln N - \ln \eta}{l}}.$$

§ 4. Детерминистская постановка задачи

Величина доверительного интервала, вычисленная согласно теореме 6.1, может оказаться завышенной. В самом деле, рассмотрим случай, когда множество, состоящее из N решающих правил, содержит правило, которое идеально решает задачу распознавания образов, т. е. правило, для которого вероятность ошибочной классификации равна нулю. Такую постановку задачи иногда называют *детерминистской* [1]). Это правило (или близкое к нему) и надо найти, используя выборку $x_1, \omega_1; \ldots; x_l, \omega_l$.

Искать такое решающее правило будем, используя метод минимизации эмпирического риска. Так как среди функций $F(x, \alpha_i)$ $(i = 1, 2, \ldots N)$ есть та, которая идеально решает задачу, то заведомо ясно, что на любой выборке $x_1, \omega_1; \ldots; x_l, \omega_l$ значение минимума эмпирического риска будет равно нулю. Этот минимум, однако, может достигаться на многих функциях. Поэтому возникает необходимость оценить вероятность того, что качество любой функции, доставляющей нуль величине эмпирического риска, будет не хуже заданного $\varkappa$.

Введем функцию

$$\bar{\theta}(z) = \begin{cases} 1, & \text{если } z = 0, \\ 0, & \text{если } z > 0. \end{cases}$$

Тогда оценка скорости равномерной сходимости частот к вероятностям по множеству событий, для которых частота ошибок равна нулю, состоит в оценке вероятности события

$$\left\{ \sup_i | P(\alpha_i) - \nu(\alpha_i) | \, \bar{\theta}(\nu(\alpha_i)) > \varkappa \right\},$$

[1]) Термин выбран неудачно, так как задача по-прежнему остается статистической. Однако этот термин широко распространен и потому будем его придерживаться.

(а не события $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$, как в теореме 6.1).

Так как число функций, на которых достигается нуль величины эмпирического риска, не превосходит N (числа всех функций в классе), то справедливо неравенство

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| \bar{\theta}(\nu(\alpha_i)) > \varkappa\} \leqslant NP_\varkappa, \quad (6.21)$$

где $P_\varkappa$ — вероятность того, что решающее правило, для которого вероятность совершить ошибку больше $\varkappa$, правильно классифицирует все элементы обучающей последовательности. Эту вероятность легко оценить:

$$P_\varkappa \leqslant (1-\varkappa)^l. \quad (6.22)$$

Подставляя оценку $P_\varkappa$ в (6.21), получаем

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| \bar{\theta}(\nu(\alpha_i)) > \varkappa\} \leqslant N(1-\varkappa)^l. \quad (6.23)$$

Для того чтобы вероятность

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| \bar{\theta}(\nu(\alpha_i)) > \varkappa\}$$

не превосходила величину η, достаточно выполнения равенства

$$N(1-\varkappa)^l = \eta. \quad (6.24)$$

Разрешим относительно l это равенство

$$l = \frac{\ln N - \ln \eta}{-\ln(1-\varkappa)}. \quad (6.25)$$

Так как для малых $\varkappa$ справедливо

$$-\ln(1-\varkappa) \approx \varkappa,$$

то формула (6.25) может быть представлена в виде

$$l = \frac{\ln N - \ln \eta}{\varkappa}.$$

В отличие от оценки (6.20), здесь знаменатель равен $\varkappa$, а не $\varkappa^2$, т. е. в детерминистской постановке достаточная длина обучающей последовательности оказывается меньшей, чем в общем случае. Разрешая (6.24) относительно $\varkappa$,

получим

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

Таким образом, справедлива теорема.

Теорема 6.2. *Если из множества решающих правил, состоящего из N элементов, выбирается такое решающее правило, которое на обучающей последовательности не совершает ни одной ошибки, то с вероятностью $1-\eta$ можно утверждать, что вероятность ошибочной классификации с помощью выбранного правила заключена в пределах*

$$0 \leqslant P \leqslant \varkappa,$$

где

$$\varkappa = \frac{\ln N - \ln \eta}{l}.$$

§ 5. Верхние оценки вероятности ошибок

Несмотря на кажущуюся простоту, теоремы 6.1, 6.2 являются чрезвычайно глубокими. По существу, дальнейшее развитие теории минимизации эмпирического риска состоит в обобщении этих теорем на случай бесконечного числа решающих правил. Основные же моменты всей будущей теории здесь уже присутствуют. Остановимся на них подробнее.

1. Теоремы 6.1 и 6.2 немедленно получаются из оценки скорости равномерной сходимости частот к вероятностям по классу событий. Теорема 6.1 основана на оценке (6.16) скорости равномерной сходимости частот к вероятностям по классу событий S_N: $A_1, \ldots, A_N$. Теорема 6.2 — на оценке скорости равномерной сходимости по более узкому классу $\left\{ \mid P(\alpha_i) - \nu(\alpha_i) \mid \bar{\theta}(\nu(\alpha_i)) > \varkappa \right\}$. Обозначим этот класс через $\hat{S}_N$.

2. В обоих случаях скорость равномерной сходимости определялась произведением двух величин; числа событий в классе и оценки вероятности того, что частота любого фиксированного события в классе уклонится больше чем на $\varkappa$ от вероятности этого события. Для событий, рассмотренных в теореме 6.1, эта вероятность не превосходит $\exp\{-\varkappa^2 l\}$, для событий же, рассмотренных в тео-

реме 6.2, аналогичная вероятность не превосходит $(1-\varkappa)^l \approx \approx \exp\{-\varkappa l\}$.

Таким образом, оценка скорости равномерной сходимости частот к вероятностям по классу событий получается из оценки скорости обычной сходимости, вытекающей из закона больших чисел, умножением на число событий в классе.

При построении теории равномерной сходимости по классу событий, состоящему из бесконечного числа элементов, такая структура оценки скорости равномерной сходимости сохранится. Однако вместо числа событий в этом случае будут использованы другие емкостные характеристики класса событий.

3. В теореме 6.1 были получены двусторонние оценки вероятности ошибочной классификации с помощью решающего правила, минимизирующего эмпирический риск.

Однако во всей дальнейшей теории роль оценки снизу незначительна. Поэтому представляет интерес получение оценки равномерного одностороннего уклонения, т. е. получение оценки величины

$$P\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\},$$

а не величины

$$P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}.$$

Вероятность события $\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$ не превосходит вероятности события $\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$. Следовательно, возможна более тонкая оценка вероятности равномерного одностороннего уклонения $P\{\sup_i (P(\alpha_i) - \nu(\alpha_i)) > \varkappa\}$, чем оценка вероятности двустороннего равномерного уклонения $P\{\sup_i |P(\alpha_i) - \nu(\alpha_i)| > \varkappa\}$. Более тонкая оценка вероятности одностороннего равномерного уклонения позволит получить лучшую оценку сверху вероятности ошибочной классификации, чем та, которая следует из теоремы 6.1.

4. Оценки скорости равномерной сходимости (6.16) и (6.23) существенно зависят от оценки вероятности уклонения частоты от вероятности для события из рассматри-

ваемого класса (S_N или $\hat{S}_N$). Для класса S_N наиболее неблагоприятное событие A — то, для которого $P(A) = 1/2$. Поэтому возможна лишь оценка (6.16). Для класса событий $\hat{S}_N$ наиболее неблагоприятное событие — то, для которого $P(A) = \varkappa$. Для оценки вероятности уклонения частоты от вероятности этого события возможна более тонкая оценка (6.22). Таким образом, оценки, полученные для событий S_N и $\hat{S}_N$, различаются так, как различаются оценки вероятности уклонения события A, для которого $P(A) = 1/2$, и события A', для которого $P(A') = \varkappa$. Это обстоятельство заставляет внимательнее отнестись к тем требованиям, которые предъявляются к величинам уклонения частот от вероятностей для различных событий в классе. Для наших целей — получения равномерной оценки риска разумно требовать не равномерного уклонения частот от вероятностей для всей событий в классе, а разрешить большее уклонение для тех событий, для которых $P(A) = 1/2$, и меньшее — для событий с вероятностью $P(A') = \varkappa$. Например, разумно оценивать равномерную относительную величину уклонения

$$\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sigma(\alpha_i)} > \varkappa\right\},$$

где $\sigma(\alpha_i) = \sqrt{P(\alpha_i)(1 - P(\alpha_i))}$, для малых $P(\alpha_i)$ справедливо: $\sigma(\alpha_i) \sim \sqrt{P(\alpha_i)}$. Найдем оценку вероятности одностороннего относительного уклонения

$$P\left\{\sup_i \frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} \tag{6.26}$$

и построим с ее помощью верхнюю оценку вероятности ошибочной классификации. Для получения оценки вероятности (6.26) воспользуемся следующим фактом (неравенство Бернштейна):

$$P\left\{\frac{P(\alpha_i) - \nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < P\left\{\frac{P(\alpha_i) - \nu(\alpha_i)}{\sigma(\alpha_i)} > \varkappa\right\} < \exp\{-x^2 l\}. \tag{6.27}$$

Из справедливости (6.27) следует, что для класса, состоящего из N событий, имеет место следующая оценка

скорости равномерной сходимости:

$$P\left\{\sup_i \frac{P(\alpha_i)-\nu(\alpha_i)}{\sqrt{P(\alpha_i)}} > \varkappa\right\} < N\exp\{-\varkappa^2 l\}. \qquad (6.28)$$

Потребуем, чтобы вероятность равномерного одностороннего относительного уклонения (6.28) не превосходила η:

$$N\exp\{-\varkappa^2 l\} = \eta.$$

Это, во всяком случае, произойдет, если

$$\varkappa = \sqrt{\frac{\ln N - \ln\eta}{l}}. \qquad (6.29)$$

Пусть условие (6.29) выполнено. Тогда с вероятностью $1-\eta$ одновременно для всех событий A_i выполняются неравенства

$$\frac{P(\alpha_i)-\nu(\alpha_i)}{\sqrt{P(\alpha_i)}} < \varkappa. \qquad (6.30)$$

Разрешая неравенства (6.30) относительно $P(\alpha_i)$, получим, что с вероятностью $1-\eta$ одновременно для всех событий класса справедливо

$$P(\alpha_i) < \frac{\varkappa^2}{2}\left(1+\sqrt{1+\frac{4\nu(\alpha_i)}{\varkappa^2}}\right)+\nu(\alpha_i). \qquad (6.31)$$

Подставляя (6.29) в (6.31), получим, что с вероятностью $1-\eta$ одновременно выполняются N неравенств

$$P(\alpha_i) < \frac{\ln N - \ln\eta}{2l}\left(1+\sqrt{1+\frac{4\nu(\alpha_i)\,l}{\ln N-\ln\eta}}\right)+\nu(\alpha_i).$$

Таким образом, мы доказали теорему.

Теорема 6.3. *Пусть множество решающих правил состоит из N элементов, и пусть для каждого правила $F(x, \alpha_i)$ частота ошибок на обучающей последовательности равна $\nu(\alpha_i)$. Тогда с вероятностью $1-\eta$ можно утверждать, что одновременно для всех решающих правил класса выполнятся оценки*

$$P(\alpha_i) < \frac{\ln N - \ln\eta}{2l}\left(1+\sqrt{1+\frac{4\nu(\alpha_i)\,l}{\ln N-\ln\eta}}\right)+\nu(\alpha_i). \qquad (6.32)$$

Замечание. Так как с вероятностью $1-\eta$ оценка (6.32) справедлива одновременно для всех правил класса, то

она выполняется и для правила $F(x, \alpha_9)$, минимизирующего эмпирический риск.

Теорема 6.3 позволяет оценить качество правила, минимизирующего эмпирический риск. При этом оценка (6.32) совпадает с оценкой теоремы 6.2, полученной для крайнего случая, когда $\nu(\alpha) = 0$, и близка к оценке теоремы 6.1 для другого крайнего случая, когда $P(\alpha^*) \approx \approx 1/2$. Точно такая же структура оценки будет иметь место и для бесконечного класса решающих правил.

§ 6. ε-сеть множества

В предыдущих параграфах было показано существование равномерной сходимости частот появления событий к их вероятностям по классу событий, состоящему из конечного числа элементов; была оценена скорость этой сходимости, и с ее помощью получена оценка качества решающего правила, минимизирующего эмпирический риск. Теперь нам предстоит обобщить полученные результаты на случай бесконечного числа событий.

Однако, вообще говоря, при бесконечном числе событий равномерной сходимости частот к вероятностям может и не существовать, например, если множество событий задается всеми открытыми подмножествами множества X, ω. В этом случае возникает ситуация, когда (см. пример в § 2) алгоритм минимизации эмпирического риска доставляет нуль величине эмпирического риска, но не способен обучаться. Поэтому проблема состоит в том, чтобы определить условия, при которых имеет место равномерная сходимость для бесконечного числа событий, оценить ее скорость и, наконец, получить верхнюю оценку вероятности ошибочной классификации для правила, минимизирующего эмпирический риск.

В математике часто возникает необходимость перенести результаты, справедливые для конечного множества элементов на случай бесконечного множества. Обычно такое обобщение оказывается возможным, если бесконечное множество, допускает покрытие *конечной ε-сетью*.

Определение. *Множество B элементов метрического пространства R называется ε-сетью множества G, если любая точка c* из G *находится на расстоянии, не превышающем* ε, *от некоторой точки $b \in B$, т. е.* $\rho(b, c) \leqslant \varepsilon$.

Говорят также, что множество G допускает покрытие конечной ε-сетью, если для любого ε найдется ε-сеть B, состоящая из конечного числа элементов.

В этом параграфе для бесконечного множества решающих правил, допускающих покрытие конечной ε-сетью, мы получим утверждения, аналогичные утверждениям теорем 6.1 и 6.3.

Итак, пусть задано бесконечное множество решающих правил $F(x, \alpha)$, на котором определена метрика $\rho(\alpha_1, \alpha_2) = \rho(F(x, \alpha_1), F(x, \alpha_2))$ и выделена конечная ε-сеть. Пусть эта конечная ε-сеть состоит из $N(\varepsilon)$ элементов. Пусть, кроме того, известно, что если два решающих правила $F(x, \alpha_1)$ и $F(x, \alpha_2)$ отстоят друг от друга на

расстоянии, не превышающем ε: $\rho(\alpha_1, \alpha_2) \leqslant \varepsilon$, то качества этих правил различаются не более чем на величину $\delta(\varepsilon)$, т. е.

$$\left| \int (\omega - F(x, \alpha_1))^2 P(x, \omega)\, dx\, d\omega - \int (\omega - F(x, \alpha_2))^2 P(x, \omega)\, dx\, d\omega \right| \leqslant \delta(\varepsilon).$$

Иначе говоря, малая вариация решающего правила приводит к малому изменению качества классификации.

В этих условиях теоремы 6.1, 6.3 допускают следующие обобщения.

Теорема 6.4. *Пусть множество решающих правил $F(x, \alpha)$ может быть покрыто конечной ε-сетью. Тогда с вероятностью $1-\eta$ качество решающего правила $F(x, \alpha_э)$, минимизирующего величину эмпирического риска, оценится неравенствами*

$$\nu(\alpha_i(\alpha_э)) - \sqrt{\frac{\ln N(\varepsilon) - \ln \eta}{l}} - \delta(\varepsilon) \leqslant P(\alpha_э) \leqslant \nu(\alpha_i(\alpha_э)) + \sqrt{\frac{\ln N(\varepsilon) - \ln \eta}{l}} + \delta(\varepsilon),$$

где $F(x, \alpha_i(\alpha_э))$ — ближайший к $F(x, \alpha_э)$ элемент ε-сети.

Теорема 6.5. *Пусть множество решающих правил $F(x, \alpha)$ может быть покрыто конечной ε-сетью. Тогда с вероятностью $1-\eta$ качество решающего правила $F(x, \alpha_э)$, минимизирующего величину эмпирического риска, оценится неравенством*

$$P(\alpha_э) < \nu(\alpha_i(\alpha_э)) + \frac{\ln N(\varepsilon) - \ln \eta}{2l}\left(1 + \sqrt{1 + \frac{4\nu(\alpha_i(\alpha_э))\, l}{\ln N(\varepsilon) - \ln \eta}}\right) + \delta(\varepsilon),$$

где $F(x, \alpha_i(\alpha_э))$ — ближайший к $F(x, \alpha_э)$ элемент ε-сети.

Замечание. Теоремы 6.4 и 6.5 справедливы для любой ε-сети, заданной априорно (до появления обучающей последовательности). В частности, величина ε, задающая ε-сеть, может быть выбрана в теореме 6.4 из условия минимума выражения

$$\sqrt{\frac{\ln N(\varepsilon) - \ln \eta}{l}} + \delta(\varepsilon),$$

а в теореме 6.5 — из условия минимума выражения

$$\frac{\ln N(\varepsilon) - \ln \eta}{2l}\left(1 + \sqrt{1 + \frac{4cl}{\ln N(\varepsilon) - \ln \eta}}\right) + \delta(\varepsilon),$$

где $0 \leqslant c \leqslant 1$ — константа (например, $c = 0{,}5$).

Доказательство теорем 6.4 и 6.5 проводится по одной и той же схеме.

1°. На множестве решающих правил $F(x, \alpha)$ задается конечная ε-сеть, состоящая из $N(\varepsilon)$ элементов

$$F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)}). \qquad (6.33)$$

Согласно теореме 6.1 (6.3) с вероятностью $1-\eta$ одновременно для всех $N(\varepsilon)$ элементов (6.33) выполнятся неравенства

$$\nu(\alpha_i) - \sqrt{\frac{\ln N(\varepsilon) - \ln\eta}{l}} < P(\alpha_i) < \nu(\alpha_i) + \sqrt{\frac{\ln N(\varepsilon) - \ln\eta}{l}} \tag{6.34}$$

$$\left(P(\alpha_i) < \frac{\ln N(\varepsilon) - \ln\eta}{2l}\left(1 + \sqrt{1 + \frac{4\nu(\alpha_i)\, l}{\ln N(\varepsilon) - \ln\eta}}\right) + \nu(\alpha_i)\right).$$

2°. Для всякого решающего правила $F(x, \alpha^*)$ (в том числе и того, которое минимизирует в $F(x, \alpha)$ величину эмпирического риска) найдется ближайший элемент ε-сети $F(x, \alpha_i(\alpha^*))$, и для него

$$|P(\alpha^*) - P(\alpha_i(\alpha^*))| \leqslant \delta(\varepsilon). \tag{6.35}$$

Из неравенств (6.34) и (6.35) следует, что для решающего правила $F(x, \alpha_i(\alpha_э))$ с вероятностью $1-\eta$ справедливо

$$\nu(\alpha_i(\alpha_э)) - \sqrt{\frac{\ln N(\varepsilon) - \ln\eta}{l}} - \delta(\varepsilon) <$$

$$< P(\alpha_э) < \nu(\alpha_i(\alpha_э)) + \sqrt{\frac{\ln N(\varepsilon) - \ln\eta}{l}} + \delta(\varepsilon),$$

$$\left(P(\alpha_э) < \frac{\ln N(\varepsilon) - \ln\eta}{2l}\left(1 + \sqrt{1 + \frac{4\nu(\alpha_i(\alpha_э))\, l}{N(\varepsilon) - \ln\eta}}\right) + \right.$$

$$\left. + \delta(\varepsilon) + \nu(\alpha_i(\alpha_э))\right).$$

Теоремы доказаны.

Итак, если множество решающих правил $F(x, \alpha)$ допускает покрытие конечной ε-сетью, а распределение $P(x, \omega)$ таково, что близким решающим правилам соответствуют близкие значения вероятности ошибочной классификации, то принципиально с ростом объема выборки метод минимизации эмпирического риска приводит к успеху [1]). При этом для каждого фиксированного ε вероятность ошибочной классификации с помощью правила, минимизирующего эмпирический риск, оценивается с помощью неравенств (6.34).

Однако, для того чтобы воспользоваться этими оценками, надо знать величину $\delta(\varepsilon)$ наибольшего уклонения качества двух решающих правил, отстоящих друг от друга на расстоянии ε. При вычислении же этой величины используется плотность $P(x)$, которая, согласно постановке задачи распознавания образов, считается неизвестной. В следующей главе при решении задачи восстановления регрессии мы найдем величину $\delta(\varepsilon)$ и получим возможность использовать оценки качества функции, выраженные через величины эмпирического риска $\delta(\varepsilon)$ и $N(\varepsilon)$. Здесь же для получения скорости равномерной сходимости частот появления событий к их вероятностям по бесконечному классу событий мы реализуем иную идею, которая в конце концов приведет к построению необходимых и

[1]) Формально это утверждение не следует из теоремы 6.4, но доказывается совершенно аналогично.

достаточных условий равномерной сходимости, получению на базе этих условий оценки скорости равномерной сходимости и, наконец, к конструктивной оценке качества решающего правила, найденного методом минимизации эмпирического риска.

§ 7. Необходимые и достаточные условия равномерной сходимости частот к вероятностям

До сих пор для получения оценок скорости равномерной сходимости, мы использовали достаточно грубые емкостные характеристики множества решающих правил (число элементов множества).

В этом параграфе мы введем более тонкую емкостную характеристику — *энтропию системы событий на выборках длины* l. С помощью этой характеристики могут быть установлены исчерпывающие, необходимые и достаточные условия равномерной сходимости частот появления событий к их вероятностям, т. е. необходимые и достаточные условия того, что для любого $\varkappa$ выполнится равенство

$$\lim_{l\to\infty} P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} = 0.$$

Итак, пусть задано множество S решающих правил $F(x, \alpha)$ и дана выборка $x_1, \ldots, x_l$. Эта выборка, вообще говоря, может быть разделена на два класса 2^l способами. Однако нас будут интересовать только те способы разделения выборки, которые могут быть реализованы с помощью правил $F(x, \alpha)$. (С помощью правила $F(x, \alpha^*)$ множество $x_1, \ldots, x_l$ делится на два подмножества — подмножество, на котором $F(x, \alpha^*) = 1$, и подмножество, на котором $F(x, \alpha^*) = 0$.)

Число таких способов разделения зависит как от класса решающих правил $F(x, \alpha)$, так и от состава выборки. Будем обозначать это число через

$$\Delta^S(x_1, \ldots, x_l).$$

Рассмотрим систему событий

$$S(\alpha) = \{x, \omega : (\omega - F(x, \alpha))^2 = 1\},$$

образованную множеством решающих правил $F(x, \alpha)$.

Пусть дана случайная независимая выборка

$$x_1, \omega_1; \ldots; x_l, \omega_l. \tag{6.36}$$

На выборке (6.36) система событий $S(\alpha)$ индуцирует $\Delta(S(\alpha); x_1, \omega_1; \ldots; x_l, \omega_l)$ различных подвыборок. Очевидно, что число этих подвыборок равно $\Delta^S(x_1, \ldots, x_l)$.

Так как $x_1, \ldots, x_l$ — случайная независимая выборка, то число разделений $\Delta^S(x_1, \ldots, x_l)$ — величина случайная.

Определение. *Назовем величину*

$$H^S(l) = M \ln \Delta(S(\alpha); x_1, \omega_1; \ldots; x_l\omega_l) = M \ln \Delta^S(x_1, \ldots, x_l)$$

энтропией системы событий $S(\alpha)$ *на выборках длины* l.

Оказывается, что для существования равномерной сходимости частот $\nu(\alpha)$ к их вероятностям $P(\alpha)$ по множеству событий $S(\alpha)$ необходимо и достаточно, чтобы с ростом объема выборки доля энтропии, приходящаяся на один элемент выборки, стремилась к нулю, т. е. чтобы последовательность

$$\frac{H^S(1)}{1}, \frac{H^S(2)}{2}, \ldots, \frac{H^S(l)}{l}$$

стремилась к нулю с ростом l. Иначе говоря, выполнялось условие

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = 0. \tag{6.37}$$

Доказательство этого утверждения приведено в монографии [12].

Как и всякие исчерпывающие условия, сформулированные необходимые и достаточные условия равномерной сходимости частот появления событий к их вероятностям используют тонкие понятия. В нашем случае таким понятием является энтропия $H^S(l)$ системы событий $S(\alpha)$ на выборках длины l, которая конструируется с помощью плотности $P(x)$. Согласно же постановке задачи распознавания образов плотность $P(x)$ неизвестна. Поэтому, для того чтобы установить возможность минимизации среднего риска путем нахождения минимума эмпирического риска, нельзя использовать необходимые и достаточные условия (6.37).

Вот почему важно получить более грубые достаточные условия, которые, во-первых, не зависели бы от свойств меры $P(x)$, а во-вторых, допускали бы оценку скорости равномерной сходимости. Такие условия могут быть найдены на основе емкостной характеристики системы

событий $S(\alpha)$, которая получается из энтропии $H^S(l)$ абстрагированием от свойств меры.

Определение. *Назовем функцию*

$$m^S(l) = \max_{x_1, \ldots, x_l} \Delta^S(x_1, \ldots, x_l),$$

где максимум берется по всем возможным выборкам длины l, функцией роста системы событий, образованной решающими правилами $F(x, \alpha)$.

Функция роста построена так, что не зависит от свойств меры $P(x)$ и для нее всегда выполняется неравенство

$$\ln m^S(l) \geqslant H^S(l). \tag{6.38}$$

Теперь, если окажется, что величина

$$\frac{\ln m^S(l)}{l}$$

с ростом l стремится к нулю, то отношение $H^S(l)/l$ в силу (6.38) и подавно устремится к нулю. Поэтому условие

$$\lim_{l \to \infty} \frac{\ln m^S(l)}{l} = 0$$

является достаточным условием равномерной сходимости частот к вероятностям. Ниже мы покажем, что функция роста легко может быть найдена для событий, заданных различными классами решающих правил $F(x, \alpha)$ и, следовательно, может быть установлен факт равномерной сходимости. Более того, как будет показано ниже, с помощью функции роста $m^S(l)$ может быть оценена и скорость равномерной сходимости.

§ 8. Свойства функции роста

Функция роста имеет простой смысл: она вычисляет максимальное число способов разделения l точек на два класса с помощью решающих правил $F(x, \alpha)$. Для функции роста справедлива замечательная теорема, которая позволяет легко ее оценить.

Теорема 6.6. *Функция роста либо тождественно равна 2^l, либо при $l > h$ мажорируется функцией*

$$m^S(l) < 1{,}5 \frac{l^h}{h!},$$

где $h+1$ — минимальный объем выборки, при котором нарушается условие $m^S(l)=2^l$.
Иначе говоря,

$$m^S(l)=\begin{cases}\text{либо } \equiv 2^l,\\ \text{либо } <1{,}5\dfrac{l^h}{h!}\end{cases}\quad (l>h).$$

Доказательство теоремы 6.6 приведено в приложении к главе.

Для того чтобы оценить функцию роста, надо показать, что либо для любого l найдутся точки $x_1, \ldots, x_l$ такие, что с помощью решающих правил $F(x, \alpha)$ их можно разбить на два класса всеми 2^l возможными способами, либо существует число h такое, что h точек можно, но никакие $h+1$ точек нельзя разбить на два класса всеми возможными способами. В первом случае функция роста — показательная, во втором случае — степенная. Число h может служить мерой разнообразия класса решающих правил.

Определение. *Будем говорить, что класс характеристических функций имеет емкость h, если справедливо неравенство*

$$m^S(l)<1{,}5\frac{l^h}{h!}\qquad (l>h). \tag{6.39}$$

В случае выполнения равенства

$$m^S(l)\equiv 2^l$$

будем говорить, что емкость класса характеристических функций $F(x, \alpha)$ бесконечна.

Нетрудно убедиться, что если емкость класса характеристических функций конечна, то всегда имеет место равномерная сходимость частот к вероятностям. В самом деле, в этом случае справедливо

$$0\leqslant \lim_{l\to\infty}\frac{\ln m^S(l)}{l}\leqslant \lim_{l\to\infty}\frac{h\ln l-\sum\limits_{i=1}^{h}\ln i}{l}=0$$

и достаточное условие выполнено.

Важную роль в дальнейшей теории играет класс линейных по параметру решающих правил:

$$F(x,\ \alpha)=\theta\left(\sum_{i=1}^{n}\alpha_i\varphi_i(x)\right);\qquad \theta(z)=\begin{cases}1, & \text{если } z\geqslant 0,\\ 0, & \text{если } z<0.\end{cases} \tag{6.40}$$

Нетрудно найти функцию роста для класса событий, заданных линейными решающими правилами (6.40). Для этого достаточно определить максимальное число точек h в пространстве размерности n, которые можно с помощью гиперплоскости разбить на два класса всеми 2^h способами. Известно, что это число равно n. Поэтому, согласно теореме 6.6, для класса линейных решающих правил (6.40) функция роста оценивается

$$m^S(l) < 1{,}5\,\frac{l^n}{n!} \qquad (l > n).$$

И, следовательно, для класса линейных решающих правил выполнены достаточные условия равномерной сходимости.

В главе II было показано, что факт равномерной сходимости частот появления событий к их вероятностям по классу событий, заданному одномерными линейными решающими правилами $F(x, \alpha) = \theta(x+\alpha)$, составляет содержание теоремы Гливенко — Кантелли, утверждающей равномерную сходимость эмпирической функции распределения к истинной.

§ 9. Оценка уклонения эмпирически оптимального решающего правила

В приложении к главе получена оценка скорости равномерной сходимости частот к вероятностям по классу событий $S(\alpha)$. Показано, что имеет место неравенство

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} < 6m^S(2l)\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \tag{6.41}$$

Оценка (6.41) имеет тот же вид, что и раньше, — она образуется произведением емкостной характеристики системы событий ($6m^S(2l)$) и оценкой вероятности того, что уклонение частоты от вероятности превзойдет $\varkappa\left(\exp\left\{-\frac{\varkappa^2 l}{4}\right\}\right)$.

Если емкость класса решающих правил бесконечна ($m^S(l) \equiv 2^l$), то оценка (6.41) травиальна, так как при всех $\varkappa$ правая часть неравенства больше единицы. Оценка (6.41) становится содержательной, когда емкость класса

решающих правил конечна:

$$m^S(l) < 1{,}5\,\frac{l^h}{h!}.$$

В этом случае она принимает вид

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\} < 9\,\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \quad (6.42)$$

С ростом l правая часть неравенства (6.42) стремится к нулю и притом тем быстрее, чем меньше емкость класса h. Потребуем, чтобы вероятность

$$P\left\{\sup_{\alpha} |P(\alpha) - \nu(\alpha)| > \varkappa\right\}$$

не превышала η. Это во всяком случае произойдет, если выполняется равенство

$$9\,\frac{(2l)^h}{h!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\} = \eta. \quad (6.43)$$

Равенство (6.43) можно разрешить относительно $\varkappa$ (используя формулу Стирлинга):

$$\varkappa = 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right) - \frac{\ln\eta}{9}}{l}}. \quad (6.44)$$

И тогда из (6.42) — (6.44) следует справедливость следующей теоремы.

Теорема 6.7. *Пусть $F(x, \alpha)$ — класс решающих правил ограниченной емкости h, и пусть $\nu(\alpha)$ — частота ошибок, вычисленная по обучающей последовательности для правила $F(x, \alpha)$. Тогда с вероятностью $1-\eta$ можно утверждать, что при $l > h$ одновременно для всех правил $F(x, \alpha)$ вероятность ошибочной классификации заключена в пределах*

$$\nu(\alpha) - 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right) - \frac{\ln\eta}{9}}{l}} <$$

$$< P(\alpha) < \nu(\alpha) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right) - \frac{\ln\eta}{9}}{l}}.$$

Замечание. Из теоремы 6.7 следует, что для правила $F(x, \alpha_э)$, минимизирующего эмпирический риск, с

вероятностью $1-\eta$ справедлива оценка сверху

$$P(\alpha_э) < \nu(\alpha_э) + 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{9}}{l}} \qquad (l>h).$$

В приложении к главе показано, что наряду с (6.41) справедлива и оценка

$$P\left\{\sup_{\alpha}\frac{P(\alpha)-\nu(\alpha)}{\sqrt{P(\alpha)}}>\varkappa\right\} < 8m^s(2l)\,e^{-\frac{\varkappa^2 l}{4}},$$

которая для класса решающих правил ограниченной емкости является нетривиальной:

$$P\left\{\sup_{\alpha}\frac{P(\alpha)-\nu(\alpha)}{\sqrt{P(\alpha)}}>\varkappa\right\} < 12\frac{(2l)^h}{h!}e^{-\frac{\varkappa^2 l}{4}}. \qquad (6.45)$$

Потребуем, чтобы правая часть неравенства равнялась η:

$$12\frac{(2l)^h}{h!}e^{-\frac{\varkappa^2 l}{4}}=\eta.$$

Это произойдет, если

$$\varkappa = 2\sqrt{\frac{\ln\frac{(2l)^h}{h!}-\frac{\ln\eta}{12}}{l}} \approx 2\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{12}}{l}}. \qquad (6.46)$$

С другой стороны, неравенство (6.45) можно переписать в виде утверждения: с вероятностью η одновременно для всех α справедливо неравенство

$$P(\alpha) < \frac{\varkappa^2}{2}\left(1+\sqrt{1+\frac{4\nu(\alpha)}{\varkappa^2}}\right)+\nu(\alpha). \qquad (6.47)$$

Из (6.46) и (6.47) следует справедливость следующей теоремы.

Теорема 6.8. *Пусть $F(x, \alpha)$ — класс решающих правил ограниченной емкости h, и пусть для каждого правила $F(x, \alpha)$ частота ошибок, вычисленная на обучающей последовательности, равна $\nu(\alpha)$. Тогда с вероятностью $1-\eta$ можно утверждать, что при $l>h$ одновременно для всех*

правил класса имеет место оценка

$$P(\alpha_э) < 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{12}}{l}\times$$
$$\times\left(1+\sqrt{1+\frac{\nu(\alpha_э)\,l}{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{12}}}\right)+\nu(\alpha_э). \qquad (6.48)$$

Замечание. Из теоремы 6.8 следует, что для правила $F(x, \alpha_э)$, минимизирующего эмпирический риск, справедлива оценка

$$P(\alpha_э) < 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{12}}{l}\times$$
$$\times\left(1+\sqrt{1+\frac{\nu(\alpha_э)\,l}{h\left(\ln\frac{2l}{h}+1\right)-\frac{\ln\eta}{12}}}\right)+\nu(\alpha_э).$$

§ 10. Замечания об оценке скорости равномерной сходимости частот к вероятностям

Итак, в этой главе мы получили оценки скорости равномерной сходимости частот к вероятностям

$$P\left\{\sup_\alpha |P(\alpha)-\nu(\alpha)| > \varkappa\right\} < \begin{cases} Ne^{-\varkappa^2 l}, \\ 6m^s(2l)\,e^{-\frac{\varkappa^2 l}{4}} \end{cases}$$

и оценки равномерного одностороннего относительного уклонения частот от вероятностей

$$P\left\{\sup_\alpha \frac{P(\alpha)-\nu(\alpha)}{\sqrt{P(\alpha)}} > \varkappa\right\} < \begin{cases} Ne^{-\varkappa^2 l}, \\ 8m^s(2l)\,e^{-\frac{\varkappa^2 l}{4}}. \end{cases}$$

С помощью этих оценок были получены теоремы 6.1, 6.3 и 6.7, 6.8, которые позволили оценить качество решающего правила, минимизирующего величину эмпирического риска.

Все полученные оценки имеют одну и ту же структуру — они состоят из двух сомножителей, один из которых оценивает вероятность соответствующего уклонения для каждого (в отдельности) события в классе, другой

характеризует разнообразие класса решающих правил. В оценке используются различные характеристики разнообразия класса решающих правил. Наиболее примитивная — число решающих правил в классе. Примитивность такой характеристики заключается в том, что она, например, никак не учитывает, состоит ли класс решающих правил из «существенно различных» правил, или же все правила в классе «эквивалентны».

Адекватная мера разнообразия класса решающих правил, с помощью которой и удается построить необходимые и достаточные условия равномерной сходимости частот к вероятностям, — энтропия системы событий, заданной решающими правилами. Однако вычислить энтропию системы событий на выборках длины l можно, лишь зная плотность $P(x)$, которая, согласно постановке задачи обучения распознаванию образов, считается неизвестной. Поэтому была введена новая мера разнообразия, которая получается из энтропии выбором наиболее неблагоприятного распределения. Эта мера разнообразия выражается через емкость класса решающих правил и легко может быть вычислена.

Различные определения меры разнообразия класса решающих правил порождают разные теоремы о качестве алгоритмов, минимизирующих эмпирический риск.

Однако во всех этих теоремах утверждается один и тот же факт: если мера разнообразия класса решающих правил мала по сравнению с объемом выборки, то метод минимизации эмпирического риска позволяет выбрать правило, близкое к наилучшему в классе.

Характерной особенностью изложенной теории минимизации эмпирического риска является полное отсутствие каких бы то ни было указаний на конструктивную возможность построения алгоритмов. Такая особенность имеет как отрицательные, так и положительные стороны.

С одной стороны, построенная теория не указывает на регулярные процедуры минимизации эмпирического риска, которые должна реализовать соответствующая программа.

С другой стороны, теория минимизации эмпирического риска весьма обща. Метод минимизации эмпирического риска может быть применен для самых различных классов решающих правил: линейных дискриминантных функций, кусочно-линейных дискриминантных функций, логи-

ческих функций определенного вида и др. Это связано с тем, что теория метода минимизации эмпирического риска отвечает на вопрос, «что надо делать», оставляя в стороне вопрос «как это сделать». Поэтому для минимизации эмпирического риска могут быть применены различные методы, в том числе и эвристические.

Применение эвристических методов в этом случае имеет теоретическое оправдание: если в классе решающих правил, емкость которого невелика по сравнению с объемом выборки, выбрать правило, которое хотя и не минимизирует эмпирический риск, но доставляет ему достаточно малую величину, то в силу доказанных теорем выбранное решающее правило будет иметь достаточно высокое качество.

Конструктивные идеи таких алгоритмов имеют наглядную геометрическую интерпретацию: в пространстве X надо построить гиперповерхность, принадлежащую заданному классу гиперповерхностей, которая, по возможности, с меньшим количеством ошибок отделит векторы обучающей последовательности одного класса от векторов обучающей последовательности другого класса.

Отнесение векторов (в том числе и не входящих в обучающую последовательность) к тому или иному классу производится в зависимости от того, по какую сторону от разделяющей гиперповерхности он находится.

Методы построения разделяющих гиперповерхностей составляют конструктивную часть теории обучения распознаванию образов. Эти методы будут изложены в главе XI.

Основные утверждения главы VI

1. Успешное решение задачи обучения распознаванию образов с помощью метода минимизации эмпирического риска может быть гарантировано в условиях существования равномерной сходимости частот к вероятностям по классу событий

$$S(\alpha) = \{x,\ \omega\colon (\omega - F(x,\ \alpha))^2 = 1\}.$$

В этом случае равномерная $\varkappa$-близость частот к вероятностям

$$\sup_{\alpha \in S(\alpha)} |P(\alpha) - \nu(\alpha)| < \varkappa$$

обеспечивает 2ε-близость качеств найденного решающего правила и наилучшего в классе.

2. Равномерная сходимость частот к вероятностям по классу событий $S(\alpha)$ имеет место всегда, когда класс состоит из конечного числа событий.

3. Проблема получения условий равномерной сходимости частот к вероятностям по классу событий, состоящему из бесконечного числа элементов, связана с введением емкостных характеристик бесконечного множества событий $S(\alpha)$, образованного бесконечным множеством решающих правил $F(x, \alpha)$.

4. Адекватной емкостной характеристикой класса событий $S(\alpha)$, с помощью которой удается сформулировать необходимые и достаточные условия равномерной сходимости частот к вероятностям, является энтропия класса событий на выборках длины l. Однако эта характеристика строится с учетом вероятностной меры $P(x)$.

Оценка энтропии класса событий для самой неблагоприятной вероятностной меры может быть получена с помощью величины h, определяющей емкость класса решающих правил.

Емкость класса решающих правил $F(x, \alpha)$ легко вычислить — она равна максимальному числу точек $x_1, \ldots, x_h$, которые правилами из $F(x, \alpha)$ делятся на два класса всеми возможными способами.

Равномерная сходимость частот к вероятностям имеет место, когда величина h ограничена.

5. Алгоритмы обучения распознаванию образов способны обучаться, если:

— емкость класса решающих правил мала по отношению к объему обучающей выборки;

— выбирается правило, которое доставляет величине эмпирического риска малое значение.

ПРИЛОЖЕНИЕ К ГЛАВЕ VI

ТЕОРИЯ РАВНОМЕРНОЙ СХОДИМОСТИ ЧАСТОТ К ВЕРОЯТНОСТЯМ [1])

§ П.1. Достаточные условия равномерной сходимости частот к вероятностям

Согласно классической теореме Бернулли частота появления некоторого события A сходится (по вероятности) в последовательности независимых испытаний к вероятности этого события. Часто, однако, возникает необходимость судить одновременно о вероятностях целого класса событий S по одной и той же выборке. При этом требуется, чтобы частоты сходились к вероятностям равномерно по всем событиям класса S. Точнее, требуется, чтобы вероятность того, что максимальное по классу уклонение частоты от вероятности превзойдет заданную сколь угодно малую положительную константу, стремилась к нулю при неограниченном увеличении числа испытаний.

Оказывается, что даже в простейших примерах равномерная сходимость может не иметь места. Поэтому нужен критерий, позволяющий судить, есть ли такая сходимость.

Пусть X — множество элементарных событий, на котором задана вероятностная мера $P(x)$. Пусть S — некоторая совокупность случайных событий, т. е. подмножеств пространства, измеримых относительно меры $P(x)$ (S включается в σ-алгебру случайных событий, но не обязательно совпадает с ней). Обозначим через $X(l)$ пространство случайных независимых выборок из X длины l.

---

[1]) Здесь изложены лишь достаточные условия равномерной сходимости частот появления событий к их вероятностям. Необходимые и достаточные условия изложены в монографии [12].

Для каждой выборки $X^l = x_1, \ldots, x_l$ и события $A \in S$ определена частота выпадания события A, равная отношению числа $n(A)$ элементов выборки, принадлежащих A, к общей длине выборки l

$$\nu^l(A) = \nu(x_1, \ldots, x_l) = \frac{n(A)}{l}.$$

Теорема Бернулли утверждает, что при фиксированном событии A уклонение частоты от вероятности стремится к нулю (по вероятности) с ростом объема выборки, т. е. для любого $\varkappa$

$$P\{|P(A) - \nu^l(A)| > \varkappa\} \xrightarrow[l\to\infty]{} 0.$$

Нас же будет интересовать максимальное по классу S уклонение частоты от вероятности:

$$\pi(l) = \sup_{A\in S} |\nu^l(A) - P(A)|.$$

Величина $\pi(l)$ является функцией точки в пространстве $X(l)$. Будем предполагать, что эта функция измерима относительно меры в $X(l)$, т. е. что $\pi(l)$ есть случайная величина. Дальнейшие теоремы посвящены оценкам вероятности события $\pi(l)$.

§ П.2. Функция роста

Пусть X — множество, S — некоторая система его подмножеств, $X^l = x_1, \ldots, x_l$ — последовательность элементов x длины l. Каждое множество $A \in S$ определяет подпоследовательность X_A этой последовательности, состоящую из тех элементов, которые принадлежат A. Будем говорить, что A индуцирует подпоследовательность X_A на последовательности X^l.

Обозначим через

$$\Delta^S(x_1, \ldots, x_l)$$

число различных подпоследовательностей X_A, индуцированных множествами $A \in S$. Очевидно, что

$$\Delta^S(x_1, \ldots, x_l) \leqslant 2^l.$$

Число $\Delta^S(x_1, \ldots, x_l)$ будем называть *индексом системы S относительно выборки $x_1, \ldots, x_l$*.

Определение индекса системы можно сформулировать и иначе. Будем считать, что $A_1 \in S$ эквивалентно $A_2 \in S$ относительно выборки $x_1, \ldots, x_l$, если $X_{A_1} = X_{A_2}$. Тогда индекс $\Delta^S(x_1, \ldots, x_l)$ есть число классов эквивалентности, на которые система S разбивается этим отношением эквивалентности.

Очевидно, эти два определения равносильны. Функцию

$$m^S(l) = \max_{x_1, \ldots, x_l} \Delta^S(x_1, \ldots, x_l), \qquad (\text{П}.1)$$

где максимум берется по всем последовательностям длины l, назовем *функцией роста системы* S. Здесь максимум всегда достигается, так как индекс $\Delta^S(x_1, \ldots, x_l)$ принимает конечное число значений.

Функция роста класса событий S обладает следующим замечательным свойством.

Теорема П.1. *Функция роста либо тождественно равна* 2^l, *либо, если это не так, мажорируется функцией* $\sum_{i=0}^{n-1} C_l^i$, *где* n — *минимальное значение* l, *при котором*

$$m^S(l) \neq 2^l.$$

Иначе говоря,

$$m^S(l) = \begin{cases} \text{либо} \equiv 2^l, \\ \text{либо} < \sum_{i=0}^{n-1} C_l^i. \end{cases} \qquad (\text{П}.2)$$

Для доказательства этого утверждения нам понадобится следующая

Лемма П.1. *Если для некоторой последовательности* $x_1, \ldots, x_l$ *и некоторого* n

$$\Delta^S(x_1, \ldots, x_l) > \sum_{i=0}^{n-1} C_l^i,$$

то существует подпоследовательность X^n *длины* n *такая, что*

$$\Delta^S(X^n) = 2^n.$$

Доказательство. Обозначим

$$\sum_{i=0}^{n-1} C_l^i = \Phi(n, l)$$

(здесь и дальше считаем, что при $i > l$, $C_l^i = 0$). Для этой функции, как легко убедиться, выполняются соотношения

$$\left.\begin{aligned} &\Phi(1,\ l) = 1, \\ &\Phi(n,\ l) = 2^l, \quad \text{если} \quad l \leqslant n+1, \\ &\Phi(n,\ l) = \Phi(n,\ l-1) + \Phi(n-1,\ l-1), \ \text{если} \ n \geqslant 2,\ l \geqslant 1. \end{aligned}\right\} \tag{П.3}$$

Эти соотношения в свою очередь однозначно определяют функцию $\Phi(n,\ l)$ при $l > 0$ и $n > 0$.

Будем доказывать лемму индукцией по l и n. Для $n = 1$ и любого $l \geqslant 1$ утверждение леммы очевидно. Действительно, в этом случае из

$$\Delta^S(x_1,\ \ldots,\ x_l) > 1$$

следует, что существует элемент последовательности x_i такой, что для некоторого $A^* \in S$ выполнится $x_i \in A^*$, а для некоторого другого $A^{**} \in S$ окажется $x_i \notin A^{**}$ и, следовательно,

$$\Delta^S(x_i) = 2.$$

Для $l < n$ утверждение леммы верно ввиду ложности посылки. Действительно, в этом случае посылка есть

$$\Delta^S(x_1,\ \ldots,\ x_l) > 2^l,$$

что невозможно, так как

$$\Delta^S(x_1,\ \ldots,\ x_l) \leqslant 2^l.$$

Наконец, допустим, что лемма верна для $n \leqslant n_0$ $(n_0 \geqslant 1)$ при всех l. Рассмотрим теперь случай $n = n_0 + 1$. Покажем, что лемма верна и в этом случае для всех l.

Зафиксируем $n = n_0 + 1$ и проведем индукцию по l. Для $l < l_0 + 1$, как указывалось, лемма верна. Предположим, что она верна для $l \leqslant l_0$, и покажем, что она справедлива для $l = l_0 + 1$. Действительно, пусть для некоторой последовательности $x_1,\ \ldots,\ x_{l_0},\ x_{l_0+1}$ справедливо условие леммы

$$\Delta^S(x_1,\ \ldots,\ x_{l_0},\ x_{l_0+1}) > \Phi(n_0+1,\ l_0+1).$$

Лемма будет доказана, если мы найдем подпоследовательность длины $n_0 + 1$: $X^{n_0+1} = x_1,\ \ldots,\ x_{n_0+1}$ такую, что

$$\Delta^S(x_1,\ \ldots,\ x_{n_0+1}) = 2^{n_0+1}.$$

Рассмотрим подпоследовательность $X^{l_0} = x_1, \ldots, x_{l_0}$. Возможны два случая:

а) $\Delta^S(x_1, \ldots, x_{l_0}) > \Phi(n_0 + 1, l_0)$,

б) $\Delta^S(x_1, \ldots, x_{l_0}) \leqslant \Phi(n_0 + 1, l_0)$.

В случае а), в силу предположения индукции, существует подпоследовательность длины $n_0 + 1$ такая, что $\Delta^S(X^{n_0+1}) = 2^{n_0+1}$, что и требуется.

Для случая б) разделим подпоследовательности последовательности X^{l_0}, индуцируемые множествами из S, на два типа. К первому типу отнесем такие подпоследовательности X^r, что на полной последовательности X^{l_0+1} событиями из S индуцируется как X^r, так и (X^r, x_{l_0+1}). Ко второму — такие X^r, что на последовательности X^{l_0+1} индуцируется либо X^r, либо (X^r, x_{l_0+1}). Обозначим число подпоследовательностей первого типа K_1, а второго типа K_2. Легко видеть, что

$$\Delta^S(x_1, \ldots, x_{l_0}) = K_1 + K_2,$$
$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = 2K_1 + K_2;$$

и, следовательно,

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = \Delta^S(x_1, \ldots, x_{l_0}) + K_1. \qquad (\text{П}.4)$$

Обозначим через S' систему всех подмножеств $A \in S$ таких, что на последовательности X^{l_0} они индуцируют подпоследовательности первого типа. Тогда, если

б′) $$K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) > \Phi(n_0, l_0),$$

то в силу предположения индукции существует подпоследовательность $X^{n_0} = x_{i_1}, \ldots, x_{i_{n_0}}$ такая, что

$$\Delta^{S'}(x_{i_1}, \ldots, x_{i_{n_0}}) = 2^{n_0} \quad (X^{n_0} \subset X^{l_0}).$$

Но тогда для последовательности $x_{i_1}, \ldots, x_{i_{n_0}}, x_{l_0+1}$ имеем

$$\Delta^{S'}(x_1, \ldots, x_{i_{n_0}}, x_{l_0+1}) = 2^{n_0+1},$$

так как для каждой подпоследовательности X^r, индуцированной на последовательности X^{n_0}, найдутся две подпоследовательности, индуцированные на X^r, x_{l_0+1}, а именно X^r и (X^r, x_{l_0+1}). Таким образом, в случае б) искомая подпоследовательность найдена.

Если же

б″) $$K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) \leqslant \Phi(n_0, l_0),$$

то получим в силу (П.4) и б)

$$\Delta^S(x_1, \ldots, x_{l_0+1}) \leqslant \Phi(n_0+1, l_0) + \Phi(n_0, l_0),$$

откуда в силу свойств (П.3) функции $\Phi(n, l)$

$$\Delta^S(x_1, \ldots, x_{l_0+1}) \leqslant \Phi(n_0+1, l_0+1),$$

что противоречит условию леммы (т. е. б″) невозможно).

Лемма доказана.

Теперь докажем теорему. Как уже отмечалось, $m^S(l) \leqslant 2^l$. Пусть $m^S(l)$ не равно тождественно 2^l, и пусть n — первое значение l, при котором $m^S(l) \neq 2^l$.

Тогда для любой выборки длины l, большей n, справедливо

$$\Delta^S(x_1, \ldots, x_l) \leqslant \Phi(n, l).$$

Действительно, в противном случае на основании утверждения леммы нашлась бы такая подвыборка $x_1, \ldots, x_n$, что

$$\Delta^S(x_1, \ldots, x_n) = 2^n. \qquad (\text{П.5})$$

Равенство же (П.5) невозможно, так как по допущению $m^S(n) \neq 2^n$.

Таким образом, функция $m^S(l)$ либо тождественно равна 2^l, либо мажорируется $\Phi(n, l)$.

Теорема доказана.

Замечание. Функция $\Phi(n, l)$ может быть оценена сверху при $n \leqslant 1$ и $l > n$ следующим образом:

$$\Phi(n, l) < 1{,}5\,\frac{l^{n-1}}{(n-1)!}. \qquad (\text{П.6})$$

Поскольку для функции $\Phi(n, l)$ выполняются соотношения (П.3), для доказательства (П.6) достаточно убедиться, что при $n \geqslant 1$ и $l > n$ справедливо неравенство

$$\frac{l^{n-1}}{(n-1)!} + \frac{l^n}{n!} \leqslant \frac{(l+1)^n}{n!} \qquad (\text{П.7})$$

и проверить (П.6) на границе, т. е. при $n = 1$ $l = n+1$.

Неравенство (П.7), очевидно, равносильно неравенству

$$l^{n-1}(l+n) - (l+1)^n \leqslant 0,$$

справедливость которого следует из формулы бинома Ньютона.

Остается проверить соотношение (П.6) на границе. При $n=1$ оно проверяется непосредственно. Далее проверим оценку при малых n и l.

$l=n+1$	2	3	4	5	6
$\Phi(n, l)$	1	4	11	26	57
$1{,}5\frac{l^{n-1}}{(n-1)!}$	1,5	4,5	12	$31\frac{1}{4}$	81

Теперь, чтобы проверить (П.6) при $n\geqslant 6$, воспользуемся формулой Стирлинга для оценки сверху $l!$

$$l!\leqslant \sqrt{2\pi l}\, l^l e^{-l+\frac{1}{12l}},$$

откуда при $l=n+1$

$$\frac{l^{n-1}}{(n-1)!}=\frac{(l-1)\,l^{(l-1)}}{l!}\geqslant \frac{l-1}{\sqrt{2\pi l}\, l}\, e^{l-\frac{1}{12l}}$$

и далее при $l\geqslant 6$

$$\frac{l^{(n-1)}}{(n-1)!}\geqslant 0{,}8\,\frac{1}{\sqrt{2\pi l}}\, e^l.$$

С другой стороны, всегда $\Phi(n, l)\leqslant 2^l$. Поэтому достаточно проверить, что при $l\geqslant 6$

$$2^l\leqslant 1{,}2\,\frac{1}{\sqrt{2\pi l}}\, e^l.$$

С ростом l (при $l>2$) это неравенство усиливается, и поэтому достаточно его проверить при $l=6$, в чем и убеждаемся непосредственно.

Итак, оказывается, что функция роста либо тождественно равна 2^l, либо при некотором n впервые нарушается равенство, т. е. $m^s(l)\neq 2^l$, и тогда функция роста мажорируется степенной функцией

$$m^s(l)<1{,}5\,\frac{l^{n-1}}{(n-1)!}.$$

Таким образом, для того чтобы оценить поведение функции роста, достаточно выяснить, каково минимальное число n такое, что ни на одной последовательности длины l система S не индуцирует все возможные подпоследовательности.

§ П.3. Основная лемма

Пусть взята выборка длины $2l$:

$$X^{2l} = x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$$

и подсчитаны частоты выпадания события $A \in S$ на первой полувыборке $x_1, \ldots, x_l$ и второй полувыборке $x_{l+1}, \ldots, x_{2l}$. Обозначим соответственно частоты через $\nu'(A)$ и $\nu''(A)$ и рассмотрим отклонение этих величин

$$\rho_A(x_1, \ldots, x_{2l}) = |\nu'(A) - \nu''(A)|.$$

Нас будет интересовать максимальное отклонение частот по всем событиям класса S:

$$\rho^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} \rho_A(x_1, \ldots, x_{2l}).$$

Введем обозначение:

$$\pi^S(x_1, \ldots, x_{2l}) = \sup_{A \in S} |\nu'(A) - P(A)|.$$

Далее будем полагать, что как $\pi^S(x_1, \ldots, x_l)$, так и $\rho^S(x_1, \ldots, x_{2l})$ — измеримые функции.

Основная лемма. *Распределения величин* $\pi^S(x_1, \ldots, x_l)$ *и* $\rho^S(x_1, \ldots, x_{2l})$ *связаны следующим соотношением*:

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} \leqslant 2P\left\{\rho^S(x_1, \ldots, ,x_{2l}) > \frac{\varkappa}{2}\right\},$$

если только $l > 2/\varkappa$.

Доказательство. По определению

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X^{(2l)}} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}),$$

где

$$\theta(z) = \begin{cases} 1, & \text{если } z > 0, \\ 0, & \text{если } z \leqslant 0. \end{cases}$$

Учитывая, что пространство $X(2l)$ выборок длины $2l$ есть прямое произведение $X_1(l)$ и $X_2(l)$ полувыборок длины l, согласно теореме Фубини [28], для любой измеримой функции $\varphi(x_1, \ldots, x_{2l})$ справедливо

$$\int_{X(2l)} \varphi(x_1, \ldots, x_{2l})\, dX^{2l} = \int_{X_1(l)} \left[\int_{X_2(l)} \varphi(x_1, \ldots, x_{2l})\, dX_2^l \right] dX_1^l.$$

Поэтому имеем

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int_{X_1(l)} dP(X_1^l) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l)$$

(во внутреннем интеграле первая полувыборка фиксируется). Обозначим через Q событие пространства $X_1(l)$

$$\{\pi^S(x_1, \ldots, x_l) > \varkappa\}$$

ограничивая область интегрирования, получим

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geqslant \int_Q dP(X_1^{2l}) \int_{X_2(l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l). \qquad (\text{П.8})$$

Оценим внутренний интеграл правой части неравенства, обозначив его через I. Здесь выборка $x_1, \ldots, x_l$ фиксирована и такова, что $\pi^S(x_1, \ldots x_l) > \varkappa$. Следовательно, существует $A^* \in S$ такое, что

$$|P(A^*) - \nu(A^*; x_1, \ldots x_l)| > \varkappa.$$

Тогда

$$I = \int_{X_2(l)} \theta\left[\sup_{A \subset S} \rho_A(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l) \geqslant \int_{X_2(l)} \theta\left[\rho_{A^*}(X^{2l}) - \frac{\varkappa}{2}\right] dP(X_2^l).$$

Пусть, например,

$$\nu'(A^*; x_1, \ldots, x_l) < P(A^*) - \varkappa$$

(совершенно аналогично рассматривается случай $\nu'(A^*) \geqslant P(A^*) + \varkappa$). Тогда для выполнения условия

$$|\nu'(A^*; x_1, \ldots, x_l) - \nu''(A^*; x_{l+1}, \ldots, x_{2l})| > \frac{\varkappa}{2}$$

достаточно потребовать, чтобы выполнялось соотношение

$$\nu''(A^*) > P(A^*) - \frac{\varkappa}{2},$$

откуда получаем

$$I \geqslant \int\limits_{X_2(l)} \theta\left[\nu''(A^*) - P(A^*) + \frac{\varkappa}{2}\right] dP(X_2^l) =$$

$$= \sum_{\frac{k}{l} > P(A^*) - \frac{\varkappa}{2}} C_l^k [P(A^*)]^k [1 - P(A^*)]^{l-k}.$$

Как известно, последняя сумма превосходит 1/2, если только $l > 2/\varkappa$. Возвращаясь к (П.8), получим для $l > 2/\varkappa$

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} \geqslant \frac{1}{2} \int\limits_Q dP(X^l) = \frac{1}{2} P\{\pi^S(X^l) > \varkappa\},$$

что и требовалось доказать.

§ П.4. Вывод достаточных условий

Справедлива

Теорема П.2. *Вероятность того, что хотя бы для одного события из класса S частота уклонится от соответствующей вероятности в эксперименте длины l более чем на* $\varkappa$, *удовлетворяет неравенству*

$$P\{\pi^S(x_1, \ldots, x_l) > \varkappa\} < 6m^S(2l)\, e^{-\frac{\varkappa^2 l}{4}}. \qquad (\text{П.9})$$

Следствие. *Для того чтобы частоты событий класса S сходились (по вероятности) к соответствующим вероятностям равномерно по классу S, достаточно существование такого конечного n, что при* $l > n$

$$m^S(l) < 1{,}5\,\frac{l^{n-1}}{(n-1)!}.$$

Доказательство. В силу основной леммы достаточно оценить величину

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int\limits_{X(2l)} \theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] dP(X^{2l}).$$

Рассмотрим отображение пространства $X(2l)$ на себя, получаемое некоторой перестановкой T_i элементов после-

довательности X^{2l}. В силу симметрии определения меры имеет место следующее равенство:

$$\int\limits_{X(2l)} f(X^{2l})\, dP(X^{2l}) = \int\limits_{X(2l)} f(T_i X^{2l})\, dP(X^{2l})$$

для любой интегрируемой функции $f(X)$.

Поэтому

$$P\left\{\rho^S(X^{2l}) > \frac{\varkappa}{2}\right\} = \int\limits_{X(2l)} \frac{\sum\limits_{i=1}^{(2l)!} \theta\left[\rho^S(T_i X^{2l}) - \frac{\varkappa}{2}\right]}{(2l)!}\, dP(X^{2l}), \tag{П.10}$$

где сумма берется по всем $(2l)!$ перестановкам.

Заметим прежде всего, что

$$\theta\left[\rho^S(X^{2l}) - \frac{\varkappa}{2}\right] = \theta\left[\sup_A |\nu'(A) - \nu''(A)| - \frac{\varkappa}{2}\right] =$$

$$= \sup_A \theta\left[|\nu'(A) - \nu''(A)| - \frac{\varkappa}{2}\right].$$

Очевидно, что если два множества A_1 и A_2 индуцируют на выборке $x_1, \ldots, x_l, x_{l+1}, \ldots, x_{2l}$ одну и ту же подвыборку, то

$$\nu'(A_1;\ T_i X^{2l}) = \nu'(A_2;\ T_i X^{2l}),$$
$$\nu''(A_1;\ T_i X^{2l}) = \nu''(A_2;\ T_i X^{2l}),$$

и, следовательно,

$$\rho_{A_1}(T_i X^{2l}) = \rho_{A_2}(T_i X^{2l})$$

для любой перестановки T_i.

Иными словами, если два события эквивалентны относительно выборки $x_1, \ldots, x_{2l}$, то уклонения частот для этих событий одинаковы при всех перестановках T_i. Поэтому, если из каждого класса эквивалентности взять по одному множеству и образовать конечную систему S', то

$$\sup_{A \in S} \rho_A(T_i X^{2l}) = \sup_{A \in S'} \rho_A(T_i X^{2l}).$$

Число событий в системе S' конечно и было обозначено через $\Delta^{S'}(x_1, \ldots, x_{2l})$. Поэтому, заменяя операцию sup

суммированием, получаем

$$\sup_{A\in S}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]=\sup_{A\in S'}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]\leqslant$$
$$\leqslant\sum_{A\in S'}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right].$$

Эти соотношения позволяют оценить подынтегральное выражение в (П.10)

$$\sup_{A\in S'}\frac{1}{(2l)!}\sum_{i=1}^{(2l)'}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]=$$
$$=\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\sup_{A\in S'}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]\leqslant$$
$$\leqslant\sum_{A\in S'}\left[\frac{\sum_{i=1}^{(2l)!}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]}{(2l)!}\right].$$

Выражение в квадратных скобках означает отношение числа порядков в выборке (при фиксированном составе), для которых

$$|\nu'(A)-\nu''(A)|>\frac{\varkappa}{2},$$

к общему числу перестановок. Легко видеть, что оно равно

$$\Gamma=\sum_k\frac{C_m^kC_{2l-m}^{l-k}}{C_{2l}^l},$$
$$k:\left\{\left|\frac{k}{l}-\frac{m-k}{l}\right|>\varkappa\right\},$$

где m равно числу элементов выборки $x_1, \ldots, x_{2l}$, принадлежащих A.

В § П.5 мы оценим величину

$$\Gamma<3\exp\left\{-\frac{\varkappa^2l}{4}\right\}.$$

Таким образом,

$$\sum_{A\subset S'}\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\theta\left[\rho_A(T_iX^{2l})-\frac{\varkappa}{2}\right]<$$

$$<\sum_{A\in S'}3\exp\left\{-\frac{\varkappa^2 l}{4}\right\}=3\,\Delta^S(x_1,\ \ldots,\ x_{2l})\exp\left\{-\frac{\varkappa^2 l}{4}\right\}\leqslant$$

$$\leqslant 3m^S(2l)\exp\left\{-\frac{\varkappa^2 l}{4}\right\}.$$

Подставляя эту оценку в интеграл (П.10), получим

$$P\left\{\rho^S(X^{2l})>\frac{\varkappa}{2}\right\}<3m^S(2l)\exp\left\{-\frac{\varkappa^2 l}{4}\right\},$$

откуда в силу основной леммы

$$P\{\pi(X^l)>\varkappa\}<6m^S(2l)\exp\left\{-\frac{\varkappa^2 l}{4}\right\}.$$

Теорема доказана.

Доказательство следствия. Пусть существует такое n, что при $l>n$

$$m^S(l)<1{,}5\,\frac{l^{n-1}}{(n-1)!}.$$

Тогда, очевидно,

$$\lim_{l\to\infty}P\{\pi^S(X^l)>\varkappa\}<9\lim_{l\to\infty}\frac{(2l)^{n-1}}{(n-1)!}\exp\left\{-\frac{\varkappa^2 l}{4}\right\}=0,$$

т. е. имеет место равномерная сходимость по вероятности.

Полученное достаточное условие не зависит от свойств распределения (единственное требование — измеримость функций π^S и ρ^S), а зависит от внутренних свойств системы S.

Замечание. Как было доказано в § П.2, если только функция $m^S(l)$ не равна тождественно 2^l, то существует n такое, что при $l>n$

$$m^S(l)<1{,}5\,\frac{l^{n-1}}{(n-1)!}.$$

Поэтому достаточное условие выполняется всегда, когда

$$m^S(l)\not\equiv 2^l.$$

§ П.5. Оценка величины Г

Оценим величину

$$\Gamma = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

где k — пробегает значения, удовлетворяющие неравенствам

$$\left|\frac{k}{l} - \frac{m-k}{l}\right| > \varkappa, \quad \max(0,\ m-l) \leqslant k \leqslant \min(m,\ l),$$

или, что то же самое, неравенствам

$$\left|k - \frac{m}{2}\right| > \frac{\varkappa l}{2}, \quad \max(0,\ m-l) \leqslant k \leqslant \min(m,\ l),$$

а l и $m \leqslant 2l$ — произвольные положительные целые числа.

Разложим Г на два слагаемых $\Gamma = \Gamma_1 + \Gamma_2$:

$$\Gamma_1 = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{где} \quad k > \frac{\varkappa l}{2} + \frac{m}{2},$$

$$\Gamma_2 = \sum_k \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \quad \text{где} \quad k < \frac{\varkappa l}{2} - \frac{m}{2}.$$

Введем обозначения

$$p(k) = \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l}, \tag{П.11}$$

$$q(k) = \frac{p(k+1)}{p(k)} = \frac{(m-k)(l-k)}{(k+1)(l+k+1-m)}, \tag{П.12}$$

где

$$\max(0,\ m-l) \leqslant k \leqslant \min(m,\ l).$$

Далее обозначим

$$s = \min(m,\ l), \quad T = \max(0,\ m-l);$$

$$d(k) = \sum_{i=k}^{s} p(i).$$

Очевидно, что имеет место соотношение

$$d(k+1) = \sum_{i=k+1}^{s} p(i) = \sum_{i=k}^{s-1} p(i+1) = \sum_{i=k}^{s-1} p(i)\, q(i). \tag{П.13}$$

Далее из (П.12) непосредственно следует, что при $i < j$ $q(i) < q(j)$, т. е. $q(i)$ монотонно убывает. Поэтому из (П.13) следует неравенство

$$d(k+1) = \sum_{i=k}^{s-1} p(i)\, q(i) < q(k) \sum_{i=k}^{s} p(i).$$

Далее, по определению $d(k)$ имеем

$$d(k+1) < q(k)\, d(k).$$

Применяя последовательно это соотношение, получим для произвольных k и j, удовлетворяющих условию $T \leqslant j < k \leqslant s$,

$$d(k) < d(j) \prod_{i=j}^{k-1} q(i).$$

Далее, поскольку $d(j) \leqslant 1$, то

$$d(k) < \prod_{i=j}^{k-1} q(i), \qquad \text{(П.14)}$$

где j — любое целое число, меньшее, чем k.

Положим

$$t = k - \frac{m-1}{2}.$$

Тогда

$$q(t) = \frac{\frac{m+1}{2} - t}{\frac{m+1}{2} + t} \cdot \frac{\left(l - \frac{m-1}{2}\right) - t}{\left(l - \frac{m-1}{2}\right) + t}.$$

При этом, очевидно, пока $T < k < s$, справедливо

$$|t| < \min\left(\frac{m+1}{2},\ l - \frac{m-1}{2}\right).$$

Для аппроксимации $q(k)$ исследуем функцию

$$F(t) = \frac{a-t}{a+t} \cdot \frac{b-t}{b+t},$$

считая, что a и b больше нуля.

При $|t|<\min(a, b)$

$$\ln F(t)=\ln(a-t)-\ln(a+t)+\ln(b-t)-\ln(b+t).$$

Далее имеем

$$\ln F(0)=0, \quad \frac{d}{dt}(\ln F(t))=-\left[\frac{2a}{a^2-t^2}+\frac{2b}{b^2-t^2}\right].$$

Отсюда следует, что при $|t|<\min(a, b)$

$$\frac{d}{dt}(\ln F(t))\leqslant -2\left[\frac{1}{a}+\frac{1}{b}\right].$$

Соответственно при $|t|<\min(a, b)$ и $t\geqslant 0$ выполнится неравенство

$$\ln F(t)\leqslant -2\left[\frac{1}{a}+\frac{1}{b}\right]t.$$

Возвращаясь к $q(t)$, получаем, что при $t\geqslant 0$

$$\ln q(t)\leqslant -2\left[\frac{2}{m+1}+\frac{2}{2l-m+1}\right]t=-8\frac{l+1}{(m+1)(2l-m+1)}t.$$

Оценим теперь

$$\ln\left(\prod_{i=j}^{k-1} q(i)\right),$$

считая, что $\frac{m-1}{2}\leqslant j\leqslant k-1$:

$$\ln\left(\prod_{i=j}^{k-1} q(i)\right)=\sum_{i=j}^{k-1}\ln q(i)\leqslant$$

$$\leqslant\frac{-8(l+1)}{(m+1)(2l-m+1)}\sum_{i=j}^{k-1}\left(i-\frac{m-1}{2}\right).$$

Возвращаясь к (П.14), получим

$$\ln d(k)<\frac{-8(l+1)}{(m+1)(2l-m+1)}\sum_{i=j}^{k-1}\left(i-\frac{m-1}{2}\right);$$

здесь j — любое число, меньшее k. Поэтому для $k>\frac{m-1}{2}$ можно положить $j=\frac{m-1}{2}$ для m нечетного и $j=m/2$ для m четного, получив более сильную оценку. Суммируя

далее арифметическую прогрессию, получим

$$\ln d(k) < \begin{cases} -\frac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\frac{m}{2}+1\right)^2 & \text{для четного } m, \\ -\frac{4(l+1)}{(m+1)(2l-m+1)}\left(k-\frac{m-1}{2}+1\right)\left(k-\frac{m-1}{2}\right) & \text{для нечетного } m. \end{cases}$$

Наконец, Γ_1 есть $d(k)$ при k первом целом таком, что

$$k-\frac{m}{2} > \frac{\varkappa^2 l}{2},$$

откуда

$$\ln \Gamma_1 < -\frac{(l+1)}{(m+1)(2l-m+1)}\varkappa^2 l^2.$$

Точно так же оценивается величина Γ_2, так как распределение (П.11) симметрично относительно точки $k=m/2$.

Таким образом,

$$\Gamma < 2\exp\left\{-\frac{(l+1)\varkappa^2 l^2}{(m+1)(2l-m+1)}\right\}. \qquad (\text{П.15})$$

Правая часть (П.15) достигает максимума при $m=l$, и, следовательно,

$$\Gamma < 2\exp\left\{-\frac{\varkappa^2 l^2}{l+1}\right\} < 3\exp\{-\varkappa^2 l\}. \qquad (\text{П.16})$$

При $\varkappa^2 l < 1$ оценка (П.16) тривиальна, поскольку левая часть неравенства не превосходит единицу, а правая всегда больше единицы.

Таким образом, оценка (П.16) справедлива при любых целых l и $\varkappa$ в пределах $0 \leqslant \varkappa \leqslant 1$.

§ П.6. Оценка вероятности равномерного относительного уклонения

В этом параграфе мы докажем теорему.

Теорема П.3. *При* $l > \frac{1}{2\varkappa^2}$ *справедлива оценка*

$$P\left\{\sup_{A\in S}\frac{P(A)-\nu(A)}{\sqrt{P(A)}} > \varkappa\right\} < 8m^S(2l)e^{-\frac{\varkappa^2 l}{4}}.$$

Доказательство. Рассмотрим два события, построенные по случайной и независимой выборке длины $2l$, событие Q_1:

$$Q_1 = \left\{ \sup_{A \in S} \frac{P(A) - \nu'(A)}{\sqrt{P(A)}} > \varkappa \right\}$$

— и событие Q_2:

$$Q_2 = \left\{ \sup_{A \in S} \frac{|\nu'(A) - \nu''(A)|}{\sqrt{\nu(A) + \frac{1}{2l}}} > \varkappa \right\},$$

где $\nu'(A)$ — частота события A, вычисленная на первой полувыборке длины l; $\nu''(A)$ — частота события A, вычисленная на второй полувыборке, $\nu(A)$ — частота события, вычисленная на выборке длины $2l$.

Доказывать теорему мы будем по следующей схеме. Сначала покажем, что справедливо неравенство

$$P(Q_1) < 4P(Q_2),$$

а затем оценим вероятность события Q_2. Итак, докажем лемму.

Лемма П.2. *При* $l > \frac{1}{2\varkappa^2}$ *справедливо неравенство*

$$P(Q_1) < 4P(Q_2). \qquad (\text{П.17})$$

Доказательство. Допустим, что событие Q_1 произошло. Это значит, что существует такое A^*, что на первой полувыборке выполнится неравенство

$$P(A^*) - \nu'(A^*) > \varkappa \sqrt{P(A^*)}.$$

Поскольку $\nu'(A) \geqslant 0$, это значит, что

$$P(A^*) > \varkappa^2.$$

Допустим, что на второй полувыборке частота выпадания события A^* превзошла вероятность $P(A^*)$:

$$\nu''(A^*) > P(A^*).$$

Вспомним еще, что $l > \frac{1}{2\varkappa^2}$. При этих условиях обязательно выполнится событие Q_2.

Действительно, оценим величину

$$\mu = \frac{|\nu'(A^*) - \nu''(A^*)|}{\sqrt{\nu(A^*) + \frac{1}{2l}}} > \frac{\nu''(A^*) - \nu'(A^*)}{\sqrt{\nu(A^*) + \frac{1}{2l}}} \qquad (\text{П.18})$$

при условии

$$\nu'(A^*) < P(A^*) - \varkappa\sqrt{P(A^*)},$$
$$\nu''(A^*) > P(A^*),$$
$$P(A^*) > \varkappa^2.$$

Для этого найдем минимум функции

$$T = \frac{x - y}{\sqrt{x + y + c}}$$

в области $0 < a \leqslant x \leqslant 1$, $0 < y \leqslant b$, $c > 0$.

Имеем

$$\frac{\partial T}{\partial x} = \frac{1}{2}\frac{x + 3y + 2c}{(x + y + c)^{3/2}} > 0,$$
$$\frac{\partial T}{\partial y} = -\frac{1}{2}\frac{3x + y + 2c}{(x + y + c)^{3/2}} < 0.$$

Следовательно, T достигает минимума в допустимой области при $x = a$, $y = b$.

Поэтому величина μ будет оценена снизу, если в (П.18) $\nu'(A^*)$ заменить на $P(A^*) - \varkappa\sqrt{P(A^*)}$, а $\nu''(A^*)$ — на $P(A^*)$. Таким образом,

$$\mu > \frac{\varkappa\sqrt{2P(A^*)}}{\sqrt{2P(A^*) - \varkappa\sqrt{P(A^*)} + \frac{1}{2l}}}.$$

Далее, поскольку $P(A^*) > \varkappa$, $2l > \frac{1}{\varkappa^2}$, имеем

$$\mu > \frac{\varkappa\sqrt{2P(A^*)}}{\sqrt{2P(A^*) - \varkappa^2 + \varkappa^2}} = \varkappa.$$

Таким образом, при выполнении Q_1, а также условий $P(A^*) \leqslant \nu''(A^*)$ и $l > \frac{1}{2\varkappa^2}$, выполняется и Q_2.

Заметим, что вторая полувыборка выбирается независимо от первой и, как известно, при $l > \frac{1}{2P(A^*)}$ частота

выпадания события A^* с вероятностью 1/4 превышает $P(A^*)$. Поэтому событие

$$\nu''(A) > P(A^*)$$

выполняется при условии выполнения Q_1 с вероятностью, большей 1/4, если только $l > \frac{1}{2\varkappa^2}$.

Таким образом, при $l > \frac{1}{2\varkappa^2}$

$$P(Q_2) > \frac{1}{4} P(Q_1).$$

Лемма доказана.

Лемма П.3. *Справедлива оценка*

$$P(Q_2) < 2m^S(2l)\, e^{-\frac{\varkappa^2 l}{4}}.$$

Доказательство. Обозначим через $R_A(X^{2l})$ величину

$$R_A(X^{2l}) = \frac{|\nu'(A) - \nu''(A)|}{\sqrt{\nu(A) + \frac{1}{2l}}}.$$

Тогда оцениваемая вероятность равна

$$P(Q_2) = \int\limits_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}).$$

Здесь интегрирование ведется в пространстве всех возможных выборок длины $2l$.

Рассмотрим теперь все возможные перестановки T_i ($i = 1, 2, \ldots, (2l)!$) последовательности $x_1, \ldots, x_{2l}$.

Для каждой такой перестановки T_i справедливо равенство

$$\int\limits_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}) = \int\limits_{X(2l)} \theta\left[\sup_{A \in S} R_A(T_i X^{2l}) - \varkappa\right] dP(X^{2l}).$$

Поэтому справедливо равенство

$$\int\limits_{X(2l)} \theta\left[\sup_{A \in S} R_A(X^{2l}) - \varkappa\right] dP(X^{2l}) = \int\limits_{X(2l)} \frac{1}{(2l)!} \sum_{i=1}^{(2l)!} \theta\left[\sup_{A \in S} R_A(T_i X^{2l}) - \varkappa\right] dP(X^{2l}).$$

Рассмотрим теперь подынтегральное выражение. Так как в нем выборка $x_1, \ldots, x_{2l}$ фиксирована, то вместо системы событий S можно рассматривать конечную систему событий S', куда входит по одному представителю из каждого класса эквивалентности.

Таким образом, справедливо равенство

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\theta\left[\sup_{A\in S}R_A(T_iX^{2l})-\varkappa\right]=$$

$$=\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\theta\left[\sup_{A\in S'}R_A(T_iX^{2l})-\varkappa\right].$$

Далее справедливо

$$\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\theta\left[\sup_{A\in S'}R_A(T_iX)^{2l}-\varkappa\right]<$$

$$<\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\sum_{A\in S'}\theta\left[R_A(T_iX^{2l})-\varkappa\right]=$$

$$=\sum_{A\in S'}\left\{\frac{1}{(2l)!}\sum_{i=1}^{(2l)!}\theta\left[R_A(T_iX^{2l})-\varkappa\right]\right\}. \quad (\text{П}.19)$$

Выражение в фигурных скобках есть вероятность уклонения частот в двух полувыборках для фиксированного события A и для данного состава полной выборки.

Эта вероятность равна

$$\Gamma=\sum_k\frac{C_m^kC_{2l-m}^{l-k}}{C_{2l}^l},$$

где m — число выпадений событий A в полной выборке, k — число выпадений событий в первой полувыборке, k пробегает значения

$$\max(0,\ m-l)\leqslant k\leqslant\min(m,\ l),$$

$$\frac{\left|\frac{k}{l}-\frac{m-k}{l}\right|}{\sqrt{\frac{m+1}{2l}}}>\varkappa.$$

Обозначим через $\varkappa'$ величину

$$\sqrt{\frac{m+1}{2l}}\,\varkappa=\varkappa'.$$

В этих обозначениях ограничения примут вид

$$\max(0,\ m-l)\leqslant k\leqslant \min(m,\ l),$$

$$\left|\frac{k}{l}-\frac{m-k}{l}\right|>\varkappa'. \qquad (\text{П.20})$$

В § П.5 была найдена оценка величины Γ при ограничениях (П.20)

$$\Gamma<2\exp\left\{-\frac{(l+1)(\varkappa')^2 l^2}{(m+1)(2l-m+1)}\right\}. \qquad (\text{П.21})$$

Выражая (П.19) через $\varkappa$, получим

$$\Gamma<2\exp\left\{-\frac{\varkappa^2(l+1)l}{2(2l-m+1)}\right.$$

Правая часть неравенства достигает максимума при $m=0$. Таким образом,

$$\Gamma<2\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \qquad (\text{П.22})$$

Подставим (П.22) в правую часть (П.19) и проведем интегрирование

$$P(Q_2)<2m^S(2l)\exp\left\{-\frac{\varkappa^2 l}{4}\right\}. \qquad (\text{П.23})$$

Лемма доказана.

Из неравенств (П.17) и (П.23) следует справедливость теоремы.

ГЛАВА VII

МЕТОД МИНИМИЗАЦИИ ЭМПИРИЧЕСКОГО РИСКА В ЗАДАЧЕ ВОССТАНОВЛЕНИЯ РЕГРЕССИИ

§ 1. О равномерной сходимости средних к математическим ожиданиям

В этой книге задача обучения распознаванию образов сформулирована как наиболее простая задача восстановления зависимостей по эмпирическим данным. Элементарность ее определяется тем, что задача сводится к минимизации функционала

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(x, y)\, dx\, dy \tag{7.1}$$

с неизвестной плотностью $P(x, y)$, по выборке

$$x_1, y_1; \ldots; x_l, y_l, \tag{7.2}$$

в ситуации, когда y принимает лишь два значения — нуль и единица, а $F(x, \alpha)$ — класс характеристических функций.

Задача восстановления регрессии считается более сложной. Она также сводится к минимизации функционала с неизвестной плотностью $P(x, y)$ по выборке (7.2), но здесь значение y может быть любым числом, а класс $F(x\ \alpha)$ принадлежит интегрируемым с квадратом функциям.

Поэтому построение теории минимизации риска (7.1) в классе не обязательно характеристических функций $F(x, \alpha)$ путем минимизации эмпирического функционала

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 \tag{7.3}$$

может рассматриваться как обобщение результатов теории, полученной в предыдущей главе, на более широкий класс функций.

В этой главе мы построим теорию восстановления регрессии методом минимизации эмпирического риска (7.2) как естественное обобщение решения задачи обучения распознаванию образов.

Возможность провести такую точку зрения представляется нам впервые. Использование параметрических методов в задачах распознавания образов (гл. III) и восстановления регрессии (гл. IV, V) не позволяло это сделать. Решение задач проводилось в условиях существенно различных моделей плотностей $P(x, y)$: в задаче обучения распознаванию образов структура плотности определялась объединением двух плотностей, а в задаче восстановления регрессии — схемой измерения с аддитивной помехой. Здесь же принцип решения задач один и тот же: поиск функции, минимизирующей (7.1), проводится путем минимизации эмпирического функционала (7.3).

В предыдущей главе были получены условия, при которых этот путь приводит к успеху для класса характеристических функций $F(x, \alpha)$. Теперь мы найдем условия, гарантирующие успех применения метода минимизации эмпирического риска, если $F(x, \alpha)$ — класс функций более общей природы.

В задаче распознавания образов функционал (7.1) для каждого фиксированного α определяет вероятность некоторого события (неправильной классификации вектора, поступающего для распознавания), а эмпирический функционал (7.3) — частоту этого события, вычисленную по обучающей последовательности. Условия применимости метода минимизации эмпирического риска здесь связаны с существованием равномерной сходимости частот появления событий к их вероятностям по классу событий.

В задаче восстановления регрессии функционал (7.1) определяет для каждого фиксированного α математическое ожидание случайной величины

$$\xi(\alpha) = (y - F(x, \alpha))^2,$$

а эмпирический функционал (7.3) — эмпирическое среднее этой случайной величины, найденное по выборке (7.2).

Выше (§ 1 гл. VI) было показано, что успех применения метода минимизации эмпирического риска может быть связан с существованием равномерной сходимости средних к математическим ожиданиям:

$$P\left\{\sup_{\alpha} |I(\alpha) - I_э(\alpha)| > \varkappa\right\} < \eta(l, \varkappa),$$
$$\lim_{l \to \infty} \eta(l, \varkappa) = 0. \tag{7.4}$$

Было показано, что при выполнении условия (7.4) с вероятностью $1-\eta$ значение функционала (7.1) в точке эмпирического минимума $F(x, \alpha_э)$ уклонится от минимального в классе $F(x, \alpha)$ значения $I(\alpha_0)$ не более чем на $2\varkappa$:

$$P\{I(\alpha_э) - I(\alpha_0) > 2\varkappa\} < \eta.$$

Таким образом, проблема сводится к отысканию условий существования равномерной сходимости средних к математическим ожиданиям и оценке скорости сходимости.

§ 2. Частный случай

Как и раньше, начнем с простого случая: множество функций $F(x, \alpha)$ состоит из конечного числа N элементов

$$F(x, \alpha_1), \ldots, F(x, \alpha_N).$$

Для этого случая справедливо неравенство

$$P\left\{\sup_i |I(\alpha_i) - I_э(\alpha_i)| > \varkappa\right\} < \sum_{i=1}^{N} P\{|I(\alpha_i) - I_э(\alpha_i)| > \varkappa\} \leqslant$$
$$\leqslant N \cdot \sup_i P\{|I(\alpha_i) - I_э(\alpha_i)| > \varkappa\}. \qquad (7.5)$$

В главе VI в аналогичной ситуации при оценке скорости равномерной сходимости частот появления событий к их вероятностям использовалась нетривиальная оценка второго сомножителя. Здесь же нетривиальная оценка сомножителя

$$\sup_i P\{|I(\alpha_i) - I_э(\alpha_i)| > \varkappa\},$$

вообще говоря, невозможна — случайная величина $I_э(\alpha_i)$ может иметь «большие выбросы», и поэтому ее отклонение от среднего $I(\alpha_i)$ может быть любым. В главе II мы уже сталкивались с ситуацией, в которой для получения гарантированной оценки математического ожидания по величине эмпирического среднего необходимо было учесть меру «возможного выброса». В частности, было показано (см. гл. II, § 2), что для этого достаточно знать либо оценку величины возможных потерь

$$\sup_{\alpha, x, y} (y - F(x, \alpha))^2 \leqslant \tau,$$

либо оценку относительной величины дисперсии потерь

$$\sup_{\alpha} \sqrt{\frac{\int (y-F(x,\ \alpha))^4 P(x,\ y)\, dx\, dy}{\left(\int (y-F(x,\ \alpha))^2 P(x,\ y)\, dx\, dy\right)^2} - 1} \leqslant \tau.$$

Таким образом, для получения оценок скорости равномерной сходимости средних к математическим ожиданиям должна быть использована априорная информация о величине возможных выбросов. Заметим, что при решении задачи обучения распознаванию образов такая проблема не возникала — согласно постановке в этой задаче величина функции потерь $(y-F(x,\ \alpha))^2$ не превышает единицу, т. е. априорная информация о выбросах содержится в самой постановке.

В этой главе мы используем оба типа априорной информации о выбросах и для каждого из них получим оценку скорости равномерной сходимости.

Наиболее простым условием, при котором возможно получение оценки скорости равномерной сходимости средних к математическим ожиданиям, является условие равномерной ограниченности потерь

$$(y-F(x,\ \alpha))^2 \leqslant \tau \tag{7.6}$$

для всех α, $x \in X$, $y \in Y$.

Пусть имеет место неравенство (7.6). Покажем, что в этом случае справедлива оценка

$$P\left\{\sup_i |I(\alpha_i) - I_э(\alpha_i)| > \varkappa\tau\right\} < 18Nle^{-\frac{\varkappa^2 l}{4}}.$$

Для получения этой оценки запишем функционалы $I(\alpha_i)$ и $I_э(\alpha_i)$ с помощью интегралов Лебега:

$$\begin{aligned} I(\alpha_i) &= \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} P\left\{(y-F(x,\ \alpha_i))^2 > \frac{j\tau}{n}\right\}, \\ I_э(\alpha_i) &= \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} \nu\left\{(y-F(x,\ \alpha_i))^2 > \frac{j\tau}{n}\right\}, \end{aligned} \tag{7.7}$$

где через $\nu\left\{(y-F(x,\ \alpha_i))^2 > \frac{j\tau}{n}\right\}$ обозначена частота события $\left\{(y-F(x,\ \alpha_i))^2 > \frac{j\tau}{n}\right\}$, вычисленная по обучающей

последовательности (7.2). Обозначим событие

$$\left\{(y - F(x,\ \alpha_i))^2 > \frac{j\tau}{n}\right\}$$

через $A_{\alpha_i, j}$. Тогда, согласно (7.7),

$$|I(\alpha_i) - I_э(\alpha_i)| \leqslant \lim_{n\to\infty} \sum_{j=1}^{n} \frac{\tau}{n} |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})| \leqslant$$

$$\leqslant \tau \sup_j |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})|.$$

Таким образом,

$$P\{|I(\alpha_i) - I_э(\alpha_i)| > \tau\varkappa\} \leqslant P\left\{\sup_j |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})| > \varkappa\right\}.$$

Рассмотрим теперь класс событий $A_{\alpha_i, \beta}$:

$$\{(y - F(x,\ \alpha_i))^2 > \beta\},$$

где β — неотрицательная величина.

Очевидно, этот класс содержит события $\{A_{\alpha_i, j}\}$, откуда следует

$$P\left\{\sup_j |P(A_{\alpha_i, j}) - \nu(A_{\alpha_i, j})| > \varkappa\right\} \leqslant$$

$$\leqslant P\left\{\sup_\beta P|(A_{\alpha_i, \beta}) - \nu(A_{\alpha_i, \beta})| > \varkappa\right\}.$$

И задача свелась к оценке равномерной сходимости частот к вероятностям по классу S_β событий $A_{\alpha_i, \beta}$ (с фиксированными значениями α_i).

Используя результаты предыдущей главы, оценим скорость равномерной сходимости частот к вероятностям по классу событий

$$S_\beta = \{x,\ y:\ (y - F(x,\ \alpha_i))^2 > \beta\}.$$

Для этого оценим функцию роста $m^{S_\beta}(l)$. Так как с помощью правил

$$\theta[(y - F(x,\ \alpha_i))^2 - \beta]$$

(α_i — фиксировано) всеми возможными способами можно делить лишь одну точку x, y, то, согласно теореме 6.6, справедливо

$$m^{S_\beta}(l) < 1{,}5^l.$$

Следовательно, используя теорему П.2 Приложения к гл. VI, получим

$$P\{|I(\alpha_i)-I_э(\alpha_i)|>\tau\varkappa\}\leqslant$$
$$\leqslant P\left\{\sup_{\beta}|P(A_{\alpha_i,\beta})-\nu(A_{\alpha_i,\beta})|>\varkappa\right\}<$$
$$<6m^{S_\beta}(2l)e^{-\frac{\varkappa^2 l}{4}}<18le^{-\frac{\varkappa^2 l}{4}}. \qquad (7.8)$$

Правая часть неравенства не зависит от параметра α. Поэтому наряду с (7.8) справедливо и более сильное утверждение

$$\sup_{\alpha} P\{|I(\alpha)-I_э(\alpha)|>\tau\varkappa\}<18le^{-\frac{\varkappa^2 l}{4}}.$$

Теперь, возвращаясь к оценке (7.5), получаем

$$P\left\{\sup_{\alpha}|I(\alpha)-I_э(\alpha)|>\tau\varkappa\right\}<18Nle^{-\frac{\varkappa^2 l}{4}}.$$

Потребуем, чтобы эта вероятность равнялась η:

$$18Nle^{-\frac{\varkappa^2 l}{4}}=\eta.$$

Для этого уклонение $\varkappa$ должно быть не меньше

$$\varkappa=2\sqrt{\frac{\ln N+\ln l-\ln\eta/18}{l}}.$$

Полученный результат может быть сформулирован в виде следующей теоремы.

Теорема 7.1. *Пусть класс $F(x,\ \alpha_l)$ состоит из N функций, для которых величины потерь $(y-F(x,\ \alpha))^2$ в области $x\in X$ и $y\in Y$ равномерно ограничены константой τ. Тогда с вероятностью $1-\eta$ можно утверждать, что одновременно для всех N функций $F(x,\ \alpha_l)$ имеет место неравенство*

$$I_э(\alpha_l)-2\tau\sqrt{\frac{\ln N+\ln l-\ln\eta/18}{l}}<I(\alpha_l)<$$
$$<I_э(\alpha_l)+2\tau\sqrt{\frac{\ln N+\ln l-\ln\eta/18}{l}}.$$

Замечание. Теорема справедлива одновременно для всех N функций, в том числе и для той $F(x\ \alpha_э)$, кото-

рая доставляет минимум величине эмпирического риска. Таким образом, имеет место неравенство

$$I(\alpha_э) < I_э(\alpha_э) + 2\tau \sqrt{\frac{\ln N + \ln l - \ln \eta/18}{l}}.$$

Итак, если функция потерь равномерно ограничена, а число функций $F(x, \alpha_l)$ в классе конечно, то имеет место равномерная сходимость средних к их математическим ожиданиям. Теорема 7.1 является прямым обобщением теоремы 6.1.

§ 3. Обобщение на класс с бесконечным числом элементов

Пусть теперь класс $F(x, \alpha)$ состоит из бесконечного числа элементов и в то же время допускает покрытие конечной ε-сетью в метрике C или метрике L_p^2. По-прежнему справедливо ограничение (7.6). Покажем, что в этом случае справедлива оценка качества правила, минимизирующего эмпирический риск, аналогичная той, которая следует из теоремы 7.1.

Теорема 7.2. *Пусть множество функций $F(x, \alpha)$ покрыто конечной ε-сетью $F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)})$. Тогда с вероятностью $1-\eta$ качество функции $F(x, \alpha_э)$, минимизирующей эмпирический риск, оценивается величиной*

$$I(\alpha_э) < I_э(\alpha_l(\alpha_э)) + 2\tau \sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/18}{l}} + 2\varepsilon \sqrt{\tau},$$

где $F(x, \alpha_l(\alpha_э))$ — ближайшая к $F(x, \alpha_э)$ функция ε-сети.

Доказательство теоремы 7.2 проводится по той же схеме, что и доказательство теоремы 6.4.

1°. На множестве функций $F(x, \alpha)$ выделим конечную ε-сеть, состоящую из $N(\varepsilon)$ элементов

$$F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)}).$$

Согласно теореме 7.1 с вероятностью $1-\eta$ одновременно для всех элементов ε-сети справедливы неравенства

$$I(\alpha_l) < I_э(\alpha_l) + 2\tau \sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/18}{l}}. \qquad (7.9)$$

2°. Оценим величину уклонения функционалов $I(\alpha_1)$ и $I(\alpha_2)$ для функций $F(x, \alpha_1)$, $F(x, \alpha_2)$, отстоящих друг

от друга не более чем на ε, т. е. найдем наименьшее $\delta(\varepsilon)$, при котором выполнится неравенство

$$|I(\alpha_1) - I(\alpha_2)| \leqslant \delta(\varepsilon),$$

если только окажутся выполненными условия

$$\begin{aligned} &\rho_L(\alpha_1, \alpha_2) = \left(\int (F(x, \alpha_1) - F(x, \alpha_2))^2 P(x)\, dx\right)^{1/2} \leqslant \varepsilon, \\ &\left(\rho_C(\alpha_1, \alpha_2) = \sup_x |F(x_1\alpha_1) - F(x_1, \alpha_2)| \leqslant \varepsilon\right). \end{aligned} \tag{7.10}$$

Для этого проведем преобразования:

$$\begin{aligned} |I(\alpha_1) - I(\alpha_2)| &= \Big|\int (y - F(x, \alpha_1))^2 P(x, y)\, dx\, dy - \\ &\quad - \int (y - F(x, \alpha_2))^2 P(x, y)\, dx\, dy\Big| = \\ &= \Big|\int (F(x, \alpha_1) - F(x, \alpha_2)) \times \\ &\quad \times (2y - F(x, \alpha_1) - F(x, \alpha_2)) P(x, y)\, dx\, dy\Big| \leqslant \\ &\leqslant \varepsilon \sqrt{\int (2y - F(x, \alpha_1) - F(x, \alpha_2))^2 P(x, y)\, dx\, dy}. \end{aligned}$$

Здесь мы воспользовались неравенством Коши и оценкой (7.10). Далее воспользуемся выпуклостью функции $(y - F(x, \alpha))^2$:

$$\begin{aligned} \int (2y - F(x, \alpha_1) - F(x, \alpha_2))^2 P(x, y)\, dx\, dy &\leqslant \\ &\leqslant 2\int (y - F(x, \alpha_1))^2 P(x, y)\, dx\, dy + \\ &\quad + 2\int (y - F(x, \alpha_2))^2 P(x, y)\, dx\, dy. \end{aligned}$$

Таким образом, получаем

$$|I(\alpha_1) - I(\alpha_2)| \leqslant \varepsilon \sqrt{2(I(\alpha_1) + I(\alpha_2))}. \tag{7.11}$$

Но, согласно условию, $I(\alpha) \leqslant \tau$. Окончательно получаем

$$|I(\alpha_1) - I(\alpha_2)| \leqslant 2\varepsilon \sqrt{\tau}. \tag{7.11a}$$

3°. Пусть теперь $F(x, \alpha_э)$ — функция, доставляющая минимум эмпирическому риску. Выберем ближайшую к $F(x, \alpha_э)$ функцию $F(x, \alpha_i(\alpha_э))$ ε-сети: $F(x, \alpha_1)$, ..., $F(x, \alpha_{N(\varepsilon)})$.

Для этой функции с вероятностью $1 - \eta$ выполнится неравенство (7.9). Усилим его, воспользовавшись оценкой

(7.11а). В результате получим

$$I(\alpha_э) < I_э(\alpha_l(\alpha_э)) + 2\varepsilon\sqrt{\tau} + 2\tau\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/18}{l}}. \tag{7.12}$$

Теорема доказана.

Замечания. Теорема справедлива для любой величины ε (заданной до момента появления выборки). Поэтому величину ε можно выбрать из условия минимума выражения

$$r(\varepsilon) = \varepsilon\sqrt{\tau} + \tau\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/18}{l}}.$$

Заметим также, что для любого множества $F(x, \alpha)$ и любого ε минимальное число элементов ε-сети, построенной в метрике L_P^2 не больше минимального числа элементов ε-сети в метрике C. Поэтому оценка (7.12) будет более точной, если ε-сеть строится в метрике L_P^2. Однако, для того чтобы задать эту метрику, надо знать плотность $P(x)$.

§ 4. Емкость множества произвольных функций

В главе VI мы ввели понятие «емкость» для множества характеристических функций. Емкость определялась максимальным числом точек $x_1, \ldots, x_h$, которые всеми возможными способами могли быть разделены на два класса с помощью заданного множества характеристических функций.

Распространим теперь понятие «емкость» на множества функций $F(x, \alpha)$ произвольной природы. Для этого введем параметрическое множество характеристических функций

$$\hat{F}(x, y; \alpha, \beta) = \theta((y - F(x, \alpha))^2 + \beta)$$

(по параметрам α, β; β — действительное число).

Определение. *Назовем емкостью множества $F(x, \alpha)$ — емкость множества характеристических функций $\hat{F}(x, y; \alpha, \beta)$.*

Таким образом, емкость множества $F(x, \alpha)$ определяет наибольшее число h пар x_i, y_i, которые всеми возможными способами можно разбить на два класса с помощью правил $\hat{F}(x, y; \alpha, \beta)$.

Емкость множества линейных по параметрам функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$$

равна $n+1$.

При таком определении емкости функция роста системы событий

$$S_{\alpha, \beta} = \{x, y: (y - F(x, \alpha))^2 > \beta\}$$

при $l > h$ оценивается величиной

$$m^{S_{\alpha, \beta}}(l) < 1{,}5 \frac{l^h}{h!}.$$

Итак, пусть емкость множества функций $F(x, \alpha)$ равна h и по-прежнему функция потерь ограничена величиной τ. В этих условиях справедлива

Теорема 7.3. *При $l > h$ с вероятностью $1 - \eta$ одновременно для всего класса функций $F(x, \alpha)$ выполняются неравенства*

$$I_э(\alpha) - 2\tau \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}} < I(\alpha) < I_э(\alpha) + 2\tau \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{9}}{l}}.$$

Доказательство. Выразим функционалы $I(\alpha)$ и $I_э(\alpha)$ через интегралы Лебега:

$$I(\alpha) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{\tau}{n} P\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\},$$

$$I_э(\alpha) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{\tau}{n} \nu\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\}.$$

Здесь $P\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\}$ означает вероятность события $\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\}$, а $\nu\left\{(y - F(x, \alpha))^2 > \frac{i\tau}{n}\right\}$ — частоту этого события, вычисленную на обучающей последовательности.

Обозначим через $A_{\alpha,\beta} \in S_{\alpha\beta}$ событие

$$\{(y - F(x,\ \alpha))^2 > \beta\}.$$

Тогда

$$|I(\alpha) - I_э(\alpha)| \leqslant \lim_{n\to\infty} \sum_{i=1}^{n} \frac{\tau}{n} |P(A_{\alpha,i}) - \nu(A_{\alpha,i})|.$$

И следовательно,

$$|I(\alpha) - I_э(\alpha)| \leqslant \tau \sup_{\beta} |P(A_{\alpha,\beta}) - \nu(A_{\alpha,\beta})|.$$

Далее следует

$$P\left\{\sup_{\alpha} |I(\alpha) - I_э(\alpha)| > \tau\varkappa\right\} \leqslant$$

$$\leqslant P\left\{\sup_{\alpha,\ \beta} |P(A_{\alpha,\beta}) - \nu(A_{\alpha,\beta})| > \varkappa\right\}.$$

Так как при $l > h$ функция роста системы событий $S_{\alpha,\beta}$ ограничена величиной $1{,}5\frac{l^h}{h!}$, то, используя теорему П.2 Приложения к гл. VI, получим

$$P\left\{\sup_{\alpha} |I(\alpha) - I_э(\alpha)| > \tau\varkappa\right\} <$$

$$< 6m^S(2l)\, e^{-\frac{\varkappa^2 l}{4}} < 9\frac{(2l)^h}{h!}\, e^{-\frac{\varkappa^2 l}{4}}. \qquad (7.13)$$

Приравняв правую часть неравенства величине η и разрешив это равенство относительно $\varkappa$, получим

$$\varkappa = 2\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\frac{\eta}{9}}{l}}. \qquad (7.14)$$

Таким образом, из (7.13) и (7.14) следует, что при $l > h$ с вероятностью $1 - \eta$ одновременно для всех функций множества $F(x,\ \alpha)$ выполняются неравенства

$$I_э(\alpha) - 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\eta/9}{l}} < I(\alpha) <$$

$$< I_э(\alpha) + 2\tau\sqrt{\frac{h\left(\ln\frac{2l}{h} + 1\right) - \ln\eta/9}{l}}.$$

Теорема доказана.

§ 5. Равномерная ограниченность отношения моментов

Пусть теперь выполнены условия

$$\sup_{\alpha} \frac{\sqrt[p]{\int (y-F(x,\alpha))^{2p} P(x,y)\,dx\,dy}}{\int (y-F(x,\alpha))^2 P(x,y)\,dx\,dy} \leqslant \tau, \quad p \geqslant 2, \qquad (7.15)$$

т. е. для всякого фиксированного $\alpha = \alpha^*$ отношение среднего p-го порядка случайной величины [1])

$$\xi(\alpha^*) = (y - F(x, \alpha^*))^2$$

к среднему первого порядка не превосходит τ.

Выполнение условия (7.15) явится тем основным требованием, которое мы будем предъявлять при решении задач восстановления зависимостей и решении некорректных задач.

Для $p=2$ это требование эквивалентно требованию равномерной ограниченности относительной величины дисперсии, рассмотренной в § 2 гл. II, причем число τ_σ, ограничивающее относительную величину дисперсии, связано с числом τ, ограничивающим относительную величину среднего второго порядка, соотношением

$$\tau = \sqrt{\tau_\sigma^2 + 1}.$$

Условие (7.15) является достаточно слабым. Так все параметрические схемы восстановления регрессии, рассмотренные в главах IV и V, удовлетворяют условию (7.15), причем τ заключено в узких пределах $1{,}35 < \tau < 2{,}45$ (см. § 3 гл. II).

Ниже мы покажем, что если наряду с (7.15) выполнится одно из следующих трех условий:

1) множество $F(x, \alpha)$ состоит из конечного числа элементов,

2) множество $F(x, \alpha)$ может быть покрыто конечной ε-сетью,

3) множество функций $F(x, \alpha)$ имеет конечную емкость,

то метод минимизации эмпирического риска позволяет решать задачи восстановления зависимостей.

---

[1]) Средним p-го порядка случайной величины ξ называется величина $\sqrt[p]{M\xi^p}$.

Итак, оценим скорость равномерной сходимости средних к математическим ожиданиям в условиях, когда выполняется (7.15), а класс функций имеет ограниченную емкостную характеристику в любом из рассмотренных определений.

§ 6. Две теоремы о равномерной сходимости

В этом параграфе мы докажем две из трех основных теорем, оценивающих скорость равномерной сходимости средних к математическим ожиданиям. Мы рассмотрим случай, когда множество функций $F(x, \alpha)$ состоит из конечного числа элементов, и случай, когда множество функций может быть покрыто конечной ε-сетью в метрике C или L_p^2.

При доказательстве обеих теорем будет существенно использован следующий факт: пусть некоторая функция $F(x, \alpha^*)$ такова, что для нее выполняется условие

$$\frac{\sqrt[p]{\int (y - F(x, \alpha^*))^{2p} P(x, y)\, dx\, dy}}{\int (y - F(x, \alpha^*))^2 P(x, y)\, dx\, dy} \leqslant \tau, \quad p \geqslant 2. \tag{7.16}$$

Тогда, если ограничение (7.16) задается для моментов $p > 2$, то справедливо неравенство

$$P\left\{\frac{I(\alpha^*) - I_э(\alpha^*)}{I(\alpha^*)} > \tau a(p) \varkappa\right\} < 24le^{-\frac{\varkappa^2 l}{4}}, \tag{7.17}$$

где

$$a(p) = \sqrt[p]{\frac{(p-1)^{p-1}}{(p-2)^{p-1} 2}}; \tag{7.18}$$

Если ограничение (7.16) задается для $p = 2$, то справедливо неравенство

$$P\left\{\frac{I(\alpha^*) - I_э(\alpha^*)}{I(\alpha^*)} > \tau\varkappa \sqrt{1 - \frac{1}{2}\ln \varkappa}\right\} < 24le^{-\frac{\varkappa^2 l}{4}}. \tag{7.19}$$

Заметим, что при $p > 3$ значение $a(p)$ в (7.17) близко к единице. Большое значение величина $a(p)$ принимает лишь, когда p близко к 2.

Эти оценки будут получены как следствие из теоремы 7.6, приведенной в § 7.

Теорема 7.4. *Пусть выполнено условие* (7.15), *а класс функций* $F(x, \alpha)$ *состоит из конечного числа* N *элементов.*

Тогда, если в ограничении (7.15) $p>2$, *то с вероятностью* $1-\eta$ *одновременно для всех функций из класса* $F(x, \alpha)$ *выполнятся неравенства*

$$I(\alpha)<\left[\frac{I_{\text{э}}(\alpha)}{1-2\tau a(p)\sqrt{\dfrac{\ln N+\ln l-\ln \eta/24}{l}}}\right]_{\infty}, \quad (7.20)$$

если же $p=2$, *то с вероятностью* $1-\eta$ *одновременно для всех функций* $F(x, \alpha)$ *выполнятся неравенства*

$$I(\alpha)<\left[\frac{I_{\text{э}}(\alpha)}{1-2\tau\sqrt{\dfrac{\ln N+\ln l-\ln \eta/24}{l}\left(1-\dfrac{1}{4}\ln 4\left(\dfrac{\ln N+\ln l-\ln \eta/24}{l}\right)\right)}}\right]_{\infty}, \quad (7.21)$$

где

$$[z]_{\infty}=\begin{cases} z, & \text{если } z\geqslant 0,\\ \infty, & \text{если } z<0.\end{cases}$$

Доказательство. Пусть в условии (7.15) $p>2$. Воспользуемся неравенством

$$P\left\{\sup_i \frac{I(\alpha_i)-I_{\text{э}}(\alpha_i)}{I(\alpha_i)}>\varkappa\tau a(p)\right\}<$$
$$<N\sup_i P\left\{\frac{I(\alpha_i)-I_{\text{э}}(\alpha_i)}{I(\alpha_i)}>\varkappa\tau a(p)\right\}. \quad (7.22)$$

Оценим второй сомножитель правой части неравенства (7.22) с помощью (7.17). Получаем оценку

$$P\left\{\sup_i \frac{I(\alpha_i)-I_{\text{э}}(\alpha_i)}{I(\alpha_i)}>\tau\varkappa a(p)\right\}<24Nle^{-\frac{\varkappa^2 l}{4}},$$

которая может быть записана в виде следующего эквивалентного утверждения: с вероятностью $1-\eta$ одновременно для всех α_i справедливы неравенства

$$I(\alpha_i)<\left[\frac{I_{\text{э}}(\alpha_i)}{1-2\tau a(p)\sqrt{\dfrac{\ln N+\ln l-\ln \eta/24}{l}}}\right]_{\infty}.$$

Первое утверждение теоремы доказано.

Аналогично для случая $p=2$ следует воспользоваться оценкой (7.19), используя которую в правой части нера-

венства (7.22) получим оценку скорости равномерной сходимости, которая в эквивалентной форме и является утверждением теоремы.

Теорема 7.5. *Пусть выполнено условие* (7.15), *и пусть множество* $F(x, \alpha)$ *может быть покрыто конечной* ε*-сетью. Тогда с вероятностью* $1-\eta$ *можно утверждать, что качество функции* $F(x, \alpha_э)$, *доставляющей минимум эмпирическому риску, оценивается величиной*

$$I(\alpha_э) < \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_э(\alpha_l(\alpha_э))}{1-T(\varepsilon)}\right]_\infty}\right)^2,$$

где обозначено: $F(x, \alpha_l(\alpha_э))$ *— ближайший к* $F(x, \alpha_э)$ *элемент* ε*-сети,*

$$T(\varepsilon) = 2\tau a(p)\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/24}{l}}, \quad \text{если} \quad p > 2;$$

$$T(\varepsilon) = 2\tau\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/24}{l}\left(1 - \frac{1}{4}\ln 4\left(\frac{\ln N(\varepsilon) + \ln l - \ln \eta/24}{l}\right)\right)},$$

$$\text{если} \quad p = 2.$$

Замечание. Теорема 7.5 справедлива для любого ε, задающего ε-сеть, выбранного априорно, т. е. до того, как реализовалась случайная выборка.

В частности, ε может быть выбрано из условия минимума выражения

$$\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{c}{1-T(\varkappa)}\right]_\infty},$$

где c — некоторая константа. Разумно в качестве c выбирать величину, близкую к минимуму функционала $I(\alpha_0)$. Априорная информация о величине $I(\alpha_0)$ используется, таким образом, для выбора подходящей величины ε.

Доказательство этой теоремы, по существу, повторяет доказательство теоремы 7.2.

1°. Выберем произвольную ε-сеть. При $p > 2$, согласно теореме 7.4, с вероятностью $1-\eta$ одновременно для всех элементов ε-сети выполнятся неравенства

$$I(\alpha_i) < \left[\frac{I_э(\alpha_i)}{1 - 2\tau a(p)\sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/24}{l}}}\right]_\infty. \qquad (7.23)$$

2°. Согласно оценке (7.11), полученной при доказательстве теоремы 7.2, значения функционалов $I(\alpha)$ для функций $F(x, \alpha_э)$ и $F(x, \alpha_i(\alpha_э))$, отстоящих друг от друга в метрике C или L_P^2 на величину, меньшую ε, уклонятся на величину, не превосходящую

$$|I(\alpha_э) - I(\alpha_i(\alpha_э))| < 2\varepsilon\sqrt{\max(I(\alpha_э), I(\alpha_i(\alpha_э)))}. \quad (7.24)$$

3°. Будем различать два случая: случай, когда $I(\alpha_э) > I(\alpha_i(\alpha_э))$, и случай, когда $I(\alpha_э) \leqslant I(\alpha_i(\alpha_э))$.

В первом случае из (7.23) и (7.24) следует, что с вероятностью $1-\eta$ справедлива оценка

$$I(\alpha_э) < \left[\frac{I_э(\alpha_i(\alpha_э))}{1-2\tau a(p)\sqrt{\frac{\ln N(\varepsilon)+\ln l-\ln \eta/24}{l}}}\right]_\infty + 2\varepsilon\sqrt{I(\alpha_э)}. \quad (7.25)$$

Во втором случае с вероятностью $1-\eta$ оценка

$$I(\alpha_э) < \left[\frac{I_э(\alpha_i(\alpha_э))}{1-2\tau a(p)\sqrt{\frac{\ln N(\varepsilon)+\ln l-\ln \eta/24}{l}}}\right]_\infty + 2\varepsilon\sqrt{I(\alpha_i(\alpha_э))}. \quad (7.25а)$$

4°. Разрешим неравенство (7.25) относительно $I(\alpha_э)$:

$$I(\alpha_э) < \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_э(\alpha_i(\alpha_э))}{1-2\tau a(p)\sqrt{\frac{\ln N(\varepsilon)+\ln l-\ln \eta/24}{l}}}\right]_\infty}\right)^2. \quad (7.26)$$

Учитывая (7.23), убеждаемся, что оценка (7.26) справедлива и в случае (7.25а).

Аналогично доказывается теорема и для $p=2$.

Замечание. Так же как и в теореме 7.2, оценка (7.26) будет меньше (меньше величина $N(\varepsilon)$), если ε-сеть строится в метрике L_P^2, т. е. в случае, когда используется информация о плотности $P(x)$.

§ 7. Теорема о равномерном относительном уклонении

Докажем третью основную теорему.

Теорема 7.6. *Пусть выполнено условие* (7.15), *а множество функций* $F(x, \alpha)$ *имеет конечную емкость* $h < l$; *тогда, если* $p > 2$, *то с вероятностью* $1-\eta$ *одновременно*

для всех функций $F(x, \alpha)$ *окажутся выполненными неравенства*

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1 - 2\tau a(p) \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \frac{\ln \eta}{12}}{l}}} \right]_\infty,$$

$$a(p) = \sqrt[p]{\left(\frac{p-1}{p-2}\right)^{p-1} \cdot \frac{1}{2}},$$

если же $p = 2$, *то с вероятностью* $1 - \eta$ *одновременно для всех функций* $F(x, \alpha)$ *выполнятся неравенства*

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1 - 2\tau \sqrt{\Gamma(h, l, \eta)\left(1 - \frac{1}{4} \ln \Gamma(h, l, \eta)\right)}} \right]_\infty,$$

где

$$\Gamma(h, l, \eta) = \frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}{l}.$$

Доказывать теорему мы будем сначала для случая, когда ограничение (7.15) задано с помощью константы $p > 2$, и затем для случая, когда константа $p = 2$.

Для доказательства выразим функционал $I(\alpha)$ через интеграл Лебега

$$I(\alpha) = \int_0^\infty P\{(y - F(x, \alpha))^2 > t\}\, dt. \qquad (7.27)$$

Заметим, что для всякого фиксированного α и любого t вероятность события $\{(y - F(x, \alpha))^2 > t\}$ выражается через функцию распределения вероятностей положительной случайной величины $\xi(\alpha) = (y - F(x, \alpha))^2$, а именно, функция распределения вероятностей случайной величины $\xi(\alpha)$:

$$\Phi(\xi(\alpha) \leqslant t) = \Phi_\alpha(t)$$

связана с вероятностью появления события $\{(y - F(x, \alpha))^2 > t\}$ соотношением

$$P\{(y - F(x, \alpha))^2 > t\} = 1 - \Phi_\alpha(t).$$

Поэтому запишем функционал (7.27) в виде

$$I(\alpha) = \int (1 - \Phi_\alpha(t))\, dt.$$

Введем в рассмотрение новый функционал

$$R(\alpha) = \int \sqrt{1 - \Phi_\alpha(t)}\, dt.$$

Легко видеть, что он больше, чем $I(\alpha)$, так как

$$1 - \Phi_\alpha(t) < \sqrt{1 - \Phi_\alpha(t)}.$$

Справедлива следующая

Лемма. *Если для каждой функции множества $F(x, \alpha)$ существует функционал $R(\alpha)$, а множество функций $F(x, \alpha)$ имеет конечную емкость $h < l$, то справедливо неравенство*

$$P\left\{\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} > \varkappa\right\} < 8n^S (2l)\, e^{-\frac{\varkappa^2 l}{4}} < 12 \frac{(2l)^h}{h!} e^{-\frac{\varkappa^2 l}{4}}. \quad (7.28)$$

Доказательство. Обозначим через $A_{\alpha,i}$ событие $\left\{(y - F(x, \alpha))^2 > \frac{i}{n}\right\}$. Рассмотрим выражение

$$\frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} = \frac{\lim\limits_{n\to\infty}\left[\sum\limits_{i=1}^{\infty} \frac{1}{n} P(A_{\alpha,i}) - \sum\limits_{i=1}^{\infty} \frac{1}{n} \nu(A_{\alpha,i})\right]}{R(\alpha)}. \quad (7.29)$$

Покажем, что если выполняется неравенство

$$\sup_{\alpha, i} \frac{P(A_{\alpha,i}) - \nu(A_{\alpha,i})}{\sqrt{P(A_{\alpha,i})}} \leqslant \varkappa, \quad (7.30)$$

то выполнится и неравенство

$$\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} \leqslant \varkappa.$$

Действительно, из (7.29) и (7.30) следует

$$\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} \leqslant \sup_\alpha \frac{\lim\limits_{n\to\infty} \varkappa \sum\limits_{i=1}^{\infty} \frac{1}{n} \sqrt{P(A_{\alpha,i})}}{R(\alpha)} = \sup_\alpha \frac{\varkappa R(\alpha)}{R(\alpha)} = \varkappa.$$

Таким образом, вероятность выполнения неравенства

$$\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} > \varkappa$$

не превышает вероятность выполнения неравенства

$$\sup_{\alpha, i} \frac{P(A_{\alpha, i}) - \nu(A_{\alpha, i})}{\sqrt{P(A_{\alpha, i})}} > \varkappa.$$

В свою очередь, согласно теореме П.3 приложения к главе VI, справедлива оценка

$$P\left\{\sup_{\alpha, i} \frac{P(A_{\alpha, i}) - \nu(A_{\alpha, i})}{\sqrt{P(A_{\alpha, i})}} > \varkappa\right\} < 8m^S(2l)e^{-\frac{\varkappa^2 l}{4}},$$

откуда следует, что

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} > \varkappa\right\} < 8m^S(2l)\, e^{-\frac{\varkappa^2 l}{4}}. \qquad (7.31)$$

Учитывая, что $m^S(l) < 1{,}5\,\frac{l^h}{h!}$, получим оценку (7.28). Лемма доказана.

Доказательство теоремы. В формулировке леммы содержалось требование: для любой функции $F(x, \alpha)$ существует функционал $R(\alpha)$. Покажем теперь, что функционал $R(\alpha)$ существует, если существует момент порядка выше второго (хотя бы и не целого) для случайной величины

$$\xi(\alpha) = (y - F(x, \alpha))^2.$$

Более того, для $p > 2$ справедливо соотношение

$$R(\alpha) < \sqrt[p]{M\xi^p(\alpha)} \cdot a(p), \quad a(p) = \sqrt[p]{\frac{(p-1)^{p-1}}{2(p-2)^{p-1}}}.$$

В самом деле, справедливо преобразование

$$M\xi^p(\alpha) = \int (y - F(x, \alpha))^{2p} P(x, y)\, dx\, dy =$$

$$= \int_0^\infty t^p\, d\Phi_\alpha(t) = p\int_0^\infty t^{p-1}(1 - \Phi_\alpha(t))\, dt.$$

С другой стороны, согласно определению,

$$R(\alpha) = \int_0^\infty \sqrt{1 - \Phi_\alpha(t)}\, dt.$$

Пусть теперь p-й момент равен величине $m_p(\alpha)$:

$$p\int_0^\infty t^{p-1}(1 - \Phi_\alpha(t))\, dt = m_p(\alpha).$$

Найдем такое распределение $\Phi_\alpha(t)$, при котором максимизируется $R(\alpha)$.

Для этого составим функцию Лагранжа

$$L(\alpha) = R(\alpha) - \lambda M\xi^p(\alpha) = \int_0^\infty \sqrt{1-\Phi_\alpha(t)}\,dt - \\ - \lambda p \int_0^\infty t^{p-1}(1-\Phi_\alpha(t))\,dt. \quad (7.32)$$

Определим такую функцию распределения вероятностей $\Phi_\alpha(t)$, на которой достигается максимум выражения $L(\alpha)$. Обозначим $z^2 = 1-\Phi_\alpha(t)$, $b = \lambda p$ и перепишем (7.32) в этих обозначениях:

$$L(\alpha) = \int_0^\infty z(1 - bzt^{p-1})\,dt. \quad (7.33)$$

Функция z, на которой достигается максимум функционала (7.33), задается условием

$$1 - 2bzt^{p-1} = 0,$$

откуда вытекает, что

$$z = \left(\frac{t_0}{t}\right)^{p-1},$$

где обозначено $t_0 = \left(\frac{1}{2b}\right)^{\frac{1}{p-1}}$.

Учитывая, что при изменении t в пределах от нуля до бесконечности функция $z(t)$ должна меняться от единицы до нуля, оптимальной будет функция

$$z(t) = \begin{cases} 1, & \text{если } t < t_0, \\ \left(\frac{t_0}{t}\right)^{p-1}, & \text{если } t \geqslant t_0. \end{cases}$$

Вычислим теперь величину $\max\limits_\alpha R(\alpha)$, учитывая, что $p > 2$:

$$\max_\alpha R(\alpha) = \int_0^\infty z(t)\,dt = t_0 + \int_0^\infty \left(\frac{t_0}{t}\right)^{p-1} dt = \frac{p-1}{p-2}\,t_0. \quad (7.34)$$

С другой стороны, выразим через t_0 константу m_p:

$$m_p = p\int_0^\infty z^2(t)\,t^{p-1}\,dt =$$

$$= p\int_0^{t_0} t^{p-1}\,dt + p\int_{t_0}^\infty \left(\frac{t_0}{t}\right)^{2p-2} t^{p-1}\,dt = 2t_0^p\left(\frac{p-1}{p-2}\right). \quad (7.35)$$

Подставляя значение t_0, найденное из (7.35), в (7.34), получим

$$\sup_\alpha \frac{R(\alpha)}{\sqrt[p]{m_p(\alpha)}} = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}} = a(p),$$

откуда следует, что если $p > 2$, то

$$R(\alpha) < \sqrt[p]{M\xi^p(\alpha)}\,a(p). \quad (7.36)$$

Используя лемму и оценку (7.36), докажем первую часть теоремы. Для этого заметим, что в условиях теоремы справедливо неравенство

$$R(\alpha) < \tau a(p)\,I(\alpha). \quad (7.37)$$

Воспользуемся оценкой (7.37) для того, чтобы усилить неравенство (7.28)

$$P\left\{\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{I(\alpha)} > \tau a(p)\,\varkappa\right\} <$$

$$< P\left\{\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} > \varkappa\right\} < 12\,\frac{(2l)^h}{h!}\,e^{-\frac{\varkappa^2 l}{4}}. \quad (7.38)$$

Первое утверждение теоремы и есть эквивалентная запись этого неравенства.

Докажем теперь вторую часть теоремы. Для этого опять рассмотрим разность

$$I(\alpha) - I_э(\alpha) = \lim_{n\to\infty} \sum_{i=1}^{\infty} \frac{1}{n}\left(P(A_{\alpha,i}) - \nu(A_{\alpha,i})\right). \quad (7.39)$$

Допустим, что для всех событий $A_{\alpha,i}$ выполняется условие

$$P(A_{\alpha,i}) - \nu(A_{\alpha,i}) \leqslant \varkappa\sqrt{P(A_{\alpha,i})}. \quad (7.40)$$

Кроме того, всегда справедливо

$$P(A_{\alpha,i}) - \nu(A_{\alpha,i}) \leqslant P(A_{\alpha,i}). \quad (7.41)$$

При вычислении суммы (7.39) воспользуемся оценкой (7.40) для тех слагаемых, для которых события $A_{\alpha,i}$ имеют вероятность, большую $\varkappa^2$, т. е. $P(A_{\alpha,i}) > \varkappa^2$. Для слагаемых, соответствующих событиям $A_{\alpha,i}$ с вероятностью $P(A_{\alpha,i}) \leqslant \varkappa^2$, будем пользоваться тривиальной оценкой (7.41). В результате получим

$$I(\alpha) - I_{\text{э}}(\alpha) \leqslant$$
$$\leqslant \varkappa \int\limits_{1-\Phi_\alpha(t)>\varkappa^2} \sqrt{1-\Phi_\alpha(t)}\, dt + \int\limits_{1-\Phi_\alpha(t)\leqslant\varkappa^2} (1-\Phi_\alpha(t))\, dt. \tag{7.42}$$

Найдем теперь максимальное значение (по $\Phi_\alpha(t)$) правой части неравенства при условии, что второй момент принимает некоторое фиксированное значение m_2, т. е.

$$2\int\limits_0^\infty t(1-\Phi_\alpha(t))^2\, dt = m_2.$$

Для этого воспользуемся методом множителей Лагранжа, обозначив

$$z^2 = 1 - \Phi_\alpha(t).$$

Таким образом, будем искать максимум выражения

$$L(\alpha) = \int\limits_{z>\varkappa} \varkappa z\, dt + \int\limits_{z\leqslant\varkappa} z^2\, dt - \lambda\int\limits_0^\infty tz^2\, dt.$$

Представим $L(\alpha)$ в виде

$$L(\alpha) = \int\limits_{z>\varkappa} (\varkappa z - \lambda t z^2)\, dt + \int\limits_{z\leqslant\varkappa} (z^2 - \lambda t z^2)\, dt,$$

в котором первое слагаемое определяет функцию $z(t)$ в области, где $z > \varkappa$, а второе слагаемое — в области, где $z < \varkappa$.

Первое слагаемое достигает абсолютного максимума, когда

$$z = \frac{\varkappa}{2\lambda t}.$$

Однако, учитывая, что z — монотонно убывающая от 1 до $\varkappa$ функция, получаем

$$z(t) = \begin{cases} 1, & \text{если} \quad 0 \leqslant t < \dfrac{\varkappa}{2\lambda}, \\ \dfrac{\varkappa}{2\lambda t}, & \text{если} \quad \dfrac{\varkappa}{2\lambda} \leqslant t < \dfrac{1}{2\lambda}. \end{cases}$$

Второе слагаемое достигает максимума в области $z \leqslant \varkappa$ на функции

$$z(t) = \begin{cases} \varkappa, & \text{если } \frac{1}{2\lambda} \leqslant t < \frac{1}{\lambda}, \\ 0, & \text{если } t \geqslant \frac{1}{\lambda}. \end{cases}$$

Таким образом, окончательно получаем

$$z(t) = \begin{cases} 1, & \text{если } 0 \leqslant t < \frac{\varkappa}{2\lambda}, \\ \frac{\varkappa}{2\lambda t}, & \text{если } \frac{\varkappa}{2\lambda} \leqslant t < \frac{1}{2\lambda}, \\ \varkappa, & \text{если } \frac{1}{2\lambda} \leqslant t < \frac{1}{\lambda}, \\ 0, & \text{если } \frac{1}{\lambda} \leqslant t < \infty. \end{cases}$$

Выразим теперь величину m_2 второго момента через множитель Лагранжа λ. Для этого вычислим величину второго момента

$$m_2 = 2\int_0^\infty tz^2\,dt = \frac{\varkappa^2}{\lambda^2}\left(1 - \frac{1}{2}\ln\varkappa\right). \tag{7.43}$$

Аналогично вычислим величину

$$I(\alpha) - I_э(\alpha) = \varkappa\int_0^{1/2\lambda} z\,dt + \int_{1/2\lambda}^\infty z^2\,dt = \frac{\varkappa^2}{\lambda}\left(1 - \frac{1}{2}\ln\varkappa\right). \tag{7.43а}$$

Из (7.43) и (7.43а) заключаем, что

$$\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{\sqrt{m_2(\alpha)}} < \varkappa\sqrt{1 - \frac{1}{2}\ln\varkappa}. \tag{7.44}$$

Таким образом, мы показали, что условие (7.40) влечет за собой неравенство (7.44). Поэтому вероятность события

$$\left\{\sup_\alpha \frac{I(\alpha) - I_э(\alpha)}{\sqrt{m_2(\alpha)}} > \varkappa\sqrt{1 - \frac{1}{2}\ln\varkappa}\right\}$$

не превосходит вероятности события

$$\left\{\sup_{\alpha, i} \frac{P(A_{\alpha, i}) - \nu(A_{\alpha, i})}{\sqrt{P(A_{\alpha, i})}} > \varkappa\right\}.$$

Согласно утверждению теоремы П.3 приложения к главе VI вероятность этого события при $l > h$ оценивается неравенством (6.45), откуда следует

$$P\left\{\sup_{\alpha}\frac{I(\alpha)-I_э(\alpha)}{\sqrt{m_2(\alpha)}} > \varkappa\sqrt{1-\frac{1}{2}\ln\varkappa}\right\} < 8m^S(2l)e^{-\frac{\varkappa^2 l}{4}} < 12\frac{(2l)^h}{h!}e^{-\frac{\varkappa^2 l}{4}}.$$

С другой стороны, по условию теоремы

$$\sqrt{m_2(\alpha)} = \sqrt{M\xi^2(\alpha)} \leqslant \tau I(\alpha).$$

Учитывая это, получаем

$$P\left\{\sup_{\alpha}\frac{I(\alpha)-I_э(\alpha)}{I(\alpha)} > \tau\varkappa\sqrt{1-\frac{1}{2}\ln\varkappa}\right\} \leqslant P\left\{\sup_{\alpha}\frac{I(\alpha)-I_э(\alpha)}{I(\alpha)} > \varkappa\tau\sqrt{1-\frac{1}{2}\ln\varkappa}\right\}.$$

Таким образом, окончательно получаем, что при $l > h$ справедливо

$$P\left\{\sup_{\alpha}\frac{I(\alpha)-I_э(\alpha)}{I(\alpha)} > \tau\varkappa\sqrt{1-\frac{1}{2}\ln\varkappa}\right\} < 8m^S(2l)e^{-\frac{\varkappa^2 l}{4}} < 12\frac{(2l)^h}{h!}e^{-\frac{\varkappa^2 l}{4}}. \quad (7.45)$$

Эквивалентная запись неравенства (7.45) и является утверждением второй части теоремы.

Замечание. При доказательстве теорем 7.4 и 7.5 мы пользовались оценками относительных уклонений (7.17) и (7.19). Эти оценки легко могут быть получены из неравенств (7.38), (7.45), если учесть, что емкость класса решающих правил $F(x, \alpha)$, образованного фиксированной функцией $F(x, \alpha^*)$, равна единице.

§ 8. Замечания о теории равномерной сходимости

Итак, мы построили теорию равномерной сходимости средних к их математическим ожиданиям. Формально теория строилась для квадратичной функции потерь.

Однако полученные результаты справедливы и для функций потерь общей природы.

Ниже мы сформулируем основные утверждения теории равномерного уклонения эмпирических оценок от средних в общей постановке.

Доказательство этих утверждений тождественно доказательствам аналогичных теорем, рассмотренных выше:

Пусть $Q(z, \alpha) > 0$ — параметрическое семейство положительных функций, удовлетворяющих следующим условиям:

1) при любом фиксированном значении параметра $\alpha^* \in \Lambda$ функции $Q(z, \alpha)$ измеримы по z;

2) множество функций $Q(z, \alpha)$ имеет конечную емкость h (характеристические функции $\theta(Q(z, \alpha) + \beta)$ имеют емкость h).

Тогда справедливы следующие утверждения о скорости равномерной сходимости эмпирических средних:

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha),$$

построенных по выборке $z_1, \ldots, z_l$ к их математическим ожиданиям:

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz.$$

Утверждение 1. *Если для функций $Q(z, \alpha)$ существует функционал*

$$R(\alpha) = \int \sqrt{1 - P\{Q(z, \alpha) \leqslant t\}}\, dt,$$

то при $l > h$ справедливо неравенство

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_э(\alpha)}{R(\alpha)} > \varkappa\right\} < 12 \frac{(2l)^h}{h!} e^{-\frac{\varkappa^2 l}{4}}. \qquad (7.46)$$

Утверждение 2. *Если для функций $Q(z, \alpha)$ существует второй момент*

$$m_2(\alpha) = \int Q^2(z, \alpha) P(z)\, dz,$$

то справедливо неравенство

$$P\left\{\sup_{\alpha} \frac{I(\alpha) - I_э(\alpha)}{\sqrt{m_2(\alpha)}} > \varkappa \sqrt{1 - \frac{1}{2} \ln \varkappa}\right\} < 12 \frac{(2l)^h}{h!} e^{-\frac{\varkappa^2 l}{4}}.$$

Утверждение 3. *Если для функций $Q(z, \alpha)$ существует p-й момент $(p > 2)$*

$$m_p(\alpha) = \int Q^p(z, \alpha) P(z)\, dz,$$

то при $l > h$ *справедливо неравенство*

$$P\left\{\sup_{\alpha}\frac{I(\alpha)-I_{\text{э}}(\alpha)}{\sqrt[p]{m_p(\alpha)}} > a(p)\,\varkappa\right\} < 12\,\frac{(2l)^h}{h!}\,e^{-\frac{\varkappa^2 l}{4}},$$

где

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}.$$

Утверждение 4. *Если окажется выполненным условие*

$$\sup_{\alpha}\frac{\sqrt[p]{m_p(\alpha)}}{I(\alpha)} \leqslant \tau \qquad (p>2),$$

то при $l > h$ *с вероятностью* $1-\eta$ *одновременно для всех* α *будет выполнено неравенство*

$$I(\alpha) < \left[\frac{I_{\text{э}}(\alpha)}{1-2\tau a(p)\dfrac{\sqrt{h\left(\ln\dfrac{2l}{h}+1\right)-\ln\dfrac{\eta}{12}}}{l}}\right]_{\infty}. \qquad (7.47)$$

Если же выполнится условие

$$\sup_{\alpha}\frac{\sqrt{m_2(\alpha)}}{I(\alpha)} \leqslant \tau,$$

то при $l > h$ *с вероятностью* $1-\eta$ *одновременно для всех* α *будет выполнено неравенство*

$$I(\alpha) < \left[\frac{I_{\text{э}}(\alpha)}{1-2\tau\sqrt{\Gamma(h,\,l,\,\eta)\left(1-\frac{1}{4}\ln\Gamma(h,\,l\,\eta)\right)}}\right]_{\infty}, \qquad (7.48)$$

где

$$\Gamma(h,\,l,\,\eta) = \frac{h\left(\ln\dfrac{2l}{h}+1\right)-\ln\dfrac{\eta}{12}}{l}.$$

В главах VIII и IX мы используем построенную теорию равномерной сходимости для конструирования экстремальных алгоритмов восстановления зависимостей в условиях ограниченной выборки, а пока лишь отметим, что если только выполнено условие (7.15) и емкость класса функций $F(x,\,\alpha)$ ограничена, то в соответствии с построенной теорией метод минимизации эмпирического риска при достаточном объеме выборки приводит к отысканию

функции, близкой к наилучшей в классе. Действительно, в этом случае знаменатель в оценках (7.47), (7.48) близок к единице, и величину среднего риска определяет величина эмпирического риска.

Основные утверждения главы VII

1. *Успешное решение задачи восстановления регрессии методом минимизации эмпирического риска может быть гарантировано в случае равномерного (или равномерного относительного) уклонения средних от их математических ожиданий.*

2. *Равномерное (или равномерное относительное) уклонение может иметь место, если существует ограничение на возможные величины потерь. Такими ограничениями являются:*

ограничение на величину потерь

$$\sup_{x,\,y,\,\alpha} (y - F(x,\,\alpha))^2 \leqslant \tau,$$

ограничение на отношение моментов

$$\sup_{\alpha} \frac{\sqrt[p]{M\,(y - F(x,\,\alpha))^{2p}}}{M\,(y - F(x,\,\alpha))^2} \leqslant \tau, \quad p \geqslant 2.$$

3. *При выполнении соответствующих ограничений и достаточном объеме выборки метод минимизации эмпирического риска приводит к успеху, если емкостная характеристика класса функций ограничена в одном из следующих смыслов:*

— множество функций состоит из конечного числа элементов,

— множество функций может быть покрыто конечной ε-сетью,

— множество функций $F(x, \alpha)$ имеет конечную емкость.

ГЛАВА VIII

МЕТОД УПОРЯДОЧЕННОЙ МИНИМИЗАЦИИ РИСКА В ЗАДАЧАХ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ

§ 1. Идея метода упорядоченной минимизации риска

До сих пор при исследовании методов восстановления зависимостей по эмпирическим данным количество данных играло второстепенную роль: принципы, которые определяли выбор искомой зависимости из заданного множества возможных зависимостей, непосредственно не учитывали объема имеющейся информации.

Начиная с этой главы, мы будем рассматривать методы восстановления, позволяющие для фиксированного объема эмпирических данных получать наилучший в определенном смысле результат. Здесь учет объема имеющейся информации, особенно в случае малой обучающей выборки

$$x_1, y_1; \ldots; x_l, y_l, \tag{8.1}$$

окажется существенным. Однако, прежде чем приступить к изложению методов восстановления зависимостей, рассчитанных на использование малой выборки, договоримся о том, какую выборку мы будем называть малой.

Определение. *Будем говорить, что для восстановления функции из заданного класса* $F(x, \alpha)$ *выборка объема* l *является малой, если отношение* l/h *мало* (*например*, $l/h < 30$), *где* h *— емкость класса функций.*

Величина l/h определяет относительный объем выборки (объем выборки на единицу емкости класса).

Заметим, что оценки величины среднего риска, полученные в главах VI и VII, зависели не от абсолютной величины объема выборки, а от относительной величины.

Основной результат главы VI состоял в том, что с вероятностью $1-\eta$ одновременно для всех характеристических функций $F(x, \alpha)$ выполняются неравенства

$$P(\alpha) < \nu(\alpha) + \Omega_1\left(\frac{l}{h}, -\frac{\ln \eta}{h}\right), \tag{8.2}$$

а основной результат главы VII — в том, что с вероятностью $1-\eta$ одновременно для всего множества произвольных функций $F(x, \alpha)$ выполняются неравенства

$$I(\alpha) < I_э(\alpha)\,\Omega_2\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right). \tag{8.3}$$

Сейчас для нас не важны ни конкретный вид слагаемого $\Omega_1\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)$, ни конкретный вид сомножителя $\Omega_2\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)$. Важно другое — полученные оценки были таковы, что с ростом $\frac{l}{h}$ величина $\Omega_1\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)$ стремилась к нулю, а величина $\Omega_2\left(\frac{l}{h}, -\frac{\ln\eta}{h}\right)$ — к единице.

Это обстоятельство и позволило обосновать метод минимизации эмпирического риска для больших выборок. Для всякого δ найдется такая величина T, что как только $\frac{l}{h} > T$, с вероятностью $1-\eta$ одновременно для всего множества $F(x, \alpha)$ выполнятся неравенства

$$P(\alpha) < \nu(\alpha) + \delta,$$

если $F(x, \alpha)$ — характеристические функции, и неравенства

$$I(\alpha) < I_э(\alpha)(1+\delta),$$

если $F(x, \alpha)$ — произвольные функции.

Поэтому малая величина эмпирического риска гарантирует (с вероятностью $1-\eta$) малую величину среднего риска.

Однако если объем обучающей выборки мал, слагаемое $\Omega_1\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)$ может значительно отличаться от нуля, а сомножитель $\Omega_2\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)$ — от единицы. В этом случае функция, доставляющая малую величину эмпирическому риску, может не обеспечить малую величину среднему риску.

Для того чтобы на малых выборках можно было достичь глубокого гарантированного минимума, необходимо учитывать не только величину эмпирического риска $\nu(\alpha_э)$ (или $I_э(\alpha_э)$), но и величину слагаемого $\Omega_1\left(\frac{l}{h}, -\frac{\ln\eta}{h}\right)$ (сомножителя $\Omega_2\left(\frac{l}{h}, -\frac{\ln\eta}{h}\right)$).

В этой главе мы рассмотрим метод минимизации риска, который, в отличие от метода минимизации эмпирического риска, минимизирует верхнюю оценку риска (8.2) или (8.3)) не по одному слагаемому $\nu(\alpha)$ (сомножителю $I_э(\alpha)$), а сразу по обоим:

$$\nu(\alpha)+\Omega_1\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right) \quad \left(I_э(\alpha)\,\Omega_2\left(\frac{l}{h}, \frac{-\ln\eta}{h}\right)\right).$$

Реализует эту идею метод, который мы будем называть *методом упорядоченной минимизации риска.*

Пусть на множестве функций $F(x, \alpha)$ задана *структура*, т. е. выделено минимальное подмножество элементов S_1, затем подмножество S_2, содержащее S_1, и, наконец, подмножество S_q, совпадающее со всем множеством:

$$S_1 \subset S_2 \subset \ldots \subset S_q. \tag{8.4}$$

Упорядочение (8.4) на множестве $F(x, \alpha)$ задано априорно (до появления обучающей последовательности).

Пусть структура задана так, что емкость h_i подмножества функций S_i меньше емкости h_{i+1} подмножества S_{i+1}, т. е.

$$h_1 < h_2 < \ldots < h_q.$$

Для каждого подмножества S_i с вероятностью $1-\eta$ справедлива оценка

$$P(\alpha_э^i) < \nu(\alpha_э^i)+\Omega_1\left(\frac{l}{h_i}, \frac{-\ln\eta}{h_i}\right), \tag{8.5}$$

если упорядочению подлежало множество характеристических функций, и с вероятностью $1-\eta$ справедлива оценка

$$I(\alpha_э^i) < I_э(\alpha_э^i)\,\Omega_2\left(\frac{l}{h_i}, \frac{-\ln\eta}{h_i}\right), \tag{8.6}$$

если упорядочивалось множество произвольных функций; $F(x, \alpha_э^i)$ — элемент, доставляющий минимум эмпирическому риску в S_i.

В выражении (8.5) ((8.6)) величина первого слагаемого (сомножителя) правой части падает с ростом i, а величина второго слагаемого (сомножителя) растет.

Метод упорядоченной минимизации риска состоит в том, чтобы найти подмножество S_*, в котором функция $F(x, \alpha_э^*)$, минимизирующая эмпирический риск, доставит минималь-

ную оценку среднему риску, и принять эту функцию за решение.

Заметим, что так как для каждого элемента S_i справедлива оценка (8.5) (оценка (8.6)), а всего в структуре q элементов, то с вероятностью $1-q\eta$ оценки справедливы одновременно для всех q функций, минимизирующих эмпирический риск (каждая в своем S_i). Поэтому найденное с помощью метода упорядоченной минимизации риска решение $F(x, \alpha_э^*)$ доставит гарантированную с вероятностью $1-q\eta$ минимальную оценку риску. Иначе говоря, неравенство

$$P(\alpha_э^*) < \nu(\alpha_э^*) + \Omega_1\left(\frac{l}{h_*}, \frac{-\ln\eta}{h_*}\right) \tag{8.7}$$

(неравенство

$$I(\alpha_э^*) < I_э(\alpha_э^*)\,\Omega_2\left(\frac{l}{h_*}, -\frac{\ln\eta}{h_*}\right)\Big) \tag{8.8}$$

будет справедливо с вероятностью $1-q\eta$.

При реализации метода упорядоченной минимизации нам будет важно, чтобы гарантия полученной оценки риска была велика (равнялась $1-\eta$). Поэтому, положив $\eta^* = \eta/q$ в (8.7) (и в (8.8)) вместо η, получим, что с вероятностью $1-\eta$ имеет место неравенство

$$P(\alpha_э^*) < \nu(\alpha_э^*) + \Omega_1\left(\frac{l}{h_*}, \frac{-\ln\eta+\ln q}{h_*}\right) \tag{8.9}$$

(неравенство

$$I(\alpha_э^*) < I_э(\alpha_э^*)\,\Omega_2\left(\frac{l}{h_*}, \frac{-\ln\eta+\ln q}{h_*}\right)\Big). \tag{8.10}$$

Для структуры, состоящей из небольшого числа элементов ($q < 20-100$), вообще говоря, полученное увеличение верхней оценки для $F(x, \alpha_э^*)$ (по отношению к (8.5) и (8.6)) незначительно. (Число q элементов структуры находится под знаком логарифма). Это означает, что при самом неблагоприятном стечении обстоятельств гарантированная величина риска для решения, полученного методом упорядоченной минимизации, может оказаться лишь на малую величину хуже гарантированной величины риска для решения, полученного методом минимизации эмпирического риска, в то время как в обычных

ситуациях, как мы увидим ниже, выигрыш от применения метода упорядоченной минимизации риска может оказаться значительным.

В дальнейшем нам удобно будет считать, что метод упорядоченной минимизации риска реализует двухуровневую процедуру минимизации: сначала на каждом элементе S_i структуры (8.4) выбирается функция $F(x, \alpha_э^i)$, минимизирующая величину эмпирического риска, а затем из q отобранных функций выбирается такая, которая доставляет величине риска гарантированный минимум.

Таким образом, при реализации метода упорядоченной минимизации риска возникают две проблемы:

1) Как задать структуру на исходном множестве функций $F(x, \alpha)$?

2) Каким должен быть алгоритм выбора второго уровня?

Задание структуры на множестве функций $F(x, \alpha)$ является неформальным моментом в реализации метода. В структуре должна быть отражена априорная информация о задачах, которая имеется у исследователя. Те функции, которые, по мнению исследователя, более вероятно приближают искомую, следует относить к классу S_i с меньшим номером. При этом чем больше имеется априорной информации, тем более узкими следует задавать классы с малыми номерами.

Задание алгоритма выбора второго уровня отражает умение оценивать качество каждого из отобранных на первом уровне решающих правил.

Ниже при построении алгоритмов выбора второго уровня мы используем оценку среднего риска (6.48)

$$P(\alpha) < \nu(\alpha) + 2\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}{l}\left(1+\sqrt{1+\frac{\nu(\alpha)}{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}}\right),$$

если выбор осуществляется среди характеристических функций (решается задача распознавания образов), и оценку (7.27)

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1-2\tau a(p)\sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{12}}{l}}}\right]_+^{\infty},$$

если выбор осуществляется среди функций произвольной природы (решается задача восстановления регрессии).

Использование этих оценок позволит получить наилучшее для заданной структуры гарантированное решение. Другая идея построения алгоритмов второго уровня связана с использованием процедуры «*скользящий контроль*».

§ 2. Оценка «скользящий контроль»

Оценим качество решающего правила $F(x, \alpha_э)$, минимизирующего на заданной последовательности

$$x_1, y_1; \ldots; x_l, y_l,$$

эмпирический риск

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 \qquad (8.11)$$

с помощью следующего приема. Исключим из обучающей последовательности первую пару x_1, y_1 и найдем функцию, минимизирующую эмпирический риск на оставшихся $l-1$ элементах обучающей последовательности. Пусть эта функция есть $F(x; \alpha(\widehat{x_1, y_1}; \ldots; x_l, y_l))$.

Здесь знак $\widehat{x_1, y_1}$ указывает на то, что из обучающей последовательности была исключена пара x_1, y_1. Вычислим величину уклонения на исключенной паре x_1, y_1:

$$(y_1 - F(x_1; \alpha(\widehat{x_1, y_1}; \ldots; x_l, y_l)))^2.$$

Затем из обучающей последовательности исключим вторую пару (первая пара остается в последовательности) и вычислим уклонение

$$(y_2 - F(x_2; \alpha(x_1, y_1; \widehat{x_2, y_2}; \ldots; x_l, y_l))^2;$$

Так вычислим уклонение для всех l пар. Образуем величину

$$T(x_1, y_1; \ldots; x_l, y_l) =$$

$$= \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i; \alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2 \qquad (8.12)$$

и примем ее за оценку качества функции $F(x, \alpha_э)$, минимизирующей эмпирический риск (8.11):

$$I(\alpha_э(x_1, y_1; \ldots; x_l, y_l)) = \int (y - F(x, \alpha_э))^2 P(x, y)\, dx\, dy \sim T(x_1, y_1; \ldots; x_l, y_l).$$

Такую процедуру оценивания назовем *«скользящий контроль»*.

Справедлива

Теорема 8.1. *Оценка «скользящий контроль» является несмещенной, т. е.*

$$MI(\alpha_э(x_1, y_1; \ldots; x_{l-1}, y_{l-1})) = MT(x_1, y_1; \ldots; x_l, y_l).$$

Доказательство. Доказательством этой теоремы служит следующая цепочка преобразований:

$$M \int \ldots \int \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha(x_1, y_1, \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2 \times$$
$$\times P(x_1, y_1) \ldots P(x_l, y_l)\, dx_1\, dy_1 \ldots dx_l\, dy_l =$$
$$= M \int \ldots \int \frac{1}{l} \sum_{i=1}^{l} \left[\int (y_i - F(x_i, \alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots \right.$$
$$\left. \ldots; x_l, y_l)))^2 P(x_i, y_i)\, dx_i\, dy_i\right] P(x_1, y_1) \ldots$$
$$\ldots P(x_{i-1}, y_{i-1}) P(x_{i+1}, y_{i+1}) \ldots P(x_l, y_l)\, dx_1\, dy_1 \ldots$$
$$\ldots dx_{i-1}\, dy_{i-1}\, dx_{i+1}\, dy_{i+1} \ldots dx_l\, dy_l =$$
$$= M \frac{1}{l} \sum_{i=1}^{l} I(\alpha(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)) =$$
$$= MI(\alpha_э(x_1, y_1; \ldots; x_{l-1}, y_{l-1})).$$

Теорема доказана.

Замечание. При доказательстве теоремы мы нигде не использовали свойства функции $F(x, \alpha)$. Поэтому процедура «скользящий контроль» определяет несмещенные оценки качества как при восстановлении характеристической функции, так и при восстановлении произвольной функциональной зависимости.

Свойство несмещенности, однако, недостаточно характеризует оценку. Необходимо знать еще и дисперсию D

оценки. Если бы дисперсия оценки T была известна, то можно было бы оценить качество решающего правила, минимизирующего эмпирический риск на выборках длины l. А именно: с вероятностью $1-\eta$ справедливо неравенство

$$MI(\alpha_{\text{э}}(x_1, y_1; \ldots; x_{l-1}, y_{l-1})) \leqslant \leqslant T(x_1, y_1; \ldots; x_l, y_l) + \sqrt{\frac{D}{\eta}}, \qquad (8.13)$$

где $1-\eta$ — надежность, с которой требуется выполнение неравенства (8.13). (Оценка (8.13) следует из теоремы 8.1 и неравенства Чебышева $P\{|MT - T| > \sqrt{D/\eta}\} < \eta$.)

Однако вычислить дисперсию оценки «скользящий контроль» в достаточно общей постановке не удается. Поэтому применение процедуры «скользящий контроль» для оценки качества алгоритмов, минимизирующих эмпирический риск, связано с гипотезой о том, что при объеме выборки, в несколько раз превосходящем величину емкости класса функции, дисперсия оценки мала (имеет порядок $1/l$, а не h/l) [1]).

§ 3. Оценка «скользящий контроль» в задаче восстановления регрессии

Покажем, что для восстановления регрессии в классе линейных по параметру функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$$

---

[1]) Для частного случая, когда $F(x, \alpha) = \sum_{i=1}^{n} \alpha_i x^i$, вектор $x = (x^1, \ldots, x^n)^{\tau}$ распределен по нормальному закону, $y = F(x, \alpha_0) + \xi$, где ξ нормально распределенная помеха, существует доказательство этого утверждения.

Гипотеза состоит в том, что такой же порядок величины дисперсии сохраняется и в общем случае: для задачи распознавания образов, когда класс $F(x, \alpha)$ имеет ограниченную емкость, а для задачи восстановления регрессии, когда класс $F(x, \alpha)$ имеет ограниченную емкость и выполняется неравенство

$$\sup_{\alpha} \frac{\sqrt{M(y - F(x, \alpha))^4}}{M(y - F(x, \alpha))^2} = \tau < \infty.$$

оценка «скользящий контроль» допускает следующее эквивалентное представление:

$$T(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{\left(y_i - f_i^{\mathrm{T}} (\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}} Y\right)^2}{\left(1 - f_i^{\mathrm{T}} (\Phi^{\mathrm{T}}\Phi)^{-1} f_i\right)^2}, \quad (8.14)$$

где обозначено

$$\Phi = \begin{Vmatrix} \varphi_1(x_1), & \ldots, & \varphi_n(x_1) \\ \cdots\cdots & \cdots & \cdots\cdots \\ \varphi_1(x_l), & \ldots, & \varphi_n(x_l) \end{Vmatrix},$$

f_i^{T} — i-я строка матрицы Φ, Y — l-мерный вектор-столбец значений y

$$Y = (y_1, \ldots, y_l)^{\mathrm{T}}.$$

Выражение $(\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}Y$ в числителе (8.14) есть оценка вектора параметров α, полученная методом наименьших квадратов по всей обучающей последовательности. Числитель в (8.14) определяет квадрат уклонения в точке x_i, а знаменатель определяет ту мультипликативную поправку, которая возникает, если оценку параметров α получать не по всей обучающей выборке, а по выборке, из которой исключена i-я пара x_i, y_i.

Представление (8.14) замечательно тем, что в нем используется лишь одно обращение матрицы, а не l, как в общей процедуре, описанной в предыдущем параграфе. Это обстоятельство делает процедуру «скользящий контроль» в вычислительном отношении не более сложной, чем вычисление невязки в методе наименьших квадратов.

Ниже при построении алгоритмов восстановления зависимостей мы будем искать решение, доставляющее не только безусловный минимум (8.11), но и условный минимум при ограничении на решение

$$\|\alpha\|^2 = \sum_{i=1}^{n} \alpha_i^2 < c.$$

Отыскание такого условного минимума — задача, эквивалентная поиску минимума функционала

$$I_{\text{э}}(\alpha : \gamma) = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha))^2 + \gamma \|\alpha\|^2, \quad (8.15)$$

где γ — положительная константа, зависящая от c (множитель Лагранжа).

Оценку качества решения $\alpha=\alpha_\gamma$, минимизирующего функционал (8.15), будем проводить также с помощью процедуры «скользящий контроль».

Найдем решения $\alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)$, минимизирующие функционал (8.15), заданный на $l-1$ паре (пара $\widehat{x_i, y_i}$ исключена, γ фиксировано), и образуем величину

$$T_\gamma(x_1, y_1; \ldots; x_l, y_l) =$$

$$= \frac{1}{l}\sum_{i=1}^{l}(y_i - F(x_i, \alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)))^2. \quad (8.16)$$

Величина T_γ и будет оценкой качества функции $F(x, \alpha_\gamma)$, минимизирующей функционал (8.15).

Эквивалентное представление (8.16) мы получим с помощью матрицы

$$A_\gamma = (\Phi^T\Phi + \gamma I), \quad I = \begin{vmatrix} 1 & & 0 \\ & 1 & \\ & \ddots & \\ 0 & & 1 \end{vmatrix}. \quad (8.17)$$

А именно,

$$T_\gamma(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l}\sum_{i=1}^{l}\frac{(y_i - f_i^T A_\gamma^{-1}\Phi^T Y)^2}{(1 - f_i^T A_\gamma^{-1} f_i)^2}. \quad (8.18)$$

При $\gamma=0$ выражение (8.18) совпадает с (8.14).

Получим представление оценки «скользящий контроль» (8.16) в виде (8.18). Обозначим

$$(A_\gamma - \| f_i \|^T \| f_i \|)^{-1} = B, \quad (8.19)$$

$\| f_i \|$ — матрица, все строки которой, кроме i-й, равны нулю. В i-й строке матрицы записан вектор f_i^T.

Тогда минимум (8.15) на обучающей последовательности без x_i, y_i достигается на векторе

$$\alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l) = B(\Phi^T - \| f_i \|^T) Y. \quad (8.20)$$

Выразим матрицу B через A_γ. Для этого перепишем (8.19) в виде

$$I = BA_\gamma - B \| f_i \|^T \| f_i \|. \quad (8.21)$$

В свою очередь из равенства (8.21) получим

$$B = A_\gamma^{-1} + B \| f_i \|^T \| f_i \| A_\gamma^{-1}. \quad (8.22)$$

Умножим левую и правую части равенства (8.22) на $\|f_i\|^{\mathrm{T}}$:

$$B\|f_i\|^{\mathrm{T}} = A_\gamma^{-1}\|f_i\|^{\mathrm{T}} + B\|f_i\|^{\mathrm{T}}\|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}}. \tag{8.23}$$

Из (8.23) получаем

$$B\|f_i\|^{\mathrm{T}} = A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}.$$

Подставим выражение $B\|f_i\|^{\mathrm{T}}$ в (8.22):

$$B = A_\gamma^{-1} + A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}. \tag{8.24}$$

Вычислим теперь $\alpha_\gamma = \alpha_\gamma(x_1, y_1; \ldots; \widehat{x_i, y_i}; \ldots; x_l, y_l)$. Согласно (8.24) имеем

$$\begin{aligned}\alpha_\gamma &= B(\Phi^{\mathrm{T}} - \|f_i\|^{\mathrm{T}})Y = \\ &= A_\gamma^{-1}\Phi^{\mathrm{T}}Y + A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}\Phi^{\mathrm{T}}Y - \\ &\quad - A_\gamma^{-1}\|f_i\|^{\mathrm{T}}Y - A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}}Y.\end{aligned}$$

Вычислим квадрат уклонения, используя равенство

$$f_i^{\mathrm{T}}A_\gamma^{-1}\|f_i\|^{\mathrm{T}}Y = f_i^{\mathrm{T}}A_\gamma^{-1}f_i y_i.$$

Получим

$$\begin{aligned}&(y_i - F(x_i, \alpha_\gamma))^2 = \\ &= \left(y_i - f_i^{\mathrm{T}}A_\gamma^{-1}\Phi^{\mathrm{T}}Y - f_i^{\mathrm{T}}A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}\|f_i\|A_\gamma^{-1}\Phi^{\mathrm{T}}Y + \right. \\ &\quad \left. + f_i^{\mathrm{T}}A_\gamma^{-1}f_i y_i + f_i^{\mathrm{T}}A_\gamma^{-1}\|f_i\|^{\mathrm{T}}(I - \|f_i\|A_\gamma^{-1}\|f_i\|^{\mathrm{T}})^{-1}f_i^{\mathrm{T}}A_\gamma^{-1}f_i y_i\right)^2 = \\ &= \left(\frac{y_i}{1 - f_i^{\mathrm{T}}A_\gamma^{-1}f_i} - \left(1 + \frac{f_i^{\mathrm{T}}A_\gamma^{-1}f_i}{1 - f_i^{\mathrm{T}}A_\gamma^{-1}f_i}\right)f_i^{\mathrm{T}}A_\gamma^{-1}\Phi^{\mathrm{T}}Y\right)^2 = \frac{(y_i - f_i^{\mathrm{T}}A_\gamma^{-1}\Phi^{\mathrm{T}}Y)^2}{(1 - f_i^{\mathrm{T}}A_\gamma^{-1}f_i)^2}.\end{aligned}$$

Таким образом, окончательно получаем

$$T(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l}\sum_{i=1}^{l}\left(\frac{y_i - f_i^{\mathrm{T}}A_\gamma^{-1}\Phi^{\mathrm{T}}Y}{1 - f_i^{\mathrm{T}}A_\gamma^{-1}f_i}\right)^2.$$

§ 4. Восстановление характеристической функции в классе линейных решающих правил

Итак, для метода упорядоченной минимизации риска мы определили критерии выбора второго уровня. Ими будут либо минимальные гарантированные оценки риска, либо минимальные оценки, полученные с помощью процедуры «скользящий контроль».

Теперь, для того чтобы задать алгоритмы упорядоченной минимизации риска, надо определить структуру на множестве функций $F(x, \alpha)$.

В этой главе мы рассмотрим некоторые способы задания структуры на множестве линейных решающих правил (в задаче распознавания образов) и на множестве линейных по параметру функций (в задаче восстановления регрессии) и построим соответствующие алгоритмы восстановления зависимостей.

Рассмотрим сначала задачу распознавания образов.

Пусть задан класс линейных решающих правил

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right).$$

Выстроим признаки $\varphi_i(x)$ в порядке уменьшения априорной вероятности того, что этот признак будет «полезен» при классификации, и определим следующую структуру линейных решающих правил:

$$S_1^1 \subset S_2^1 \subset \ldots \subset S_n^1. \tag{8.25}$$

Класс S_1^1 состоит из тех правил, у которых может отличаться от нуля лишь параметр α_1. Класс S_2^1 состоит из правил, у которых могут отличаться от нуля два параметра α_1 и α_2 и т. д. Такое упорядочение имеет следующий смысл. В первый класс попадают те правила, которые при распознавании используют лишь первый признак, во второй класс попадают правила, которые используют первые два признака, и т. д. Показатель емкости каждого из этих классов, как было установлено в главе VI, равен i, где i — число используемых признаков.

Для такой структуры метод упорядоченной минимизации риска состоит в том, чтобы найти решающее правило $F(x, \alpha_i^*)$, которое минимизирует по i и $F(x, \alpha) \in S_i^1$ функционал

$$R_1(\alpha, i) = \\ = 2\frac{i\left(\ln\frac{2l}{i}+1\right)-\ln\frac{\eta}{12}}{l}\left(1+\sqrt{1+\frac{\nu(\alpha)\,l}{i\left(\ln\frac{2l}{i}+1\right)-\ln\frac{\eta}{12}}}\right)+ \\ +\nu(\alpha). \tag{8.26}$$

С достоверностью $1-n\eta$ вероятность ошибочной классификации с помощью найденного решающего правила не

превосходит достигнутого минимума (8.26), т. е.

$$P\{P(\alpha_э^*) < R_1(\alpha_э^*,\ S_*^1)\} > 1 - n\eta \qquad (n < l).$$

Рассмотренный способ задания структуры на классе линейных решающих правил требует априорной ранжировки признаков. Это не всегда просто сделать.

Поэтому мы определим еще одну структуру, для задания которой априорная ранжировка признаков не нужна. Будем включать в класс S_i^2 те решающие правила, которые для классификации используют не более i признаков, т. е. рассмотрим структуру

$$S_1^2 \subset \ldots \subset S_n^2. \tag{8.27}$$

Структура (8.32) построена так, что $S_p^1 \subset S_p^2$. Очевидно, что функция роста $m^{S_p^2}(l)$ оценивается через функцию роста $m^{S_p^1}(l)$

$$m^{S_p^2}(l) \leqslant C_n^p m^{S_p'}(l) < 1{,}5 C_n^p \frac{l^p}{p!}. \tag{8.28}$$

Таким образом, метод упорядоченной минимизации риска на структуре (8.27) приведет к выбору функции $F(x, \alpha_э^*)$, минимизирующей по i и $F(x, \alpha) \in S_i$ функционал

$$R_2(\alpha,\ i) = 2\frac{i\left(\ln\frac{2l}{i}+1\right)+\ln C_n^i-\ln\frac{\eta}{12}}{l}\times$$

$$\times\left(1+\sqrt{1+\frac{\nu(\alpha)\,l}{i\left(\ln\frac{2l}{i}+1\right)+\ln C_n^i-\ln\frac{\eta}{12}}}\right)+\nu(\alpha). \tag{8.29}$$

Для найденного решения $F(x, \alpha_э^*)$ справедливо

$$P\{P(\alpha_э^*) < R_2(\alpha_э^*,\ S_*^2)\} > 1 - n\eta \qquad (n < l).$$

Для обоих типов структур в качестве алгоритма выбора второго уровня может быть также использована процедура «скользящий контроль».

Таким образом, при решении задачи обучения распознаванию образов в классе линейных решающих правил рекомендации метода упорядоченной минимизации риска состоят в том, чтобы выбрать экстремальное подпространство признаков (которое может быть разным как по составу, так и по числу признаков в зависимости от того, ранжирована ли система признаков или нет) и построить

в нем решающее правило, минимизирующее эмпирический риск.

Выбор экстремального пространства признаков в условиях малых выборок позволяет существенно увеличить вероятность правильной классификации экзаменационной (не участвовавшей в обучении) выборки. Возможный выигрыш иллюстрирует табл. 1, полученная при решении задач медицинской дифференциальной диагностики.

Таблица 1

№ задачи	Объем выборки	Исходная размерность пространства признаков	Размерность экстремального пространства	Вероятность ошибки	
				в исходном пространстве	в экстремальном пространстве
1	114	84	56	0,21	0,14
2	108	92	47	0,18	0,11
3	131	112	51	0,22	0,10
4	240	134	65	0,13	0,07
5	360	196	82	0,15	0,07

В столбцах таблицы указаны номер задачи, объем выборки, исходная размерность бинарного пространства признаков, размерность экстремального пространства признаков, вероятность ошибочной классификации в исходном и экстремальном пространствах. Задачи решались с помощью алгоритмов, приведенных в главе XI.

§ 5. Восстановление регрессии в классе полиномов

Задача определения числа членов разложения по ранжированной системе функций является одной из центральных в теории регрессии.

Частная ее постановка составляет задачу *восстановления полиномиальной регрессии*. Суть задачи заключается в следующем: пусть статистическая модель, связывающая величину y с переменной x, есть

$$y = R(x) + \xi, \tag{8.30}$$

где $R(x)$ — полином неизвестной степени, а ξ — помеха, не зависящая от x, математическое ожидание которой равно нулю, а дисперсия ограничена.

Требуется, наблюдая пары

$$x_1, y_1; \ldots; x_l, y_l,$$

восстановить полином $R^*(x)$, близкий к $R(x)$. Близость понимается в смысле метрики L_p^2:

$$\rho_L(R(x), R^*(x)) = \left(\int (R(x) - R^*(x))^2 P(x)\, dx\right)^{1/2},$$

где $P(x)$ — та же самая плотность, согласно которой выбирались значения x.

Существует традиционный путь решения этой задачи: сначала определить степень n искомого полинома $R(x)$, а затем в классе функций, разложимых по системе n ортонормальных с весом $P(x)$ полиномов, восстановить регрессию. Таким образом, основное содержание проблемы здесь сводится к определению степени полиномиальной регрессии.

Определение степени полиномиальной регрессии осуществляется с помощью стандартных приемов математической статистики. Наиболее просто реализуются эти приемы в схеме Гаусса — Маркова, т. е. в условиях, когда величины x фиксированы (см. § 2 гл. V). Пусть они равны $x_1, \ldots, x_l$. В этом случае, не ограничивая общности, будем считать, что функция $R(x)$ разложима по системе ортонормальных на $x_1, \ldots, x_l$ полиномов $R_i(x)$:

$$\frac{1}{l}\sum_{i=1}^{l} R_p(x_i) R_q(x_i) = \begin{cases} 1, & \text{если } p=q, \\ 0, & \text{если } p \neq q. \end{cases}$$

Система ортонормальных полиномов замечательна тем, что с ее помощью регрессия $R(x)$ может быть представлена в виде

$$R(x) = \sum_{p=1}^{n} \alpha_p^0 R_p(x),$$

где

$$\alpha_p^0 = M\frac{1}{l}\sum_{i=1}^{l} R_p(x_i)\, y_i.$$

Оценка параметров $\hat{\alpha}_p$, вычисленная с помощью метода наименьших квадратов, оказывается равной

$$\hat{\alpha}_p = \frac{1}{l}\sum_{i=1}^{l} R_p(x_i)\, y_i. \tag{8.31}$$

Таким образом, проблема определения степени регрессии заключается в том, чтобы на основании информации о величинах $\hat{\alpha}_1, \ldots, \hat{\alpha}_n$

принять (или опровергнуть) гипотезу о том, что $\alpha_i^0 = 0$ ($i = 1, 2, \ldots, n$).

Заметим, что если помеха ξ в (8.30) распределена по нормальному закону $N(0, \sigma^2)$ с нулевым средним и дисперсией σ^2, то случайная величина $\hat{\alpha}_p$ оказывается распределенной тоже по нормальному закону со средним α_p^0 и дисперсией $\sigma_1^2 = \sigma^2/l$. В этом случае для $\alpha_p^0 = 0$ величина

$$(\hat{\alpha}_p)^2 = \left(\frac{1}{l} \sum_{i=1}^{l} R_p(x_i)\, y_i \right)^2 \tag{8.32}$$

распределена согласно $\sigma_1^2\chi^2$-распределению с одной степенью свободы. Если бы дисперсия помехи была бы нам известна, то для проверки гипотезы $M\hat{\alpha}_p = 0$ можно было бы воспользоваться распределением

$$\left(\frac{\hat{\alpha}_p}{\sigma_1} \right)^2 \sim \chi_1^2.$$

В том случае, когда величина $(\hat{\alpha}_p/\sigma_1)^2$ больше $\varkappa(\eta)$ (величина $\varkappa(\eta)$ определяется из условия $P\{\chi_1^2 > \varkappa(\eta)\} = \eta$), для заданного уровня значимости η гипотеза $M\hat{\alpha}_p = \alpha_p^0 = 0$ отвергается, в противном случае эта гипотеза принимается.

Однако на практике дисперсия σ^2 помехи ξ неизвестна. Поэтому наряду с (8.32) рассматривается статистика

$$\pi^2 = \sum_{i=1}^{l} y_i^2 - \sum_{i=1}^{r} (\hat{\alpha}_i)^2. \tag{8.33}$$

Если, начиная с $i = r+1$, коэффициенты $\alpha_i^0 = 0$ ($i = r+1, \ldots, l$), то статистика (8.33) распределена согласно $\sigma_1^2\chi^2$-распределению с $\nu = l - r - 1$ степенями свободы.

Образуем статистику $\zeta = \nu\hat{\alpha}_p^2/\pi^2$. Эта статистика подчинена распределению Фишера ($F_{1\nu}$-распределение)

$$\zeta = \frac{\nu\chi_1^2}{\chi_\nu^2} \sim F_{1\nu}. \tag{8.34}$$

$F_{1\nu}$-распределение табулировано. Таблицы $F_{1\nu}$-распределения даны во всех практических руководствах по статистике. Таким образом, используя статистику $\nu\hat{\alpha}_p^2/\pi^2$ можно для заданного уровня значимости η определить, приемлема ли гипотеза $\alpha_i^0 = 0$: для этого достаточно проверить выполнение неравенства

$$\zeta > \varkappa(\eta).$$

При решении практических задач нет необходимости строить ортонормальную на $x_1, \ldots, x_l$ систему полиномов. Нетрудно убедиться, что статистика

$$\zeta = \frac{R_r - R_{r+1}}{R_{r+1}} (l - r - 1),$$

где R_r — невязка (величина минимума эмпирического риска), вычисленная для полиномов степени r, также распределена согласно $F_{1,\, l-r-1}$-распределению Фишера.

Таким образом, в случае нормально распределенной помехи ξ, используя невязки $R_1, \ldots, R_{l-1}$, вычисленные для полиномов степени $1, r, \ldots, l-1$, можно с помощью F-критерия Фишера (8.34) установить степень полиномиальной регрессии.

Однако классическая схема восстановления полиномиальной регрессии — выяснение истинной степени регрессии и приближение к регрессии в классе полиномов этой степени приводит к успеху лишь при использовании больших выборок. Только для достаточно больших объемов выборки можно утверждать, что наилучшее приближение будет достигнуто на функции, минимизирующей эмпирический риск в классе полиномов, степень которых равна истинной степени регрессии. Для малых выборок вопрос о том, какова наиболее подходящая степень приближения, остается открытым [1]).

Ниже мы применим метод упорядоченной минимизации для решения этой задачи, но прежде, чем приступить к построению соответствующих алгоритмов, обратим внимание читателя на то, что по существу задача будет решаться в более общей постановке, чем классическая. Мы не будем полагать, что регрессия есть полином — она может быть любой интегрируемой с квадратом функцией, но приближать регрессию будем полиномом. Требуется в этих условиях отыскать подходящее приближение.

Итак, будем решать задачу методом упорядоченной минимизации риска. Для этого зададим структуру на множестве полиномов. Заметим, что уже в самой постановке задачи содержится указание на особенность задания структуры

$$S_1 \subset \ldots \subset S_n. \tag{8.35}$$

Множество S_p состоит из полиномов, степень которых не превосходит p. Такое упорядочение полиномов является «естественным» (но не единственным). Оно соответствует упорядочению по числу членов разложения ряда,

---

[1]) На малых выборках реализация классической схемы может приводить к парадоксальным ситуациям: чем более мощный критерий используется для установления степени регрессии, тем хуже может оказаться окончательный результат.

составленного из элементов

$$1,\ x,\ x^2,\ \ldots,\ x^n\ldots, \tag{8.36}$$

расположенных в порядке возрастания степени n. Однако возможен другой порядок расположения элементов ряда, например следующий:

$$x^5,\ 1,\ x^4,\ x^3,\ x^7,\ \ldots \tag{8.37}$$

Упорядочение полиномов в соответствии с разложением по первым p членам ряда (8.37) приведет к иному заданию структуры на множестве полиномиальных зависимостей.

Итак, рассмотрим структуру (8.35), заданную разложением по первым членам ряда, ранжированного согласно (8.36).

Пусть, кроме того, известно, что выполнено ограничение [1])

$$\sup_{\alpha} \frac{\sqrt[p]{M\,(y-F\,(x,\ \alpha))^{2p}}}{M\,(y-F\,(x,\ \alpha))^2} \leqslant \tau \qquad (p>2).$$

Тогда, согласно теореме 7.6, с вероятностью $1-\eta$ одновременно для всех полиномов степени $r-1$ (всех полиномов $F(x,\ \alpha)$, принадлежащих S_r) выполнится неравенство

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1-2\tau a(p)\sqrt{\dfrac{r\left(\ln\dfrac{2l}{r}+1\right)-\ln\dfrac{\eta}{12}}{l}}} \right]_{\infty}. \tag{8.38}$$

Неравенство (8.38) выполняется и для полинома $F(x,\ \alpha_э^r)$, минимизирующего на S_r эмпирический риск.

Выберем в качестве приближения к регрессии функцию, минимизирующую эмпирический риск на таком элементе структуры S_*, для которого достигается минимум правой части оценки (8.38).

---

[1]) Как уже отмечалось, знание оценки τ — требование существенно более слабое, чем знание типа плотности помехи, необходимое для восстановления полинома регрессии классическими методами (см. выше текст, набранный петитом).

Пусть минимум достигается на функции $F(x, \alpha_э^*)$ и равен $R(\alpha_э^*, S_*)$. Тогда справедливо утверждение

$$P\{I(\alpha_э^*) < R(\alpha_э^*, S_*)\} > 1 - n\eta.$$

Использование метода упорядоченной минимизации риска для восстановления полиномиальной регрессии в условиях малой выборки весьма эффективно на практике.

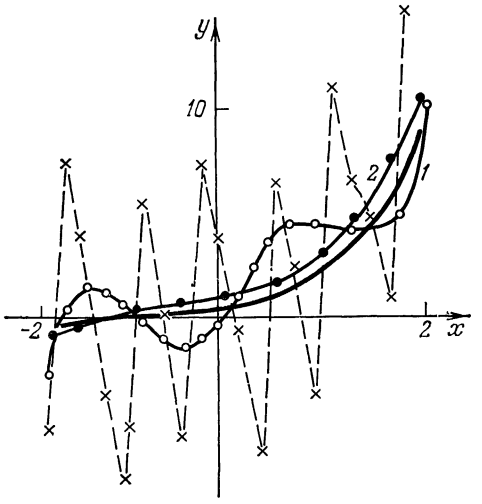

Рис. 6.

На рис. 6 показан результат восстановления регрессии, заданной полиномом пятой степени на отрезке [—2, 2]. Восстановление проводилось по измерениям функции, осуществленным в 20 случайно взятых точках интервала [—2, 2]. Измерение осуществлялось с помехой, распределенной равномерно на интервале $[-a, +a]$, где a — максимальное значение регрессии на интервале [—2, 2]. На рисунке показаны эмпирические данные (крестики) и регрессия (жирная кривая). Наилучшее приближение в классе полиномов пятой степени — белые точки, кривая *1*, приближение, полученное методом упорядоченной минимизации риска — полином четвертой степени, черные точки, кривая *2*. Видно, что кривая *2* лучше приближает регрессию,

чем кривая *1*. На рис. 7 приведен пример восстановления неполиномиальной регрессии (жирная линия) в классе полиномов (тонкая линия) по 20 измерениям (крестики).

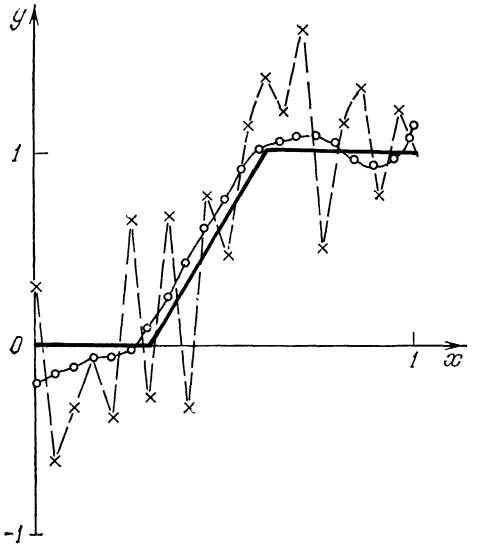

Рис. 7.

Функции восстанавливались с помощью алгоритма 12-1, приведенного в главе XII.

§ 6. Восстановление регрессии в классе линейных по параметрам функций

Рассмотрим класс линейных по параметрам функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x). \qquad (8.39)$$

Существуют две идеи задания структуры на этом классе:

1) упорядочение функций по числу членов разложения;

2) упорядочение функций по норме вектора параметров α (норме функций в L_p^2 для ортонормальной по мере $P(x)$ системы $\varphi_1(x), \ldots, \varphi_n(x)$).

Построим на этих структурах алгоритмы упорядоченной минимизации риска, использующие в качестве критерия выбора второго уровня оценку «скользящий контроль».

Упорядочение по числу членов разложения. Пусть дана априорно ранжированная система функций

$$\varphi_1(x), \ldots, \varphi_n(x). \tag{8.40}$$

Зададим на множестве функций $F(x, \alpha)$ структуру

$$S_1 \subset \ldots \subset S_n, \tag{8.41}$$

где элемент структуры S_i содержит лишь такие функции, которые могут быть разложены по первым i членам ряда (8.40).

В этом случае метод упорядоченной минимизации состоит в определении такого подпространства $\varphi_1(x), \ldots, \varphi_r(x), 0, \ldots, 0$ исходного пространства $\varphi_1(x), \ldots, \varphi_n(x)$, на котором достигается минимум величины

$$T_r(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{\left(y_i - (f_i^r)^{\mathrm{T}} (\Phi_r^{\mathrm{T}}\Phi_r)^{-1}\Phi_r^{\mathrm{T}}Y\right)^2}{\left(1 - (f_i^r)^{\mathrm{T}} (\Phi_r^{\mathrm{T}}\Phi_r)^{-1} f_i^r\right)^2}. \tag{8.42}$$

Функция $F(x, \alpha_э^*)$, минимизирующая эмпирический риск в S_r (вектор параметров $\alpha_э^* = (\Phi_r^{\mathrm{T}}\Phi_r)^{-1}\Phi_r^{\mathrm{T}}Y$), считается наилучшим приближением к регрессии.

В формуле (8.42) обозначено: (f_i^r) — вектор $(\varphi_1(x_i), \ldots, \varphi_r(x_i), 0, \ldots, 0)^{\mathrm{T}}$, Φ_r — матрица, строки которой равны

$$(f_i^r)^{\mathrm{T}} = (\varphi_1(x_i), \ldots, \varphi_r(x_i), 0, \ldots, 0).$$

Упорядочение по величинам нормы вектора параметров. Рассмотрим систему упорядоченных множеств

$$S_1 \subset \ldots \subset S_q \tag{8.43}$$

таких, что подмножества S_i содержат лишь функции $F(x, \alpha)$, для которых выполняются условия

$$\sum_{j=1}^{n} \alpha_j^2 \leqslant c_i. \tag{8.44}$$

Величины c_i растут с увеличением номера i

$$0 < c_1 < c_2 < \ldots < c_q < \infty.$$

В соответствие величинам c_l может быть поставлен монотонно убывающий ряд положительных величин γ_i (множителей Лагранжа)

$$\gamma_1 > \gamma_2 > \ldots > \gamma_q = 0,$$

такой, что задача минимизации эмпирического риска на множестве S_r оказывается эквивалентной минимизации функционала

$$I_{\gamma_r}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{j=1}^{n}\alpha_j\varphi_j(x_i)\right)^2 + \gamma_r\sum_{j=1}^{n}\alpha_j^2. \quad (8.45)$$

В этом случае двухуровневая схема метода упорядоченной минимизации риска состоит в том, чтобы на первом уровне отобрать q функций $F(x, \alpha_э(\gamma_r))$, минимизирующих для различных величин γ_r функционал (8.45), а на втором уровне среди q отобранных функций определить такую, которая доставляет минимум оценке «скользящий контроль». Иначе говоря, при таком способе задания структуры метод упорядоченной минимизации состоит в том, чтобы определить γ_r, на котором достигается минимум выражения

$$T_{\gamma_r}(x_1, y_1; \ldots; x_l, y_l) = \frac{1}{l}\sum_{i=1}^{l}\frac{\left(y_i - f_i^{\mathrm{T}} A_{\gamma_r}^{-1}\Phi^{\mathrm{T}}Y\right)^2}{\left(1 - f_i^{\mathrm{T}}A_{\gamma_r}^{-1}f_i\right)^2}, \quad (8.46)$$

где

$$A_{\gamma_r} = (\Phi^{\mathrm{T}}\Phi + \gamma_r I),$$

и найти функцию $F(x, \alpha^*)$, минимизирующую при этом γ_r функционал (8.45). Эту функцию задает вектор параметров $\alpha^* = A_{\gamma_r}^{-1}\Phi^{\mathrm{T}}Y$.

Наконец, рассмотрим комбинированную структуру на множестве линейных по параметрам функций $F(x, \alpha)$. Сначала упорядочим функции по числу членов разложения (8.40), а затем каждое подмножество S_p, состоящее из функций, разложимых по p членам, упорядочим по величинам нормы вектора параметров (8.44).

Таким образом, рассмотрим систему множеств

$$\begin{array}{ccccc} S_{10} & \subset & S_{20} & \subset \ldots \subset & S_{q0} \\ \cup & & \cup & & \cup \\ S_{11} & \subset & S_{21} & \subset \ldots \subset & S_{q1} \\ \cup & & \cup & & \cup \\ \vdots & & \vdots & & \vdots \\ \cup & & \cup & & \cup \\ S_{1n_1} & \subset & S_{2n_2} & \subset \ldots \subset & S_{q,\,n_q}. \end{array} \tag{8.47}$$

Элемент S_{pr} есть подмножество, состоящее из функций, разложенных по p членам ряда таких, что выполнено неравенство

$$\sum_{i=1}^{p} \alpha_i^2 < c_r.$$

Метод упорядоченной минимизации здесь состоит в определении такой пары p, γ_r, для которой оценка качества алгоритма, минимизирующего эмпирический функционал

$$I_{\gamma_r}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{p} \alpha_j \varphi_j(x_i) \right)^2 + \gamma_r \sum_{j=1}^{p} \alpha_j^2,$$

полученная с помощью процедуры «скользящий контроль», будет минимальной.

В вычислительном отношении это означает, что необходимо найти такую пару p, γ_r, на которой будет достигаться минимум выражения

$$T_{\gamma_r, p}(x_1, y_1; \ldots, x_l, y_l) = \frac{1}{l} \sum_{i=1}^{l} \frac{\left(y_i - (f_i^p)^{\mathrm{T}} A_{\gamma_r, p}^{-1} \Phi_p^{\mathrm{T}} Y\right)^2}{\left(1 - (f_i^p)^{\mathrm{T}} A_{\gamma_r, p}^{-1} f_i^p\right)^2}, \tag{8.48}$$

$$A_{\gamma_r, p} = (\Phi_p^{\mathrm{T}} \Phi_p + \gamma_r I),$$

и определить функцию $F(x, \alpha^*)$ (вектор параметров $\alpha^* = A_{\gamma_r, p}^{-1} \Phi_p^{\mathrm{T}} Y$).

Итак, мы рассмотрели алгоритмы упорядоченной минимизации риска, использующие в качестве критерия выбора второго уровня процедуру «скользящий контроль».

Реализация этих алгоритмов восстановления регрессии в классе линейных по параметру функций оказалась не многим сложнее, чем реализация метода наименьших квад-

ратов. Однако, вообще говоря, рассмотренные алгоритмы являются эвристическими — они основаны на гипотезе о том, что дисперсия оценки «скользящий контроль» мала. На практике при восстановлении регрессии эти алгоритмы дают хорошие и устойчивые результаты, если объем выборки в несколько раз (3—4 раза) больше размерности пространства. Создание же алгоритмов упорядоченной минимизации риска для объемов выборки, соизмеримых (или меньших) размерности вектора параметров α связано с оценками вероятности равномерного относительного уклонения средних от их математических ожиданий.

§ 7. Восстановление регрессии в классе линейных по параметрам функций (продолжение)

Как и в предыдущем параграфе, рассмотрим три типа структур на классе линейных по параметрам функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x):$$

а) структуру, образованную по числу членов разложения;

б) структуру, образованную по величине нормы вектора параметров α (нормы $F(x, \alpha)$ в метрике L_P^2 для ортогональной по мере $P(x)$ системы $\varphi_1(x) \ldots, \varphi_n(x)$;

в) комбинированную структуру, образованную и по числу членов разложения и по величине нормы функции $F(x, \alpha)$.

Ниже для этих структур мы построим метод упорядоченной минимизации риска, основанный на оценках вероятности равномерного относительного уклонения средних от математических ожиданий.

1. Пусть задана структура а). Тогда, согласно теореме 7.6, с вероятностью $1-\eta$ одновременно для всех функций элемента S_r структуры (множество S_r содержит функции $F(x, \alpha)$, разложимые по первым r членам) выполняется неравенство

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1 - 2\tau a(p) \sqrt{\frac{r\left(\ln \frac{2l}{r} + 1\right) - \ln \frac{\eta}{12}}{l}}} \right]_{\infty}. \tag{8.49}$$

Так как неравенство с вероятностью $1-\eta$ справедливо одновременно для всех функций из S_r, то оно с вероятностью $1-\eta$ выполнится и для функции $F(x, \alpha_э)$, минимизирующей на S_r эмпирический риск. Выберем теперь такой элемент структуры S_* и в нем функцию, минимизирующую эмпирический риск, для которых достигается минимальная величина оценки (8.49). Найденная функция $F(x, \alpha_э^*)$ для структуры а) определяет минимальную гарантированную (с вероятностью $1-\eta \cdot n$, $n < l$ — число элементов структуры) величину риска.

2. Рассмотрим теперь структуру б):

$$S_1 \subset \ldots \subset S_n. \tag{8.50}$$

Здесь S_r — множество функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x),$$

для которых выполняется соотношение

$$\sum_{i=1}^{n} \alpha_i^2 < c_r.$$

Выделим на множестве S_r конечную ε-сеть $S_\varepsilon = \{F(x, \alpha_1), \ldots, F(x, \alpha_{N(\varepsilon)})\}$, состоящую из $N(\varepsilon)$ элементов. Согласно теореме 7.5 с вероятностью $1-\eta$ для функции $F(x, \alpha_э)$, минимизирующей на обучающей последовательности величину эмпирического риска, справедлива оценка

$$I(\alpha_э) < \left(\varepsilon + \sqrt{\varepsilon^2 + \left[\frac{I_э(\alpha_i(\alpha_э))}{1-T(\varepsilon)}\right]_\infty}\right)^2, \tag{8.51}$$

где

$$T(\varepsilon) = 2\tau a(p) \sqrt{\frac{\ln N(\varepsilon) + \ln l - \ln \eta/24}{l}}.$$

В оценке (8.51) $F(x, \alpha_i(\alpha_э))$ — ближайший к $F(x, \alpha_э)$ элемент ε-сети. Таким образом, для функции, минимизирующей на элементе S_r структуры (8.50) эмпирический риск, может быть вычислена гарантированная оценка величины среднего риска. Выберем такую функцию (такой элемент структуры), для которой эта оценка минимальная.

3. Рассмотрим структуру в), каждый элемент которой $S_{q,r}$ определяется как числом членов разложения

$$F(x, \alpha) = \sum_{i=1}^{q} \alpha_i \varphi_i(x),$$

так и нормой функций

$$\sum_{i=1}^{q} \alpha_i^2 \leqslant c_r.$$

Построим метод упорядоченной минимизации риска на этой структуре. Для оценки качества функции, минимизирующей в $S_{q,r}$ эмпирический риск, также используем оценку (8.51). В результате выберем такой элемент $S_{q,r}$ структуры и такую в нем функцию, для которых оценка минимальна.

Для того чтобы строить алгоритмы упорядоченной минимизации риска на структурах б) и в), надо уметь вычислять емкость ε-сети.

§ 8. Селекция обучающей последовательности

В этом параграфе мы рассмотрим идею *селекции обучающей последовательности*: исключение из обучающей последовательности нескольких элементов с тем, чтобы с помощью оставшегося множества найти функцию, доставляющую меньшую гарантированную величину среднему риску.

Заметим, что для задачи распознавания образов селекция обучающей последовательности не имеет смысла: решения, получаемые минимизацией эмпирического риска по всей обучающей выборке и по обучающей подвыборке, полученной исключением минимального числа элементов с тем, чтобы подвыборка могла быть разделена безошибочно, достигаются на одном и том же решающем правиле. Это обстоятельство является следствием того, что функция потерь $(\omega - F(x, \alpha))^2$ в задаче распознавания образов принимает только два значения — нуль и единица.

В задаче восстановления регрессии функция потерь может принимать любые положительные значения, и поэтому исключение некоторых элементов x, y может суще-

ственно изменить как само решение, так и оценку качества полученного решения.

Итак, пусть задана обучающая последовательность

$$x_1,\ y_1;\ \ldots;\ x_l, y_l. \tag{8.52}$$

Рассмотрим одновременно $H_l^t = \sum_{m=0}^{t} C_l^m$ различных задач восстановления функциональной зависимости по эмпирическим данным

$$x_1,\ y_1;\ \ldots;\ \widehat{x_i,\ y_i};\ \ldots;\ \widehat{x_j,\ y_j};\ \ldots;\ x_l,\ y_l.$$

Выражение $\widehat{x_i,\ y_i}$ означает, что из обучающей последовательности (8.52) исключен элемент x_i, y_i. Задачи различаются лишь тем, что в каждой из них функциональная зависимость восстанавливается по своей обучающей выборке, полученной из (8.52) исключением не более t элементов. (Из (8.52) может быть образовано C_l^m различных подвыборок, состоящих из $l-m$ элементов. Всего, таким образом, возможно $H_l^t = \sum_{m=0}^{t} C_l^m$ различных задач.)

Согласно теореме 7.6 для каждой из H_l^t задач с вероятностью $1-\eta$ справедливо неравенство

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1-2\tau a(p) \sqrt{\dfrac{\ln m^S (2(l-t_i)) - \ln \eta/8}{l-t_i}}} \right]_\infty ,$$

где $t_i \leqslant t$ — число исключенных векторов в i-й задаче. Следовательно, одновременно для всех H_l^t задач с вероятностью $1-\eta$ справедливы неравенства

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1-2\tau a(p) \sqrt{\dfrac{\ln m^S (2(l-t_i)) + \ln H_l^t - \ln \eta/8}{l-t_i}}} \right]_\infty . \tag{8.53}$$

Будем теперь искать минимум правой части (8.53) не только по элементам S_r структуры и функции $F(x, \alpha) \subset \subset S_r$, но и по всем H_l^t задачам.

Иначе говоря, будем минимизировать функционал

$$I_э\left(\alpha,\ \widehat{x_{r_1}},\ \widehat{y_{r_1}};\ \ldots;\ \widehat{x_{r_{t_l}}},\ \widehat{y_{r_{t_l}}}\right) = \left[\frac{\frac{1}{l-t_i}\sum_{j=1}^{l}{}^{(ti)}\,(y_j - F(x_j, \alpha))^2}{1-2\tau a(p)\sqrt{\frac{\ln m^{S_r}(2(l-t_i)) + \ln H_l^t - \ln \eta/8}{l-t_i}}}\right]_{\infty}, \quad (8.54)$$

по элементам $F(x, \alpha) \in S_r$ и $\widehat{x_{r_1}}, \widehat{y_{r_1}}; \ldots; \widehat{x_{r_{t_i}}}, \widehat{y_{r_{t_i}}}$; здесь знак $\sum_{j=1}^{l}{}^{(t_i)}$ указывает на то, что $t_i \leqslant t$ элементов не суммируются.

Перебором по t (обычно $t = 1, 2, 3, 4$) найдем наименьшую величину (8.54). Она определяет гарантированную (с вероятностью $1 - \eta t$) величину среднего риска.

Таким образом, при поиске наилучшего гарантированного решения, кроме оптимизации по структуре и функциям, принадлежащим элементу структуры, возможна дополнительная оптимизация по выбору обучающего подмножества из заданного обучающего множества (8.52). В условиях малой выборки подбор обучающего подмножества из заданного множества весьма часто оказывается полезным на практике.

§ 9. Несколько общих замечаний

В этой главе был сформулирован новый принцип минимизации риска в условиях малых выборок.

Оказалось, что если на допустимом множестве решений задать структуру, то появляется возможность дополнительной оптимизации по элементам структуры. Нужно лишь, чтобы структура была задана априорно.

Другая дополнительная возможность минимизации риска по эмпирическим данным появляется за счет селекции обучающей выборки.

В этой главе мы применили метод упорядоченной минимизации риска для решения задач распознавания образов и восстановления регрессии, а идею селекции

выборки для задачи восстановления регрессии (вследствие примитивности функции потерь в задаче распознавания образов селекция выборки не приводит к уменьшению гарантированной оценки риска).

Возникает вопрос, сколь общими являются метод упорядоченной минимизации риска и метод селекции выборки.

Очевидно, что метод упорядоченной минимизации риска применим для решения любой задачи минимизации риска, для которой может быть получена оценка равномерного или равномерного относительного уклонения эмпирических средних от математических ожиданий (см. гл. VII, § 8). В этом случае при минимизации функционала

$$I(\alpha) = \int Q(z, \alpha) P(z)\, dz$$

по эмпирическим данным $z_1, \ldots, z_l$ на множестве Λ значений параметров α задается структура

$$\Lambda_1 \subset \Lambda_2 \subset \ldots \subset \Lambda_q.$$

На каждом элементе этой структуры Λ_l находится значение параметра $\alpha_э^i \in \Lambda_l$, минимизирующее эмпирический риск

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha),$$

а затем с помощью оценок, приведенных в § 8 гл. VII, из q найденных параметров выбирается такой $\alpha_э^*$, который доставляет гарантированный минимум величине среднего риска.

Селекция выборки также может быть проведена, когда существует равномерная оценка среднего риска.

Основные утверждения главы VIII

1. В условиях малой выборки минимизацию риска целесообразно проводить методом упорядоченной минимизации.

Для этого на множестве функций надо задать структуру, а затем найти такой элемент структуры и в нем такую функцию, которые доставляют наименьшую гарантированную оценку риску.

2. Для оценки величины риска могут быть использованы равномерные оценки среднего риска по величине эмпирического риска и оценка процедуры «скользящий контроль».

3. В классической теории регрессии уже рассматривалась задача определения подходящего элемента структуры — задача восстановления полиномиальной регрессии. Однако основной принцип решения этой задачи — определение модели искомой функции (степени полинома регрессии) и восстановление регрессии в рамках этой модели, оправдан лишь для больших выборок и в ситуации, когда заданное множество функций содержит регрессию.

3. Возможны различные алгоритмы упорядоченной минимизации риска. Они определяются разными способами задания структуры на множестве функций.

4. При восстановлении регрессии на малых выборках отыскание решения с меньшей гарантированной оценкой риска может быть получено за счет селекции выборки.

ГЛАВА IX

РЕШЕНИЕ НЕКОРРЕКТНЫХ ЗАДАЧ ИНТЕРПРЕТАЦИИ ИЗМЕРЕНИЙ МЕТОДОМ УПОРЯДОЧЕННОЙ МИНИМИЗАЦИИ РИСКА

§ 1. Некорректные задачи интерпретации результатов косвенных экспериментов

Пусть в классе $f(t, \alpha)$ $(a \leqslant t \leqslant b)$ надо восстановить функциональную зависимость $f(t, \alpha_0) = f(t)$ (здесь $f(t)$ принадлежит множеству $f(t, \alpha)$). И пусть ситуация такова, что нельзя непосредственно измерять значения функции $f(t)$, но можно измерять значения другой функции $F(x)$ $(a \leqslant x \leqslant b)$, связанной с искомой операторным уравнением

$$Af(t) = F(x). \tag{9.1}$$

Оператор A осуществляет непрерывное взаимно однозначное отображение элементов $f(t, \alpha)$ пространства E_1 в элементы $F(x, \alpha)$ пространства E_2.

Пусть проведены измерения функции $F(x)$:

$$x_1, y_1; \ldots; x_l, y_l. \tag{9.2}$$

Пара x_l, y_l означает, что в точке x_l измеренное значение функции $F(x_l)$ оказалось равным y_l.

Требуется, зная оператор A и измерения (9.2), восстановить в $f(x, \alpha)$ функцию $f(t) = f(x, \alpha_0)$. При этом допускается, что задача решения операторного уравнения (9.1) может быть некорректно поставленной.

Восстанавливать функцию $f(t)$ будем в ситуации, когда:

1) значения функции $F(x)$ измеряются с аддитивной помехой

$$y_l = F(x_l) + \xi, \quad M\xi = 0, \quad M\xi^2 = \sigma^2 < \infty,$$

не зависящей от x;

2) точки x_l, в которых проводятся измерения, определяются случайно и независимо согласно некоторой не обращающейся в нуль на $[a, b]$ плотности. Ниже будем считать, что плотность равномерная.

В главе I было показано, что функция $f(t, \alpha_0)$, которая является прообразом в E_1 регрессии $F(x, \alpha_0)$ из про-

странства E_2, т. е. прообразом точки минимума функционала

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(y \mid x)\, dy\, dx, \tag{9.3}$$

совпадает с решением уравнения (9.1)

Однако найти регрессию по выборке фиксированного объема — задача нереальная. Можно надеяться лишь на то, что удастся найти функцию $F(x, \hat{\alpha})$, близкую (в метрике пространства E_2) к регрессии, и тогда за решение уравнения (9.1) можно будет принять прообраз $f(t, \hat{\alpha})$ этой функции в пространстве E_1. Такая идея, вообще говоря, не всегда приводит к успеху: несостоятельность ее заключается в том, что в случае, когда уравнение (9.1) определяет некорректно поставленную задачу, близким образам в E_2 могут (но не обязательно должны) соответствовать далекие в E_1 прообразы.

В нашем случае это означает, что не все методы минимизации риска в пространстве образов могут быть использованы для решения задачи интерпретации результатов косвенных экспериментов и что, возможно, существуют такие способы минимизации риска, которые указывают лишь на элементы $F(x, \hat{\alpha})$ пространства E_2, являющиеся образами функций, близких к искомому решению. Эти способы минимизации риска и следует применять для решения некорректных задач интерпретации измерений (конечно, если указанные способы минимизации риска вообще существуют).

Ниже мы покажем, что при определенных условиях алгоритмы упорядоченной минимизации риска могут быть использованы для решения некорректных задач измерений. Мы покажем, что с увеличением числа измерений последовательность решений, получаемых согласно методу упорядоченной минимизации риска, сходится к искомой функции $f(t)$.

§ 2. Определение понятия сходимость

Пусть в E_1 выбрана мера близости функции $\rho_{E_1}(f(t, \alpha_1), f(t, \alpha_2)) = \rho_{E_1}(\alpha_1, \alpha_2)$ и зафиксирован алгоритм восстановления зависимости $f(t) = f(t, \alpha_0)$ по косвенным экспериментам

$$x_1, y_1; \ldots; x_l, y_l. \tag{9.4}$$

Тогда для каждой конкретной реализации (9.4) может быть найдена функция $f(t, \hat{\alpha}_l)$ (вектор параметров $\hat{\alpha}_l = \alpha(x_1, y_1; \ldots; x_l, y_l)$) и, таким образом, получена последовательность

$$\hat{\alpha}_1, \ldots, \hat{\alpha}_l, \ldots \tag{9.5}$$

Эта последовательность определяет последовательность чисел

$$\rho_{E_1}(\hat{\alpha}_1, \alpha_0), \ldots, \rho_{E_1}(\hat{\alpha}_l, \alpha_0), \ldots, \tag{9.6}$$

задающих расстояния от параметров $\hat{\alpha}_l$ до α_0. Как (9.5), так и (9.6) являются случайными последовательностями, которые порождаются алгоритмом A восстановления зависимости $f(t)$ и реализацией (9.4). Исследование алгоритмов восстановления зависимостей сводится, таким образом, к исследованию сходимости последовательности (9.6).

Существуют различные понятия сходимости случайных последовательностей. В этой главе мы используем два понятия: сходимость по вероятности и сходимость с вероятностью единица (почти наверное).

Определение 1. *Последовательность случайных величин* $\xi_1, \ldots, \xi_l \ldots$ *сходится к величине* ξ_0 *по вероятности, если, каково бы ни было* $\varepsilon > 0$, *вероятность выполнения неравенства*

$$|\xi_l - \xi_0| < \varepsilon$$

при $l \to \infty$ *стремится к единице, т. е.*

$$\lim_{l \to \infty} P\{|\xi_l - \xi_0| < \varepsilon\} = 1.$$

Факт сходимости по вероятности записывается так: $\xi \xrightarrow{P} \xi_0$.

Определение 2. *Последовательность случайных величин* $\xi_1, \ldots, \xi_l, \ldots$ *сходится к величине* ξ_0 *с вероятностью единица, если, каково бы ни было* $\varepsilon > 0$, *вероятность выполнения неравенства*

$$\sup_{i \geqslant l} |\xi_i - \xi_0| < \varepsilon$$

стремится к единице при $l \to \infty$, *т. е.*

$$\lim_{l \to \infty} P\left\{\sup_{i \geqslant l} |\xi_i - \xi_0| < \varepsilon\right\} = 1.$$

Сходимость с вероятностью единица (почти наверное) принято обозначать так: $\xi \xrightarrow{\text{п. н.}} \xi_0$.

Приведенные определения отражают различные требования к понятию сходимости.

В первом случае событие $\{|\xi_l - \xi_0| < \varepsilon\}$ выделяет множество последовательностей, для которых выполняется условие $|\xi_l - \xi_0| < \varepsilon$ для заданного фиксированного l. При этом каждая последовательность с ростом l может то удовлетворять этому условию, то не удовлетворять ему. Сходимость по вероятности есть в некотором смысле «слабая» сходимость — она не дает никаких гарантий того, что каждая конкретная реализация $\xi_1, \ldots, \xi_l$ сходится в обычном смысле.

Напротив, сходимость с вероятностью единица есть понятие «сильной» сходимости. Оно означает, что почти все реализации сходятся в обычном смысле. Сходимость почти наверное может быть определена еще и так.

Определение 2а. *Последовательность случайных величин* $\xi_1, \ldots, \xi_l, \ldots$ *сходится с вероятностью единица к* ξ_0, *если мера множества реализаций, для которых существует предел*

$$\lim_{l\to\infty} \xi_l = \xi_0,$$

равна единице, т. е.

$$P\left\{\lim_{l\to\infty} \xi_l = \xi_0\right\} = 1.$$

Легко проверить, что из сходимости с вероятностью единица следует сходимость по вероятности. В самом деле, так как для любого l справедливо неравенство

$$P\{|\xi_l - \xi_0| < \varepsilon\} \geqslant P\left\{\sup_{i\geqslant l} |\xi_i - \xi_0| < \varepsilon\right\},$$

то из условия

$$\lim_{l\to\infty} P\left\{\sup_{i\geqslant l} |\xi_i - \xi_0| < \varepsilon\right\} = 1$$

следует

$$\lim_{l\to\infty} P\{|\xi_l - \xi_0| < \varepsilon\} = 1.$$

Обратное утверждение, вообще говоря, неверно. Дополнительные условия, при которых из сходимости по вероятности вытекает сходимость с вероятностью единица, определяет следующая лемма.

Лемма (Борель — Кантелли). *Если для случайной последовательности* $\xi_1, \ldots, \xi_l, \ldots$ *найдется такое* ξ_0, *что для любого* $\varepsilon > 0$ *окажется выполненным неравенство*

$$\sum_{i=1}^{\infty} P\{|\xi_i - \xi_0| \geqslant \varepsilon\} < \infty, \tag{9.7}$$

то последовательность $\xi_1, \ldots, \xi_l, \ldots$ *сходится к* ξ_0 *с вероятностью единица.*

Доказательство. Обозначим через E_l^r событие, состоящее в том, что выполняется неравенство

$$|\xi_l - \xi_0| > \frac{1}{r} \qquad (r - \text{целое число}).$$

Рассмотрим событие S_l^r, состоящее в том, что выполнится хотя бы одно из событий $E_l^r, E_{l+1}^r, \ldots, E_{l+i}^r, \ldots$

$$S_l^r = \bigcup_{i=0}^{\infty} E_{l+i}^r.$$

Оценим вероятность этого события

$$P\{S_l^r\} < \sum_{i=0}^{\infty} P\{E_{l+i}^r\} = \sum_{i=l+1}^{\infty} P\left\{|\xi_i - \xi_0| > \frac{1}{r}\right\}.$$

Так как в силу условий леммы ряд (9.7) сходится, то

$$\lim_{l \to \infty} P\{S_l^r\} = 0. \tag{9.8}$$

Рассмотрим теперь событие S^r:

$$S^r = \bigcap_{l=1}^{\infty} S_l^r.$$

Из того, что событие S^r влечет за собой любое из событий S_l^r, в силу (9.8) получаем

$$P\{S^r\} = 0. \tag{9.9}$$

Наконец, положим $S = \bigcup_{r=1}^{\infty} S^r$. Как нетрудно установить, это событие означает, что найдется такое r, что для каждого l ($l = 1, 2, \ldots$) хотя бы при одном i ($i = i(l)$)

будут выполняться неравенства

$$|\xi_{l+l} - \xi_0| > \frac{1}{r}.$$

Так как

$$P\{S\} \leqslant \sum_{r=1}^{\infty} P\{S^r\},$$

то в силу (9.9) $P\{S\} = 0$, что и требовалось доказать.

§ 3. Теоремы об интерпретации результатов косвенных экспериментов

Пусть A — линейный вполне непрерывный оператор, действующий из пространства L_2 в пространство C, A^* — оператор, сопряженный к A. Тогда оператор A^*A будет также вполне непрерывным оператором. Пусть

$$\lambda_1^2 \geqslant \lambda_2^2 \geqslant \ldots \geqslant \lambda_m^2 \geqslant \ldots$$

полная система его собственных чисел, а

$$\varphi_1(t), \ldots, \varphi_m(t) \ldots \tag{9.10}$$

— полная ортонормированная система его собственных элементов.

Рассмотрим также оператор AA^*. Он имеет ту же самую систему собственных чисел, которой соответствует полная ортонормированная система собственных элементов

$$\psi_1(x), \ldots, \psi_m(x), \ldots \tag{9.11}$$

Для элементов (9.10) и (9.11) выполняются равенства

$$\begin{aligned} A\varphi_p(t) &= \lambda_p \psi_p(x), \\ A^*\psi_p(x) &= \lambda_p \varphi_p(t). \end{aligned} \quad p = 1, 2, \ldots$$

Решение операторного уравнения (9.1) разложимо в ряд по системе функций (9.10)

$$f(t) = \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t). \tag{9.12}$$

Будем рассматривать в качестве приближения к решению (9.12) функцию

$$f_l(t, \alpha_э) = \sum_{p=1}^{n(l)} \alpha_э^p \varphi_p(t), \tag{9.13}$$

где $n(l)$ — подходящее число членов разложения (оно будет определено ниже), $\alpha_э = (\alpha_э^1, \ldots, \alpha_э^{n(l)})^T$ — вектор параметров, доставляющий минимум функционалу

$$I_э(\alpha) = \frac{1}{l}\sum_{j=1}^{l}\left(y_j - \sum_{p=1}^{n(l)}\lambda_p\alpha_p\psi_p(x_j)\right)^2. \qquad (9.14)$$

Оказывается, что при определенных предположениях относительно решения (9.12) существует такой закон $n(l)$, что с ростом объема выборки получаемые приближения стремятся по вероятности к решению операторного уравнения (9.1)

Справедливы следующие две теоремы.

Теорема 9.1. *Пусть существует единственное решение операторного уравнения* (9.1). *Тогда, если только функция* $n(l)$ *удовлетворяет условиям:*

$$1) \qquad n(l) \xrightarrow[l\to\infty]{} \infty, \qquad (9.15)$$

$$2) \qquad \frac{1}{\lambda_{n(l)}^2}\sqrt{\frac{n(l)\ln l}{l}} \xrightarrow[l\to\infty]{} 0, \qquad (9.16)$$

то последовательность приближений $f_l(t, \alpha_э)$ *с ростом* l *сходится по вероятности к* $f(t)$ *в метрике* L_2.

Потребуем дополнительно, чтобы оператор A^* действовал из L_2 в C.

Теорема 9.2. *Пусть решение операторного уравнения* (9.1) *таково, что выполняются условия*

$$\sup_t\left|\sum_{p=m}^{\infty}\alpha_p^0\varphi_p(t)\right| = T(m),$$

$$T(m) \xrightarrow[m\to\infty]{} 0. \qquad (9.17)$$

Тогда выполнение условий (9.15), (9.16) *обеспечивает сходимость по вероятности функций* $f_l(t, \alpha_э)$ *к* $f(t)$ *в метрике* C.

Таким образом, теоремы 9.1 и 9.2 утверждают, что если приближать решение операторного уравнения (9.1) разложением по конечному числу собственных функций самосопряженного оператора A^*A, то при правильном (удовлетворяющем условиям (9.16)) выборе числа членов разложения метод минимизации эмпирического риска

(9.14) обеспечит с ростом объема выборки сходимость по вероятности получаемых решений к искомому.

Ниже мы покажем, что при определенных условиях выбор $n(l)$ может быть осуществлен минимизацией правой части (8.49). Тем самым будет показано, что стандартная процедура метода упорядоченной минимизации риска, рассмотренная в главе VIII, приводит к построению последовательности функций, сходящейся к решению операторного уравнения (9.1).

Будем полагать, что помеха ξ_i при измерении функции в правой части операторного уравнения (9.1) определяется симметричной функцией плотности вероятностей $P(\xi)$ и подчиняется условию

$$\tau = \frac{\sqrt[3]{M\xi^6}}{M\xi^2} < \infty. \tag{9.18}$$

Пусть выполнено неравенство

$$\sup_{\alpha} \frac{\sqrt[3]{M\left(y - \sum_{p=1}^{k} \lambda_p \alpha_p \psi_p(x)\right)^6}}{M\left(y - \sum_{p=1}^{k} \lambda_p \alpha_p \psi_p(x)\right)^2} \leqslant \frac{\text{const}}{\lambda_k^2} = \tau_k. \tag{9.19}$$

Неравенство (9.19) следует из (9.18), если

$$y_j = F(x_j, \alpha_0) + \xi_j = \sum_{p=1}^{k} \lambda_p \alpha_p^0 \psi_p(x_j) + \xi_j.$$

В этом случае

$$\left(y_j - \sum_{p=1}^{k} \lambda_p \alpha_p \psi_p(x_j)\right)^2 = \left(\xi_j - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x_j)\right)^2, \tag{9.20}$$

где

$$\beta_p = \alpha_p - \alpha_p^0.$$

Из (9.19) и (9.20) следует

$$T_k = \sup_{\alpha} \frac{\sqrt[3]{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^6}}{M\left(\xi - \sum_{p=1}^{k} \lambda_p \beta_p \psi_p(x)\right)^2}. \tag{9.21}$$

Оценим теперь отдельно знаменатель и числитель правой части равенства (9.21)

$$M\left(\xi-\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^2=\sigma^2+\sum_{p=1}^{k}\lambda_p^2\beta_p^2=\sigma^2+B, \qquad (9.22)$$

где обозначено

$$B=\sum_{p=1}^{k}\lambda_p^2\beta_p^2;$$

$$M\left(\xi-\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^6=M\xi^6+15\,M\xi^4\,M\left(\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^2+$$
$$+15\,M\xi^2\,M\left(\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^4+M\left(\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^6. \qquad (9.23)$$

Справедлива оценка

$$\sup_x\left(\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^2\leqslant\text{const}\,\frac{B}{\lambda_k^2}. \qquad (9.24)$$

Подставляя (9.24) в (9.23) и учитывая, что из условия $\frac{\sqrt[3]{M\xi^6}}{M\xi^2}<\text{const}$ следует $\frac{\sqrt{M\xi^4}}{M\xi^2}<\text{const}$, получим

$$M\left(\xi-\sum_{p=1}^{k}\lambda_p\beta_p\psi_p(x)\right)^6<$$
$$<\text{const}\,[\sigma^6+15\,\sigma^4(B\lambda_k^{-2})+15\,\sigma^2(B\lambda_k^{-2})^2+(B\lambda_k^{-2})^3]<$$
$$<\text{const}\,[\sigma^2+B\lambda_k^{-2}]^3. \qquad (9.25)$$

Подставляя (9.22) и (9.25) в (9.21), окончательно получим

$$T_k\leqslant\text{const}\,\frac{\sigma^2+B\lambda_k^{-2}}{\sigma^2+B}<\frac{\text{const}}{\lambda_k^2}.$$

Итак, будем полагать, что выполнено неравенство (9.19). Согласно же теореме 7.6 в этом случае с вероятностью $1-\eta$ одновременно для всех функций, разложимых по $n\,(n<l)$, собственным векторам системы (9.11) выполнится неравенство

$$I(\alpha)<\left[\frac{I_э(\alpha)}{1-2\sqrt[3]{2}\,\tau_n\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)-\ln\frac{\eta}{12}}{l}}}\right]_\infty. \qquad (9.26)$$

Для каждого объема выборки l будем использовать такое число членов разложения $n\,(l)$, чтобы, во-первых, выпол-

нилось ограничение

$$n(l) < l^{1-\delta}, \tag{9.27}$$

где δ — любая малая величина, а, во-вторых, правая часть неравенства (9.26) достигала минимума. (Здесь появилось дополнительное условие, согласно которому число членов разложения с ростом объема выборки l растет не быстрее $l^{1-\delta}$.)

При таком модифицированном способе определения числа членов разложения оказывается выполненным требование (9.16) теорем 9.1 и 9.2.

Иначе говоря, справедлива

Теорема 9.3. *Пусть решение операторного уравнения* (9.1) *удовлетворяет условию*

$$\left\| \sum_{i=1}^{\infty} \alpha_p^0 \varphi_p(l) \right\|_{L_2} < \infty \tag{9.28}$$

и выполнены условия (9.19) *и* (9.27). *Тогда с помощью упорядоченной минимизации оценки* (9.26) *определяется такое число членов разложения, что с вероятностью единица оказываются выполненными условия:*

1) $n(l) \xrightarrow[l \to \infty]{} \infty$,

2) $\frac{1}{\lambda^2_{n(l)}} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0$.

Итак, теоремы 9.1 и 9.2 указывают на класс алгоритмов, который обеспечивает сходимость последовательности получаемых функций к решению операторного уравнения, а теорема 9.3 утверждает, что метод упорядоченной минимизации риска, заданный с помощью оценки (9.26) на структуре, образованной системой собственных функций, принадлежит этому классу.

В § 6 будут приведены примеры, показывающие эффективность применения метода упорядоченной минимизации риска при интерпретации результатов косвенных экспериментов. Однако хотелось бы здесь отметить, что успехи применения метода при решении некорректных задач интерпретации измерений, вероятно, определяются не тем, что последовательность получаемых решений сходится к искомому решению операторного уравнения (9.1),

а тем, что для каждого конечного l он определяет решение, обладающее экстремальным свойством: образ решения в E_2 доставляет гарантированный минимум величине среднего риска.

§ 4. Доказательство теорем

Докажем сформулированные теоремы.

1. Доказательство теоремы 9.1. Итак, пусть выполнены условия теоремы 9.1. Обозначим через

$$f_l(t, \alpha_э) = \sum_{p=1}^{n(l)} \alpha_э^p \varphi_p(t)$$

прообраз функции

$$F_l(x, \alpha_э) = \sum_{p=1}^{n(l)} \lambda_p \alpha_э^p \psi_p(x),$$

минимизирующей величину эмпирического риска

$$I_э(\alpha) = \frac{1}{l} \sum_{j=1}^{l} \left(y_j - \sum_{p=1}^{n(l)} \lambda_p \alpha_p \psi_p(x_j) \right)^2. \qquad (9.29)$$

Нашей целью является доказательство того, что $f_l(t, \alpha_э)$ сходится по вероятности к $f(t)$ в метрике L_2, или, что то же самое, последовательность случайных величин

$$\nu(l) = \int \left(\sum_{p=1}^{n(l)} \alpha_э^p \varphi_p(t) - \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t) \right)^2 dt \qquad (9.30)$$

стремится по вероятности к нулю с ростом l.

Заметим, что

$$\nu(l) = \sum_{p=1}^{n(l)} \beta_p^2 + \sum_{p=n(l)+1}^{\infty} (\alpha_p^0)^2 = T_1(n(l)) + T_2(n(l)),$$

где $\beta_p = \alpha_э^p - \alpha_p^0$.

Так как решение принадлежит L_2, то с ростом $n(l)$ последовательность $T_2(n(l))$ стремится к нулю. Поэтому для доказательства теоремы 9.1 достаточно показать, что

$$T_1(n(l)) \xrightarrow[l \to \infty]{} 0.$$

Оценим величину

$$T_1(n(l)) = \sum_{p=1}^{n(l)} \beta_p^2. \tag{9.31}$$

Для этого определим вектор $\beta = (\beta_1, \ldots, \beta_{n(l)})^T$, на котором достигается минимум эмпирического риска.

Перепишем (9.29) в виде

$$I_э(\beta) = \frac{1}{l}\sum_{j=1}^{l} \hat{y}_j^2 - 2\sum_{p=1}^{n(l)} \lambda_p \beta_p G_p + \\ + \sum_{p,\,q=1}^{n(l)} \lambda_p \beta_p \lambda_q \beta_q \sum_{j=1}^{l} \frac{\psi_p(x_j)\,\psi_q(x_j)}{l}, \tag{9.32}$$

где обозначено

$$G_p = \frac{1}{l}\sum_{j=1}^{l} \hat{y}_j \psi_p(x_j), \qquad \hat{y}_j = \xi_j + \sum_{p=n+1}^{\infty} \lambda_p \alpha_p^0 \psi_p(x_j).$$

Обозначим через $\|K\|$ ковариационную матрицу, элементы которой K_{pq} равны

$$K_{pq} = \frac{1}{l}\sum_{i=1}^{l} \psi_p(x_i)\,\psi_q(x_i),$$

а через $G - n$-мерный вектор с координатами $G_1, \ldots, G_n$. Тогда вектор $\gamma = (\beta_1\lambda_1, \ldots, \beta_n\lambda_n)^T$, доставляющий минимум (9.32), находится так:

$$\gamma = \|K\|^{-1} G.$$

Поэтому справедлива оценка

$$|\gamma|^2 = |\|K\|^{-1} G|^2 \leqslant |\|K\|^{-1}|^2 |G|^2. \tag{9.33}$$

С другой стороны, справедливо

$$|\gamma|^2 = \sum_{p=1}^{n(l)} (\beta_p\lambda_p)^2 > \lambda_{n(l)}^2 \sum_{p=1}^{n(l)} \beta_p^2 = \lambda_{n(l)}^2 T_1(n(l)). \tag{9.34}$$

Из неравенств (9.33) и (9.34) получаем

$$T_1(n(l)) < \frac{1}{\lambda_{n(l)}^2} |\|K\|^{-1}|^2 |G|^2. \tag{9.35}$$

Таким образом, для доказательства теоремы достаточно оценить сверху норму матрицы $\|K\|^{-1}$ и норму вектора G.

Оценим $\|\|K\|^{-1}\|$. Для этого заметим, что норма матрицы $\|K\|$ не превосходит $\mu_{\max}$, где $\mu_{\max}$ — наибольшее собственное число матрицы, а норма матрицы $\|K\|^{-1}$ не превосходит $1/\mu_{\min}$, где $\mu_{\min}$ — наименьшее собственное число матрицы $\|K\|$.

Оценим снизу величину $\mu_{\min}$. Для этого рассмотрим положительно определенную квадратичную форму

$$F_n(x, \gamma) = \left(\sum_{p=1}^{n} \gamma_p \psi_p(x)\right)^2,$$

которую будем исследовать в области $\sum_{p=1}^{n} \gamma_p^2 \leqslant 1$. Так как вполне непрерывный оператор A из L_2 в C ограничен, $\|A\| < L$, то справедливо неравенство

$$\sup_x \left|\sum_{p=1}^{n} \lambda_p \gamma_p \psi_p(x)\right| \leqslant \|A\| \left\|\sum_{p=1}^{n} \gamma_p \varphi_p(t)\right\| < L \sqrt{\sum_{p=1}^{n} \gamma_p^2},$$

откуда вытекает, что в области $\sum_{p=1}^{n} \gamma_p^2 \leqslant 1$ выполняется неравенство

$$\sup_{x\gamma} \left|\sum_{p=1}^{n} \gamma_p \psi_p(x)\right| < L \sqrt{\sum_{p=1}^{n} \frac{\gamma_p^2}{\lambda_p^2}} \leqslant \frac{L}{\lambda_n}$$

и, следовательно,

$$\sup_x F_n(x, \gamma) < \frac{L^2}{\lambda_n^2}. \tag{9.36}$$

Рассмотрим теперь выражение

$$\frac{1}{l} \sum_{i=1}^{l} F_n(x_i, \gamma) = \frac{1}{l} \sum_{i=1}^{l} \left(\sum_{p=1}^{n} \gamma_p \psi_p(x_i)\right)^2.$$

Заметим, что

$$MF_n(x, \gamma) = \sum_{p=1}^{n} \gamma_p^2, \tag{9.37}$$

$$\frac{1}{l} \sum_{i=1}^{l} F_n(x_i, \gamma) = \sum_{p, q=1}^{n} \gamma_p \gamma_q K_{pq}.$$

С помощью преобразования поворота перейдем к новой дважды ортогональной системе функций $\psi_1'(x), \ldots, \psi_n'(x)$ такой, что

$$MF_n(x,\ \gamma') = \sum_{p=1}^{n} (\gamma_p')^2, \tag{9.38}$$

$$\frac{1}{l}\sum_{i=1}^{l} F_n(x_i,\ \gamma') = \sum_{p=1}^{n} \mu_p (\gamma_p')^2,$$

где $\mu_1, \ldots, \mu_n$ — собственные числа матрицы $\|K\|$.

Для оценки собственных чисел воспользуемся теоремой о равномерной сходимости средних к их математическим ожиданиям для класса ограниченных функций (теорема 7.3).

Так как функции $F(x, \gamma')$ при $\|\gamma'\| \leqslant 1$ ограничены величиной L^2/λ_n^2, то справедлива оценка (см. § 3 гл. VII)

$$P\left\{\sup_{\gamma'}\left|MF_n(x,\ \gamma') - \frac{1}{l}\sum_{i=1}^{l} F_n(x_i,\ \gamma')\right| > \varkappa\frac{L^2}{\lambda_n^2}\right\} < 9\frac{(2l)^n}{n!}e^{-\frac{\varkappa^2 l}{4}},$$

или, учитывая (9.38), получаем

$$P\left\{\sup_{\gamma'}\left|\sum_{p=1}^{n}(\gamma_p')^2(1-\mu_p)\right| > \varkappa\frac{L^2}{\lambda_n^2}\right\} < 9\frac{(2l)^n}{n!}e^{-\frac{\varkappa^2 l}{4}}. \tag{9.39}$$

Потребуем, чтобы вероятность не превосходила $9/\ln l$. Для этого достаточно, чтобы $\varkappa$ было не меньше величины

$$\varkappa^* = \frac{2L^2}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)+\ln\ln l}{l}}. \tag{9.40}$$

Из (9.39) и (9.40) следует, что с вероятностью $1-\frac{9}{\ln l}$ все собственные числа $\mu_1, \ldots, \mu_n$ находятся в интервале

$$1-\varkappa^* \leqslant \mu_i \leqslant 1+\varkappa^*, \tag{9.41}$$

откуда заключаем, что с вероятностью $1-\frac{9}{\ln l}$ выполняется неравенство

$$\mu > 1-\varkappa^*. \tag{9.42}$$

Подставляя (9.42) в (9.35), получим, что с вероятностью $1-\frac{9}{\ln l}$ справедливо

$$T_1(n) < \frac{|G|^2}{\lambda_n^2\left(1-\frac{2L^2}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)+\ln\ln l}{l}}\right)^2}. \quad (9.43)$$

Нам осталось оценить величину $|G|^2$:

$$|G|^2 = \sum_{p=1}^{n} G_p^2 = \sum_{p=1}^{n} \frac{1}{l^2}\left(\sum_{i=1}^{l} \hat{y}_i \psi_p(x_i)\right)^2.$$

Для этого найдем математическое ожидание

$$M|G|^2 = M\sum_{p=1}^{n} G_p^2 \leqslant \frac{\sigma^2+\|A\|^2 T_2(0)}{l} n = R\cdot\frac{n}{l},$$

где T и R — константы, не зависящие от l и n. Для оценки величины $|G|$ используем неравенство Чебышева для первого момента положительной величины ξ:

$$P\{\xi > \varepsilon\} < \frac{M\xi}{\varepsilon},$$

где потребуем $\varepsilon = \frac{R\cdot n\ln l}{l}$. А так как $M|G|^2 < \frac{Rn}{l}$, то получаем

$$P\left\{|G|^2 > \frac{Rn\ln l}{l}\right\} < \frac{1}{\ln l}.$$

Таким образом, с вероятностью $1-\frac{1}{\ln l}$

$$|G|^2 \leqslant \frac{Rn\ln l}{l}. \quad (9.44)$$

Подставляя (9.44) в (9.43), получаем что для достаточно больших l с вероятностью $1-\frac{10}{\ln l}$ выполняется неравенство

$$T_1(n) < c\frac{n\ln l}{l\lambda_n^2\left(1-\frac{c}{\lambda_n^2}\sqrt{\frac{n\ln l}{l}}\right)^2}, \quad (9.45)$$

где c — некоторая константа.

Из неравенства (9.45) следует, что $T_1(n(l))$ стремится по вероятности к нулю при

$$\frac{1}{\lambda^2_{n(l)}}\sqrt{\frac{n(l)\ln l}{l}} \xrightarrow[l\to\infty]{} 0.$$

Теорема доказана.

2. Доказательство теоремы 9.2. Пусть теперь решение операторного уравнения (9.1) подчинено дополнительному условию

$$\sup_t \left| \sum_{p=n}^{\infty} \alpha_p \varphi_p(t) \right| \xrightarrow[n\to\infty]{} 0. \tag{9.46}$$

Покажем, что в этом случае условия

$$n(l) \xrightarrow[l\to\infty]{} \infty,$$

$$\frac{1}{\lambda^2_{n(l)}}\sqrt{\frac{n(l)\ln l}{l}} \xrightarrow[l\to\infty]{} 0 \tag{9.47}$$

являются достаточными условиями того, что последовательность решений $f_l(t, \alpha_э)$ сходится по вероятности к решению операторного уравнения (9.1) в метрике C. Обозначим через

$$\nu(l) = \sup_t \left| \sum_{p=1}^{\infty} \alpha_p^0 \varphi_p(t) - \sum_{p=1}^{n(l)} \alpha_э^p \varphi_p(t) \right|,$$

где $\alpha_э = (\alpha_э^1, \ldots, \alpha_э^{n(l)})^T$ — вектор, доставляющий минимум (9.29). Нашей целью является доказательство того, что

$$\nu(l) \xrightarrow[l\to\infty]{p} 0.$$

Заметим, что

$$\nu(l) \leqslant \sup_t \left| \sum_{p=1}^{n(l)} \beta_p \varphi_p(t) \right| + \sup_t \left| \sum_{p=n(l)+1}^{\infty} \alpha_p^0 \varphi_p(t) \right|, \tag{9.48}$$

где обозначено $\beta_p = \alpha_p^0 - \alpha_l^э$.

Так как по условию теоремы (9.46) второе слагаемое суммы (9.48) стремится к нулю с ростом l, то достаточно показать, что

$$T_э(n(l)) = \sup_t \left| \sum_{p=1}^{n(l)} \beta_p \varphi_p(t) \right| \xrightarrow[l\to\infty]{} 0. \tag{9.49}$$

Для доказательства этого факта воспользуемся оценкой

$$T_3^2(n(l)) < \frac{\text{const}}{\lambda_n^2} \sum_{p=1}^{n} \beta_p^2, \tag{9.50}$$

справедливость которой вытекает из ограниченности оператора A^*.

При доказательстве теоремы 9.1 было показано, что с вероятностью $1 - \frac{10}{\ln l}$ справедлива оценка

$$T_1(n) = \sum_{p=1}^{n} \beta_p^2 < \frac{\text{const}\, n \ln l}{l\lambda_n^2 \left(1 - \frac{\text{const}}{\lambda_n^2} \sqrt{\frac{n \ln l}{l}}\right)^2}.$$

Подставляя ее в (9.50), получим, что с вероятностью $1 - \frac{10}{\ln l}$ справедлива оценка

$$T_3^2(n(l)) < \frac{\frac{\text{const}}{\lambda_n^4} \frac{n \ln l}{l}}{\left(1 - \frac{\text{const}}{\lambda_n^2} \sqrt{\frac{n \ln l}{l}}\right)^2}. \tag{9.51}$$

Из оценки (9.51) следует, что $T_3^2(n)$ стремится по вероятности к нулю, если

$$\frac{1}{\lambda_{n(l)}^2} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0.$$

Теорема 9.2 доказана.

3. Доказательство теоремы 9.3. Пусть число $n(l)$ членов разложения решения операторного уравнения определяется минимальным значением оценки (9.26). Покажем, что если решение операторного уравнения $f(t)$ удовлетворяет условию

$$\left\| \sum_{p=1}^{\infty} \alpha_p^0 \varphi_0(t) \right\|_{L_2} < \infty, \tag{9.52}$$

то рассмотренный алгоритм задания числа членов разложения удовлетворяет условиям:

1) $$n(l) \xrightarrow[l \to \infty]{} \infty, \tag{9.53}$$

2) $$\frac{1}{\lambda_n^2} \sqrt{\frac{n(l) \ln l}{l}} \xrightarrow[l \to \infty]{} 0. \tag{9.54}$$

Покажем справедливость условия (9.53). Предположим противное. Пусть $\alpha_n^0 \neq 0$, $r < n$, а вместе с тем для любого $l > l_0$ выполняется неравенство

$$\frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{r}\lambda_p\alpha_э^p\psi_p(x_i)\right)^2}{1-\frac{c}{\lambda_r^2}\sqrt{\frac{r\left(\ln\frac{2l}{r}+1\right)-\ln\frac{\eta}{12}}{l}}} <$$

$$< \frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{n}\lambda_p\alpha_э^p\psi_p(x_i)\right)^2}{1-\frac{c}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)-\ln\frac{\eta}{12}}{l}}}. \qquad (9.55)$$

Согласно теореме 7.6 для достаточно больших l с вероятностью $1-\eta$ справедливо

$$I(\alpha_э, r) < \frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{p=1}^{n}\lambda_p\alpha_э^p\psi_p(x_i)\right)^2}{1-\frac{c}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)-\ln\frac{\eta}{12}}{l}}}.$$

Представим величину $I(\alpha_э, r)$ в виде

$$I(\alpha_э, r) = M\left(y - \sum_{p=1}^{r}\lambda_p\alpha_э^p\psi_p(x)\right)^2 =$$

$$= M\left(\xi + \Delta(x, r) - \sum_{p=1}^{r}\beta_p\lambda_p\psi_p(x)\right)^2,$$

где

$$\Delta(x, r) = \sum_{p=r+1}^{\infty}\lambda_p\alpha_p^0\psi_p(x), \quad \beta_p = \alpha_э^p - \alpha_p^0,$$

и оценим ее:

$$I(\alpha_э, r) > I(\alpha_0, r) = \sigma^2 + M\Delta^2(x, r) \geqslant \sigma^2 + \sum_{p=r+1}^{\infty}(\alpha_p^0\lambda_p)^2.$$

Таким образом, с вероятностью $1-\eta$ должна быть справедлива оценка

$$\sigma^2+\sum_{p=r+1}^{\infty}(\alpha_p^0\lambda_p)^2<\frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i-\sum_{p=1}^{n}\alpha_э^p\lambda_p\psi_p(x_i)\right)^2}{1-\frac{c}{\lambda_n^2}\sqrt{\frac{n\left(\ln\frac{2l}{n}+1\right)-\ln\frac{\eta}{12}}{l}}}. \quad (9.56)$$

Преобразуем и оценим выражение в числителе правой части (9.56):

$$I_э(\alpha_э,\ n)=\frac{1}{l}\sum_{i=1}^{l}\left(y_i-\sum_{p=1}^{n}\lambda_p\alpha_э^p\psi_p(x_i)\right)^2=$$

$$=\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i+\Delta(x_i,\ n)-\sum_{p=1}^{n}\lambda_p\beta_p\psi_p(x_i)\right)^2\leqslant$$

$$\leqslant\frac{1}{l}\sum_{i=1}^{l}(\xi_i+\Delta(x_i,\ n))^2=$$

$$=\frac{1}{l}\sum_{i=1}^{l}\xi_i^2+\frac{1}{l}\sum_{i=1}^{l}\Delta^2(x_i,\ n)+\frac{2}{l}\sum_{i=1}^{l}\xi_i\Delta(x_i,\ n).$$

Заметим, что, согласно закону больших чисел,

$$\frac{1}{l}\sum_{i=1}^{l}\xi_i^2\xrightarrow[l\to\infty]{}\sigma^2;\quad\frac{1}{l}\sum_{i=1}^{l}\xi_i\Delta(x_i,\ n)\xrightarrow[l\to\infty]{}0,$$

$$\frac{1}{l}\sum_{i=1}^{l}\Delta^2(x_in)\xrightarrow[l\to\infty]{}\sum_{p=n+1}^{\infty}(\lambda_p\alpha_p^0)^2.$$

Поэтому для достаточно больших l с вероятностью $1-\eta$ должно выполниться неравенство

$$\sigma^2+\sum_{p=r+1}^{\infty}(\lambda_p\alpha_p^0)^2<\sigma^2+\sum_{p=n+1}^{\infty}(\lambda_p\alpha_p^0)^2. \quad (9.57)$$

Однако для $r<n$ неравенство (9.57) неверно с вероятностью единица. Полученное противоречие доказывает справедливость условия (9.53).

Покажем теперь, что справедливо и условие (9.54). Для этого заметим, что всегда выполняются неравенства

$$\frac{1}{l}\sum_{i=1}^{l}(\xi_i+\Delta(x_i,\ n))^2>\min_{\alpha} I_э(\alpha,\ n)>$$

$$>\min_{\alpha,\ \gamma}\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i-\sum_{p=1}^{n}\lambda_p\alpha_p\psi_p(x_i)-\gamma\Delta(x_i,\ n)\right)^2. \quad (9.58)$$

Найдем математическое ожидание левой части неравенства (9.58)

$$M\left\{\frac{1}{l}\sum_{i=1}^{l}(\xi_i+\Delta(x_i,\ n))^2\right\}=\sigma^2+\sum_{p=r+1}^{\infty}(\alpha_p\lambda_p)^2=\sigma^2+T_1(n).$$

Заметим, что при фиксированном числе n справедливо

$$\frac{1}{\lambda_n^2}\sqrt{\frac{n\ln l}{l}}\xrightarrow[l\to\infty]{}0.$$

Поэтому выполняется неравенство

$$\lim_{l\to\infty}\frac{I_э(\alpha_э,\ r)}{1-\dfrac{\text{const}}{\lambda_r^2}\sqrt{\dfrac{r\left(\ln\dfrac{2l}{r}\right)-\ln\dfrac{\eta}{12}}{l}}}<\sigma^2+T_1(r). \quad (9.59)$$

Так как неравенство (9.59) справедливо для любого r, выполнено условие $T_1(r)\xrightarrow[r\to\infty]{}0$ и условие $n(l)\xrightarrow[l\to\infty]{}\infty$, то имеет место неравенство

$$\lim_{l\to\infty}\min_{r<l^{1-\delta}}\frac{I_э(\alpha_э,\ r)}{1-\dfrac{\text{const}}{\lambda_r^2}\sqrt{\dfrac{r\ln l}{l}}}\leqslant\sigma^2.$$

С другой стороны, воспользуемся оценками среднего и дисперсии:

$$MI_э(\alpha_э,\ \gamma_э,\ r)=M\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i-\sum_{p=1}^{l}\lambda_p\alpha_p^э\psi_p(x_i)-\gamma_э\Delta(x_i,\ r)\right)^2=$$

$$=\sigma^2\left(1-\frac{r+1}{l}\right), \quad (9.60)$$

$$D[I_э(\alpha_э,\ \gamma_э,\ r)]=M(I_э(\alpha_э,\ \gamma_э,\ r)-MI_э(\alpha_э,\ \gamma_э,\ r))^2<$$

$$<\frac{M\xi^4+\sigma^4}{l^2}(r+1)=R\frac{r+1}{l^2}, \quad (9.61)$$

$\alpha_э$, $\gamma_э$ — значения параметров, доставляющих минимум $I_э(\alpha, \gamma, r)$.

Справедливость этих оценок мы покажем ниже.

Используем неравенство Чебышева

$$P\left\{\left|I_э(\alpha_э, \gamma_э; n(l)) - \sigma^2\left(1 - \frac{n(l)+1}{l^2}\right)\right| > \varepsilon\right\} < \frac{R}{l^2\varepsilon^2}(n(l)+1).$$

Согласно же условию теоремы $n(l) < l^{1-\delta}$. Поэтому

$$\sum_{l=1}^{\infty} P\left\{\left|\sigma^2\left(1 - \frac{n(l)+1}{l}\right) - I_э(\alpha_э, \gamma_э, n(l))\right| > \varepsilon\right\} <$$

$$< R\sum_{l}^{\infty} \frac{l^{1-\delta}+1}{l^2\varepsilon^2} < \infty,$$

и, следовательно, согласно лемме Бореля — Кантелли (см. § 2), с вероятностью единица имеет место сходимость

$$\lim_{l\to\infty} I_э(\alpha_э, \gamma_э, n(l)) = \sigma^2.$$

Таким образом, с вероятностью единица выполняются неравенство

$$\lim_{l\to\infty} \min_{n(l)} \frac{I_э(\alpha_э, n(l))}{1 - \frac{\text{const}}{\lambda^2_{n(l)}}\sqrt{\frac{n(l)\ln l}{l}}} \leqslant \sigma^2$$

и равенство

$$\lim_{l\to\infty} I_э(\alpha_э, \gamma_э, n(l)) = \sigma^2,$$

откуда следует, что с вероятностью единица

$$\lim_{l\to\infty} \frac{\text{const}}{\lambda^2_{n(l)}}\sqrt{\frac{n(l)\ln l}{l}} = 0. \qquad (9.62)$$

Выражение (9.62) и составляет содержание теоремы 9.3. При доказательстве теоремы мы использовали равенство (9.60) и неравенство (9.61). Получим их:

$$MI_э(\alpha_э, \gamma_э, r) = M\frac{1}{l}\sum_{i=1}^{l}\left(\xi_i - \sum_{p=1}^{r}\lambda_p\alpha_p^э\psi_p(x_i) - \gamma_э\Delta(x, r)\right)^2.$$

С помощью преобразования поворота перейдем к системе координат $\psi'_1(x), \ldots, \psi'_r(x), \psi'_{r+1}(x)$ такой, что

$$\frac{1}{l}\sum_{i=1}^{l}\psi'_p(x_i)\psi'_q(x_i)=\begin{cases}\mu_p, & \text{если } p=q,\\ 0, & \text{если } p\neq q.\end{cases}$$

В этой системе координат

$$I_э(\alpha_э, \gamma_э, r)=\frac{1}{l}\sum_{i=1}^{l}\xi_i^2-\sum_{p=1}^{r+1}\frac{G_p^2}{\mu_p},$$

где обозначено

$$G_p=\frac{1}{l}\sum_{i=1}^{l}\xi_i\psi'_p(x_i).$$

Таким образом, получим

$$MI_э(\alpha_э, \gamma_э, r)=$$

$$=\sigma^2-\sum_{p=1}^{r+1}M\sum_{i,j=1}^{l}\frac{\xi_i\xi_j\psi'_p(x_i)\psi'_p(x_j)}{l^2\mu_p}=\sigma^2\left(1-\frac{r+1}{l}\right),$$

$$D[I_э(\alpha_э, \gamma_э, r)]=$$

$$=\sum_{p=1}^{r+1}\left[M\left(\frac{G_p^2}{\mu_p}\right)^2-\left(M\frac{G_p^2}{\mu_p}\right)^2\right]\leqslant\frac{r+1}{l}R.$$

Теорема доказана.

§ 5. Методы полиномиального и кусочно-полиномиального приближений

Итак, с помощью метода упорядоченной минимизации риска может быть получена последовательность приближений, сходящаяся с ростом числа измерений к искомому решению операторного уравнения.

Однако сходимость гарантируется лишь при условии, что приближения ищутся в виде разложений по собственным функциям оператора A^*A. Отыскание же собственных функций оператора A^*A — задача не всегда простая. Поэтому хотелось бы заменить разложение решения по собственным функциям оператора разложением по другой системе функций.

В этом параграфе мы рассмотрим два типа приближений — полиномиальные и кусочно-полиномиальные.

Основное свойство полиномиальных приближений, сформулированное в теореме Вейерштрасса, состоит в том, что любая непрерывная на отрезке $[a, b]$ функция может быть сколь угодно точно приближена в равномерной метрике полиномом. В этой книге в качестве приближения к функции $y(x)$ выбирается функция, минимизирующая в $F(x, \alpha)$ функционал

$$I(\alpha) = \int (y(x) - F(x, \alpha))^2 dx. \tag{9.63}$$

Возникает вопрос: для всякой ли непрерывной функции $y(x)$ последовательность

$$F(x, \alpha_1^0), \ldots, F(x, \alpha_r^0), \ldots \tag{9.64}$$

полиномов степени $r = 0, 1, 2, \ldots$, каждый из которых доставляет минимум (9.63) в классе полиномов соответствующей степени, сходится к $y(x)$ в равномерной метрике?

Оказывается, нет, не для всякой. Известна теорема Лозинского — Харшиладзе [30], согласно которой существует такая непрерывная функция $y(x)$, к которой последовательность (9.64) равномерно не сходится.

Таким образом, идея минимизации среднеквадратичного уклонения для получения равномерного полиномиального приближения непрерывной функции оказывается неприемлемой. Из этого факта немедленно вытекает, что получение равномерного приближения к регрессии путем минимизации среднего риска невозможно, если приближение ведется в классе полиномов.

Возможность построения равномерного приближения к регрессии в схеме минимизации среднего риска связана с *кусочно-полиномиальными приближениями* или, как их еще называют, *сплайн-приближениями*.

Рассмотрим кусочно-полиномиальные приближения функции на отрезке $[a, b]$. Разобьем отрезок $[a, b]$ на N частей точками $a = a_0, a_1, \ldots, a_{N+1} = b$. На каждом интервале $[a_i, a_{i+1}]$ будем приближать функцию $y(x)$ полиномом фиксированной степени m. Таким образом, функция приближается с помощью $N+1$ кусков полиномов (каждый для своего интервала). Полиномы выбираются так, чтобы в точках $a_1, \ldots, a_N$ полученное приближение было непрерывно вместе со своей $m-1$ производной. Назовем такое

кусочно-полиномиальное приближение сплайнами степени m, сопряженными на сетке $(a_1, \ldots, a_N)$. Будем считать, что точки сопряжения фиксированы и заданы равномерно на $[a, b]$ $\left(a_i = a_0 + \Delta i,\ \Delta = \frac{b-a}{N}\right)$.

Обозначим через $V_N^m(x, \alpha)$ класс сплайнов степени m с N сопряжениями, заданный на равномерной сетке, а через $V_N^m(x, \alpha_э)$ сплайн, доставляющий минимум величине эмпирического функционала

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - V_N^m(x_i\alpha))^2. \tag{9.65}$$

Пусть теперь определено условие, связывающее число сопряжений N с объемом выборки l, а именно $N = N(l)$.

Рассмотрим последовательность сплайнов

$$V_{N(1)}^m(x, \alpha_э), \ldots, V_{N(l)}^m(x, \alpha_э) \ldots \tag{9.66}$$

степени m, имеющих $N(1), \ldots, N(l), \ldots$ сопряжений и минимизирующих эмпирический риск на выборке $i = 1, 2, \ldots$ $\ldots l \ldots$ (выборка образована случайно и независимо согласно плотности $P(x, y) = P(y \mid x) P(x)$).

Справедлива

Теорема 9.4 (Михальский). *Пусть регрессия определяется непрерывной функцией $y(x)$. Тогда последовательность* (9.66) *с вероятностью единица сходится в равномерной метрике к регрессии $y(x)$, если только плотность $P(x)$ абсолютно непрерывна относительно равномерной и выполнены условия*

$$N(l) \xrightarrow[l\to\infty]{} \infty,$$

$$\frac{N^4(l) \ln l}{l} \xrightarrow[l\to\infty]{} 0.$$

Если же, кроме того, будут выполнены более сильные условия

$$\frac{N^{2(2+p)}(l) \ln l}{l} \xrightarrow[l\to\infty]{} 0, \tag{9.67}$$

а регрессия $y(x)$ непрерывна вместе со своими p производными, то последовательность

$$[V_{N(1)}^m(x, \alpha_э)]^{(p)}, \ldots, [V_{N(l)}^m(x, \alpha_э)]^{(p)} \ldots,$$

составленная из p-х производных сплайнов (9.66), *сходится с вероятностью единица в равномерной метрике к функции* $y^{(p)}(x)$, *являющейся p-й производной регрессии.*

Замечание. Из теоремы следует, что выполнение условия (9.67) гарантирует восстановление в классе сплайнов p-й непрерывной производной функции $F(x)$ по значениям этой функции, измеренным в l случайно выбранных точках (l — достаточно большое число), т. е. отыскание приближенного решения интегрального уравнения

$$\int_a^b \frac{(x-t)^{p-1}}{(p-1)!}\,\theta(x-t)\,f(t)\,dt = F(x) - \sum_{k=0}^{p-1} \frac{F^{(k)}(a)}{k!}$$

по измерениям $y_i = F(x_i) + \xi_i$ $(i = 1, 2, \ldots, l)$.

Ниже при интерпретации результатов экспериментов мы будем искать решение в разложении по сплайнам.

§ 6. Методы решения некорректных задач измерения

В этом параграфе мы приведем примеры использования метода упорядоченной минимизации риска для восстановления решения линейного операторного уравнения

$$Af(t) = F(x) \tag{9.68}$$

по эмпирическим данным $x_1, y_1; \ldots; x_l, y_l$ $(y_i = F(x_i) + \xi_i$, x — случайная величина, распределенная по равномерному закону на $[a, b]$). Восстановление проводится в классе сплайнов.

В главе XII будет показано, что любой сплайн $V_N^m(t, \alpha)$ порядка m с N сопряжениями представим как линейная комбинация системы $N+m+1$ фундаментальных сплайнов степени m с N сопряжениями

$$\pi_1(t), \ldots, \pi_{N+m+1}(t). \tag{9.69}$$

Иначе говоря, справедливо

$$V_N^m(t, \alpha) = \sum_{i=1}^{N+m+1} \alpha_i \pi_i(t),$$

где $\alpha = (\alpha_1, \ldots, \alpha_{N+m+1})$ — коэффициенты, задающие конкретные кусочно-полиномиальные приближения в классе сплайнов степени m с N сопряжениями.

При построении сплайн-приближения решения уравнения (9.68) проблема состоит в том, чтобы определить, во-первых, подходящее число N точек сопряжения сплайна, а во-вторых, коэффициенты разложения $\alpha_1, \ldots, \alpha_{N+m+1}$.

Рассмотрим образы фундаментальной системы (9.69) в E_2

$$\mu_1(x) = A\pi_1(t), \ \ldots, \ \mu_{N+m+1}(x) = A\pi_{N+m+1}(t)$$

и примем в качестве решения операторного уравнения (9.68) такой сплайн $V_N^m(t, \alpha^*)$, образ которого $F(x, \alpha^*)$ гарантирует малую величину риску:

$$I(\alpha) = \int (y - F(x, \alpha))^2 P(y \mid x)\, dy\, dx.$$

Согласно теореме 7.6 с вероятностью $1-\eta$ одновременно для всех сплайнов с N сопряжениями выполнится неравенство

$$I(\alpha) < \left[\frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i - \sum_{j=1}^{N+m+1}\alpha_j\mu_j(x_i)\right)^2}{1 - 2\sqrt[3]{2}\,\tau_N\sqrt{\frac{(N+m+1)\left(\ln\frac{2l}{N+m+1}+1\right) - \ln\frac{\eta}{12}}{l}}} \right]_{\infty}.$$

В качестве решения операторного уравнения выберем такую сплайн-функцию (т. е. такое число сопряжений N и такое α), для которой достигается минимум правой части этого неравенства. Несмотря на то, что сходимость приближений, найденных методом упорядоченной минимизации риска, к решению операторного уравнения доказана лишь для разложения по собственным функциям, примеры успешного решения практических задач интерпретации результатов косвенных экспериментов в классе сплайнов позволяют рекомендовать и это разложение для решения интегральных уравнений Фредгольма I рода.

1. Задача ядерной спектроскопии. На вход измерительного прибора поступает энергия, распределенная по частоте $f(t)$ (t — частота). На выходе прибора наблюдается экспериментальный спектр $F(x)$. Связь между входом и выходом задается уравнением

$$\int_a^b \left[1 - \frac{t}{x}\right]_+ f(t)\, dt = F(x),$$

где a, b — границы излучаемого спектра

$$[z]_+ = \begin{cases} z, & \text{если } z \geqslant 0, \\ 0, & \text{если } z < 0. \end{cases}$$

Требуется по наблюдениям восстановить $f(t)$.

На рис. 8 показаны измерения функции $F(x)$ (каждый второй замер). Всего было проведено 40 измерений. Измерения осуществля-

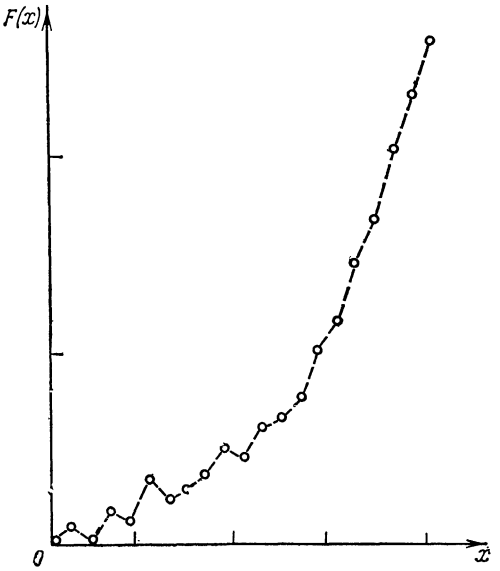

Рис. 8.

лись с равномерно распределенной помехой, заданной на интервале $[-c, c]$. Величина c бралась равной 2% от максимума $F(x)$.

На рис. 9 показан истинный спектр (жирная линия) и сплайн-приближение, полученное методом упорядоченной минимизации риска.

2. Обратная задача гравиметрии. Интегральное уравнение

$$\frac{2}{(\rho_1-\rho_2)\pi}\int_a^b \frac{Hf(t)}{H^2+(x-t)^2}\,dt = F(x)$$

описывает аномалию силы тяжести на поверхности Земли, созданную массой плотности ρ_1, отделенную от окружающей среды с плотностью ρ_2, границей $f(t)$, H — глубина залегания массы, вызывающей аномалию. Требуется по измерениям аномалий $F(t)$ восстановить границу $f(t)$.

На рис. 10 показано истинное (жирная линия) и полученное методом упорядоченной минимизации риска сплайн-приближение (тон-

кая линия). Решение получено по 40 измерениям, проводимым с равномерной помехой, амплитуда которой составила 12% от максимума $F(x)$.

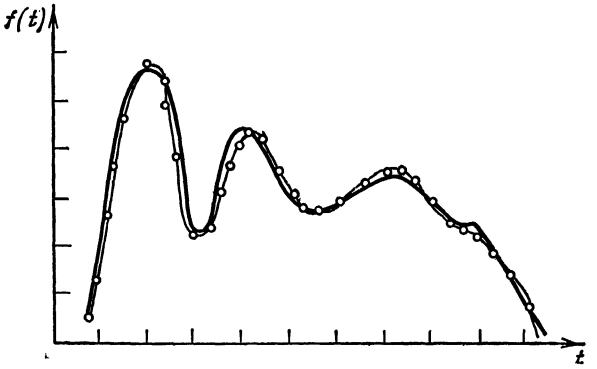

Рис. 9.

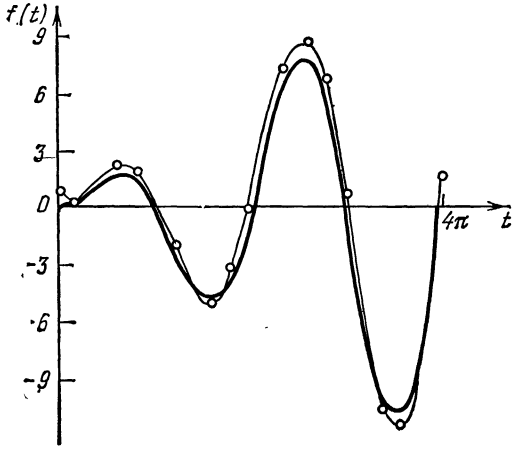

Рис. 10.

3. Задача восстановления производных. Задача восстановления n-й производной в классе непрерывных функций сводится к решению следующего интегрального уравнения:

$$\int_a^b \frac{[x-t]_+^{n-1}}{(n-1)!} f(t)\,dt = F(x) - \sum_{j=0}^{n-1} \frac{F^{(j)}(a)}{j!}.$$

Ниже приведены решения этой задачи для $n=1$, 2, 3 в случае, когда функция $F(x)$ измерена в 40 точках.

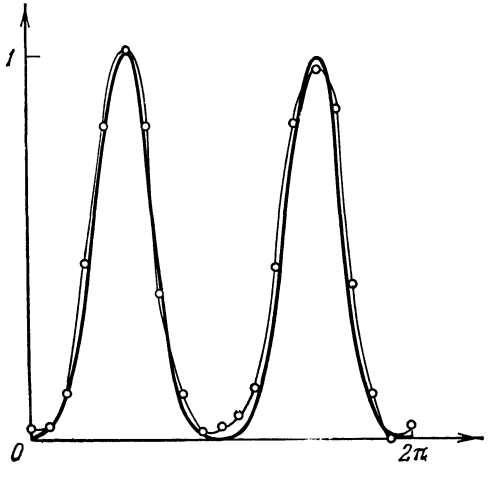

Рис. 11.

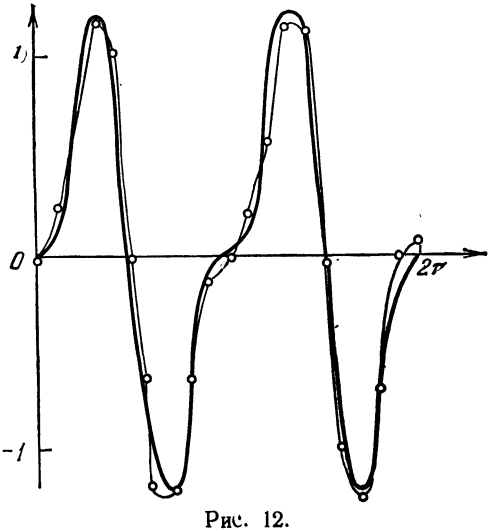

Рис. 12.

На рис. 11 показана функция $F(t)$ (жирная линия) и ее измерения (показан каждый второй замер функции). Измерения функции

проводились с помехой, распределенной согласно равномерному закону с амплитудой 5% от максимума $F(x)$.

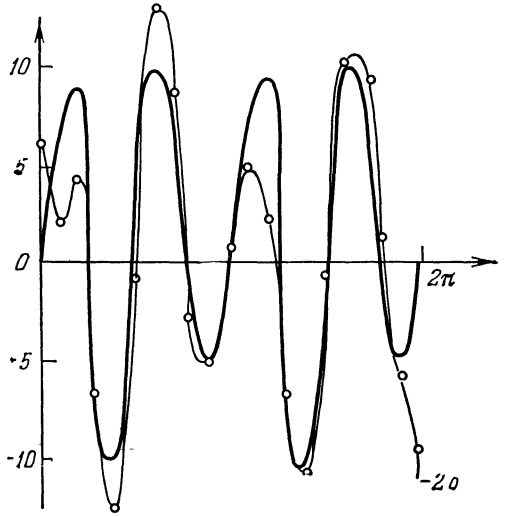

Рис. 13.

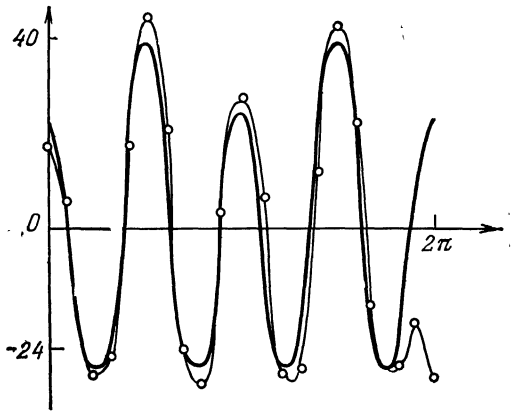

Рис. 14.

На рис. 12, 13, 14 показаны первая, вторая и третья производные функции $F(x)$ (жирные линии) и соответствующие сплайн-приближения, найденные методом упорядоченной минимизации риска

Примеры решались с помощью алгоритма 12-3, приведенного в главе XII.

§ 7. Проблема восстановления плотности распределения вероятностей

В главе II проблема восстановления плотности распределения вероятностей в классе непрерывных на $[a, b]$ функций была связана с решением некорректной задачи численного дифференцирования. Согласно определению плотность распределения вероятностей $f(t)$ есть производная от функции распределения вероятностей $F(x) = P(t \leqslant x)$, т. е. является решением уравнения

$$\int_a^b \theta(x-t) f(t)\, dt = F(x). \tag{9.70}$$

Поэтому задачу восстановления плотности $f(t)$ по эмпирическим данным $t_1, \ldots, t_l$ следует рассматривать как задачу приближенного решения интегрального уравнения (9.70), правая часть которого задана неточно: вместо функции распределения $F(x)$ дана ее оценка

$$F_l(x) = \frac{1}{l} \sum_{i=1}^{l} \theta(x - t_i).$$

Будем решать эту задачу методом регуляризации (см. приложение к гл. I). Выпишем функционал

$$R_{\gamma_l}(f, F_l) = \rho^2_{E_2}(Af, F_l) + \gamma_l \Omega(f), \tag{9.71}$$

где A — оператор уравнения (9.70), γ_l — константа регуляризации $\gamma_l \to 0$ при $l \to \infty$.

Рассмотрим последовательность элементов

$$f_1^{\gamma_1}(t), \ldots, f_l^{\gamma_l}(t), \ldots, \tag{9.72}$$

минимизирующих (9.71) при $l \to \infty$.

Эта последовательность является случайной, так как она образована с помощью случайных функций $F_l(x)$.

В приложении к главе доказана теорема, утверждающая, что если искомое решение операторного уравнения принадлежит некоторому компакту $\Omega(f) \leqslant c$, то для любых ν и μ существует такое $n(\mu, \nu)$, что, начиная с $l > n(\mu, \nu)$ для всех элементов (9.72) выполнится неравенство

$$P\{\rho_{E_1}(f_l^{\gamma_l}, f) > \nu\} \leqslant P\{\rho^2_{E_2}(F_l, F) > \mu\gamma_l\}. \tag{9.73}$$

Воспользуемся этим неравенством для определения условий, обеспечивающих сходимость последовательности (9.72) к искомой плотности.

Рассмотрим асимптотическую оценку скорости сходимости эмпирической функции распределения к истинной (оценка Колмогорова — Смирнова)

$$P\left\{\sup_x |F_l(x) - F(x)| > \varepsilon\right\} < 2e^{-2\varepsilon^2 l}. \qquad (9.74)$$

Пусть теперь $\rho_{E_2}(F_l, F) = \sup_x |F_l(x) - F(x)|$. Тогда из (9.73) и (9.74) получим

$$P\left\{\rho_{E_1}\left(f_l^{\gamma_l}, f\right) > \nu\right\} < 2e^{-\mu l \gamma_l}.$$

Из этого неравенства следует, что для того, чтобы последовательность (9.72) сходилась по вероятности в метрике пространства E_1 к истинной плотности, достаточно, чтобы

$$\begin{aligned} &\gamma_l \xrightarrow[l\to\infty]{} 0, \\ &l\gamma_l \xrightarrow[l\to\infty]{} \infty, \end{aligned} \qquad (9.75)$$

а для того, чтобы последовательность сходилась с вероятностью единица, согласно лемме Бореля — Кантелли достаточно, чтобы хотя бы для одного μ выполнилось неравенство

$$\sum_{l=1}^{\infty} e^{-\mu l \gamma_l} < \infty. \qquad (9.76)$$

Используя в (9.71) различные стабилизирующие функционалы $\Omega(f)$, можно получать оценки $f_l^{\gamma_l}$, сходящиеся к искомой плотности в различных метриках.

Итак, мы установили, что если плотность принадлежит компакту $\Omega(f) \leqslant c$, то можно подобрать такие γ_l, чтобы последовательность (9.72) сходилась к искомой плотности.

Требование о принадлежности искомой плотности компакту можно снять.

В приложении к главе показано, что можно получать последовательность решений, сходящуюся к непрерывной плотности, — достаточно в качестве стабилизирующего функционала взять $\Omega(f) = \|f\|^2$, где $\|f\|$ — норма гильбер-

това пространства. Но теперь условие (9.75) должно быть выполненным для любого положительного μ, а последовательность сходится к решению в метрике L_2.

Таким образом, методы восстановления плотности связаны с решением некорректных задач численного дифференцирования [1]).

Ниже мы используем метод упорядоченной минимизации риска для решения задачи восстановления плотности. Однако прежде чем приступить к изложению соответствующих результатов, следует заметить, что существуют классические непараметрические методы восстановления плотности (например, метод Парзена), которые, казалось бы, позволяют обойти решение некорректной задачи. Однако при более внимательном анализе оказывается, что все они содержат константу, определение которой — проблема, полностью эквивалентная определению константы регуляризации γ_l при решении некорректных задач.

§ 8. Восстановление плотности методом Парзена

Идея метода Парзена восстановления плотности состоит в следующем. Справедливо тождество

$$P(x) = \int \delta(x-t)\, P(t)\, dt.$$

Рассмотрим некоторую параметрическую последовательность функций, сходящуюся к $\delta(x)$:

$$\frac{1}{h_1} K\left(\frac{x}{h_1}\right), \ldots, \frac{1}{h_l} K\left(\frac{x}{h_l}\right); \quad \lim_{l\to\infty} \frac{1}{h_l} K\left(\frac{x}{h_l}\right) = \delta(x).$$

Такая последовательность существует. Например, она может быть следующей:

$$\lim_{h\to 0} \frac{1}{\sqrt{2\pi}\, h} e^{-\frac{x^2}{2h^2}} = \delta(x).$$

---

[1]) Можно показать, что если в (9.71) выбирать $\gamma_l = (\ln\ln l)/l$, то асимптотическая скорость сходимости получаемых решений к плотности, имеющей n производных, для больших n имеет порядок близкий к $(\ln\ln l/l)^{1/2}$.

А именно справедливо:

$$P\left\{\overline{\lim_{l\to\infty}} (l/\ln\ln l)^{n/[2(n+1)]} \sup_t |f_l^{\gamma_l}(t) - f(t)| \leqslant \text{const}\right\} = 1.$$

Для всякой непрерывной плотности $P(x)$ существует такая величина h, что замена в подынтегральном выражении $\delta(x)$ на функцию $\frac{1}{h}K\left(\frac{x}{h}\right)$ мало повлияет на результат, т. е.

$$P(x)=\int \delta(x-t)\,P(t)\,dt \approx \int \frac{1}{h}K\left(\frac{x-t}{h}\right)P(t)\,dt.$$

Заменим теперь математическое ожидание величиной среднего по выборке. При достаточно большом объеме выборки это также мало повлияет на результат

$$P(x)\approx \int \frac{1}{h}K\left(\frac{x-t}{h}\right)P(t)\,dt \approx \frac{1}{l}\sum_{i=1}^{l}\frac{1}{h}K\left(\frac{x-t_i}{h}\right).$$

Выражение в правой части и используется как формула для оценки плотности:

$$\hat{P}_l(x)=\frac{1}{l}\sum_{i=1}^{l}\frac{1}{h}K\left(\frac{x-t_i}{h}\right). \tag{9.77}$$

Проблема же состоит в том, чтобы установить:

1) Каким должен быть закон образования величины h, чтобы с ростом объема выборки оценка стремилась к истинной плотности вероятностей?

2) Как выбирать константу h, если объем выборки ограничен?

Ответа на второй вопрос нет. Что же касается асимптотических свойств метода, то в 1962 г. Парзен, а в 1965 г. Надарая получили условия, обеспечивающие сходимость оценки (9.77) к искомой равномерно непрерывной плотности. Оказывается, для сходимости в метрике C последовательности (9.77) по вероятности к искомой плотности достаточно, чтобы

$$h_l \xrightarrow[l\to\infty]{} 0, \quad lh_l^2 \xrightarrow[l\to\infty]{} \infty \tag{9.78}$$

(результат Парзена), а для сходимости с вероятностью единица достаточно, чтобы при любом положительном μ сходился ряд

$$\sum_{l=1}^{\infty} e^{-\mu h_l^2 l} < \infty \tag{9.79}$$

(результат Надарая).

Заметим, что требования (9.78), (9.79) к выбору констант h_l оказались тождественными требованиям (9.75), (9.76) к выбору констант регуляризации при решении некорректной задачи (9.70). А это означает, что, обойдя постановку некорректной задачи, по-существу, не удается избежать трудностей, связанных с ее решением[1]).

§ 9. Восстановление плотности методом упорядоченной минимизации риска

Итак, будем решать уравнение (9.70), где вместо $F(x)$ используется эмпирическая оценка $F_l(x)$. Заметим, что $F_l(x)$ является случайной функцией, значения которой в разных точках x_i^* и x_j^* коррелированы.

Коэффициент ковариации K_{ij} случайных величин $y_i = F_l(x_i^*)$ и $y_j = F_l(x_j^*)$ равен

$$K_{ij} = \begin{cases} \frac{1}{l} F_i(1-F_j), & \text{если } x_i^* \leqslant x_j^*, \\ \frac{1}{l} F_j(1-F_i), & \text{если } x_j^* < x_i^*. \end{cases} \tag{9.80}$$

Здесь обозначено $F_i = F(x_i^*)$, $F_j = F(x_j^*)$.

Рассмотрим N величин $y_1, \ldots, y_N$, образованных с помощью функции $F_l(x)$ и N случайных чисел $x_1^*, \ldots, x_N^*$, полученных согласно равномерной на $[a, b]$ плотности вероятностей: $y_i = F_l(x_i^*)$ $i = 1, 2, \ldots, N$.

Так как случайные величины y_i и y_j коррелированы, то применять непосредственно метод упорядоченной минимизации для восстановления плотности вероятностей нельзя.

Поэтому применим к случайному вектору $Y = (y_1, \ldots, y_N)^T$ следующее линейное преобразование:

$$Z = BY, \quad B^T B = K^{-1},$$

где K — матрица ковариаций с элементами (9.80).

Известно, что с помощью этого преобразования образуется случайный вектор z, компоненты которого некор-

---

[1]) Известные оценки скорости сходимости для метода Парзена [100] имеют порядок $(\ln l/l)^{2/5}$.

Эта скорость меньше чем та, которая получена при восстановлении достаточно гладкой плотности методом численного дифференцирования (см. сноску к стр. 323).

релированы и имеют единичную дисперсию (можно показать, что в нашем случае компоненты независимы).

Поэтому можно рассматривать каждую компоненту z_i вектора z как реализацию случайной величины при условии x_i^* в независимой выборке объема N.

Обозначим через F_α вектор с координатами $F(x_1^*, \alpha), \ldots, F(x_N^*, \alpha)$. Преобразование B переводит вектор F_α в вектор $W = BF_\alpha$, компоненты которого будем рассматривать как значения некоторой функции $W(x, \alpha)$ в точках $x_1^*, \ldots, x_N^*$.

Будем теперь решать задачу минимизации функционала

$$\hat{I}(\alpha) = \int (z - W(x, \alpha))^2 P(z \mid x)\, dz\, dx.$$

Для решения этой задачи применим метод упорядоченной минимизации риска.

Таким образом, имеем

$$\hat{I}(\alpha) < \left[\frac{\frac{1}{N}\sum_{i=1}^{N} (z_i - W(x_i^*, \alpha))^2}{1 - 2\sqrt[3]{2}\,\tau \sqrt{\frac{\ln m^s(2N) - \ln\frac{\eta}{8}}{N}}} \right]_\infty .$$

Числитель правой части может быть представлен в виде

$$\frac{1}{N}\sum_{i=1}^{N} (z_i - W(x_i^*, \alpha))^2 = \frac{1}{N}(Y - F_\alpha)^{\mathrm{T}} B^{\mathrm{T}} B (Y - F_\alpha) =$$

$$= \frac{1}{N}(Y - F_\alpha)^{\mathrm{T}} K^{-1} (Y - F_\alpha).$$

Выпишем теперь окончательно функционал, минимизация которого по классам функций $S_1 \subset \ldots \subset S_N$ и по всевозможным функциям $F(x, \alpha)$ в каждом из классов определит оценку плотности распределения вероятностей:

$$R(\alpha) = \left[\frac{\frac{1}{N}(Y - F_\alpha)^{\mathrm{T}} K^{-1} (Y - F_\alpha)}{1 - 2\sqrt[3]{2}\,\tau \sqrt{\frac{\ln m^{S_p}(2N) - \ln\frac{\eta}{8}}{N}}} \right]_\infty , \qquad (9.81)$$

где

$$F_\alpha = (F(x_1^*, \alpha), \ldots, F(x_N^*, \alpha))^T,$$

$$F(x_i^*, \alpha) = \int_{-\infty}^{x_i^*} f(t, \alpha)\, dt,$$

$f(t, \alpha)$ — функции, принадлежащие S_p.

Для того чтобы воспользоваться соотношением (9.81), надо знать обратную ковариационную матрицу. Эта матрица может быть найдена аналитически [1]). Непосредственной проверкой соотношения $KK^{-1} = I$ читатель может убедиться, что следующая матрица является обратной к K:

$$K^{-1} = \left\| \begin{array}{cccc} \frac{lF_2}{F_1(F_2-F_1)} & -\frac{l}{F_2-F_1} & 0 & 0 \\ -\frac{l}{F_2-F_1} & \frac{l(F_3-F_1)}{(F_3-F_2)(F_2-F_1)} & -\frac{l}{F_3-F_2} & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \frac{-l}{F_{N-1}-F_{N-2}} & \frac{l(F_N-F_{N-2})}{F_N-F_{N-1}} \times & -\frac{l}{F_N-F_{N-1}} \\ & & \times(F_{N-1}-F_{N-2})^{-1} & \\ 0 & 0 & -\frac{l}{F_N-F_{N-1}} & \frac{l(1-F_{N-1})}{(1-F_N)(F_N-F_{N-1})} \end{array} \right\|. \tag{9.82}$$

Однако матрица (9.82) выражается через неизвестную нам функцию распределения случайной величины $F(x)$, производную от которой и требуется найти.

Итак, оказалось, что для того, чтобы по выборке восстановить плотность вероятностей, необходимо иметь априорную информацию о плотности (знать ковариационную матрицу (9.82)). В этом и сказывается принципиальная трудность проблемы восстановления плотности вероятностей в широком классе функций.

Вместо матрицы (9.82), однако, в (9.81) можно использовать ее оценку, где значения матрицы определяются по предварительно восстановленной непрерывной функции $F_э(x)$. Заметим, что задача восстановления функции $F(x)$ проще задачи восстановления плотности (согласно теореме Гливенко — Кантелли эмпирическая функция распределе-

[1]) Будем полагать, что искомая плотность на (a, b) не обращается в нуль.

ния сходится к истинной в равномерной метрике). Таким образом, алгоритм восстановления плотности вероятностей состоит из двух этапов: предварительного оценивания матрицы (9.82) и нахождения плотности [1]).

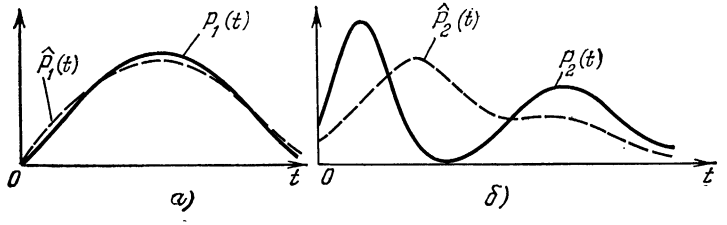

Рис. 15.

Такой двухэтапный метод восстановления плотности, видимо, является принципиально необходимым при восстановлении плотности на практике.

На рис. 15 и 16 показано применение метода Парзена для восстановления одно- и двумодальной плотности (рис. *а* и *б*) по выборке объема $l = 50$.

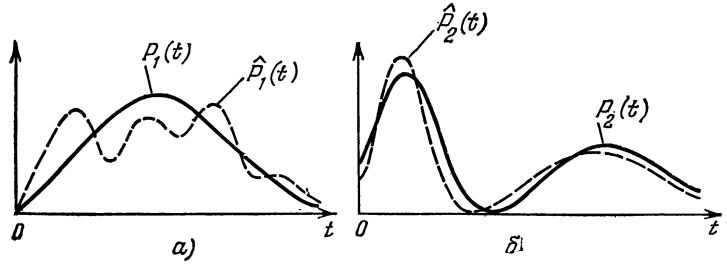

Рис. 16.

Показано, что при тех значениях параметра h, при которых удовлетворительно восстанавливается одномодальная плотность, плохо восстанавливается двумодальная плотность и наоборот: значения параметра, подобранные

[1]) Можно указать такое соотношение между объемом выборки l и числом N точек, в которых определяется значение функции $F_э(x)$, при котором в двухэтапном методе восстанавливаемая функция стремится к искомой плотности с ростом объема выборки.

для восстановления двумодальной плотности не дают удовлетворительного результата при восстановлении одномодальной плотности.

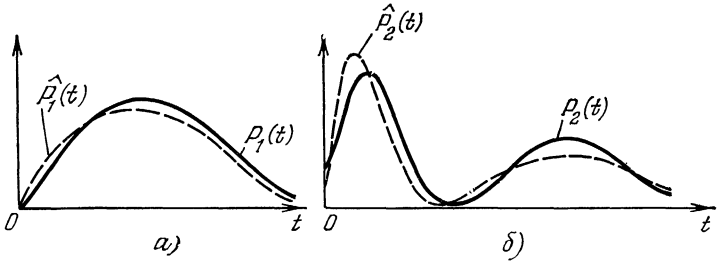

Рис. 17.

Применение рассмотренного двухэтапного метода упорядоченной минимизации (в классе сплайнов) в тех же

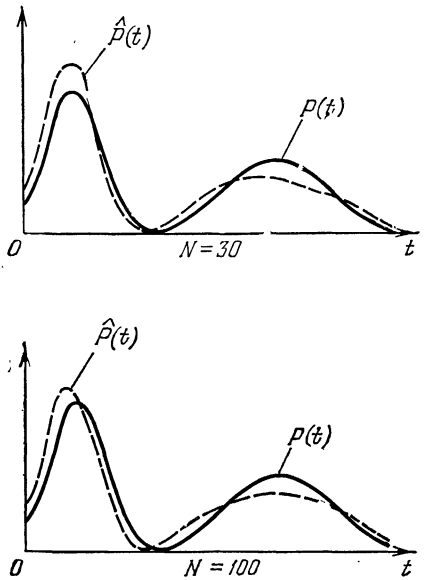

Рис. 18.

условиях позволяет получить удовлетворительные решения (рис. 17). При вычислении использовались значения поли-

гона $F_э(x)$ в $N=50$ случайно выбранных точках. Изменение N в достаточно широких пределах ($N=30$, $N=100$) несущественно влияет на результат (рис. 18). Плотности восстанавливались с помощью модифицированного алгоритма 12-3 (см. гл. XII).

Основные утверждения главы IX

1. Интерпретация результатов косвенных экспериментов в условиях, когда искомая функция связана с измеряемой операторным уравнением

$$Af=F,$$

образующим некорректно поставленную задачу, возможна лишь при проведении достаточно большого числа измерений.

2. С помощью метода упорядоченной минимизации риска может быть получена последовательность решений, сходящаяся с ростом числа измерений к искомой функции, если структура на множестве возможных решений задана разложением функций по собственным элементам самосопряженного оператора A^*A, упорядоченным в порядке убывания собственных чисел.

3. Задание другой структуры на множестве функций (например, связанной с разложением по полиномам), вообще говоря, не обеспечивает сходимости получаемой последовательности решений к искомому. Однако для задачи восстановления регрессии и ее производных сходимость в равномерной метрике последовательности получаемых решений к искомому возможна, если структура образована кусочно-полиномиальными зависимостями (сплайн-функциями), упорядоченными по числу точек сопряжения.

4. К задаче численного дифференцирования функции, заданной измерениями в N случайно выбранных точках приводится задача восстановления плотности вероятностей в классе гладких функций.

Особенность ее состоит в том, что ошибки измерения коррелированы. Конструктивные алгоритмы восстановления плотности методом упорядоченной минимизации риска связаны с возможностью учета ковариационной матрицы ошибок измерения. Эти алгоритмы восстановления плотности на практике более точны, чем алгоритмы Парзена.

ПРИЛОЖЕНИЕ К ГЛАВЕ IX

СТАТИСТИЧЕСКАЯ ТЕОРИЯ РЕГУЛЯРИЗАЦИИ

Пусть дано операторное уравнение

$$Af = F, \tag{П.1}$$

заданное непрерывным оператором A, осуществляющим взаимно однозначное отображение элементов метрического пространства E_1 в элементы метрического пространства E_2. Пусть $\Omega(f)$ — функционал такой, что:

1) решение уравнения (П.1) принадлежит $D(\Omega)$ — области определения функционала $\Omega(f)$;

2) функционал $\Omega(f)$ принимает в $D(\Omega)$ вещественные неотрицательные значения;

3) множества $\mathcal{M}_c = \{x : \Omega(f) < c\}$ $c > 0$ являются компактами.

Рассмотрим F_l — случайные функции и $f_l^{\gamma_l}$ — элементы, минимизирующие функционал

$$R_{\gamma_l}(f, F_l) = \rho^2_{E_2}(Af, F_l) + \gamma_l \Omega(f). \tag{П.2}$$

Пусть $\gamma_l \to 0$ при $l \to \infty$.

В этих условиях справедливы следующие две теоремы, которые являются стохастическим аналогом теорем А. Н. Тихонова (см. приложение к гл. I, теоремы П.1, П.2).

Теорема П.1. *Для любых положительных* ν *и* μ *найдется такое число* $n(\mu, \nu)$, *что для всех* $l > n(\mu, \nu)$ *выполнятся неравенства*

$$P\{\rho^2_{E_1}(f_l^{\gamma_l}, f) > \nu\} \leqslant P\{\rho^2_{E_2}(F_l, F) > \mu\gamma_l\}.$$

Теорема П.2. *Пусть* E_1 — *гильбертово пространство,* A — *линейный оператор,* $\Omega(f) = \|f\|^2$, *тогда для всякого* ε *найдется такой номер* $n(\varepsilon)$, *что при всех* $l > n(\varepsilon)$ *будут*

выполнены неравенства

$$P\left\{\|f_l^{\gamma_l}-f\|^2>\varepsilon\right\}\leqslant 2P\left\{\rho_{E_2}^2(F_l,\ F)>\frac{\varepsilon}{2}\gamma_l\right\}.$$

Доказательство теоремы П.1. По определению для любого l справедлива цепочка неравенств [1])

$$\gamma_l\Omega\left(f_l^{\gamma_l}\right)\leqslant R_{\gamma_l}\left(f_l^{\gamma_l},\ F_l\right)\leqslant R_{\gamma_l}\left(f,\ F_l\right)=$$
$$=\rho_2^2(Af,\ F_l)+\gamma_l\Omega(f)=\rho_2^2(F,\ F_l)+\gamma_l\Omega(f). \quad (\text{П.3})$$

Иначе говоря, справедливо

$$\Omega\left(f_l^{\gamma_l}\right)\leqslant\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}. \quad (\text{П.3а})$$

Кроме того, очевидно,

$$\rho_2^2\left(Af_l^{\gamma_l},\ F_l\right)\leqslant R_{\gamma_l}\left(f_l^{\gamma_l},\ F_l\right). \quad (\text{П.4})$$

Используя (П.3) и (П.4), получим неравенства

$$\rho_2\left(Af_l^{\gamma_l},\ F\right)\leqslant\rho_2\left(Af_l^{\gamma_l},\ F_l\right)+\rho_2(F_l,\ F)\leqslant$$
$$\leqslant\rho_2(F_l,\ F)+\sqrt{\rho_2^2(F_l,\ F)+\gamma_l\Omega(f)}. \quad (\text{П.5})$$

Далее, для любых $\nu>0$ и $c>\Omega(f)$ справедливо равенство

$$P\left\{\rho_1\left(f_l^{\gamma_l},\ f\right)\leqslant\nu\right\}=P\left\{\rho_1\left(f_l^{\gamma_l},\ f\right)\leqslant\nu\mid\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}\leqslant c\right\}\times$$
$$\times P\left\{\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}\leqslant c\right\}+P\left\{\rho_1\left(f_l^{\gamma_l},\ f\right)\leqslant\right.$$
$$\left.\leqslant\nu\mid\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}>c\right\}P\left\{\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}>c\right\}. \quad (\text{П.6})$$

Пусть теперь выполнится условие

$$\Omega(f)+\frac{\rho_2^2(F_l,\ F)}{\gamma_l}\leqslant c.$$

Тогда из (П.3а) следует, что справедливо неравенство $\Omega\left(f_l^{\gamma_l}\right)\leqslant c$, т. е. $f_l^{\gamma_l}$ принадлежит компакту. Согласно же лемме о непрерывности обратного оператора A на компакте (приложение к гл. I) получаем, что найдется та-

[1]) Здесь и далее для упрощения записи положим $\rho_{E_i}=\rho_i$.

кое δ, что как только выполнится неравенство $\rho_2\left(Af_l^{\gamma_l}, F\right) \leqslant \leqslant \delta$, окажется выполненным неравенство

$$\rho_1\left(f_l^{\gamma_l}, f\right) \leqslant \nu.$$

Отсюда следует, что для достаточно больших l

$$P\left\{\rho_1\left(f_l^{\gamma_l}, f\right) \leqslant \nu \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c\right\} \geqslant$$
$$\geqslant P\left\{\rho_2\left(Af_l^{\gamma_l}, F_l\right) \leqslant \delta \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c\right\}. \quad (\text{П.7})$$

Заметим теперь, что, согласно (П.5), в области

$$\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c$$

выполняется неравенство

$$\rho_2\left(Af_l^{\gamma_l}, F\right) \leqslant \sqrt{\gamma_l(c - \Omega(f))} + \sqrt{\gamma_l(c - \Omega(f)) + \gamma_l\Omega(f)} =$$
$$= \sqrt{\gamma_l}\left(\sqrt{c - \Omega(f)} + \sqrt{c}\right).$$

Так как $\gamma_l \to 0$ при $l \to \infty$, то, каково бы ни было δ, начиная с некоторого n, для всех $l > n$ выполнится равенство

$$P\left\{\rho_2\left(Af_l^{\gamma_l}, F\right) \leqslant \delta \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c\right\} = 1.$$

А так как справедливо (П.7), то для $l > n$ выполнится равенство

$$P\left\{\rho_1\left(f_l^{\gamma_l}, f\right) \leqslant \nu \mid \Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c\right\} = 1.$$

Таким образом, из (П.6) получим, что для любого $\nu > 0$ найдется такое n, что при $l > n$ выполнится неравенство

$$P\left\{\rho_1\left(f_l^{\gamma_l}, f\right) \leqslant \nu\right\} > P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} \leqslant c\right\},$$

а следовательно, и неравенство

$$P\left\{\rho_1\left(f_l^{\gamma_l}, f\right) > \nu\right\} \leqslant P\left\{\Omega(f) + \frac{\rho_2^2(F_l, F)}{\gamma_l} > c\right\}. \quad (\text{П.8})$$

Учитывая, что $c > \Omega(f)$, и вводя обозначения $\mu = c - \Omega(f)$, из (П.8) получим утверждение теоремы

$$P\left\{\rho_1\left(f_l^{\gamma_l},\ f\right) > \nu\right\} \leqslant P\left\{\rho_2^2(F_l,\ F) > \mu\gamma_l\right\}. \qquad (\text{П.9})$$

Теорема доказана.

Доказательство теоремы П.2.

1. Любой замкнутый ограниченный шар гильбертова пространства (т. е. множество векторов вида: $\{f: \|f - f_0\| \leqslant d\}$) является слабо компактным.

Поэтому относительно слабой компактности в пространстве E_1 мы находимся в условиях теоремы 1. Следовательно, для любых положительных ν и μ найдется такой номер $n = n(\mu, \nu)$, что при $l > n(\mu, \nu)$

$$P\left\{\left|\varphi\left(f_l^{\gamma_l}\right) - \varphi(f)\right| > \nu\right\} \leqslant P\left\{\rho_2^2(F_l,\ F) > \gamma_l\mu\right\},$$

где $\varphi(\cdot)$ — произвольный непрерывный линейный функционал, например проекция f на элемент q:

$$\varphi(f) = \int q(t) f(t)\, dt = (q \cdot f).$$

2. Согласно определению нормы в гильбертовом пространстве имеем

$$\left\|f_l^{\gamma_l} - f\right\|^2 = \left(f_l^{\gamma_l} - f,\ f_l^{\gamma_l} - f\right) = \\ = \left\|f_l^{\gamma_l}\right\|^2 - \|f\|^2 + 2\left(f,\ f - f_l^{\gamma_l}\right). \qquad (\text{П.10})$$

Воспользовавшись неравенством

$$P\{a + b > \varepsilon\} \leqslant P\left\{a > \frac{\varepsilon}{2}\right\} + P\left\{b > \frac{\varepsilon}{2}\right\},$$

из (П.10) получаем

$$P\left\{\left\|f_l^{\gamma_l} - f\right\|^2 > \varepsilon\right\} \leqslant \\ \leqslant P\left\{\left\|f_l^{\gamma_l}\right\|^2 - \|f\|^2 > \frac{\varepsilon}{2}\right\} + P\left\{2\left(f,\ f - f_l^{\gamma_l}\right) > \frac{\varepsilon}{2}\right\}.$$

Для оценки первого слагаемого правой части воспользуемся неравенством (П.3а), где учтем, что $\Omega(f) = \|f\|^2$. Получаем

$$\left\|f_l^{\gamma_l}\right\|^2 \leqslant \|f\|^2 + \frac{\rho_2^2(F_l,\ F)}{\gamma_l}.$$

Таким образом,

$$P\left\{\left\|f_l^{\gamma_l}\right\|^2-\|f\|>\frac{\varepsilon}{2}\right\}\leqslant P\left\{\frac{\rho_2^2(F_l,\ F)}{\gamma_l}>\frac{\varepsilon}{2}\right\}. \qquad (\text{П.11})$$

Второе слагаемое оценим с помощью (П.9), положив $\mu = \varepsilon/2$

$$P\left\{\left(f,\ f-f_l^{\gamma_l}\right)>\frac{\varepsilon}{4}\right\}\leqslant P\left\{\rho^2(F_l,\ F)>\frac{\varepsilon}{2}\gamma_l\right\}. \qquad (\text{П.12})$$

Объединяя оценки (П.11) и (П.12), получим утверждение теоремы:

$$P\left\{\left\|f_l^{\gamma_l}-f\right\|^2>\varepsilon\right\}\leqslant 2P\left\{\rho_2^2(F_l,\ F)>\frac{\varepsilon}{2}\gamma_l\right\}.$$

ГЛАВА X

ВОССТАНОВЛЕНИЕ ЗНАЧЕНИЙ ФУНКЦИИ В ЗАДАННЫХ ТОЧКАХ

§ 1. Схема минимизации суммарного риска

В условиях малых выборок

$$x_1,\ y_1;\ \ldots;\ x_l,\ y_l \tag{10.1}$$

целесообразно различать две задачи восстановления:

1. Восстановление в классе $F(x,\ \alpha)$ функциональной зависимости $y=f(x)$.

2. Восстановление в $F(x,\ \alpha)$ значений функции $y=f(x)$ в заданных точках

$$x_{l+1},\ \ldots,\ x_{l+k}. \tag{10.2}$$

Казалось бы, в задаче восстановления значений функции $y=f(x)$ в заданных точках (10.2) нет глубокого содержания. Существует «естественный» путь ее решения — восстановить по имеющимся эмпирическим данным (10.1) функциональную зависимость $y=F(x,\ \alpha^*)$ и с ее помощью определить значения функции в точках (10.2)

$$y_i=F(x_i,\ \alpha^*) \qquad (i=l+1,\ \ldots,\ l+k),$$

т. е. получить решение второй задачи, используя решение первой.

Однако такой путь восстановления значений функции часто не является лучшим, ведь здесь решение сравнительно простой задачи — восстановление k чисел (значений функции) ставится в зависимость от решения значительно более сложной задачи — восстановления функции (континуума чисел, содержащих эти k чисел).

Проблема как раз и заключается в том, чтобы в условиях дефицита информации использовать информацию для решения нужной нам, а не более общей задачи. Не исключено, что имеющегося объема информации может оказаться достаточно, чтобы удовлетворительно восстановить k чисел, но может не хватить для того, чтобы восстановить функцию во всей области ее определения.

Следует заметить, что на практике большей частью возникает потребность в определении значений функции в заданных точках, а не самой функциональной зависимости. Как правило, (а в задаче распознавания образов всегда) функциональная зависимость используется лишь для того, чтобы определить значение функции в некоторых нужных нам точках.

Итак, будем различать две постановки задачи восстановления: восстановление функции и *восстановление значений функции в заданных точках.*

В главе I мы формализовали постановку задачи восстановления функциональной зависимости с помощью схемы минимизации среднего риска. В этом параграфе мы формализуем постановку задачи восстановления значений функции в заданных точках с помощью схемы, которую будем называть схемой *минимизации суммарного риска.*

Считается, что задано множество

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}, \tag{10.3}$$

состоящее из $l+k$ векторов (*полная выборка векторов*). Существует функция $y=f(x)$, которая ставит в соответствие каждому вектору x множества (10.3) число y.

Таким образом, для $l+k$ векторов (10.3) определено $l+k$ значений

$$y_1, \ldots, y_l, y_{l+1}, \ldots, y_{l+k}. \tag{10.4}$$

Из множества (10.3) случайно отбираются l векторов x_i, для которых указываются соответствующие реализации y_i.

Образованное множество пар

$$x_1, y_1; \ldots; x_l, y_l \tag{10.5}$$

по аналогии с задачей восстановления функции будем называть *обучающей выборкой.*

Множество векторов

$$x_{l+1}, \ldots, x_{l+k} \tag{10.6}$$

будем называть *рабочей выборкой.*

Требуется по элементам обучающей и рабочей выборок среди заданного множества функций $F(x, \alpha)$ (этому множеству вовсе не обязана принадлежать $f(x)$) найти такую функцию $F(x, \alpha^*)$, которая с заданной вероятностью $1-\eta$ минимизирует суммарный риск прогноза значений

функции $y_i = f(x_i)$ на элементах рабочей выборки, т. е. с заданной вероятностью $1-\eta$ доставляет функционалу

$$I_\Sigma(\alpha) = \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha))^2 \qquad (10.7)$$

значение, близкое к минимальному.

Назовем такую постановку задачи восстановления значений функции в заданных точках *постановкой I* и рассмотрим еще одну постановку этой задачи — *постановку II*.

Пусть на множестве пар XY задано распределение вероятностей $P(x, y)$. Из этого множества в соответствии с $P(x, y)$ случайно и независимо выбирается l пар

$$x_1, y_1; \ldots; x_l, y_l,$$

образующих обучающую последовательность. Затем точно так же выбираются еще k пар

$$x_{l+1}, y_{l+1}; \ldots; x_{l+k}, y_{l+k}.$$

Требуется найти алгоритм A, который по обучающей последовательности $x_1, y_1; \ldots; x_l, y_l$ и рабочей выборке $x_{l+1}, \ldots, x_{l+k}$ выбирал бы в $F(x, \alpha)$ такую функцию

$$F(x, \alpha_A(x_1, y_1; \ldots; x_l, y_l; x_{l+1}, \ldots, x_{l+k})),$$

которая доставляет функционалу

$$I_r(A) = \int \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha_A(x_1, y_1; \ldots \\ \ldots; x_l, y_l; x_{l+1}, \ldots, x_{l+k}))^2 P(x_1, y_1) \ldots \\ \ldots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \ldots dx_{l+k}\, dy_{l+k}$$

значение, близкое к минимальному.

Справедлива следующая теорема о взаимосвязи этих постановок.

Теорема 10. *Если для некоторого алгоритма A доказано, что в постановке I с вероятностью $1-\eta$ уклонение между риском на обучающей и рабочей выборках не зависит от состава полной выборки и не превосходит $\varkappa$, то с той же вероятностью в постановке II уклонение между аналогичными величинами рисков не превосходит $\varkappa$.*

Доказательство. Обозначим

$$C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) = \\ = \left| \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i, \alpha_A))^2 - \frac{1}{k} \sum_{i=l+1}^{l+k} (y_i - F(x_i, \alpha_A))^2 \right|.$$

Рассмотрим вторую постановку задачи и вычислим вероятность уклониться от нуля более чем на $\varkappa$ величине $C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k})$:

$$P = \int_{XY} \theta\left[C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) - \varkappa\right] P(x_1, y_1) \ldots$$
$$\ldots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \ldots dx_{l+k}\, dy_{l+k},$$

где

$$\theta(z) = \begin{cases} 1, & z \geqslant 0, \\ 0, & z < 0. \end{cases}$$

Пусть T_p $(p = 1, 2, \ldots, (l+k)!)$ — оператор перестановки выборки $x_1, y_1; \ldots; x_{l+k}, y_{l+k}$. Тогда справедливо равенство

$$P = \int_{XY} \theta\left[C_A(x_1, y_1; \ldots; x_{l+k}, y_{l+k}) - \varkappa\right] P(x_1, y_1) \ldots$$
$$\ldots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \ldots dx_{l+k}\, dy_{l+k} =$$
$$= \int_{XY} \left\{ \frac{1}{(l+k)!} \sum_{p=1}^{(l+k)!} \theta\left[C_A(T_p(x_1, y_1; \ldots; x_{l+k}, y_{l+k})) - \varkappa\right] \right\} \times$$
$$\times P(x_1, y_1) \ldots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \ldots dx_{l+k}\, dy_{l+k}.$$

Выражение, заключенное в фигурные скобки, есть величина, оцениваемая в постановке I. Пусть она не превосходит $1 - \eta$. Тогда получим $P \leqslant \int_{XY} (1-\eta) P(x_1, y_1) \ldots P(x_{l+k}, y_{l+k})\, dx_1\, dy_1 \ldots dx_{l+k}\, dy_{l+k} = 1 - \eta$.

Теорема доказана.

В дальнейшем мы будем рассматривать задачу восстановления значений функции в заданных точках в постановке I. Однако с помощью теоремы 9.1 все полученные результаты могут быть перенесены и на случай постановки II.

В этой главе используется терминология, связанная с восстановлением значений функции. Однако все полученные результаты верны и для более общего случая, когда реализации (10.4) определяются не функцией $y = f(x)$, а условной плотностью $P(y \mid x)$, и требуется по случайным реализациям в одних точках (10.5) прогнозировать с помощью функции из $F(x, \alpha)$ реализации в других (10.6).

§ 2. Метод упорядоченной минимизации суммарного риска

Решать задачу восстановления значений функции в заданных точках будем с помощью метода упорядоченной минимизации риска.

В следующих двух параграфах мы получим оценки скорости равномерного относительного уклонения средних в двух подвыборках.

С помощью этих оценок построим равномерные по классу $F(x, \alpha)$ оценки суммарного риска по величинам эмпирического риска, аналогичные тем, которые использовались в предыдущих главах при построении упорядоченной минимизации среднего риска.

Мы установим, что для множества характеристических функций емкости h (в задаче распознавания образов) с вероятностью $1-\eta$ имеет место оценка вида

$$\nu_\Sigma(\alpha) < \nu(\alpha) + \Omega_1^*(l, k, h, -\ln\eta) \qquad (10.8)$$

(для этой задачи приняты обозначения $I_\Sigma(\alpha) = \nu_\Sigma(\alpha)$, $I_э(\alpha) = \nu(\alpha)$), а для множества функций емкости h произвольной природы с вероятностью $1-\eta$ имеет место оценка вида

$$I_\Sigma(\alpha) < I_э(\alpha)\,\Omega_2^*(l, k, h, -\ln\eta). \qquad (10.9)$$

Теперь, если на классе функций $F(x, \alpha)$ задать структуру

$$S_1 \subset \ldots \subset S_q,$$

то можно, минимизируя правую часть неравенства (10.8) (неравенства (10.9)), отыскать такой элемент S_* и такую функцию $F(x, \alpha_э^*)$, на которых достигается гарантированный минимум оценки суммарного риска. С помощью функции $F(x, \alpha_э^*)$ вычисляются значения $y_i = F(x_i, \alpha_э^*)$ в точках рабочей выборки. Внешне эта схема ничем не отличается от рассмотренной в главе VIII.

Однако в схеме упорядоченной минимизации суммарного риска есть особенность, которая и определяет разницу в решениях задач восстановления функции и восстановления значений функции в заданных точках.

Дело в том, что упорядочение функций класса $F(x, \alpha)$ должно быть проведено априорно. Это требование имеет разный смысл для восстановления функции и восстановления значений функции.

Для задачи восстановления функции оно означает, что необходимо, зная класс функций $F(x, \alpha)$ и область определения функции, задать структуру на $F(x, \alpha)$.

Для задачи восстановления значений функции это требование означает, что необходимо задать структуру на $F(x, \alpha)$, зная класс функций $F(x, \alpha)$ и полную выборку

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}. \qquad (10.10)$$

Разница заключается в том, что для полной выборки (10.10) множество функций $F(x, \alpha)$ распадается на множество *классов эквивалентности*. Это множество может быть изучено, и структура на $F(x, \alpha)$ может быть задана на классах эквивалентности, образуя более содержательный принцип упорядочения, чем при восстановлении функции.

Например, множество характеристических функций на полной выборке (10.10) распадается на конечное число классов эквивалентности. Две характеристические функции эквивалентны на полной выборке, если они одинаково делят эту выборку на две подвыборки (принимают на (10.10) одни и те же значения). В этом случае имеет смысл задавать структуру не на исходном множестве функции, а на конечном числе классов эквивалентности.

Ниже при восстановлении значений функции в заданных точках мы рассмотрим три разные идеи определения и упорядочения классов эквивалентности и каждую из них реализуем как для восстановления значений характеристической функции, так и для восстановления значения функции произвольной природы.

Однако прежде получим оценки, которые составят основу метода упорядоченной минимизации суммарного риска.

§ 3. Оценка равномерного относительного уклонения частот в двух подвыборках

В этом параграфе мы докажем теорему о равномерном относительном уклонении частот в двух подвыборках. В задаче минимизации суммарного риска в классе характеристических функций эта теорема играет ту же роль, которую в задаче минимизации среднего риска играет теорема о равномерном относительном уклонении частот от вероятностей. Для того чтобы сформулировать теорему, введем функцию $\Gamma_{l,k}(\varkappa)$.

Пусть дано множество

$$x_1, \ldots, x_{l+k},$$

состоящее из элементов двух типов: m элементов типа a и $l+k-m$ элементов типа b.

Выберем из этого множества наудачу l элементов. Вероятность того, что среди выбранных элементов окажется r элементов типа a, равна величине

$$P(r,\ l+k,\ l,\ m)=\frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^l}. \qquad (10.11)$$

Таким образом, с вероятностью (10.11) частота элементов типа a в отобранной группе составит r/l и, следовательно, в оставшейся группе $(m-r)/k$.

Вероятность того, что частота элементов a в первой группе уклонится от частоты элементов a во второй группе на величину, большую $\varkappa$, равна

$$P\left\{\left|\frac{r}{l}-\frac{m-r}{k}\right|>\varkappa\right\}=\sum_r \frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^l}=\Gamma_{l,k}(\varkappa,\ m),$$

где суммирование ведется по тем значениям r, для которых

$$\left|\frac{r}{l}-\frac{m-r}{k}\right|>\varkappa,\quad \max(0,\ m-k)\leqslant r\leqslant \min(m,\ l).$$

Определим функцию

$$\Gamma_{l,k}(\varkappa)=\max_m \Gamma_{l,k}\left(\sqrt{\frac{m}{l+k}}\,\varkappa,\ m\right).$$

Функция $\Gamma_{l,k}(\varkappa)$ легко может быть табулирована на ЦВМ.

Обозначим теперь через $\nu_0(\alpha)$ частоту ошибок классификации множества $x_1, \ldots, x_{l+k}$ с помощью решающего правила $F(x,\ \alpha)$. Очевидно, что

$$\nu_0(\alpha)=\frac{k}{l+k}\nu_\Sigma(\alpha)+\frac{l}{l+k}\nu(\alpha). \qquad (10.12)$$

Справедлива теорема о равномерном относительном уклонении частот в двух подвыборках.

Теорема 10.2. *Пусть класс решающих правил $F(x,\alpha)$ имеет емкость $h<l+k$. Тогда вероятность того, что относительная величина уклонения хотя бы для одного правила из $F(x,\ \alpha)$ превзойдет $\varkappa$, оценивается величиной*

$$P\left\{\sup_\alpha \frac{|\nu(\alpha)-\nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}}>\varkappa\right\}<1{,}5\,\frac{(l+k)^h}{h!}\Gamma_{l,k}(\varkappa). \qquad (10.13)$$

Здесь принято: $\frac{|\nu(\alpha)-\nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}}=0$, если $\nu(\alpha)=\nu_\Sigma(\alpha)=\nu_0(\alpha)=0$.

Доказательство. Заметим, что число классов эквивалентности на полной выборке не превосходит $N=m^S(l+k)$. Поэтому справедливо

$$P\left\{\sup_\alpha \frac{|\nu(\alpha)-\nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}}>\varkappa\right\}<N\sup_\alpha P\left\{\frac{|\nu(\alpha)-\nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}}>\varkappa\right\}.$$

При $h<l+k$ первый сомножитель правой части оценивается величиной $1{,}5\frac{(l+k)^h}{h!}$, второй сомножитель оценивается функцией $\Gamma_{l,k}(\varkappa)$. Действительно,

$$P\left\{\frac{|\nu(\alpha)-\nu_\Sigma(\alpha)|}{\sqrt{\nu_0(\alpha)}}>\varkappa\right\}=P\left\{|\nu(\alpha)-\nu_\Sigma(\alpha)|>\varkappa\sqrt{\nu_0(\alpha)}\right\}=$$
$$=P\left\{|\nu(\alpha)-\nu_\Sigma(\alpha)|>\varkappa\sqrt{\frac{m}{l+k}}\right\}=\Gamma_{l,k}\left(\varkappa\sqrt{\frac{m}{l+k}},\ m\right)$$

и, согласно определению,

$$\Gamma_{l,k}\left(\sqrt{\frac{m}{l+k}}\,\varkappa,\ m\right)\leqslant\Gamma_{l,k}(\varkappa).$$

Теорема доказана.

В дальнейшем нам понадобится равномерная по $F(x,\alpha)$ оценка частоты ошибок на рабочей выборке. Выведем ее, используя теорему 10.2. Ограничим правую часть (10.13) величиной η. Получим неравенство

$$h\left(\ln\frac{l+k}{h}+1\right)+\ln\Gamma_{l,k}(\varkappa)\leqslant\ln\frac{\eta}{1{,}5},$$

наименьшее решение которого обозначим $\varkappa_*$.

Учитывая (10.12), из (10.13) заключаем, что с вероятностью $1-\eta$ для всех α справедливо неравенство

$$\nu_\Sigma(\alpha)<\nu(\alpha)+\frac{k\varkappa_*^2}{2(l+k)}+\varkappa_*\sqrt{\nu(\alpha)+\left(\frac{k\varkappa_*}{2(l+k)}\right)^2}. \qquad (10.14)$$

Этим неравенством мы и будем пользоваться при построении алгоритмов упорядоченной минимизации риска в классе характеристических функций.

§ 4. Оценка равномерного относительного уклонения средних в двух подвыборках

При получении оценки равномерного относительного уклонения средних в двух подвыборках будем полагать, что на полной выборке

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k} \tag{10.15}$$

для множества произвольных функций $F(x, \alpha)$ выполняется условие

$$\sup_{\alpha} \frac{\sqrt[p]{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^{2p}}}{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^2} \leqslant \tau \qquad (p>2), \tag{10.16}$$

где y_i — значения реализации (10.4).

Условие (10.16) выражает априорную информацию о возможных выбросах на полной выборке (10.15). Это условие аналогично условию, рассмотренному в § 6 гл. VII.

Так же, как и в главе VII, введем функцию

$$R_1(\alpha) = \int \sqrt{\nu\{(y - F(x, \alpha))^2 > t\}}\, dt,$$

где $\nu\{(y - F(x, \alpha))^2 > t\}$ — отношение числа точек полной выборки (10.15), для которых на реализации (10.4) выполнено условие $(y - F(x, \alpha))^2 > t$, ко всем $l+k$ точкам. Для функции $R_1(\alpha)$, так же как и для функции $R(\alpha)$ (см. главу VII) справедливо соотношение

$$R_1(\alpha) < a(p) \sqrt[p]{\frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^{2p}}, \tag{10.17}$$

где

$$a(p) = \sqrt[p]{\frac{1}{2}\left(\frac{p-1}{p-2}\right)^{p-1}}.$$

Обозначим

$$I(\alpha) = \frac{1}{l+k}\sum_{i=1}^{l+k}(y_i - F(x_i, \alpha))^2 =$$

$$= \frac{l}{l+k} I_{\text{э}}(\alpha) + \frac{k}{l+k} I_{\Sigma}(\alpha). \tag{10.18}$$

Справедлива следующая

Теорема 10.3. *Пусть выполнено условие* (10.16) *и класс функций имеет емкость* $h < l+k$. *Тогда справедлива оценка*

$$P\left\{\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{э}(\alpha)|}{\bar{I}(\alpha)}>\tau a(p)\varkappa\right\}<$$

$$<1{,}5\,\frac{(l+k)^h}{h!}\,\Gamma_{l,k}(\varkappa). \qquad (10.19)$$

Доказательство. Для доказательства теоремы мы воспользуемся утверждением теоремы 10.2, согласно которому справедлива оценка

$$P\left\{\sup_{\alpha}\frac{|\nu(A_{\alpha,\beta})-\nu_{\Sigma}(A_{\alpha,\beta})|}{\sqrt{\nu_0(A_{\alpha,\beta})}}>\varkappa\right\}<1{,}5\,\frac{(l+k)^h}{h!}\,\Gamma_{l,k}(\varkappa), \qquad (10.20)$$

где $\nu(A_{\alpha,\beta})$ — частота события $\{(y-F(x,\ \alpha))^2>\beta\}$, вычисленная по обучающей последовательности, $\nu_{\Sigma}(A_{\alpha,\beta})$ — частота события $\{(y-F(x,\ \alpha))^2>\beta\}$, вычисленная на рабочей выборке по реализации (10.4), $\nu_0(A_{\alpha,\beta})$ — частота события $\{(y-F(x,\ \alpha))^2>\beta\}$, вычисленная на полной выборке (10.15) по реализации (10.4).

Покажем, что из справедливости неравенства (10.20) следует справедливость неравенства

$$P\left\{\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{э}(\alpha)|}{R_1(\alpha)}>\varkappa\right\}<1{,}5\,\frac{(l+k)^h}{h!}\,\Gamma_{l,k}(\varkappa). \qquad (10.21)$$

Для этого запишем в виде интеграла Лебега выражение

$$\sup_{\alpha}\frac{|I_{\Sigma}(\alpha)-I_{э}(\alpha)|}{R_1(\alpha)}\leqslant\sup_{\alpha}\lim_{n\to\infty}R,$$

где

$$R=\frac{\sum\limits_{i=1}^{\infty}\frac{1}{n}\left|\nu_{\Sigma}\left\{(y-F(x,\ \alpha))^2>\frac{i}{n}\right\}-\nu_{э}\left\{(y-F(x,\ \alpha))^2>\frac{i}{n}\right\}\right|}{R_1(\alpha)}.$$

Пусть теперь имеет место неравенство

$$\frac{\left|\nu_{\Sigma}\left\{(y-F(x,\ \alpha))^2>\frac{i}{n}\right\}-\nu_{э}\left\{(y-F(x,\ \alpha))^2>\frac{i}{n}\right\}\right|}{\sqrt{\nu\left\{(y-F(x,\ \alpha))^2>\frac{i}{n}\right\}}}\leqslant\varkappa.$$

Тогда

$$\frac{|I_\Sigma(\alpha)-I_э(\alpha)|}{R_1(\alpha)} \leqslant \lim_{n\to\infty} \frac{\varkappa \sum_{i=1}^{\infty} \frac{1}{n} \sqrt{\nu\left\{(y-F(x,\alpha))^2 > \frac{i}{n}\right\}}}{R_1(\alpha)} = \varkappa.$$

Таким образом, из справедливости (10.20) следует выполнение (10.21).

Для доказательства теоремы нам осталось воспользоваться неравенствами (10.17) и (10.21). Действительно,

$$P\left\{\sup_\alpha \frac{|I_\Sigma(\alpha)-I_э(\alpha)|}{I(\alpha)} > \tau a(p)\varkappa\right\} \leqslant$$

$$\leqslant P\left\{\sup_\alpha \frac{|I_\Sigma(\alpha)-I_э(\alpha)|}{R_1(\alpha)} > \varkappa\right\} \leqslant 1{,}5\frac{(l+k)^h}{h!}\Gamma_{l,k}(\varkappa).$$

Теорема доказана.

Получим теперь равномерную оценку риска на рабочей выборке. Для этого ограничим правую часть (10.19) величиной η. В результате получим неравенство

$$h\left(\ln\frac{l+k}{h}+1\right)+\ln\Gamma_{l,k}(\varkappa) \leqslant \ln\frac{\eta}{1{,}5}, \qquad (10.22)$$

наименьшее решение которого обозначим $\varkappa_*$.

Учитывая представление (10.18) из (10.19), получаем, что с вероятностью $1-\eta$ справедливо неравенство

$$I_\Sigma(\alpha) < \left[\frac{1+\tau a(p)\frac{l}{l+k}\varkappa_*}{1-\tau a(p)\frac{k}{l+k}\varkappa_*}\right]_\infty I_э(\alpha), \qquad (10.23)$$

где

$$[z]_\infty = \begin{cases} z, & \text{если } z \geqslant 0, \\ \infty, & \text{если } z < 0. \end{cases}$$

Неравенство (10.23) мы и используем при построении алгоритмов упорядоченной минимизации суммарного риска. Ниже мы ограничимся лишь линейным по параметрам классом функций

$$F(x,\ \alpha) = \sum_{i=1}^{n-1} \alpha_i \varphi_i(x) + \alpha_0.$$

Емкость этого класса функций равна n.

§ 5. Восстановление значений характеристической функции в классе линейных решающих правил

Пусть дана полная выборка

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k}. \tag{10.24}$$

На полной выборке множество решающих правил распадается на конечное число N классов эквивалентности $F_1, \ldots, F_N$. Два решающих правила $F(x, \hat{\alpha})$ и $F(x, \hat{\hat{\alpha}})$ попадают в один класс эквивалентности, если они одинаково делят выборку (10.24) на две подвыборки.

Всего возможно $\Delta^S(x_1, \ldots, x_{l+k})$ разделений выборки (10.24) на два класса с помощью правил $F(x, \alpha)$, и, следовательно, существует $\Delta^S(x_1, \ldots, x_{l+k})$ классов эквивалентности.

Согласно определению (см. § 7 гл. VI)

$$\Delta^S(x_1, \ldots, x_l) \leqslant m^S(l).$$

Для линейных решающих правил в пространстве размерности n справедлива оценка (см. § 8 гл. VI)

$$m^S(l+k) < 1{,}5\,\frac{(l+k)^n}{n!}.$$

Таким образом, на полной выборке (10.24) множество линейных решающих правил $F(x, \alpha)$ распадается на $N \leqslant 1{,}5\,\frac{(l+k)^n}{n!}$ классов эквивалентности $F_1, \ldots, F_N$.

Заметим, что классы эквивалентности не равнозначны по объему. Одни из них содержат «больше» решающих правил, другие — «меньше». Поставим в соответствие каждому классу эквивалентности величину, характеризующую долю линейных решающих правил, принадлежащих этому классу по отношению ко всем линейным решающим правилам. Такую величину можно сконструировать. Действительно, поставим в соответствие каждой функции

$$F(x, \alpha) = \theta\left(\sum_{i=1}^{n} \alpha_i \varphi_i(x)\right)$$

направляющий вектор (рис. 19)

$$\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}; \quad \|\alpha\| = 1.$$

Тогда множеству всех гиперплоскостей соответствует в пространстве параметров α единичная сфера, каждому классу эквивалентности F_i соответствует на поверхности сферы своя область. (Множество из N классов эквивалентности разбивают сферу на N областей.) Отношение площади, соответствующей области $\mathscr{L}_i$, к площади сферы $\mathscr{L}$ и характеризует долю функций из класса эквивалентности по отношению ко всем возможным линейным решающим правилам.

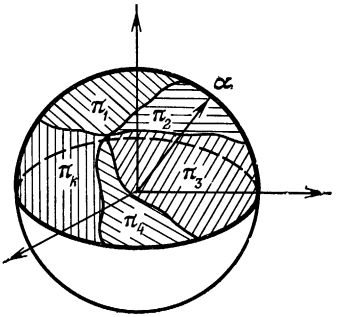

Рис. 19.

Упорядочим теперь классы эквивалентности по убыванию величин $\pi_i = \mathscr{L}_i/\mathscr{L}$ и введем следующую структуру:

$$S_1 \subset S_2 \subset \ldots \subset S_q, \quad (10.25)$$

где элементу S_p принадлежат лишь те классы эквивалентности, для которых

$$\frac{\mathscr{L}_i}{\mathscr{L}} > c_p \qquad (c_1 > c_2 > \ldots > c_q = 0).$$

Таким образом, мы построили структуру, каждый элемент S_p которой обладает экстремальным свойством: для данного числа классов эквивалентности он содержит максимальную долю всех решающих правил. Однако на практике вычислить величину $\mathscr{L}_i/\mathscr{L}$ и, следовательно, построить структуру (10.25) трудно. Поэтому мы рассмотрим другую характеристику объема класса эквивалентности, которая близка по смыслу к $\mathscr{L}_i/\mathscr{L}$ и может быть найдена на практике.

Обозначим через ρ_p величину расстояния между выпуклыми оболочками векторов полной выборки, отнесенных решающими правилами из F_p к разным классам, и поставим в соответствие классу эквивалентности F_p число

$$\pi(F_p) = \frac{\rho_p}{D}, \quad (10.26)$$

где $D/2$ — радиус минимальной сферы, содержащей множество (10.24), т. е.

$$\frac{D}{2} = \min_{x^*} \max_{x_1, \ldots, x_{l+k}} \| x_i - x^* \|.$$

Зададим на множестве классов эквивалентности структуру

$$S_1 \subset S_2 \subset \ldots \subset S_n, \tag{10.27}$$

где в S_d включены лишь такие классы эквивалентности F_i, для которых

$$\begin{aligned} &\pi^2(F_i) > \frac{1}{d-1}, \quad \text{если} \quad d < n, \\ &\pi^2(F_i) \geqslant 0, \quad \text{если} \quad d = n, \quad d \geqslant 2. \end{aligned} \tag{10.28}$$

Множество S_1 в (10.27) пусто.

Для того чтобы построить метод упорядоченной минимизации суммарного риска на структуре (10.27), оценим число N_d классов эквивалентности, принадлежащих элементу структуры S_d.

Справедлива

Лемма. *Число N_d классов эквивалентности в S_d оценивается величиной*

$$N_p < 1{,}5\,\frac{(l+k)^d}{d!}, \tag{10.29}$$

где d — минимум двух величин:

$$d = \min\left(n, \left[\frac{D^2}{\rho^2}\right] + 1\right), \tag{10.30}$$

n — размерность пространства, $[a]$ — целая часть числа a.

Доказательство. Заметим, что число N_d равно максимальному числу разбиений выборки

$$x_1, \ldots, x_{l+k}$$

на два таких класса, что расстояние между их выпуклыми оболочками больше $D/\sqrt{d-1}$, т. е.

$$\rho > \frac{D}{\sqrt{d-1}} = \rho_d. \tag{10.31}$$

Число таких решающих правил не превосходит

$$m^S(l+k) < 1{,}5\,\frac{(l+k)^r}{r!},$$

где r — максимальное число точек выборки, для которых любое разбиение на два класса удовлетворяет условию (10.31). Заметим, что если условие (10.31) выпол-

няется, то разбиение заведомо осуществляется с помощью гиперплоскости, поэтому заведомо

$$r \leqslant n,$$

где n — размерность пространства.

Пусть теперь задано r точек

$$x_1, \ldots, x_r$$

и 2^r всевозможных разбиений этих точек на два подмножества

$$T_1, \ldots, T_{2^r}.$$

Обозначим через $\rho_p(T_i)$ расстояние между выпуклыми оболочками векторов, принадлежащих разным подмножествам при разбиении T_i.

Тот факт, что условие (10.31) выполняется для всякого T_i, можно записать как

$$\min_i \rho(T_i) > \rho_d.$$

Тогда число r не превосходит максимального числа векторов, при котором еще выполняется неравенство

$$H(r) = \max_{x_1, \ldots, x_r} \min_i \rho(T_i) \geqslant \frac{D}{\sqrt{d-1}} \rho_d. \qquad (10.32)$$

Из соображений симметрии следует, что максимум r достигается, когда векторы $x_1, \ldots, x_r$ располагаются в вершинах правильного $(r-1)$-мерного симплекса, вписанного в шар радиуса $D/2$, а T_i — разбиение на два подсимплекса размерности $\frac{r}{2} - 1$ для четных r и два подсимплекса размерности $(r-1)/2$ и $(r-3)/2$ для нечетных r. При этом путем элементарных расчетов может быть найдено, что

$$H(r) = \begin{cases} \dfrac{D}{\sqrt{r-1}} & \text{для четных } r \\ \dfrac{D}{\sqrt{r-1}} \sqrt{\dfrac{r^2}{r^2-1}} & \text{для нечетных } r. \end{cases}$$

Для $r \geqslant 10$ величины

$$\frac{1}{\sqrt{r-1}} \quad \text{и} \quad \sqrt{\frac{r^2}{r^2-1}} \cdot \frac{1}{\sqrt{r-1}}$$

близки (различаются менее чем на 0,001). Поэтому примем

$$H(r)=\frac{D}{\sqrt{r-1}}. \tag{10.33}$$

(Оценкой сверху $H(r)$ было бы выражение

$$H(r)\leqslant \frac{D}{\sqrt{r-1{,}01}} \qquad (r>10).)$$

Из неравенств (10.32) и (10.33) следует, что для целых r

$$r<\left[\frac{D^2}{\rho^2}\right]+1.$$

Окончательно, учитывая, что разбиение осуществляется гиперплоскостью, т. е. $r\leqslant n$, получаем

$$d\leqslant \min\left(\left[\frac{D^2}{\rho^2}\right]+1,\ n\right). \tag{10.34}$$

И следовательно, согласно теореме 6.6

$$N_d<1{,}5\,\frac{(l+k)^d}{d!}.$$

Лемма доказана.

Из теоремы 10.2 и леммы следует, что с вероятностью $1-\eta$ одновременно для всех решающих правил из S_d будет выполнено неравенство

$$\nu_\Sigma(\alpha)<\nu(\alpha)+\frac{k\varkappa_*^2}{2(l+k)}+ \\ +\varkappa_3\sqrt{\nu(\alpha)+\left(\frac{k\varkappa_*}{2(l+k)}\right)^2}=R(\alpha,\ d), \tag{10.35}$$

где $\varkappa_*$ — наименьшее решение неравенства

$$d\left(\ln\frac{l+k}{d}+1\right)+\ln\Gamma_{l,k}(\varkappa)\leqslant \ln\frac{\eta}{1{,}5}.$$

Метод упорядоченной минимизации суммарного риска состоит в том, чтобы индексировать рабочую выборку с помощью правила $F(x,\ \alpha_э^*)$, которое минимизирует по d и α функционал (10.35). Пусть минимум равен $R(\alpha_э^*,\ d_*)$. Для такой индексации справедливо утверждение

$$P\{\nu_\Sigma(\alpha_э^*)<R(\alpha_э^*,\ d_*)\}>1-n\eta.$$

В главе XI мы приведем описание алгоритмов, минимизирующих суммарный риск в классе линейных решаю-

щих правил. Здесь же рассмотрим пример, иллюстрирующий разницу в решении задачи классификации векторов рабочей выборки методом минимизации суммарного риска и с помощью решающего правила минимизирующего эмпирический риск на обучающей последовательности.

На рис. 20 векторы первого класса обучающей последовательности обозначены крестиками, векторы второго класса — кружочками. Черными точками показаны векторы экзаменационной выборки.

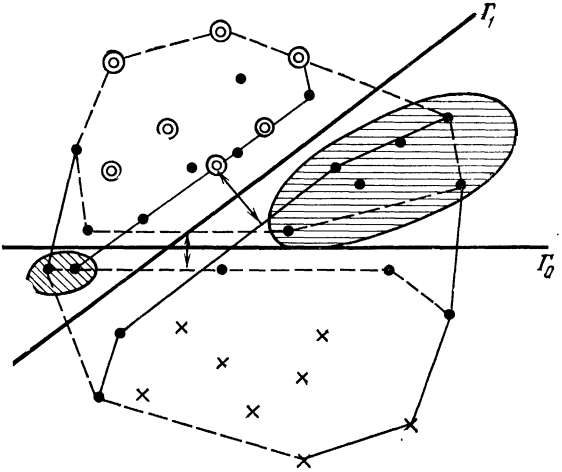

Рис. 20.

Решение этой задачи в схеме минимизации среднего риска заключается в том, чтобы построить разделяющую гиперплоскость, гарантирующую минимальную вероятность ошибки. Пусть решение выбирается среди гиперплоскостей, безошибочно делящих векторы обучающей последовательности. В этом случае минимальную гарантированную вероятность ошибок обеспечит оптимальная разделяющая гиперплоскость (такая, которая наиболее удалена от элементов обучающей последовательности). Те векторы, которые лежат по разные стороны гиперплоскости Γ_0, относятся к различным классам. Таково решение задачи методом минимизации среднего риска. Решение задачи методом минимизации суммарного риска опреде-

ляется гиперплоскостью Γ_1, максимизирующей расстояние между выпуклыми оболочками разделяемых множеств. Векторы, находящиеся по одну сторону гиперплоскости, относятся к первому классу, а по другую сторону гиперплоскости — ко второму классу.

На рисунке выделены (заштрихованы) точки рабочей выборки, которые классифицируются гиперплоскостями Γ_0 и Γ_1 по-разному.

§ 6. Селекция выборки для восстановления значений характеристической функции

Итак, решение задачи восстановления значений характеристической функции в заданных точках, полученное методом упорядоченной минимизации суммарного риска, может приводить к результатам, отличным от тех, которые следуют из классификации векторов рабочей выборки

$$x_{l+1}, \ldots, x_{l+k} \tag{10.36}$$

решающим правилом $F(x, \alpha_э)$, минимизирующим эмпирический риск на элементах обучающей последовательности

$$x_1, \omega_1; \ldots; x_l, \omega_l. \tag{10.37}$$

Эффект этот был получен потому, что полная выборка

$$x_1, \ldots, x_l, x_{l+1}, \ldots, x_{l+k} \tag{10.38}$$

состояла из сравнительно небольшого числа элементов, расположение которых в пространстве могло быть изучено и связано с конкретным способом упорядочения класса решающих правил $F(x, \alpha)$.

Способ упорядочения и определил разницу в классификации. Таким образом, геометрия векторов полной выборки (10.38) предопределила возможность более точного решения задачи восстановления значений функции в заданных точках.

Если это так, то возникает вопрос, нельзя ли, исключив из полной выборки (10.38) несколько элементов (изменив геометрию расположения векторов полной выборки в пространстве), так повлиять на задание структуры на классе решающих правил, чтобы увеличить

гарантированное число правильных классификаций элементов рабочей выборки? Оказывается, можно [1]).

Реализуем идею *селекции полной выборки*. Рассмотрим наряду с множеством X векторов полной выборки $H_l^t = \sum_{p=0}^{t} C_{l+k}^p$ различных подмножеств $X_1, \ldots, X_{H_{l+k}^t}$, полученных из (10.38) исключением не более t векторов. Пусть теперь на исходном множестве векторов (10.38) определена обучающая последовательность (10.37) и рабочая выборка (10.36). Обучающая и рабочая выборки индуцируют на каждом из множеств $X_1, \ldots, X_{H_{l+k}^t}$ свою обучающую и рабочую подвыборки.

Рассмотрим H_{l+k}^t задач восстановления значений функции в заданных точках. Каждая из этих задач определяется обучающей последовательностью

$$x_1, \omega_1, \ldots, \widehat{x_i, \omega_i}, \ldots, \widehat{x_j, \omega_j}, \ldots, x_l, \omega_l$$

и рабочей выборкой

$$x_{l+1}, \ldots, \hat{x}_{l+\tau}, \ldots, x_{l+k}$$

($\hat{x}$ означает, что элемент x исключен из последовательности).

Для каждой задачи в соответствии с ее полной выборкой

$$x_1, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, \hat{x}_{l+\tau}, \ldots, x_{l+k}$$

определим классы эквивалентности линейных решающих правил. Зададим структуру на классах эквивалентности, используя принцип упорядочения по относительным расстояниям, рассмотренный в предыдущем параграфе.

Из теоремы 10.2 и леммы следует, что с вероятностью $1-\eta$ в каждой задаче (в отдельности) для правила $F(x, \alpha_э^d)$, минимизирующего эмпирический риск в S_d справедливо неравенство

$$\nu_\Sigma(\alpha_э^d) < \nu(\alpha_э^d) + \frac{(k-k_и)\varkappa_*^2}{2(l+k-t_и)} + \varkappa_* \sqrt{\nu(\alpha_э^d) + \left[\frac{(k-k_и)\varkappa_*}{2(l+k-t_и)}\right]^2}, \quad (10.39)$$

[1]) Заметим, что при восстановлении характеристической функции селекция обучающей выборки не приводит к уменьшению оценки минимального гарантированного риска.

где $\varkappa_*$ — корень уравнения

$$d\left(\ln\frac{l+k-t_и}{d}+1\right)+\ln\Gamma_{l-l_и,\ k-k_и}(\varkappa)=\ln\frac{\eta}{1{,}5}. \quad (10.40)$$

В выражениях (10.39) и (10.40) использованы обозначения: $l_и$ — число исключенных элементов обучающей последовательности, $k_и$ — число исключенных элементов рабочей выборки $l_и+k_и=t_и$.

Одновременно для d-х элементов структур всех H^t_{l+k} задач с вероятностью $1-\eta$ выполнятся неравенства

$$\nu^{(i)}_{\Sigma}(\alpha^d_э)<\nu^{(i)}(\alpha^d_э)+\frac{(k-k^{(i)}_и)}{2(l+k-t_i)}(\varkappa^{(i)}_*)^2+$$
$$+\varkappa^{(i)}_*\sqrt{\nu^{(i)}(\alpha^d_э)+\left[\frac{(k-k^{(i)}_и)\varkappa^{(i)}_*}{2(l+k-t_i)}\right]^2}, \quad (10.41)$$

где $\varkappa^{(i)}_*$ — наименьшие решения неравенств

$$d\left(\ln\frac{l+k-t_i}{d}+1\right)+\ln H^t_{l+k}+\ln\Gamma_{l-l^{(i)}_и,\ k-k^{(i)}_и}(\varkappa^{(i)})\leqslant$$
$$\leqslant\ln\frac{\eta}{1{,}5}, \quad (10.42)$$

i меняется от 1 до H^t_{l+k}.

В выражениях (10.41), (10.42) использованы обозначения $l^{(i)}_и$, $k^{(i)}_и$ — число элементов обучающей и рабочей выборок, исключенных из (10.37) и (10.36) при образовании i-й задачи $l^{(i)}_и+k^{(i)}_и=t_i$; $\nu^{(i)}_{\Sigma}(\alpha^d_э)$, $\nu^{(i)}(\alpha^d_э)$ — частоты ошибочной классификации рабочей и обучающей выборок в i-й задаче.

Умножим каждое из неравенств (10.41) на величину $k-k^{(i)}_и$. В результате для каждой задачи получим оценку числа ошибок m_i на $k-k^{(i)}_и$ элементах ее рабочей выборки

$$m_i<\left[\nu^{(i)}(\alpha^d_э)+\frac{(k-k^{(i)}_и)}{2(l+k-t_i)}(\varkappa^{(i)}_*)^2+\right.$$
$$\left.+\varkappa^{(i)}_*\sqrt{\nu^{(i)}(\alpha^d_э)+\left[\frac{(k-k^{(i)}_и)}{2(l+k-t_i)}\varkappa^{(i)}_*\right]^2}\right](k-k^{(i)}_и). \quad (10.43)$$

Если бы число исключенных из рабочей выборки векторов для всех задач было одинаковым и равным $k_и$, то наилучшее гарантированное решение задачи классификации $k-k_и$ векторов рабочей выборки определялось бы

тем неравенством (той задачей), для которого величина, оценивающая число ошибок на $k-k_и$ элементах рабочей выборки наименьшая.

Однако число векторов, исключенных из рабочей выборки для разных задач разное. Поэтому будем считать наилучшим решением то, которое максимизирует число правильных классификаций элементов рабочей выборки, т. е. определяется той задачей, для которой достигает минимума величина [1])

$$R(d,\ i)=\left[\nu^{(i)}(\alpha_э^d)+\frac{\left(k-k_и^{(i)}\right)}{2\,(l+k-t_i)}\left(\varkappa_*^{(i)}\right)^2+\right.$$

$$\left.+\varkappa_*^{(i)}\sqrt{\nu^{(i)}(\alpha_э^d)+\left(\frac{\left(k-k_и^{(i)}\right)\varkappa_*^{(i)}}{2\,(l+k-t_i)}\right)^2}\right]\left(k-k_и^{(i)}\right)+k_и^{(i)},\quad (10.44)$$

определяющая число ошибок плюс число исключенных векторов рабочей выборки.

Теперь перебором по d и t найдем векторы, которые следует исключить, чтобы гарантировать наибольшее число правильно классифицированных векторов рабочей выборки. Задача минимизации по d и t функционала (10.44) достаточно трудна в вычислительном отношении. Точное ее решение требует большого перебора вариантов. Однако использование некоторых эвристических приемов позволяет найти удовлетворительное решение в приемлемое время. Подробно об алгоритмах упорядоченной минимизации суммарного риска см. главу XI.

Заметим, что при селекции полной выборки подбираются как элементы обучающей, так и элементы рабочей выборок.

Селекция элементов рабочей выборки позволяет за счет отказа от классификации некоторых элементов увеличить общее число правильно классифицируемых векторов.

До сих пор мы исходили из того, что пространство, в котором строится структура, фиксировано. Однако процедура упорядочения по относительным расстояниям может быть проведена в любом подпространстве E_m исходного пространства E_n. При этом минимальное значение соответствующей оценки будет достигнуто не обязательно в исходном пространстве E_n.

[1]) Здесь можно ввести различные цены за ошибку и отказ от классификации элементов рабочей выборки.

Это обстоятельство открывает возможность достичь еще более глубокого минимума оценки риска за счет дополнительной минимизации по подпространствам.

§ 7. Восстановление значений произвольной функции в классе линейных по параметрам функций

Распространим теперь методы восстановления значений характеристических функций, рассмотренные в предыдущих параграфах, на восстановление значений произвольной функции в классе линейных по параметру функций. Для этого определим классы эквивалентности линейных (по параметрам) функций на полной выборке, зададим на этих классах структуру и реализуем метод упорядоченной минимизации риска.

Пусть дана полная выборка

$$x_1, \ldots, x_l, \quad x_{l+1}, \ldots, x_{l+k} \tag{10.45}$$

и множество линейных (по параметрам) функций $F(x, \alpha)$. Поставим в соответствие каждой функции $F(x, \alpha^*)$ этого множества однопараметрическое (по параметру β) семейство решающих правил

$$F_{\alpha^*}(\beta) = \theta(F(x, \alpha^*) + \beta). \tag{10.46}$$

При изменении параметра β от $-\infty$ до ∞ однопараметрическое семейство решающих правил (10.46) образует последовательность дихотомий (разделений на два класса) множества векторов (10.45) от дихотомии, у которой первый класс пуст, а второй класс содержит все множество (10.45)

$$[\varnothing; \ \{x_1, \ldots, x_{l+k}\}]$$

(при $\beta = -\infty$), до дихотомии, у которой первый класс состоит из всего множества векторов (10.45), а второй класс пуст

$$[\{x_1, \ldots, x_{l+k}\}; \ \varnothing]$$

(при $\beta = +\infty$).

Таким образом, для каждой функции $F(x, \alpha)$ может быть получена последовательность дихотомий

$$[\varnothing, \{x_1, \ldots, x_{l+k}\}]; \quad [\{x_{i_1}, \ldots, x_{i_l}\}; \{x_{j_1}, \ldots, x_{j_k}\}]; \ldots$$
$$\ldots; [\{x_1, \ldots, x_{l+k}\}, \varnothing]. \tag{10.47}$$

В соответствии с этой последовательностью дихотомий разделим множество функций $F(x, \alpha)$ на конечное число классов эквивалентности. Две функции $F(x, \hat{\alpha})$ и $F(x, \hat{\hat{\alpha}})$ попадают в один класс эквивалентности F_i, если они образуют одну и ту же последовательность дихотомий (10.47).

Поставим теперь в соответствие каждому классу эквивалентности число $\pi(F_i)$, равное доле всех функций, принадлежащих этому классу, и затем упорядочим классы эквивалентности в порядке убывания $\pi(F_i)$:

$$F_1, F_2, \ldots, F_N, \quad \pi(F_1) \geqslant \pi(F_2) \geqslant \ldots \geqslant \pi(F_N). \quad (10.48)$$

Используя упорядоченность (10.48), можно построить структуру на классах эквивалентности

$$S_1 \subset S_2 \subset \ldots \subset S_n.$$

Элементу S_r принадлежат те классы эквивалентности, для которых $\pi(F_r) > c_r$.

Для множества линейных функций доля функций, принадлежащих классу эквивалентности, может быть определена так же, как определялась доля линейных решающих правил. Поставим в соответствие каждой линейной функции вектор направляющих косинусов.

Тогда множеству всех функций соответствует поверхность единичной сферы в пространстве размерностей n, а каждому классу эквивалентности соответствует на этой сфере своя область (см. рис. 19).

Отношение площади выделенной области к площади поверхности сферы и определяет долю функций из класса эквивалентности среди всего множества функций.

На практике, однако, трудно непосредственно вычислить характеристику $\hat{\pi}(F_i)$. Поэтому, так же как и раньше, рассмотрим другую характеристику объема класса эквивалентности.

Для каждой функции $F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$ определим направляющий вектор $\hat{\alpha} = \alpha / \|\alpha\|$.

Каждый класс эквивалентности F_m будем характеризовать числом

$$\rho_m = \min_{x_i, x_j} \sup_{\alpha} \left| (x_i - x_j)^{\mathrm{T}} \frac{\alpha}{\|\alpha\|} \right| \qquad (i \neq j),$$

где минимум берется по всем векторам полной выборки, а супремум — по всем направляющим векторам данного класса эквивалентности.

Построим теперь следующую структуру:

$$S_1 \subset \ldots \subset S_n.$$

К d-му элементу S_d структуры отнесем те функции, для которых выполняется соотношение

$$\hat{\pi}^2(F) = \left[\frac{\rho}{D}\right]^2 > \frac{1}{d-1}, \quad \text{если} \quad d < n,$$

$$\hat{\pi}^2(F) = \left[\frac{\rho}{D}\right]^2 \geqslant 0, \qquad \text{если} \quad d \geqslant n,$$

где D — минимальный диаметр сферы, содержащей множество $(x_1, \ldots, x_{l+k})$. Используя лемму, можно, как и в § 5, показать, что емкость функций S_d-го элемента структуры равна d, где

$$d = \min\left(\left[\frac{D^2}{\rho^2}\right] + 1,\ n\right).$$

Метод упорядоченной минимизации на этой структуре состоит в том, чтобы найти такой элемент S_*, а в нем такую функцию $F(x, \alpha_э^*)$, для которой достигался бы минимум правой части неравенства

$$I_\Sigma(\alpha) < \left[\frac{1 + \tau a(p)\dfrac{l}{l+k}\varkappa_*}{1 - \tau a(p)\dfrac{k}{l+k}\varkappa_*}\right]_\infty I_э(\alpha), \qquad (10.49)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$d\left(\ln\frac{l+k}{d} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leqslant \ln\frac{\eta}{1{,}5}.$$

Первый сомножитель правой части неравенства (10.49) зависит лишь от того, в каком порядке проецируются векторы полной выборки на вектор направлений выбранной линейной функции, второй — от величины эмпирического риска.

Пусть минимум правой части (10.49) равен $R(\alpha_э^*, d^*)$. Тогда справедливо утверждение

$$P\{I_\Sigma(\alpha_э^*) < R(\alpha_э^*, d_*)\} > 1 - n\eta.$$

§ 8. Селекция выборки для восстановления значений произвольной функции

В главе VIII мы установили, что при восстановлении нехарактеристической функции селекция обучающей последовательности может привести к отысканию функции с меньшей гарантированной величиной среднего риска.

В схеме минимизации суммарного риска селекция полной выборки может привести к еще большему эффекту.

Для линейных по параметрам функций этот дополнительный эффект возникает потому, что исключение части векторов полной выборки $x_1, \ldots, x_{l+k}$ меняет геометрию расположения векторов, вследствие чего может быть осуществлено более содержательное упорядочение класса функций $F(x, \alpha)$.

Итак, пусть задана полная выборка

$$x_1, \ldots, x_{l+k}. \tag{10.50}$$

Рассмотрим H^t_{l+k} различных подмножеств $X_1, \ldots, X_{H^t_l}$, каждое из которых получено исключением из (10.50) не более t элементов.

В дальнейшем будем предполагать, что для всех подмножеств выполняются условия (10.16). Пусть теперь на исходном множестве (10.50) определена обучающая последовательность

$$x_1, y_1; \ldots; x_l, y_l \tag{10.51}$$

и рабочая выборка

$$x_{l+1}, \ldots, x_{l+k}. \tag{10.52}$$

Обучающая и рабочая выборки (10.51) и (10.52) индуцируют на каждом из подмножеств свою обучающую и рабочую выборки.

Рассмотрим H^t_{l+k} задач восстановления значений функции в заданных точках.

Для каждой задачи r $(r = 1, 2, \ldots, H^t_{l+k})$ рассмотренным выше способом зададим свою структуру на классе линейных функций

$$S^r_1 \subset \ldots \subset S^r_n.$$

Получим, что с вероятностью $1-\eta$ для каждой задачи (в отдельности) справедлива оценка

$$I_{\Sigma}^{(r)}(\alpha_{э}) < \left[\frac{1+\tau a(p)\frac{l-l_{п}^{r}}{l+k-t_{r}}\varkappa_{*}^{r}}{1-\tau a(p)\frac{k-k_{и}^{r}}{l+k-t_{r}}\varkappa_{*}^{2}}\right]_{\infty} I_{э}^{(r)}(\alpha_{э}),$$

где $F(x,\ \alpha_{э}) \subset S_{d}^{r}$ — функция, минимизирующая эмпирический риск на обучающей последовательности этой задачи (индекс (r) указывает на то, что суммарный и эмпирический риск вычисляются по элементам, принадлежащим подмножеству $X^{(r)}$), $\varkappa_{*}^{r}$ — наименьшее решение неравенства

$$d\left(\ln\frac{l+k-t_{r}}{d}+1\right)+\ln\Gamma_{l-l_{и}^{r},\ k-k_{и}^{r}}(\varkappa^{r}) \leqslant \ln\frac{\eta}{1,5}.$$

Здесь обозначено: $l-l_{и}^{r}$ — длина обучающей последовательности в задаче r; $k-k_{и}^{r}$ — длина рабочей выборки в задаче $l_{и}^{r}+k_{и}^{r}=t_{r}$. С вероятностью $1-\eta$ одновременно для S_{d} элементов всех H_{l+k}^{t} задач будут выполняться неравенства

$$I_{\Sigma}^{(r)}(\alpha) < \left[\frac{1+\tau a(p)\frac{l-l_{и}^{r}}{l+k-t_{r}}\varkappa_{*}^{r}}{1-\tau a(p)\frac{k-k_{и}^{r}}{l+k-t_{r}}\varkappa_{*}^{r}}\right]_{\infty} I_{э}^{(r)}(\alpha),$$

где, в отличие от предыдущего случая, $\varkappa_{*}^{r}$ — наименьшие решения неравенств

$$d\left(\ln\frac{l+k-t_{r}}{d}+1\right)+\ln\Gamma_{l-l_{и}^{r},\ k-k_{и}^{r}}(\varkappa)+\ln H_{l+k}^{t} \leqslant \ln\frac{\eta}{1,5}.$$

Выберем такую задачу, для которой оценка величины суммарного риска минимальна.

Наконец, перебором по d и по t (на практике $t < \leqslant 5 \div 10$) найдем наилучшее решение.

§ 9. Восстановление значений характеристической функции в классе кусочно-линейных решающих правил

Рассмотрим вторую идею задания классов эквивалентности.

Пусть конечное множество векторов X, образующих полную выборку, разбито на d подмножеств

$$X_1, \ldots, X_d, \quad UX_i = X, \quad X_i \cap X_j = 0 \qquad (i \neq j).$$

Поставим в соответствие этому разбиению множество кусочно-линейных решающих правил, где на каждом подмножестве X_l определено линейное решающее правило. Таким образом, рассмотрим параметрическое семейство решающих правил

$$\theta(L(x, \alpha_1), \ldots, L(x, \alpha_d)) = \begin{cases} L(x, \alpha_1), & \text{если} \quad x \in X_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ L(x, \alpha_d), & \text{если} \quad x \in X_d, \end{cases} \tag{10.53}$$

где $L(x, \alpha_r)$ — линейное решающее правило ($L(x, \alpha_r) = \theta\left(\sum_{i=1}^{n} \alpha_i^r \varphi_i(x)\right)$).

Емкость такого кусочно-линейного класса решающих правил равна

$$h = nd.$$

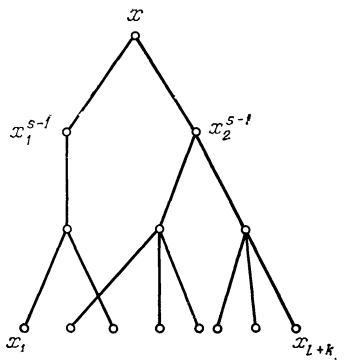

Рис. 21.

Заметим, что задание класса решающих правил (10.53) определяется способом разбиения полной выборки X на d подмножеств. Для того чтобы определить нужное разбиение множества X на d подмножеств, изучим геометрию этого множества с точки зрения его *таксонной структуры* (см. приложение к главе). Для этого построим дерево (рис. 21). На нижнем s-м уровне дерева каждый элемент множества X образуют подмножество x_i. Элементы $s-1$-го уровня образованы объединением элементов s-го уровня.

Каждый элемент s-го уровня входит лишь в один элемент $s-1$-го уровня.

Самый высокий — первый уровень состоит из одного подмножества X, объединяющего все множество элементов генеральной совокупности. Дерево построено так, что на каждом уровне p выполнено соотношение

$$\bigcup_{i=1} X_i^d = X; \quad X_i^d \bigcap X_j^d = 0 \quad (i \neq j); \; d = 1, 2, \ldots, s \quad (10.54)$$

$$(ns < l + k).$$

Поставим в соответствие каждому уровню дерева семейство кусочно-линейных решающих правил S_d, построенное согласно (10.54).

При таком способе построения классов кусочно-линейных решающих правил таксонная структура полной выборки (10.54) определит конкретную структуру на классе кусочно-линейных решающих правил

$$S_1 \subset \ldots \subset S_s.$$

На этой структуре может быть реализован метод упорядоченной минимизации суммарного риска, т. е. найден элемент S_d структуры, для которого метод минимизации эмпирического риска обеспечит наименьшую оценку суммарного риска

$$\nu_\Sigma(\alpha_э) < \nu(\alpha_э^*) + \frac{k\varkappa_*^2}{2(l+k)} + \varkappa_* \sqrt{\nu(\alpha_э^*) + \left[\frac{k\varkappa_*}{2(l+k)}\right]^2} = R^*,$$

где $\varkappa_*$ — наименьшее решение неравенства

$$np\left(\ln\frac{l+k}{p} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leqslant \ln\frac{\eta}{1,5}.$$

С помощью найденного правила $F(x, \alpha_э^*)$ классифицируются элементы рабочей выборки. Для полученной классификации справедливо

$$P\{\nu_\Sigma(\alpha_э^*) < R^*\} \leqslant 1 - s\eta.$$

Методы построения таксонной структуры рассмотрены в приложении к главе.

§ 10. Восстановление значений произвольной функции в классе кусочно-линейных функций

Структура, аналогичная рассмотренной в предыдущем параграфе, может быть задана и на множестве кусочно-линейных функций.

Для этого рассмотрим ту же самую таксонную структуру множества $x_1, \ldots, x_{l+k}$ и определим такой элемент структуры (т. е. такое разбиение множества $x_1, \ldots, x_{l+k}$ на таксоны), при котором решение задачи минимизации эмпирического риска в классе линейных решающих правил отдельно для каждого таксона обеспечили бы минимальное гарантированное число величины суммарного риска.

При реализации метода упорядоченной минимизации суммарного риска в задаче восстановления зависимостей воспользуемся той же самой идеей: для каждого элемента заданной таксонной структуры S_d получим минимальную гарантированную оценку величины суммарного риска

$$I_\Sigma(\alpha_э) < \left[\frac{1+\tau a(p)\frac{l}{l+k}\varkappa_*}{1-\tau a(p)\frac{k}{l+k}\varkappa_*}\right] I_э(\alpha_э),$$

где $F(x, \alpha_э)$ — функция, минимизирующая в S_d эмпирический риск, $\varkappa_*$ — наименьшее решение неравенства

$$nd\left(\ln\frac{l+k}{nd}+1\right)+\ln\Gamma_{l,k}(\varkappa)\leqslant\ln\frac{\eta}{1,5}$$

(здесь n — размерность пространства $nd < l+k$), а затем выберем такой элемент структуры S_* и в нем такую функцию $F(x, \alpha_э^*)$, на которых достигается минимальная гарантированная оценка величины суммарного риска для заданной структуры. Значения функции $F(x, \alpha_э^*)$ в точках рабочей выборки и будут восстановленными значениями функции.

§ 11. Локальные алгоритмы восстановления значений характеристической функции

Наконец, рассмотрим третью идею построения алгоритмов восстановления значений функции.

Определим для каждого вектора x^* полной выборки систему окрестностей:

$$(x^*)_1,\ (x^*, x_{i_1})_2,\ \ldots,\ (x^*, x_{i_1}, x_{i_\tau})_r,\ \ldots,\ (x_1, \ldots, x_{l+k})_q.$$

Таким образом, задается $l+k$ систем окрестностей, своя для каждого вектора полной выборки:

$$\begin{aligned}
&1) \quad (x_1)_1 \in (x_1,\ x_{i_1})_2 \in \ldots \in (x_1,\ \ldots,\ x_{l+k})_q;\\
&2) \quad (x_2)_1 \in (x_2,\ x_{i_2},\ x_{i_3})_2 \in \ldots \in (x_1,\ \ldots,\ x_{l+k})_q;\\
&l+k) \quad (x_{l+k})_1 \in (x_{l+k},\ x_{i_{l+k}})_2 \in \ldots \in (x_1,\ \ldots,\ x_{l+k})_q.
\end{aligned} \tag{10.55}$$

Пусть теперь произошло разделение множества X на обучающую и рабочую выборки.

Рассмотрим произвольную окрестность X_i^r точки x_i, содержащую как элементы обучающей, так и элементы рабочей выборок. Согласно теореме 10.2 с вероятностью $1-\eta$ можно утверждать, что одновременно для всех линейных решающих правил будет выполнено неравенство

$$\nu_\Sigma^{(r)}(\alpha) < \nu^{(r)}(\alpha) + \frac{k_r(\varkappa_*^{(r)})^2}{2(l_r+k_r)} + \varkappa_*^{(r)} \sqrt{\nu^r(\alpha) + \left[\frac{k_r \varkappa_*^{(r)}}{2(l_r+k_r)}\right]^2},$$

где $\nu_\Sigma^{(r)}(\alpha)$ есть величина суммарного риска классификации элементов из окрестности X_i^r с помощью решающего правила $F(x, \alpha)$, а $\nu^{(r)}(\alpha)$ — величина эмпирического риска, вычисленная для правила $F(x, \alpha)$ по элементам обучающей последовательности, принадлежащим окрестности X_i^r, $\varkappa_*^r$ — наименьшее решение неравенства

$$n\left(\ln\frac{l_r+k_r}{n}+1\right)+\ln\Gamma_{l_r,\,k_r}(\varkappa) \leqslant \ln\frac{\eta}{1,5},$$

n — размерность пространства X.

В неравенстве l_r и k_r — число элементов обучающей и рабочей выборок из окрестности X_i^r. Пусть $F(x, \alpha_э)$ — решающее правило, минимизирующее на обучающей последовательности из X_i^r величину эмпирического риска.

С вероятностью $1-\eta$ для элементов из $X_i^{(r)}$ справедлива оценка

$$\nu_\Sigma^{(r)}(\alpha_э) < \nu^{(r)}(\alpha_э) + \frac{k_r(\varkappa_*^{(r)})^2}{2(l_r+k_r)} + \varkappa_*^{(r)} \sqrt{\nu^{(r)}(\alpha_э) + \left[\frac{k_r \varkappa_*^{(r)}}{2(l_r+k_r)}\right]^2} = R_i(r).$$

Найдем теперь такую окрестность точки x_i, для которой достигается минимум (по r) величины $R_i(r)$. Пусть мини-

мум достигается на окрестности X_i^{τ}, а $\omega_{i_1}, \ldots, \omega_{i_s}$ — полученная классификация векторов рабочей выборки из этой окрестности. Очевидно, что с вероятностью $1-\eta q$ эта классификация содержит меньше $R_l(\tau)\, k_\tau = R_l$ ошибок.

Аналогично могут быть найдены решения для окрестностей всех векторов генеральной совокупности. В результате получим табл. 1.

Таблица 1

Окрестности точки	Классификация векторов					Оценка величины суммарного риска
	x_{l+1}	$\ldots$	x_{l+j}	$\ldots$	x_{l+k}	
x_1	ω_1^1	$\ldots$	—	$\ldots$	ω_{l+k}^1	R_1
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
x_s	—	$\ldots$	ω_{l+j}^s	$\ldots$	—	R_s
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
x_{l+k}	—	$\ldots$	—	$\ldots$	ω_{l+k}^{l+k}	R_{l+k}

В первом столбце таблицы указаны векторы, задающие систему окрестностей, затем наилучшая по данной системе окрестностей классификация векторов и, наконец, гарантированная оценка числа ошибок классификации.

Заметим, что одни и те же векторы рабочей выборки принадлежат окрестностям различных векторов, а классификация некоторых векторов рабочей выборки, данная в разных строках второго столбца таблицы, может не совпадать.

Обозначим через $\omega_1^*, \ldots, \omega_{l+k}^*$ истинную классификацию векторов рабочей выборки $x_{l+1}, \ldots, x_{l+k}$.

Тогда содержание таблицы может быть переписано в виде

$$\begin{aligned} &\sum_i^{(1)} |\, \omega_{l+i}^* - \omega_{l+i} \,| < R_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ &\sum_i^{(l+k)} |\, \omega_{l+i}^* - \omega_{l+i} \,| < R_{l+k}. \end{aligned} \tag{10.56}$$

Здесь $\sum^{(r)}$ означает, что суммирование ведется лишь по классификациям тех векторов рабочей выборки, которые принадлежат выбранной окрестности точки x_r.

Каждое из неравенств (10.56) выполняется с вероятностью $1-q\eta$. Следовательно, система совместна (все не-

равенства выполняются одновременно) с вероятностью, большей $1-q(l+k)\eta$.

Рассмотрим множество Ω векторов $\bar{w}=(\bar{\omega}_{l+1}, \ldots, \hat{\omega}_{l+k})$ решений системы неравенств (8.56). В принципе в качестве окончательного вектора классификации может быть выбран любой вектор из Ω. Однако целесообразнее в подобных случаях выбирать такое решение, которое обладает некоторыми дополнительными экстремальными свойствами.

Найдем среди векторов Ω минимаксный $-w_m$, т. е. наименее удаленный от самого далекого вектора из допустимого множества Ω:

$$w_m = \arg \min_{w \in \Omega} \max_{\bar{w} \in \Omega} | \bar{w} - w |.$$

Вектор w_m мы и примем за окончательное решение задачи классификации векторов рабочей выборки.

В этом алгоритме задание системы окрестностей векторов полной выборки позволило определить для каждого вектора x_i оптимальную окрестность для построения линейного решающего правила. Полученное правило использовалось лишь для классификации векторов, принадлежащих оптимальной окрестности. Такие алгоритмы иногда называют *локальными*.

На практике используются разные идеи задания окрестностей. В частности, окрестность X_i^r вектора x_i может быть определена по метрической близости (множество X_i^r содержит векторы полной выборки, для которых $\| x_i - x \| \leqslant c$, где c — константа. Набор констант $c_1 < \ldots < c_t$ определяет систему окрестностей).

§ 12. Локальные алгоритмы восстановления значений произвольной функции

По схеме предыдущего параграфа могут быть немедленно построены локальные алгоритмы восстановления значений функции произвольной природы.

Образуем систему окрестностей векторов полной выборки,

$$1) \quad (x_1)_1 \in (x_1, x_{i_1})_2 \in \ldots \in (x_1, \ldots, x_{l+k})_q,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$(l+k) \quad (x_{l+k})_1 \in (x_{l+k}, x_{i_{l+k}})_2 \in \ldots \in (x_1, \ldots, x_{l+k})_q.$$

Пусть произошло разделение множества векторов полной выборки на элементы, принадлежащие обучающей и рабочей выборкам. Рассмотрим систему окрестностей точки x_i:

$$X_i^1 \subset X_i^2 \subset \ldots \subset X_i^q,$$

$$X_i^r = (x_i,\ x_{i_2},\ \ldots,\ x_{i_p})_r.$$

Для каждого множества X_i^r с помощью алгоритмов восстановления линейной функции могут быть найдены как сами значения функции, так и гарантированная оценка величины суммарного риска

$$I_\Sigma^{(r)}(\alpha_э) < \left[\frac{1+\tau a(p)\frac{l_r}{l_r+k_r}\varkappa_*}{1-\tau a(p)\frac{k_r}{l_r+k_r}\varkappa_*}\right]_\infty I_э^{(r)}(\alpha_э), \qquad (10.57)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$n\left(\ln\frac{l_r+k_r}{n}+1\right)+\ln\Gamma_{l_r,\ k_r}(\varkappa) \leqslant \ln\frac{\eta}{1{,}5}. \qquad (10.58)$$

В неравенстве (10.58) l_r, k_r — число элементов обучающей последовательности и рабочей выборок, принадлежащих X_i^r.

Выберем такую окрестность точки x_i и такую функцию $F(x, \alpha_m^*)$, для которой оценка (10.57) минимальна. Пусть k_r^* — число элементов рабочей выборки из этой окрестности.

Для найденных с помощью функции $F(x, \alpha_э^*)$ значений y_j из этой окрестности с вероятностью $1-q\eta$ справедливо неравенство

$$\frac{1}{k_r^*}\sum_{k_r^*}(y_j - F(x_j,\ \alpha_э^*))^2 < \varkappa_r. \qquad (10.59)$$

В выражении (10.59) суммирование ведется по тем векторам x рабочей выборки, которые принадлежат оптимальной окрестности; y — истинные (но неизвестные нам) значения функциональной зависимости в точках рабочей выборки, $F(x_i, \alpha_э^*)$ — вычисленные значения.

Итак, для каждой точки x_i (а всего их $l+k$ — по числу векторов полной выборки) с вероятностью $1-\eta$ справедливо неравенство (10.59). Поэтому с вероятностью

$1-(l+k)q\eta$ выполнятся одновременно все $l+k$ неравенств

$$\begin{gathered}\frac{1}{k_l^*}\sum_{k_l^*}(y_j-F(x_j,\ \alpha_3^*(1)))^2<\varkappa_1,\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ \frac{1}{k_{l+k}^*}\sum_{k_{l+k}^*}(y_j-F(x_j,\ \alpha_3^*(N)))^2<\varkappa_{l+k}.\end{gathered}\tag{10.60}$$

Рассмотрим допустимое множество $\{Y\}$ решений $(y_{l+1}, \ldots, y_{l+k})$ системы (10.60). С вероятностью $1-q\times$ $\times(l+k)\eta$ это множество непусто. Выберем в качестве ответа такое решение Y^*, которое наименее удалено от самой далекой точки в $\{Y\}$ (минимаксное решение), т. е. такой k-мерный вектор Y^*, для которого справедливо

$$Y^*=\arg\min_{Y\in\{Y\}}\max_{\bar{Y}\in\{Y\}}\rho(Y,\ \bar{Y}).$$

§ 13. Замечания о восстановлении значений функции

1. Итак, построение методов восстановления значений функции в заданных точках оказалось связанным с различными способами изучения расположения $l+k$ векторов в n-мерном пространстве.

В этой главе мы изучали геометрию $l+k$ векторов полной выборки с трех различных точек зрения. Мы исследовали:

— структуру линейных разделений векторов полной выборки,

— таксонную структуру векторов полной выборки,

— структуру окрестностей элементов полной выборки.

Описание структуры полной выборки с каждой из этих точек зрения порождает свой метод восстановления значений функции в заданных точках.

Приведенные способы изучения геометрии полной выборки не исчерпывают всех возможных путей изучения взаимного расположения $l+k$ векторов в n-мерном пространстве. Возможны и другие способы и каждый из них может служить основой для построения метода восстановления значений функции.

2. В этой главе при восстановлении значений функции в заданных точках мы считали, что риск определяется квадратичной функцией потерь

$$(y - F(x, \alpha))^2.$$

Однако все полученные здесь результаты могут быть перенесены и на случай функции потерь более общей природы

$$\Phi(y - F(x, \alpha)).$$

3. Эффект от непосредственного восстановления значений функции в заданных точках по сравнению с традиционными методами: восстановлением по обучающей последовательности функции и вычислением ее значений — тем больше, чем меньше объем обучающей выборки.

Этот эффект иллюстрирует табл. 2, полученная на материале решения задач медицинской дифференциальной диагностики методом распознавания образов.

Таблица 2

№	l_0	l_p	m	$\hat{m}$	№	l_0	l_p	m	$\hat{m}$
1	12	21	6	3	6	49	35	13	9
2	24	21	5	2	7	42	35	10	6
3	23	10	3	1	8	52	35	12	8
4	27	14	6	3	9	65	46	14	8
5	29	28	9	3	10	33	57	18	5

В первом столбце таблицы указан номер эксперимента, во втором — длина обучающей последовательности, в третьем — длина рабочей выборки, в четвертом столбце указано число ошибок классификации рабочей выборки с помощью линейного решающего правила, минимизирующего эмпирический риск (метод обобщенного портрета см. гл. XI), в пятом столбце — число ошибок классификации элементов рабочей выборки методом минимизации суммарного риска. Исходная размерность пространства бинарных признаков в этих задачах была равна 60. Задачи решались с помощью алгоритмов, приведенных в главе XI.

Основные утверждения главы X

1. Существуют две разные задачи восстановления: восстановление функции и восстановление значений функции в заданных точках.

В тех случаях, когда нужно прогнозировать значения функции в заданных точках (а не установить модель явления) постановка задачи восстановления значений функции является более адекватной.

2. Решение задачи восстановления значений функции в заданных точках проводится методом упорядоченной минимизации риска. Однако здесь упорядочение осуществляется на классах эквивалентности функций, которые определяются структурой полной выборки векторов x.

3. Возможны различные способы изучения структуры полной выборки, каждый из них индуцирует свой алгоритм восстановления значений функции в заданных точках.

4. Дополнительный эффект увеличения точности прогнозирования может быть получен за счет селекции полной выборки. В частности, селекция полной выборки позволяет за счет отказа от прогнозирования значений функции в некоторых заданных точках увеличить точность прогноза значений функции в остальных точках.

5. Для малых выборок метод непосредственного восстановления значений функции в заданных точках является на практике более точным, чем традиционные.

ПРИЛОЖЕНИЕ К ГЛАВЕ X

ЗАДАЧА ТАКСОНОМИИ

§ П.1. Задача классификации объектов

Пусть требуется разбить множество объектов

$$X = x_1, \ldots, x_l \tag{П.1}$$

на такие подмножества

$$X_1, \ldots, X_m, \tag{П.2}$$

чтобы были выполнены следующие два условия:

1) подмножества не пересекались, т. е.

$$X_i \cap X_j = 0 \qquad (i \neq j); \tag{П.3}$$

2) любой элемент из (П.1) попадал в одно из подмножеств (П.2), т. е.

$$\bigcup_{i=1}^{m} X_i = X, \tag{П.4}$$

и при этом каждое подмножество состояло лишь из «наиболее похожих элементов».

Иначе говоря, требуется при выполнении ограничений (П.3), (П.4) минимизировать некоторый функционал, заданный на множестве всех разбиений множества X и отражающий понятие качества разделения множества X.

Подмножества элементов $X_1, \ldots, X_m$, являющиеся оптимальным решением такой задачи, называются *таксонами*, а сама задача разделения множества X на подмножества — *задачей таксономии*.

Таким образом, проблема состоит в том, чтобы выписать функционал, отражающий наши представления

о качестве разбиения множества, и найти разбиение, доставляющее минимум этому функционалу.

Проблема построения функционала является неформальной — каждый исследователь определяет свое понимание оптимального решения. Тем не менее существует «естественное» определение качества решения для частной постановки задачи таксономии — разбиения исходного множества X на заранее указанное число m таксонов $X_1, \ldots, X_m$.

В этом случае определяется число $d(X_i)$, которое ставит в соответствие каждому подмножеству X_i степень близости его объектов. С помощью величин $d(X_i)$ образуется функционал

$$d = \sum_{i=1}^{m} d(X_i). \qquad (\text{П.5})$$

В теории таксономии приняты следующие характеристики близости объектов множества X_i:

1) средний квадрат уклонения от центра тяжести множеств

$$d_1(X_i) = \frac{1}{l_i} \sum_{j=1}^{l_i} (x_{ij} - \bar{x}_i)^{\mathrm{T}} (x_{ij} - \bar{x}_i),$$

где l_i — число элементов множества X_i, $\bar{x}_i = \frac{1}{l_i} \sum_{j=1}^{l_i} x_{ij}$ — центр тяжести множества X_i; x_{ij} — элементы множества X_i,

2) средний квадрат уклонения между элементами множества

$$d_2(X_i) = \frac{1}{l_i(l_i-1)} \sum_{\substack{j,\ t \\ j>t}} (x_{ij} - x_{it})^{\mathrm{T}} (x_{ij} - x_{it}),$$

3) величина определителя матрицы рассеяния векторов множества

$$|d_3(X_i)| = \left| \frac{1}{l} \sum_{j=1}^{l} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^{\mathrm{T}} \right|.$$

С помощью этих характеристик образуются функционалы

$$d_1 = \sum_{i=1}^{m} d_1(X_i),$$

$$d_2 = \sum_{i=1}^{m} d_2(X_i), \qquad (\text{П.6})$$

$$d_3 = \left| \sum_{i=1}^{m} d_3(X_i) \right|.$$

Здесь $\left| \sum_{i=1}^{m} d_3(X_i) \right|$ — определитель матрицы $\sum_{i=1} d_3(X_i)$.

Существуют и другие функционалы, отражающие различное понимание качества решения частной задачи таксономии (число таксонов задано заранее). Эти функционалы приведены в работах [1, 2, 19].

Что же касается общей задачи таксономии (где число таксонов заранее неизвестно), то здесь нет достаточно широко принятых определений качества таксономии.

§ П.2. Алгоритмы таксономии

Задача минимизации функционала (П.5) на множестве возможных разделений l объектов на m групп является задачей дискретного программирования: всего возможно

$$N(l, m) = \sum_{i=0}^{m} (-1)^i C_m^i (m-i)^l$$

различных разбиений l объектов на m групп так, чтобы ни одна из групп не была пустой. Необходимо среди $N(l, m)$ разбиений выбрать то, которое минимизирует функционал (П.5).

Точное решение этой задачи требует большого числа вычислений (соизмеримых по объему с величиной $N(l, m)$). Поэтому для решения задачи таксономии приняты эвристические приемы, и в частности следующий прием. На множестве векторов $X(x_1, \ldots, x_l)$ строятся две последовательности $\tilde{X}_l = \tilde{x}_1, \ldots, \tilde{x}_l$ и $Q = q_1, \ldots, q_l$ по следующему индуктивному правилу:

1) Вначале выбирается любой элемент из X, например x_1, и полагается $\tilde{x}_1 = x_1$, $q_1 = 0$. Вектор x_1 исключается из множества X, образуя множество M_1 ($M_1 = X/x_1$).

2) Пусть к $(t+1)$-му шагу из исходного множества отобрано t векторов и построены последовательности

$$\tilde{X}_t = \tilde{x}_1, \ \ldots, \ \tilde{x}_t,$$
$$Q_t = q_1, \ \ldots, \ q_t,$$

а оставшиеся векторы объединены в множество M_t. Тогда на $(t+1)$-м шаге к последовательности $\tilde{X}_t$ добавляется ближайший из M_t вектор, т. е. такой вектор $x = \tilde{x}_{t+1}$, на котором достигается минимум

$$q = \min_{x_i \in M_t} \rho(x_i, \ \tilde{X}_t).$$

Этот вектор добавляется к построенной последовательности, образуя последовательность $\tilde{X}_{t+1}$, а соответствующая величина q_{t+1} добавляется к Q_t, образуя новую последовательность Q_{t+1}. Из множества M_t исключается вектор $\tilde{x}_{t+1}$, образуя множество M_{t+1}.

Так продолжается до тех пор, пока все векторы из X не будут упорядочены.

3) Используя последовательность $Q_t = q_1, \ \ldots, \ q_t$, можно разбить последовательность $\tilde{X}_t = \tilde{x}_1, \ \ldots, \ \tilde{x}_t$ на m таксонов. Для этого выберем такое число q^*, чтобы только $m-1$ значений последовательности $Q_t = q_1, \ \ldots, \ q_t$ превосходили q^*. Пусть $q_{i_1}, \ q_{i_2}, \ \ldots, \ q_{i_{m-1}}$ — соответствующие значения последовательности Q_{t+1}.

Тогда векторы из последовательности $\tilde{X}_t$ с номерами $1 \div i_1$, образуют первый таксон; векторы с номерами $i_1 \div i_2$ — второй, и т. д. Всего m таксонов.

Для того чтобы задать конкретный алгоритм, нужно определить понятие расстояния от точки x до множества X. Обычно используют следующую метрику $\rho(x, \ X)$: расстояние от x до X определяется величиной расстояния от x до ближайшего из X вектора.

Что же касается расстояния между двумя векторами x_a и x_b, элементами множества X, то наряду с обычной евклидовой метрикой в таксономии используются и специальные метрики.

В частности, используется *метрика Танимото*, определяющая близость двух множеств, множества X и множества Y:

$$\rho_{\text{т}}(X,\ Y) = \frac{n_x + n_y - 2n_{xy}}{n_x + n_y - n_{xy}},$$

где n_x — число объектов множества X, n_y — число объектов множества Y, n_{xy} — число объектов, которые входят одновременно в оба множества.

С помощью метрики Танимото определяется близость множеств объектов $X^{\delta}(x_a)$ из X, попадающих в δ-окрестность точки x_a и множество объектов $X^{\delta}(x_b)$, попадающих в δ-окрестность точки x_b (δ — заданный параметр).

Таким образом,

$$\rho(x_a,\ x_b) = \rho_{\text{т}}(X^{\delta}(x_a);\ X^{\delta}(x_b)).$$

Такая метрика более соответствует исследованию «геометрии векторов в целом».

ГЛАВА XI

АЛГОРИТМЫ ОБУЧЕНИЯ РАСПОЗНАВАНИЮ ОБРАЗОВ

§ 1. Замечания об алгоритмах

В первых десяти главах книги была изложена теория восстановления зависимостей по эмпирическим данным.

Были рассмотрены классические методы восстановления зависимостей (гл. III, IV, V). Они эффективны в условиях, когда искомая зависимость принадлежит заданному классу, и гарантируют отыскание удовлетворительного решения при достаточно большом объеме обучающей выборки.

На практике же мы не уверены ни в том, что искомая зависимость принадлежит классу функций, в котором ведется восстановление, ни в том, что объем выборки достаточен для нахождения хорошего приближения.

Поэтому были развиты методы минимизации риска, которые не требуют знания модели искомой зависимости и ориентированы на использование выборок ограниченного объема (гл. VI — X).

Последние две главы книги посвящены вопросам создания алгоритмов восстановления.

В этой главе мы рассмотрим алгоритмы обучения распознаванию образов. Алгоритмы основаны на использовании оценок равномерного относительного уклонения частот от вероятностей, которые справедливы при любой вероятностной мере $P(x, \omega)$ (в том числе и наиболее неблагоприятной).

Обычно, когда дело доходит до построения алгоритмов, основанных на некоторой теории, оказывается, что развитая теория является все-таки грубым приближением к реальности.

Как правило, эта «грубость» компенсируется тем, что при построении алгоритмов теории не следуют буквально. Авторы привносят в алгоритмы свое понимание реальной действительности, которое не поддается формализации. Так и в нашем случае.

На практике нет оснований думать, что реализуется наиболее неблагоприятное распределение $P(x, \omega)$. Поэтому оценки, которые следуют из общей теории, в реальной ситуации могут оказаться завышенными. Как же учесть, что мы собираемся иметь дело с реальными законами распределения вероятностей, а не с наиболее неблагоприятным? Ответ на этот вопрос и определяет степень нашего неформального отношения к построенной теории.

Неформальное отношение к теории при построении алгоритмов обучения распознаванию образов в этой главе состоит в том, что при восстановлении характеристических функций мы будем считать, что наряду с оценкой

$$P(\alpha) < \nu_э(\alpha) + \\ + 2\frac{h\left(\ln\frac{2l}{h}+1\right) - \ln\frac{\eta}{12}}{l}\left(1 + \sqrt{1 + \frac{\nu_э(\alpha)\, l}{h\left(\ln\frac{2l}{h}+1\right) - \ln\frac{\eta}{12}}}\right) \quad (11.1)$$

справедлива оценка, отличающаяся от (11.1) константами

$$P(\alpha) < \nu_э(\alpha) + \\ + \frac{h\left(\ln\frac{l}{h}+1\right) - \ln\eta}{2l}\left(1 + \sqrt{1 + \frac{\nu_э(\alpha)\, l}{h\left(\ln\frac{l}{h}+1\right) - \ln\eta}}\right), \quad (11.2)$$

а при восстановлении значений характеристической функции наряду с оценкой

$$\nu_\Sigma(\alpha) < \nu_э(\alpha) + \frac{k}{2(l+k)}\varkappa_*^2 + \varkappa_* \sqrt{\nu_э(\alpha) + \left[\frac{k\varkappa_*}{2(l+k)}\right]^2}, \quad (11.3)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$h\left(\ln\frac{l+k}{h}+1\right) + \ln\Gamma_{l,k}(\varkappa) \leqslant \ln\frac{\eta}{1{,}5},$$

справедлива оценка

$$\nu_\Sigma(\alpha) < \nu_э(\alpha) + \frac{k}{2(l+k)}\varkappa_*^2 + \varkappa_* \sqrt{\nu_э(\alpha) + \left[\frac{k\varkappa_*}{2(l+k)}\right]^2}, \quad (11.4)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$h\left(\ln\frac{l+k}{h}+1\right) + \ln\Gamma_{l,k}(\varkappa) \leqslant \ln\eta. \quad (11.5)$$

Эта оценка также отличается от (11.3) константами. Кроме того, мы перенесем некоторые факты, имеющие место для задачи восстановления значений функций, на задачу восстановления функции.

§ 2. Построение разделяющих гиперплоскостей

Основой создания алгоритмов обучения распознаванию образов в классе линейных решающих правил являются алгоритмы построения гиперплоскости, разделяющей два конечных множества векторов: множество векторов

$$X_a : x_1, \ldots, x_a \tag{11.6}$$

и множество векторов

$$\bar{X}_b : \bar{x}_1, \ldots, \bar{x}_b. \tag{11.7}$$

Задача сводится к отысканию вектора ψ, для которого выполняются неравенства

$$\begin{aligned} x_i^{\mathrm{T}}\psi &\geqslant 1, \quad \text{если} \quad x_i \in X_a, \\ \bar{x}_j^{\mathrm{T}}\psi &\leqslant k, \quad \text{если} \quad \bar{x}_j \in \bar{X}_b, \end{aligned} \qquad k < 1. \tag{11.8}$$

Очевидно, если существует вектор ψ, для которого выполняются неравенства (11.8), то имеется множество векторов ψ, удовлетворяющих (11.8). Будем искать среди них минимальный по модулю вектор. Этот вектор был назван *обобщенным портретом* [12].

Минимизация квадратичной формы

$$\|\psi\|^2 = \psi^{\mathrm{T}}\psi \tag{11.9}$$

при ограничениях (11.8) является задачей квадратичного программирования.

Необходимые и достаточные условия минимума (11.9) при ограничениях (11.8) определяются теоремой Куна — Таккера.

Теорема 11.1 (Кун — Таккер). *Пусть заданы дифференцируемая выпуклая функция $F(x)$ и линейные функции $f_i(x)$ $(i = 1, 2, \ldots, l)$. Пусть x_0 доставляет минимум $F(x)$ при ограничениях*

$$f_i(x) \geqslant 0. \tag{11.10}$$

Тогда существуют такие $\lambda_i \geqslant 0$, удовлетворяющие условиям

$$\lambda_i f_i(x_0) = 0, \tag{11.11}$$

что справедливо равенство

$$\nabla F(x_0) = \sum_{i=1}^{l} \lambda_i \nabla f_i(x_0) \tag{11.12}$$

(∇ — *знак градиента*).

И обратно, если для некоторой точки x_0 *выполняются условия* (11.10) *и можно найти числа* $\lambda_i \geqslant 0$, *удовлетворяющие условиям* (11.11) *и* (11.12), *то в точке* x_0 *достигается условный минимум* $F(x)$ *при ограничениях* (11.10).

Доказательство теоремы Куна — Таккера приводится во всех руководствах по выпуклому программированию (например, [65]).

Применим теорему Куна — Таккера для нашего случая минимизации (11.9) при ограничениях (11.8).

Теорема 11.2. *Минимальный по модулю вектор* ψ, *удовлетворяющий* (11.8) (*обобщенный портрет*), *представим в виде*

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j, \tag{11.13}$$

$$\alpha_i \geqslant 0, \quad \beta_j \geqslant 0,$$

причем

$$\begin{aligned} &\alpha_i [x_i^{\mathrm{T}} \psi - 1] = 0, \quad i = 1, 2, \ldots, a, \\ &\beta_j [k - x_j^{\mathrm{T}} \psi] = 0, \quad j = 1, 2, \ldots, b. \end{aligned} \tag{11.14}$$

Среди всех векторов ψ, *удовлетворяющих* (11.8), *вектор* ψ, *представимый в виде* (11.13), (11.14), *является минимальным по модулю.*

Доказательство теоремы немедленно следует из теоремы Куна — Таккера.

Назовем векторы x_i, $\bar{x}_j$, для которых выполняются условия

$$\begin{aligned} &x_i^{\mathrm{T}} \psi = 1, \quad x_i \in X_a, \\ &\bar{x}_j^{\mathrm{T}} \psi = k, \quad \bar{x}_j \in \bar{X}_b, \end{aligned} \tag{11.15}$$

крайними векторами. Согласно теореме 11.2 обобщенный портрет разложим с ненулевыми весами лишь по системе крайних векторов.

Рассмотрим теперь *двойственную задачу*, решение которой эквивалентно построению обобщенного портрета. Введем

пространство параметров $E_{\alpha\beta}$ и рассмотрим функцию

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i - k \sum_{j=1}^{b} \beta_j - \frac{1}{2} \psi^{\mathrm{T}} \psi, \qquad (11.16)$$

где вектор ψ есть

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j.$$

Покажем, что точка α_0, β_0 — максимума функции $W(\alpha, \beta)$ в положительном квадранте $\alpha_i \geqslant 0$, $\beta_j \geqslant 0$ определяет обобщенный портрет.

Действительно, необходимыми и достаточными условиями максимума функции $W(\alpha, \beta)$ в точке α_0, β_0 являются условия

$$\frac{\partial W(\alpha_0, \beta_0)}{\partial \alpha_i} = \begin{cases} 0, & \text{если } \alpha_i^0 > 0, \\ \leqslant 0, & \text{если } \alpha_i^0 = 0, \end{cases} \quad i = 1, 2, \ldots, a,$$

$$\frac{\partial W(\alpha_0, \beta_0)}{\partial \beta_j} = \begin{cases} 0, & \text{если } \beta_j^0 > 0, \\ \leqslant 0, & \text{если } \beta_j^0 = 0, \end{cases} \quad j = 1, 2, \ldots, b.$$

Выпишем эти условия, обозначив

$$\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j.$$

Получим

$$\begin{aligned} 1 - x_i^{\mathrm{T}} \psi_0 &= \begin{cases} 0, & \text{если } \alpha_i^0 > 0, \\ \leqslant 0, & \text{если } \alpha_i^0 = 0, \end{cases} \quad i = 1, 2, \ldots, a, \\ \bar{x}_j^{\mathrm{T}} \psi - k &= \begin{cases} 0, & \text{если } \beta_j^0 > 0, \\ \leqslant 0, & \text{если } \beta_j^0 = 0, \end{cases} \quad j = 1, \ldots, b. \end{aligned} \qquad (11.17)$$

Условия (11.17) могут быть переписаны в виде неравенств

$$\begin{aligned} x_i^{\mathrm{T}} \psi_0 &\geqslant 1, \quad i = 1, 2, \ldots, a, \\ \bar{x}_j^{\mathrm{T}} \psi_0 &\leqslant k, \quad j = 1, 2, \ldots, b \end{aligned} \qquad (11.18)$$

и равенств

$$\alpha_i^0 (1 - x_i^{\mathrm{T}} \psi_0) = 0, \quad i = 1, 2, \ldots, a,$$
$$\beta_j^0 (\bar{x}_j^{\mathrm{T}} \psi_0 - k) = 0, \quad j = 1, 2, \ldots, b.$$

Согласно же утверждению теоремы 11.2 эти условия определяют обобщенный портрет.

Итак, задача построения гиперплоскости, разделяющей два множества векторов, свелась к отысканию максимума функции $W(\alpha, \beta)$ в положительном квадранте.

Ниже мы рассмотрим методы минимизации квадратичной формы $W(\alpha, \beta)$ в положительном квадранте, но прежде установим следующий важный факт.

Теорема 11.3. *Если разделяющая гиперплоскость существует (существует вектор ψ_0, для которого выполняются неравенства* (11.18)), *то максимум функции $W(\alpha, \beta)$ в положительном квадранте равен половине квадрата модуля обобщенного портрета*

$$W(\alpha_0, \beta_0) = \frac{\|\psi_0\|^2}{2}. \tag{11.19}$$

Доказательство. Действительно, согласно теореме 11.2

$$\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j.$$

Поэтому

$$\|\psi_0\|^2 = \psi_0^{\mathrm{T}}\psi_0 = \sum_{i=1}^{a} \alpha_i^0 x_i^{\mathrm{T}}\psi - \sum_{j=1}^{b} \beta_j^0 \bar{x}_j^{\mathrm{T}}\psi_0$$

и, учитывая (11.15), получаем

$$\|\psi_0\|^2 = \sum_{i=1}^{a} \alpha_i^0 - k \sum_{j=1}^{b} \beta_j.$$

Таким образом,

$$W(\alpha_0, \beta_0) = \sum_{i=1}^{a} \alpha_i^0 - k \sum_{j=1}^{b} \beta_j^0 - \frac{1}{2}\psi_0^{\mathrm{T}}\psi_0 = \frac{\|\psi_0\|^2}{2}.$$

Теорема доказана.

Из теоремы 11.3 вытекает важное для построения алгоритмов распознавания следствие.

Следствие. *Если среди крайних векторов обобщенного портрета ψ_0 есть векторы обоих классов, то имеет место оценка*

$$\rho(\psi_0) \geqslant \frac{1-k}{\sqrt{2W(\alpha, \beta)}}, \tag{11.20}$$

где $\rho(\psi_0)$ — *расстояние между проекциями множеств* $x_1, \ldots, x_a$ *и* $\bar{x}_1, \ldots, \bar{x}_b$ *на направление обобщенного портрета.*

При этом равенство в оценке (11.20) *достигается в точке* $\alpha = \alpha_0$, $\beta = \beta_0$.

Доказательство. В силу теоремы 11.3

$$\sqrt{2W(\alpha_0, \beta_0)} = \|\psi_0\|.$$

Далее, в силу условия следствия найдутся такие векторы множества, что

$$\begin{aligned} x_i^{\mathrm{T}} \frac{\psi_0}{\|\psi_0\|} &= \frac{1}{\|\psi_0\|}, \\ \bar{x}_i^{\mathrm{T}} \frac{\psi_0}{\|\psi_0\|} &= \frac{k}{\|\psi_0\|}. \end{aligned} \tag{11.21}$$

Поэтому расстояние между проекциями векторов, для которых выполнялось (11.21), равно

$$\rho(\psi_0) = \frac{1-k}{\|\psi_0\|} = \frac{1-k}{\sqrt{2W(\alpha_0, \beta_0)}}.$$

Учитывая, что $W(\alpha, \beta) \geqslant W(\alpha_0, \beta_0)$, получаем неравенство (11.20).

Это следствие используется для построения критерия неразделимости векторов. В самом деле, будем считать, что два конечных множества векторов не могут быть разделимы гиперплоскостью, если расстояние между проекциями на направление обобщенного портрета меньше ρ_0. А это значит, что не существует разделимости, если найдутся такие $\alpha^* > 0$, $\beta^* > 0$, что

$$W(\alpha^*, \beta^*) > \frac{(1-k)^2}{2\rho_0^2} = W_0.$$

Таким образом, при построении обобщенного портрета проблема состоит в том, чтобы найти максимум отрицательно определенной квадратичной формы $W(\alpha, \beta)$ в положительном квадранте $\alpha > 0$, $\beta > 0$ или установить, что максимум функции $W(\alpha, \beta)$ превосходит величину W_0. Последнее означает, что построение обобщенного портрета невозможно.

§ 3. Алгоритмы максимизации квадратичной формы

Одним из наиболее эффективных алгоритмов максимизации отрицательно определенной квадратичной формы является метод сопряженных градиентов. С его помощью удается достичь максимума за n шагов (n — размерность формы). В этом параграфе мы рассмотрим алгоритмы максимизации отрицательно определенной квадратичной формы в положительном квадранте, построенные на основе модификации метода сопряженных градиентов.

Теория метода сопряженных градиентов описана во многих руководствах по поиску максимума функции (например, [12,80]).

Рассмотрим сначала метод сопряженных градиентов для максимизации квадратичной формы:

$$F(y) = b^{\mathrm{T}}y - y^{\mathrm{T}}Ay,$$

где A — положительно определенная матрица, b — вектор, y — вектор.

Согласно методу сопряженных градиентов поиск максимума функции начинается из произвольной точки $y_0 = y(0)$. Первый шаг делается в направлении градиента функции $F(y)$ в точке $y(0)$. Обозначим градиент функции в точке $y(0)$ через $g(1)$, а направление движения из точки $y(0)$ через $z(1)$. Таким образом,

$$z(1) = g(1).$$

Шаг делается по направлению $z(1)$ до достижения максимума. Нетрудно убедиться, что максимум по направлению $z(1)$ задается выражением

$$y(1) = y(0) + \frac{z^{\mathrm{T}}(1)\, g(1)}{z^{\mathrm{T}}(1)\, Az(1)}\, z(1).$$

Начиная со второго шага, направление движения определяется вектором

$$z(t+1) = g(t+1) + \frac{\|g(t+1)\|^2}{\|g(t)\|^2}\, z(t), \qquad (11.22)$$

где $g(t+1)$ и $g(t)$ — градиенты функции $F(y)$ в точках $y(t+1)$ и $y(t)$, $z(t)$ — направление движения в точке $y(t-1)$. Движение по направлению $z(t)$ ведется до достижения условного максимума. Этот максимум достигается

в точке

$$y(t) = y(t-1) + h(t)\, z(t), \tag{11.23}$$

где величина

$$h(z) = \frac{z^{\mathrm{T}}(t)\, g(t)}{z^{\mathrm{T}}(t)\, Az(t)}$$

определяет шаг движения.

Формулы (11.22), (11.23) задают, таким образом, алгоритм поиска максимума квадратичной функции $F(y)$.

Для вычисления максимума функции в положительном квадранте используем модифицированный метод сопряженных градиентов. Модификация метода направлена на то, чтобы ограничить область поиска положительным квадрантом. Определим функцию

$$\hat{g}_i(y) = \begin{cases} \dfrac{\partial F(y)}{\partial y_i}, & \text{если} \quad y_i \neq 0 \quad \text{или} \quad \dfrac{\partial F(y)}{\partial y_i} > 0, \\ 0, & \text{если} \quad y_i = 0 \quad \text{и} \quad \dfrac{\partial F(y)}{\partial y_i} \leqslant 0. \end{cases} \tag{11.24}$$

Вектор $\hat{g}(y) = (\hat{g}_1(y), \ldots, \hat{g}_n(y))^{\mathrm{T}}$, есть условный градиент функции $F(y)$ на множестве $y \geqslant 0$.

Будем совершать восхождения к максимуму, используя формулы (11.22), (11.23), где $g(y)$ заменено на $\hat{g}(y)$. Движение начинается из произвольной точки положительного квадранта и продолжается до момента выхода на ограничение в точке y_0. Тогда снова начинается восхождение по методу сопряженных градиентов, но из точки y_0. Поиск максимума заканчивается, когда выполнятся неравенства

$$|\hat{g}_i(y)| \leqslant \varepsilon \qquad (i = 1, 2, \ldots, n).$$

Для того чтобы траектория не вышла за пределы положительного квадранта, величина шага $\hat{h}(t)$ выбирается из условия минимума двух величин $h(t)$ и $h^*(t)$, где

$$h^*(t) = \min_i \frac{y_i(t)}{|z_i(t+1)|}.$$

При вычислении $h^*(t)$ минимум определяется лишь по тем координатам y_i, для которых $z_i < 0$. Если же все $z_i \geqslant 0$, то шаг равен $h(t)$.

Важной особенностью такого метода поиска максимума функции $F(y)$ в положительном квадранте является то,

что он допускает последовательную процедуру поиска. Пусть пространство E_n имеет координаты

$$y_1, \ldots, y_k, y_{k+1}, \ldots, y_n.$$

Можно сначала найти условный максимум функции $F(y)$ при ограничениях

$$y_1 \geqslant 0, \ldots, y_k \geqslant 0, \quad y_{k+1} = 0, \ldots, y_n = 0,$$

а затем, используя найденную точку максимума как начальную, найти максимум $F(y)$ в области

$$y_1 \geqslant 0, \ldots, y_n \geqslant 0.$$

В нашем случае при поиске максимума функции

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i - k \sum_{j=1}^{b} \beta_j - \frac{1}{2} \psi^{\mathrm{T}} \psi,$$

$$\psi = \sum_{i=1}^{a} \alpha_i x_i - \sum_{j=1}^{b} \beta_j \bar{x}_j,$$

в положительном квадранте, условный градиент есть вектор с координатами

$$\dot{\alpha}_i = \begin{cases} \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i}, & \text{если } \alpha_i \geqslant 0 \text{ или } \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i} > 0, \\ 0, & \text{если } \alpha_i = 0 \text{ и } \dfrac{\partial W(\alpha, \beta)}{\partial \alpha_i} \leqslant 0, \end{cases} \quad i = 1, 2, \ldots, a,$$

$$\dot{\beta}_j = \begin{cases} \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j}, & \text{если } \beta_j \geqslant 0 \text{ или } \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j} > 0, \\ 0, & \text{если } \beta_j = 0 \text{ и } \dfrac{\partial W(\alpha, \beta)}{\partial \beta_j} \leqslant 0, \end{cases} \quad j = 1, 2, \ldots, b.$$

Обозначим составляющие вектора $z(t)$, задающего направление движения на t-м шаге, через $\bar{\alpha}$, $\bar{\beta}$.

Согласно (11.22) имеют место соотношения

$$\begin{aligned} \bar{\alpha}(t+1) &= \dot{\alpha}(t+1) + \delta(t+1)\,\bar{\alpha}(t), \\ \bar{\beta}(t+1) &= \dot{\beta}(t+1) + \delta(t+1)\,\bar{\beta}(t), \end{aligned} \tag{11.25}$$

где (см. (11.22))

$$\delta(t)=\frac{\sum\limits_{i=1}^{a}\dot{\alpha}_i^2(t+1)+\sum\limits_{j=1}^{b}\dot{\beta}_j^2(t+1)}{\sum\limits_{i=1}^{a}\dot{\alpha}_i^2(t)+\sum\limits_{j=1}^{b}\dot{\beta}_j^2(t)}. \qquad (11.26)$$

При вычислении шага по формуле (11.23) необходимо вычислить величину $z^{\mathrm{T}}Az$.

В нашем случае

$$z^{\mathrm{T}}Az=\left(\sum_{i=1}^{a}\bar{\alpha}_i x_i-\sum_{j=1}^{b}\bar{\beta}_j\bar{x}_j\right)^2=\bar{\psi}^{\mathrm{T}}\bar{\psi},$$

где обозначено

$$\bar{\psi}=\sum_{i=1}^{a}\bar{\alpha}_i x_i-\sum_{j=1}^{b}\bar{\beta}_j\bar{x}_j.$$

Таким образом, используя методы сопряженного градиента, можно либо найти гиперплоскость, разделяющую два множества векторов: множество $x_1, \ldots, x_a$ и множество $\bar{x}_1, \ldots, \bar{x}_b$ (найти максимум функции $W(\alpha, \beta)$ в положительном квадранте), либо установить, что разделяющей гиперплоскости не существует (установить на очередном шаге, что $W(\alpha, \beta)>\frac{(1-k)^2}{2\rho^2}$, ρ — заданный параметр).

§ 4. Методы построения оптимальной разделяющей гиперплоскости

При создании алгоритмов восстановления зависимостей одним из важных моментов является построение *оптимальной разделяющей гиперплоскости.*

Оптимальной разделяющей гиперплоскостью называется такая гиперплоскость, которая, разделяя два множества векторов: $x_1, \ldots, x_a$ и $\bar{x}_1, \ldots, \bar{x}_b$, максимально от них удалена.

Формально это означает, что оптимальная разделяющая гиперплоскость задается такой парой: единичным вектором φ, и числом c, для которых при выполнении

неравенств

$$x_i^{\mathrm{T}}\varphi \geqslant c, \quad i=1, 2, \ldots, a,$$
$$\bar{x}_j^{\mathrm{T}}\varphi < c, \quad j=1, 2, \ldots, b,$$

где

$$c = \frac{c_1(\varphi)+c_2(\varphi)}{2}, \quad c_1(\varphi) = \min_i x_i^{\mathrm{T}}\varphi, \quad c_2(\varphi) = \max_j \bar{x}_j^{\mathrm{T}}\varphi,$$

достигается максимум выражения

$$\Pi(\varphi) = c_1(\varphi) - c_2(\varphi).$$

Для построения оптимальной разделяющей гиперплоскости рассмотрим всевозможные разности

$$y_{ij} = x_i - \bar{x}_j; \quad x_i \in X_a, \quad \bar{x}_j \in \overline{X}_b.$$

Вектор $\varphi_{\text{опт}}$ обладает свойством

$$\min_{i,j} y_{ij}^{\mathrm{T}}\varphi_{\text{опт}} = \max_{\|\psi\|=1} \min_{i,j} y_{ij}^{\mathrm{T}} \frac{\psi}{\|\psi\|},$$

поэтому он коллинеарен минимальному по модулю вектору ψ, для которого выполняются неравенства

$$y_{ij}^{\mathrm{T}}\psi \geqslant 1, \quad i=1, 2, \ldots, a, \quad j=1, 2, \ldots, b.$$

Иными словами, вектор φ коллинеарен обобщенному портрету ψ класса $\{y_{ij}\}$ при пустом втором классе.

Отыскать же обобщенный портрет можно, максимизируя квадратичную форму

$$W(\alpha) = \sum_{i,j} \alpha_{ij} - \frac{1}{2}\psi^{\mathrm{T}}\psi,$$
$$\psi = \sum_{i,j} \alpha_{ij} y_{ij}$$

в положительном квадранте $\alpha_{ij} \geqslant 0$.

Число векторов y_{ij} обычно достаточно велико. Поэтому непосредственное построение обобщенного портрета ψ затруднительно. Воспользуемся следующей итеративной процедурой.

1. Берется произвольная пара $y_1 = x_1 - \bar{x}_1$. Образуется класс Y_1, состоящий всего из одного вектора $x_1 - \bar{x}_1$. Строится обобщенный портрет этого класса (при пустом втором классе).

2. Пусть на t-м шаге построены класс Y_t векторов $x_i - \bar{x}_i$ и его обобщенный портрет ψ_t. В обучающей последовательности находится такой вектор $x_{i_{t+1}}$, что

$$x^{\mathrm{T}}_{i_{t+1}}\psi_t = \min_{x_i \in X_a} x_i^{\mathrm{T}}\psi_t,$$

и такой вектор $\bar{x}_{j_{t+1}}$, что

$$\bar{x}^{\mathrm{T}}_{j_{t+1}}\psi_t = \max_{\bar{x}_i \in \bar{X}_b} \bar{x}_j^{\mathrm{T}}\psi_t.$$

Образуется вектор

$$y_{t+1} = x_{i_{t+1}} - \bar{x}_{j_{t+1}}.$$

3. Если окажется, что

$$y^{\mathrm{T}}_{t+1}\psi_t < 1 - \varepsilon$$

(ε — параметр алгоритма), то класс Y_t пополняется вектором y_{t+1}. Находится обобщенный портрет ψ_{t+1} образовавшегося класса Y_{t+1}, и процесс продолжается дальше.

Если же выполнется неравенство

$$y^{\mathrm{T}}_{t+1}\psi_t \geqslant 1 - \varepsilon,$$

то процесс заканчивается и за оптимальную разделяющую гиперплоскость принимается гиперплоскость

$$x^{\mathrm{T}}\psi_t = \frac{\min\limits_i x_i^{\mathrm{T}}\psi + \max\limits_j \bar{x}_j^{\mathrm{T}}\psi}{2}.$$

Одновременно с процессом построения гиперплоскости проверяется условие

$$W(\alpha) > \frac{2}{\rho^2}.$$

Если хотя бы однажды оно выполнится, то построение гиперплоскости оканчивается. В этом случае считается, что разделение векторов обучающей последовательности невозможно.

При реализации этой процедуры удобно на каждой итерации при образовании класса Y_t удалять те векторы y_i, которые входили в разложение обобщенного портрета ψ с нулевым весом. Уменьшение числа векторов в Y_t позволяет сократить время построения обобщенного портрета ψ.

Рассмотренные алгоритмы построения разделяющей гиперплоскости мы используем при создании комплекса программ обучения распознаванию образов. Однако прежде остановимся на вопросе представления информации в задачах распознавания.

§ 5. Алгоритм экстремального разбиения значений признака на градации

В задачах обучения распознаванию образов приняты два способа представления информации — непрерывный и дискретный.

При непрерывном способе представления информации координаты вектора x могут принимать любые значения. При дискретном способе каждая координата вектора принимает фиксированное число значений. Дискретным способом удобно кодировать качественные признаки. Например, следующие характеристики в задачах медицинской дифференциальной диагностики: «бледность кожных покровов не выражена», «выражена умеренно», «сильно выражена», могут иметь коды 100, 010, 001.

Однако в задачах обучения распознаванию образов принято дискретно кодировать не только признаки, отражающие качественную характеристику объекта, но и признаки, принимающие числовые значения.

При этом пользуются следующим способом представления информации. Весь диапазон значений параметра разбивается на ряд градаций. Единицей кодируется j-й разряд кода, если значение параметра принадлежит j-й градации, если же значение параметра не принадлежит j-й градации, то в j-м разряде ставится нуль.

Пример. Пусть значение параметра x^i принадлежит отрезку $[-5, 8]$ и этот отрезок разбивается на пять градаций:

$$x^i < 0; \quad 0 \leqslant x^i < 2; \quad 2 \leqslant x^i < 4; \quad 4 \leqslant x^i < 5; \quad x^i \geqslant 5.$$

Кодом 10 000 обозначаются величины $x^i < 0$, кодом 01000 — величины $0 \leqslant x^i < 2$, кодом 00100 — величины $2 \leqslant x^i < 4$, кодом 00010 — величины $4 \leqslant x^i < 5$ и, наконец, кодом 00001 — величины $x^i \geqslant 5$.

Рассмотренный способ представления информации замечателен не только тем, что позволяет компактно записывать информацию (для приведенного примера вместо одной ячейки памяти в ЦВМ — пять разрядов ячейки). Дискре-

тизация координат вектора есть нелинейная операция, с помощью которой вектор x переводится в бинарный вектор x' с большим числом координат.

Использование большого числа градации при кодировке параметра эквивалентно использованию более разнообразного класса разделяющих поверхностей в пространстве E_n, чем линейные. Однако, как было установлено в главе VIII, чрезмерно большая емкость класса решающих правил при ограниченном объеме обучающей последовательности недопустима, и поэтому возникает проблема экстремальной разбивки на градации непрерывных признаков.

В этом параграфе будет приведен алгоритм экстремального разбиения значений признака на градации. Принцип, который реализует алгоритм, заключается в следующем: необходимо так разбить значения параметра на конечное число градаций, чтобы оценка неопределенности (энтропии) при классификации с помощью этого признака была минимальной (или близка к минимальной).

Итак, пусть признак (координата) x может принимать значения из интервала $c \leqslant x \leqslant C$, и пусть вектор, обладающий этим признаком, принадлежит одному из K классов. Пусть существуют условные вероятности принадлежности к каждому классу

$$P(1 \mid x), \ldots, P(K \mid x).$$

Для каждого фиксированного значения признака x может быть определена мера неопределенности (энтропия) принадлежности к тому или иному классу

$$H(x) = -\sum_{i=1}^{K} P(i \mid x) \ln P(i \mid x).$$

Среднее значение по мере $P(x)$ энтропии вычисляется так:

$$H = \int H(x) P(x)\, dx.$$

Пусть теперь параметр x разбит на τ градаций, т. е. принимает одно из τ значений $c(1), \ldots, c(\tau)$. Тогда средняя энтропия может быть записана в виде

$$H(\tau) = -\sum_{j=1}^{\tau} \sum_{i=1}^{K} P(i \mid x_j) \ln P(i \mid x_j) P(x_j). \quad (11.27)$$

Воспользуемся формулой Байеса

$$P(i \mid x_j) = \frac{P(x_j \mid i)\, P(i)}{P(x_j)}. \qquad (11.28)$$

Подставим (11.28) в (11.27)

$$H(\tau) = -\sum_{j=1}^{\tau} \sum_{i=1}^{K} P(x_j \mid i)\, P(i) \ln \frac{P(x_j \mid i)\, P(i)}{P(x_j)}. \qquad (11.29)$$

Для того чтобы оценить энтропию (11.29), необходимо оценить вероятности $P(x_j \mid i)$, $P(i)$, $P(x_j)$ по обучающей последовательности. Воспользуемся байесовыми оценками (см. § 6 гл. III):

$$H(\tau) = -\sum_{j=1}^{\tau} \sum_{i=1}^{K} \frac{m_j(i)+1}{l(i)+\tau} \frac{l(i)+1}{l+K} \ln \left[\frac{m_j(i)+1}{l(i)+\tau} \cdot \frac{l(i)+1}{\sum_{i=1}^{K} m_j(i)+1} \cdot \frac{l+\tau}{l+K} \right],$$

где $l(i)$ — число элементов i-го класса в обучающей выборке, $m_j(i)$ — число векторов i-го класса, у которых $x = x_j$, l — длина выборки.

Реализация сформулированного принципа состоит в таком подборе разбиения на градации интервала $c \leqslant x \leqslant C$, чтобы обеспечить минимум значения $H(\tau)$.

Алгоритм удобно реализовать в следующей форме: сначала разбить интервал на достаточно большое число градаций, а затем, «склеивая» соседние градации (и тем самым уменьшая число градаций τ), добиться минимизации $H(\tau)$ по τ. Для этого сначала склеивают такую пару соседних градаций, чтобы величина $H(\tau - 1)$ уменьшилась на наибольшую величину. Затем среди оставшихся $\tau - 1$ градации «склеивают» две соседние, чтобы минимизировать $H(\tau - 1)$ и т. д. Пусть минимум достигается при разбивке на τ^* градаций.

Можно оценить количество информации о принадлежности к классу $J(\tau^*)$, которые доставляют сведения о значениях параметра

$$J(\tau^*) = H_{\text{апр}} - H(\tau^*),$$

где

$$H_{\text{апр}} = -\sum_{i=1}^{K} \frac{l(i)+1}{l+K} \ln \frac{l(i)+1}{l+K}.$$

Часто разумно продолжать «склеивать» градации и после достижения минимума по τ функции $H(\tau^*)$, но лишь до тех пор, пока величина $J(\tau^*)$ не уменьшится в $(1-\delta)$ раз (δ — параметр алгоритма, обычно $\delta \approx 0{,}05$).

§ 6. Алгоритмы построения разделяющей гиперплоскости

В этом параграфе мы рассмотрим два алгоритма построения разделяющей гиперплоскости — алгоритм 11-1 и алгоритм 11-2.

Алгоритм 11-1 предназначен для построения гиперплоскости, разделяющей два конечных множества векторов, или выяснения того факта, что безошибочное линейное разделение векторов невозможно.

Этот алгоритм имеет две модификации. В одной из них он отыскивает обобщенный портрет при заданном параметре k, в другой — находит оптимальную разделяющую гиперплоскость. Алгоритм строит гиперплоскость, решая двойственную задачу — максимизируя квадратичную форму в положительном квадранте, как это было описано в §§ 3 и 4.

Модификация 1. При этой модификации алгоритм строит для заданного k обобщенный портрет.

Часто, однако, длина обучающей последовательности настолько велика, что обработка сразу всего материала обучения приводит к необходимости решать двойственную задачу слишком большой размерности.

Поэтому обработка обучающей последовательности ведется итеративно. Обучающая последовательность разбивается на m групп по p элементов в каждой группе (последняя группа может быть неполной).

Затем строится обобщенный портрет для векторов обучающей последовательности, попавших в первую группу. (Хорошо, когда в первую группу попадают векторы, принадлежащие как первому, так и второму классу).

Обобщенный портрет строится путем максимизации соответствующей квадратичной формы $W(\alpha, \beta)$ в положительном квадранте методом сопряженных градиентов (см. § 3).

В результате максимизации квадратичной формы либо будет найден обобщенный портрет, либо установлено, что разделение выделенной группы векторов невозможно $\left(W(\alpha, \beta) > \frac{(1-k)^2}{2\rho^2}\right)$.

Пусть по первой группе найден обобщенный портрет ψ_1. Формируется рабочая группа векторов, которая состоит из векторов, участвующих в разложении обобщенного портрета ψ_1 с ненулевым весом, и тех векторов первых двух групп, для которых выполняются неравенства

$$\begin{aligned} x_i^{\mathrm{T}}\psi_1 &> 1-\delta, \\ \bar{x}_j^{\mathrm{T}}\psi_1 &< k+\delta, \end{aligned} \tag{11.30}$$

где δ — параметр алгоритма $\left(0 < \delta < \frac{1-k}{2}\right)$.

Если в первых двух группах нет таких векторов, то рабочая группа векторов формируется с участием третьей группы. Если и в третьей группе нет векторов, удовлетворяющих неравенствам (11.30), то рассматривается четвертая группа, и т. д.

Если окажется, что во всех m группах нет векторов, удовлетворяющих (11.30), то ψ_1 есть обобщенный портрет всей обучающей последовательности.

По сформированной рабочей группе строится вторая итерация обобщенного портрета — ищется максимум соответствующей квадратичной формы или устанавливается, что значение максимума больше заданной величины, т. е. разделение векторов рабочей группы невозможно.

Пусть ψ_2 — обобщенный портрет, найденный во второй итерации. Формируется новая рабочая группа, состоящая из тех векторов, которые участвуют в разложении ψ_2 с ненулевым весом, и тех векторов первых трех (а возможно и большего числа) групп, для которых выполняются неравенства

$$\begin{aligned} x_i^{\mathrm{T}}\psi_2 &> 1-\delta, \\ \bar{x}_j^{\mathrm{T}}\psi_2 &< k+\delta. \end{aligned}$$

Строится новый обобщенный портрет ψ_3 — формируется новая рабочая группа и т. д. Так продолжается до тех пор, пока либо однажды окажется, что в формируемую группу не добавлен ни один вектор, а это и означает, что обобщенный портрет построен, либо не будет установлено, что безошибочное разделение векторов обучающей последовательности с помощью гиперплоскости невозможно.

Модификация 2. При этой модификации алгоритм строит оптимальную разделяющую гиперплоскость. Оптимальная разделяющая гиперплоскость также строится итерациями.

Для первой итерации формируется рабочая группа векторов, состоящая из l_1 векторов $x_1, \ldots, x_{l_1}$ обучающей последовательности, принадлежащих первому классу, и l_1 векторов $\bar{x}_1, \ldots, \bar{x}_{l_1}$ обучающей последовательности, принадлежащих второму классу. По этим векторам строится l векторов $y_i = x_i - \bar{x}_i$, для которых ищется обобщенный портрет (одного класса при пустом втором классе). Обобщенный портрет ищется путем минимизации квадратичной формы $W(\alpha)$ в положительном квадранте (см. § 4).

Пусть в результате первой итерации найден обобщенный портрет ψ. Для получения второй итерации образуем рабочую группу разностей Y_2. Для этого исключим из рабочей группы разностей Y_1 те пары, которые входили в разложение ψ_1 с нулевым весом, и найдем среди векторов обучающей последовательности такие векторы x_i и $\bar{x}_j$, для которых достигаются экстремальные значения

$$x_*^{\mathrm{T}}\psi_1 = \min_i x_i^{\mathrm{T}}\psi_1,$$

$$\bar{x}_*^{\mathrm{T}}\psi_1 = \max_j \bar{x}_j^{\mathrm{T}}\psi_1.$$

Если при этом окажется, что выполнены неравенства

$$\begin{aligned} x_*^{\mathrm{T}}\psi_1 &\geqslant \min_x x^{\mathrm{T}}\psi_1 - \delta_1, \\ \bar{x}_*^{\mathrm{T}}\psi_1 &\leqslant \max_x \bar{x}^{\mathrm{T}}\psi_1 + \delta_2, \end{aligned} \tag{11.31}$$

где в правых частях неравенства минимум и максимум вычисляются только по векторам обучающей последовательности, входящим в рабочую группу, δ_1, δ_2 параметры алгоритма $\left(\text{обычно } \delta_1 = 0{,}1\left(\min_i x_i^{\mathrm{T}}\psi\right),\ \delta_2 = 0{,}1\left(\max_j \bar{x}_j^{\mathrm{T}}\psi\right)\right)$,

то пара вектор ψ и число $\frac{\min\limits_i x_i^T\psi + \max\limits_j \bar{x}_j^T\psi}{2} = c$ задает оптимальную разделяющую гиперплоскость. Если хотя бы одно из двух неравенств (11.31) не выполнится, то пара x^*, $\bar{x}^*$ добавляется к рабочей группе векторов и строится новая итерация оптимальной разделяющей гиперплоскости.

Так продолжается до тех пор, пока либо однажды не выполнятся оба неравенства, либо окажется, что разделение невозможно ($W(\alpha) > 2/\rho^2$; ρ — заданная величина).

Таким образом, с помощью алгоритма 11-1 удается либо построить разделяющую гиперплоскость, либо установить, что безошибочное разделение векторов обучающей последовательности невозможно.

Согласно же оценке (11.2), если в пространстве размерности n удастся построить гиперплоскость, безошибочно делящую l векторов обучающей последовательности, то с вероятностью $1-\eta$ можно утверждать, что вероятность ошибочных классификаций с помощью построенной гиперплоскости будет меньше

$$P < \frac{n\left(\ln\frac{l}{n}+1\right)-\ln\eta}{l}.$$

Алгоритм 11-2 предназначен для построения гиперплоскости, разделяющей два множества векторов с минимальным числом ошибок.

Задача построения разделяющей гиперплоскости, минимизирующей число неправильно классифицируемых векторов, принципиально может быть решена, коль скоро решена задача построения разделяющей гиперплоскости, но точное ее решение требует большого перебора вариантов. Поэтому используем эвристический прием, позволяющий сократить перебор.

Алгоритм 11-2 использует следующий эвристический прием: из множества векторов обучающей последовательности исключается один элемент, «наиболее препятствующий разделению», затем, если разделение невозможно, из оставшегося множества исключается еще один элемент и т. д.

Специфика алгоритма заключается в определении элемента, «наиболее препятствующего разделению». В каче-

стве такого элемента при построении обобщенного портрета определяется вектор x_i (или $\bar{x}_j$), который в момент «останова» доставлял наибольший вклад в величину $W(\alpha, \beta)$

$$W(\alpha, \beta) = \sum_{i=1}^{a} \alpha_i \left(1 - \frac{1}{2} x_i^{\mathrm{T}} \psi\right) + \sum_{j=1}^{b} \beta_j \left(-k + \frac{1}{2} \bar{x}_j^{\mathrm{T}} \psi\right),$$

т. е. вектор $x_i(\bar{x}_j)$, для которого достигается максимум величины

$$\alpha_i \left(1 - \frac{1}{2} x_i^{\mathrm{T}} \psi\right), \quad (\beta_j (-k + x_j^{\mathrm{T}} \psi)).$$

Программа 11-2 исключает из обучающей последовательности вектор, «наиболее препятствующий разделению», делит оставшееся множество векторов и, если разделение все еще невозможно, исключает очередной вектор, делит оставшееся множество и т. д.

В конце концов, исключив m векторов, программа 11-2 разделит оставшееся множество векторов, построив разделяющую гиперплоскость $x^{\mathrm{T}}\psi = c$. Согласно же оценке (11.2) вероятность правильной классификации с помощью построенной гиперплоскости оценится величиной

$$P < \frac{n\left(\ln \frac{l}{n} + 1\right) - \ln \eta}{2l} \left(1 + \sqrt{1 + \frac{m}{n\left(\ln \frac{l}{n} + 1\right) - \ln \eta}}\right) + \frac{m}{l}.$$

§ 7. Построение разделяющей гиперплоскости в экстремальном пространстве признаков

Алгоритмы 11-1 и 11-2 реализуют идею минимизации эмпирического риска. С их помощью в классе линейных решающих правил отыскивается правило, минимизирующее число ошибок на обучающей последовательности. Начиная с алгоритма 11-3 мы будем строить алгоритмы обучения распознаванию образов, реализующие метод упорядоченной минимизации риска. В этом параграфе мы рассмотрим алгоритмы упорядоченной минимизации риска в классе линейных решающих правил.

Пусть структура класса линейных решающих правил задается по числу используемых при распознавании признаков (см. гл. VIII) исходного пространства призна-

ков E_n. В этом случае одновременно для всех решающих правил имеет место оценка

$$P(\alpha) < \nu(\alpha) + \frac{(n-r)\left(\ln\frac{l}{n-r}+1\right)+\ln C_n^r-\ln\eta}{2l}\times$$
$$\times\left(1+\sqrt{1+\frac{\nu(\alpha)\,l}{(n-r)\left(\ln\frac{l}{n-r}+1\right)+\ln C_n^r-\ln\eta}}\right), \quad (11.32)$$

где $(n-r)$ — размерность подпространства исходного пространства признаков, в котором строится линейное решающее правило. Таким образом, проблема заключается в том, чтобы, выбрав подходящее подпространство E_{n-r} и построив в нем линейное решающее правило, минимизирующее эмпирический риск, добиться минимальной величины оценки среднего риска.

В фиксированном пространстве E_{n-r} поиск решающего правила, минимизирующего эмпирический риск, осуществляется с помощью алгоритма 11-2.

Алгоритм 11-3 отыскивает в пространстве признаков такое подпространство, в котором гарантируется наилучшее решение задачи с помощью линейных решающих правил (иногда такое подпространство признаков называют информативным).

Принципиально отыскание информативного подпространства может быть осуществлено с помощью полного перебора по всем 2^n подпространствам исходного пространства признаков. Для каждого подпространства с помощью алгоритма 11-2 может быть найдено правило, минимизирующее эмпирический риск, а с помощью оценки (11.32) выбрано такое подпространство и такое решающее правило в нем, для которых гарантируется наименьшая оценка вероятности ошибочной классификации.

Однако осуществление полного перебора по всем подпространствам при вычислениях на ЦВМ требует чрезвычайно больших затрат машинного времени. Поэтому в алгоритме 11-3 используется стандартный эвристический прием последовательного улучшения оценки. Сначала отыскивается такой признак из n возможных признаков, исключение которого в наибольшей степени уменьшает оценку (11.32). Этот признак исключается. Затем из оставшихся $n-1$ признаков по тому же правилу снова исклю-

чается признак и т. д. Исключение признаков продолжается до тех пор, пока оценка вероятности ошибочной классификации не достигнет минимума.

Процедуру исключения признаков удобно проводить по следующей схеме.

1. Найдем решающее правило, минимизирующее эмпирический риск в исходном пространстве E_n, и вычислим для него оценку (11.32).

Пусть это правило построено с исключением из обучающей последовательности m векторов и соответствует решению двойственной задачи (см. §§ 3 и 4), при $\alpha = \alpha^*$, $\beta = \beta^*$.

2. Выберем в качестве рабочей группы те векторы обучающей последовательности $x(\bar{x})$, для которых $\alpha(\beta)$ отличны от нуля, а в качестве начальных значений при максимизации функции $W(\alpha, \beta)$ — соответствующие значения α^*, β^*, и будем последовательно искать в пространствах E_{n-1} решающие правила, делящие обучающую последовательность, состоящую из $l-m$ векторов (без исключенных в п. 1). Для каждого найденного правила вычисляется оценка. Выберем то решение, которое экстремизирует оценку.

3. Если найденная величина оценки меньше той, которая была получена для исходного пространства, то в качестве исходного рассматривается пространство E_{n-1} и производятся операции п.2. Если же экстремальная оценка больше той, которая была получена для E_{n-1}, то пространство E_n считается информативным, а найденное в нем линейное решающее правило, минимизирующее эмпирический риск, — наилучшим.

§ 8. Построение кусочно-линейной разделяющей поверхности

В этом разделе мы перенесем некоторые факты, справедливые для восстановления значений функции, на восстановление функции. Рассмотрим методы построения кусочно-линейных разделяющих поверхностей. При построении кусочно-линейной разделяющей поверхности пространство векторов X делится на p непересекающихся областей, в каждой из которых строится своя разделяющая гиперплоскость. Классификация векторов с помощью

построенной системы гиперплоскостей производится так: выясняется, какой области пространства принадлежит данный вектор, а затем он относится к тому или иному классу с помощью разделяющей гиперплоскости, построенной для классификации векторов данной области.

Таким образом, для заданной системы разделения пространства X на p областей алгоритм построения кусочно-линейной разделяющей поверхности состоит в том, чтобы в каждой области $X^{(i)}$ пространства X построить разделяющую гиперплоскость, минимизирующую число ошибок классификации на векторах обучающей последовательности, принадлежащих этой области. Построение гиперплоскости, минимизирующей число ошибок классификации, может быть осуществлено с помощью алгоритма 11-2. Таким образом, проблема состоит в том, чтобы задать разделения пространства X на p непересекающихся областей.

Согласно теории упорядоченной минимизации риска структура пространства должна быть задана априори, т. е. задано пространство X, разделение этого пространства на две области, разделение на три области и т. п.

При построении кусочно-линейного решающего правила необходимо определить такой элемент структуры, на котором достигается минимум оценки (11.2), где $h = np$, n — размерность пространства ($np < l$).

В этом параграфе при построении кусочно-линейных решающих правил мы отойдем от требования теории — априорного задания структуры. Будем строить структуру, используя векторы x обучающей последовательности

$$x_1, \ldots, x_l = X_l. \tag{11.33}$$

На множестве (11.33) мы определим таксонную структуру, т. е. различные способы разделения конечного множества (11.33) на подмножества, а для каждого разделения множества на подмножества

$$X_{l_1}, \ldots, X_{l_p};$$

$$\bigcup_{i=1}^{p} X_{l_i} = X_l, \quad X_{l_i} \bigcap_{i \neq j} X_{l_j} = \emptyset,$$

определим разбиение пространства X на p областей $\Gamma_1, \ldots, \Gamma_p$, по правилу:

вектор x относится к области Γ_i, если из p чисел

$$\rho(x, X_{l_1}), \ldots, \rho(x, X_{l_p})$$

число $\rho(x, X_{l_i})$ — наименьшее. Величина $\rho(x, X_{l_q})$ имеет смысл расстояния от точки x до множества X_{l_q}. Расстояние $\rho(x, X_{l_q})$ может пониматься в разных смыслах, например как расстояние от x до ближайшего элемента X_{l_q}.

Алгоритм 11-4 использует алгоритм построения таксонной структуры на конечном множестве векторов $x_1, \ldots, x_l$, описанный в приложении к главе X. Алгоритм строит последовательность векторов путем перенумерации исходной последовательности.

1. В качестве первого вектора x берется произвольный вектор исходного множества. Этот вектор из исходной выборки удаляется.

2. Пусть построена последовательность $X_t = x_1, \ldots, x_t$ и эти векторы удалены из исходной последовательности. В оставшемся множестве векторов находится вектор, ближайший к множеству X_t, т. е. такой, для которого величина

$$\rho = \rho(x^*, X_t) = \min_{x \in X_t} x^{\mathrm{T}} X_t$$

достигает минимума. Этот вектор удаляется из исходной последовательности и добавляется к строящейся.

3. Так продолжается до тех пор, пока вся последовательность (11.33) не будет переупорядочена.

4. Одновременно с построением новой последовательности определим последовательность чисел

$$\rho_t = \min_{x \in X_t} \rho(x^{\mathrm{T}}, X_{t-1}) \quad (t = 1, 2, \ldots, l).$$

Эти числа равны расстоянию между t-м членом x_t новой последовательности и первыми $t-1$ ее членами.

Примем следующее определение таксона, задавшись некоторым числом c. Две точки x_i и x_j относятся к одному таксону, если существует такая последовательность

$$x_i, \ldots, x_p, \ldots, x_s, \ldots, x_j,$$

состоящая из векторов исходной выборки, что расстояние между любыми двумя идущими подряд векторами этой последовательности меньше c. С помощью переупорядоченной выборки таксоны находятся так. Отыскиваются значения t, равные $t_1, \ldots, t_k$, для которых $\rho_t > c$. Точки последовательности

$$(x_1, \ldots, x_{t_1-1}), (x_{t_1}, \ldots, x_{t_2-1}), \ldots, (x_{t_k}, \ldots, x_{t_k+1}) \quad (11.34)$$

и образуют искомые таксоны.

Меняя константу c, получим требуемую таксонную структуру.

При построении кусочно-линейного решающего правила выбирается такой элемент таксонной структуры (11.34), на котором кусочно-линейное решающее правило $F(x, \alpha)$, минимизирующее эмпирический риск, доставит минимум функционалу

$$R(\alpha) = \nu(\alpha) + \frac{np\left(\ln\frac{l}{np}+1\right)-\ln\eta}{2l}\left(1+\sqrt{1+\frac{\nu(\alpha)\,l}{np\left(\ln\frac{l}{np}+1\right)-\ln\eta}}\right).$$

§ 9. Алгоритмы восстановления значений функции в классе линейных решающих правил

Рассмотрим алгоритмы восстановления значений функции в классе линейных решающих правил. Согласно постановке этой задачи наряду с обучающей последовательностью

$$x_1, \omega_1, \ldots, x_l, \omega_l \quad (11.35)$$

задается рабочая выборка

$$x_{l+1}, \ldots, x_{l+k}.$$

Требуется с помощью линейного решающего правила $F(x, \alpha)$ так индексировать точки рабочей выборки, чтобы минимизировать число ошибок классификации.

Рассмотрим сначала частную постановку. Будем считать, что индексация точек рабочей выборки должна быть проведена с помощью линейного решающего правила,

которое безошибочно делит векторы обучающей последовательности.

Решение этой частной задачи состоит в том, чтобы так индексировать точки рабочей выборки первым и вторым классом, чтобы расстояние между выпуклыми оболочками множества векторов обучающей и рабочей выборок первого класса и множества векторов обучающей и рабочей выборок второго класса было максимальным.

Принципиально эта задача решается так. Существует 2^k различных способов индексации рабочей выборки длины k. Для каждого способа индексации с помощью алгоритма 11-1 может быть определено расстояние между выпуклыми оболочками векторов различных классов. С помощью алгоритма 11-1 будет найдена оптимальная разделяющая гиперплоскость и с ее помощью вычислена величина ρ либо будет установлено, что разделяющей гиперплоскости не существует, т. е. $\rho < \rho_0$.

Перебором по всем способам индексаций может быть найдена такая индексация, при которой достигает максимума расстояние между выпуклыми оболочками векторов первого и второго класса.

Такая схема решения задачи опирается на полный перебор всех вариантов индексации векторов рабочей выборки. Для очень малых длин рабочей выборки ($2 \div 8$) этот путь допустим. Однако уже при объеме выборки $k = 10 \div 20$ перебор становится настолько большим, что реализация его оказывается невозможной. В этом случае приходится применять методы, направленные на сокращение перебора.

Выше при выборе информативного пространства признаков для сокращения перебора мы пользовались эвристическим методом последовательного улучшения оценки. Этот же принцип можно использовать и для отыскания наилучшего варианта индексации векторов рабочей выборки. Идею последовательного улучшения оценок при индексации векторов рабочей выборки реализует *алгоритм 11-5*.

Реализуется следующая последовательность действий:

1) По элементам обучающей последовательности строится оптимальная разделяющая гиперплоскость.

2) Точки рабочей выборки индексируются в соответствии с положением относительно этой гиперплоскости.

3) Строится оптимальная гиперплоскость, разделяющая всю выборку, в которой точки рабочей выборки взяты с индексацией, полученной в п. 2).

4) Поочередно меняется индексация точек рабочей выборки и выясняется, для какой из точек произойдет наибольшее увеличение ρ.

5) Если для найденной точки x_{l+l} такое увеличение положительно, то перейдем к новой индексации, в которой изменено значение ω_{l+l}, и снова будем действовать в соответствии с п. 4). Если же любое изменение индексации не приводит к увеличению ρ, то считается, что оптимальная индексация найдена. Более глубокий минимум может быть достигнут с помощью методов сокращения полного перебора (например, метода ветвей и границ [29, 97]).

При проведении индексации векторов рабочей выборки дальнейшее уменьшение числа ошибок классификации с помощью линейных решающих правил может быть достигнуто за счет селекции выборки и отыскания информативного пространства признаков.

Алгоритм 11-6 минимизирует в классе линейных решающих правил, безошибочно делящих обучающую выборку, функционал

$$R = \frac{\varkappa_*^2 (k - k_и)^2}{l + k - t} + k_и, \tag{11.36}$$

где $\varkappa_*$ — наименьшее решение неравенства [1])

$$d\left(\ln \frac{l + k - t}{d} + 1\right) + \ln \Gamma_{l - l_и,\ k - k_и}(\varkappa) + \ln C_{l+k}^{t} \leqslant \ln \eta, \tag{11.37}$$

t — общее число исключенных векторов, $l_и$ — число исключенных векторов обучающей выборки, $k_и$ — число исключенных векторов рабочей выборки,

$$d = \min\left(\left[\frac{D^2}{\rho^2}\right] + 1,\ n\right). \tag{11.38}$$

В (11.38) расстояние между выпуклыми оболочками ρ определяется для множества векторов, из которого исключено t элементов, за величину D принимается диаметр этого множества.

[1]) В неравенстве (11.37) $\ln C_{l+k}^{t}$ используется вместо слагаемого $\ln H_{l+k}^{t}$ (см. § 6 гл. X). При $t < \frac{l+k}{2}$, $\ln C_{l+k}^{t} \approx \ln H_{l+k}^{t}$.

Требуется исключить такое число векторов t и так индексировать оставшиеся векторы рабочей выборки, чтобы векторы первого и второго классов были разделимы гиперплоскостью и, кроме того, достигался минимум функционала (11.36).

Как обычно в подобных случаях, отыскание точного минимума функционала требует полного перебора по всем способам индексации и всем возможным способам исключения векторов. Поэтому при поиске минимума (11.36) используется метод последовательного уменьшения значения (11.36).

Сначала отыскивается наилучшее решение на всех элементах обучающей и рабочей выборки, а затем делается попытка максимально улучшить полученную оценку, исключив один элемент. Если такая попытка оказывается удачной, то предпринимается попытка уменьшить (11.36), исключив еще один элемент, и т. д.

Процесс исключения векторов продолжается до тех пор, пока в результате его уменьшается величина оценки (11.36).

§ 10. Алгоритмы восстановления значений функции в классе кусочно-линейных решающих правил

Алгоритм 11-7 определения значений функции с помощью кусочно-линейных решающих правил построен по схеме:

1. На полной выборке x определяется таксонная структура. Алгоритм построения таксонной структуры тот же, что и раньше.

2. Каждый элемент таксонной структуры S_r задает разбиение полного множества X на подмножества $X_1^r, \ldots, X_2^r$. Для каждого подмножества X_i^r решается задача классификации элементов рабочей выборки, принадлежащей X_i^r с помощью линейных решающих правил. Для этого используются алгоритмы предыдущего параграфа. Далее по формулам

$$R_i = \nu_i(\alpha_э) + \frac{k_i}{2(l_i+k_i)}\varkappa_i^2 + \varkappa_i\sqrt{\nu_i(\alpha_э) + \left(\frac{k_i\varkappa_i}{2(l_i+k_i)}\right)^2},$$

$$d_i\left(\ln\frac{l_i+k_i}{d_i}+1\right) + \ln\Gamma_{l_ik_i}(\varkappa) = \ln\eta,$$

где l_i — число элементов обучающей последовательности, принадлежащих X_i^r, k_i — число векторов рабочей выборки, принадлежащих X_i^r, определяется частота ошибок классификации рабочей выборки с помощью линейного решающего правила.

3. По формуле

$$R = \frac{1}{k} \sum_{i=1}^{r} R_i k_i \tag{11.39}$$

определяется частота ошибок классификации по всем элементам рабочей выборки.

4. Перебором по r находится такой элемент структуры, а вместе с ним и то решение, для которого оценка (11.39) минимальна.

Наконец, рассмотрим *алгоритм 11-8*, реализующий идеи локальных приближений. Пусть задана полная выборка

$$x_1, \ldots, x_{l+k}. \tag{11.40}$$

Для каждого вектора x, принадлежащего полной выборке (11.40), перенумеруем векторы последовательности (11.40) в соответствии с их расстоянием до этого вектора. Таким образом, задается $l+k$ систем окрестностей, своя для каждого вектора полной выборки (11.40)

$$1) \quad (x_1)_1 \subset (x_1, x_{i_1})_2 \subset \ldots \subset (x_1, \ldots, x_{l+k})_q;$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$(l+k) \quad (x_{l+k})_1 \subset (x_{l+k}, x_{i_{l+k}}, x_{j_{l+k}})_2 \subset \ldots \subset (x_1, \ldots, x_{l+k})_q.$$

Будем рассматривать систему окрестностей точки (x_i). Пусть X_i^r — некоторая окрестность, содержащая l_i точек обучающей выборки и k_i точек рабочей выборки.

Тогда частоту правильной классификации векторов рабочей выборки, принадлежащих этой окрестности, с помощью линейного решающего правила, минимизирующего эмпирический риск, в $X_r(x_p)$ оценим с помощью величины

$$R_i(r) = \nu(\alpha_9) + \frac{k_i \varkappa_i^2}{2(l_i+k_i)} + \varkappa_i \sqrt{\nu(\alpha_9) + \left[\frac{k_i \varkappa_i}{2(l_i+k_i)}\right]^2}, \tag{11.41}$$

где $\varkappa_i$ — наименьшее решение неравенства

$$d_i\left(\ln\frac{l_i+k_i}{d_i}+1\right)+\ln\Gamma_{l_i,\,k_i}(\varkappa)\leqslant\ln\eta.$$

Определим такую окрестность, а вместе с ней такую индексацию векторов рабочей выборки, для которой достигается минимум выражения (11.41). Пусть минимум равен R_i.

Таким образом, для каждого вектора x_i полной выборки (11.40) может быть указана классификация некоторых (попавших в окрестность) векторов рабочей выборки и получена оценка $m_i=R_ik_i$ числа ошибок классификации.

Рассмотрим табл. 1.

Таблица 1

Окрестности точек	Классификация точек					R_i
	x_{l+1}	. . .	x_{l+p}	. . .	x_{l+k}	
x_1	—	. . .	ω^1_{l+p}	. . .	ω^1_{l+k}	m_1
⋮	. . .	. . .	. . .	. . .	. . .	⋮
x_{l+1}	ω^{l+1}_{l+1}	. . .	—	. . .	—	m_{l+1}
⋮	. . .	. . .	. . .	. . .	. . .	⋮
x_{l+k}	—	. . .	ω^{l+k}_{l+p}	. . .	ω^{l+k}_{l+k}	m_{l+k}

Прочерк в столбце таблицы означает, что соответствующий вектор рабочей выборки не принадлежит окрестности, для которой проведена индексация.

Теория локальных алгоритмов предписывает индексацию с помощью минимаксного допустимого решения (см. § 10 гл. VIII). Отыскание такого минимаксного вектора индексаций есть задача целочисленного программирования, решение которой сопряжено с большим объемом вычислений.

Поэтому при реализации алгоритма 11-8 на ЦВМ в качестве окончательной индексации вектора x_{l+i} выбирается та, которая совпадает с большинством индексаций по столбцу таблицы.

ГЛАВА XII

АЛГОРИТМЫ ВОССТАНОВЛЕНИЯ НЕХАРАКТЕРИСТИЧЕСКИХ ФУНКЦИЙ

§ 1. Замечания об алгоритмах

В этой главе мы рассмотрим алгоритмы восстановления нехарактеристических функций.

Так же как и раньше, будем различать две задачи восстановления: восстановление функциональной зависимости и восстановление значений функции в заданных точках.

В основу рассматриваемых здесь алгоритмов положены две оценки, полученные в главах VIII и X. Оценка

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1 - 2\tau a(p) \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln \frac{\eta}{12}}{l}}} \right]_\infty , \qquad (12.1)$$

связывающая величину среднего риска $I(\alpha)$ с величиной эмпирического риска $I_э(\alpha)$, и оценка

$$I_\Sigma(\alpha) < \left[\frac{1 + \tau a(p) \frac{l}{l+k} \varkappa_*}{1 - \tau a(p) \frac{k}{l+k} \varkappa_*} \right]_\infty I_э(\alpha), \qquad (12.2)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$h\left(\ln \frac{l+k}{h} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leqslant \ln \frac{\eta}{1{,}5}, \qquad (12.3)$$

связывающая величину $I_\Sigma(\alpha)$ суммарного риска в точках рабочей выборки с величиной эмпирического риска $I_э(\alpha)$.

В отличие от аналогичных оценок, используемых при восстановлении характеристических зависимостей, оценки (12.1) и (12.2) содержат свободный параметр τ. Согласно теории этот параметр определяет статистическую особенность задачи (допустимую величину возможного выброса), и его значение должно быть известно заранее.

В этой главе мы используем оценки, предназначенные для «реальных» ситуаций, в которых зададим конкретные величины констант. Будем использовать оценку

$$I(\alpha) < \left[\frac{I_э(\alpha)}{1 - \sqrt{\frac{h\left(\ln \frac{l}{h} + 1\right) - \ln \eta}{l}}} \right]_\infty \tag{12.4}$$

для восстановления функциональной зависимости и оценку

$$I_\Sigma(\alpha) < \left[\frac{1 + \frac{l}{l+k} \varkappa_*}{1 - \frac{k}{l+k} \varkappa_*} \right]_\infty I_э(\alpha), \tag{12.5}$$

где $\varkappa_*$ — наименьшее решение неравенства

$$h\left(\ln \frac{l+k}{h} + 1\right) + \ln \Gamma_{l,k}(\varkappa) \leqslant \ln \eta \tag{12.6}$$

для восстановления значений функции в заданных точках.

§ 2. Алгоритм восстановления регрессии в классе полиномов

Рассмотрим алгоритмы восстановления одномерной функциональной зависимости по эмпирическим данным

$$x_1, y_1; \ldots; x_l, y_l$$

в классе линейных по параметрам функций

$$F(x, \alpha) = \sum_{i=1}^{n} \alpha_i \varphi_i(x).$$

Будем считать, что функции

$$\varphi_1(x), \ldots, \varphi_n(x) \tag{12.7}$$

априорно ранжированы, т. е. задана структура

$$S_1 \subset S_2 \subset \ldots \subset S_n, \tag{12.8}$$

элемент которой S_p есть множество функций

$$F(x, \alpha) = \sum_{i=1}^{p} \alpha_i \varphi_i(x).$$

В этом случае проблема сводится к отысканию такого элемента S_p структуры (12.8) и функции $F(x, \alpha_э)$, минимизирующей в S_p эмпирический риск, для которых достигается минимум функционала

$$R(p) = \left[\frac{I_э(\alpha_э)}{1 - \sqrt{\frac{p\left(\ln \frac{l}{p} + 1\right) - \ln \eta}{l}}} \right]_\infty \qquad (p < l).$$

Минимум эмпирического риска

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{p} \alpha_j \varphi_j(x_i) \right)^2 \qquad (12.9)$$

в S_p вычисляется с помощью стандартных приемов линейной алгебры: вектор параметров $\alpha_э = (\alpha_1^э, \ldots, \alpha_p^э)^T$ равен

$$\alpha_э = (\Phi_p^T \Phi_p)^{-1} \Phi_p^T Y, \qquad (12.10)$$

где Y — вектор значений $y_1, \ldots, y_l$; Φ_p — матрица,

$$\Phi_p = \begin{Vmatrix} \varphi_1(x_1) & \ldots & \varphi_p(x_1) \\ \ldots & \ldots & \ldots \\ \varphi_1(x_l) & \ldots & \varphi_p(x_l) \end{Vmatrix}. \qquad (12.11)$$

Проблема обращения матрицы типа $(\Phi^T\Phi)$ изучена достаточно подробно. (См., например, [63], [59].) В качестве алгоритма обращения матрицы может быть использован любой из рекомендованных там алгоритмов.

Таким образом, единственная проблема, которая возникает при реализации рассмотренной схемы, — определить, какую систему функций (12.7) использовать.

Восстанавливать функцию будем в классе полиномов, т. е. положим, что $\varphi_p(x)$ есть полином степени $p-1$:

$$\varphi_p(x) = \sum_{s=1}^{p} \beta_s x^{s-1}.$$

С принципиальной точки зрения безразлично, как заданы полиномы $\varphi_p(x)$ (лишь бы коэффициенты при старших степенях были отличны от нуля). Поэтому часто считают, что $\varphi_p(x) = x^{p-1}$. С вычислительной же точки зрения удобно выбирать систему ортонормальных на точках обучающей последовательности $x_1, \ldots, x_l$ функций

$\varphi_p(x)$, т. е. таких, для которых

$$\frac{1}{l}\sum_{i=1}^{l}\varphi_p(x_i)\,\varphi_q(x_i)=\begin{cases}1, & \text{если } p=q,\\ 0, & \text{если } p\neq q.\end{cases}$$

Для такой системы функций матрица $(\Phi^{\mathrm{T}}\Phi)$ единичная и вектор параметров $\alpha_{\text{э}}$ вычисляется без использования операции обращения матрицы

$$\alpha_{\text{э}}=\Phi^{\mathrm{T}}Y.$$

Итак, определен *алгоритм 12-1* восстановления одномерных функциональных зависимостей в классе полиномов.

В главе IX было установлено, что приближение к искомой функции в классе полиномов можно гарантировать лишь в интегральном смысле, в то время как в классе кусочно-полиномиальных зависимостей можно добиться не только интегрального приближения, но и равномерного приближения на всем отрезке определения функции. Оказывается (это будет показано в следующем параграфе), что в вычислительном отношении задача построения приближения функции в классе кусочно-полиномиальных зависимостей немногим сложнее приближения в классе полиномов.

§ 3. Фундаментальные сплайны

Пусть отрезок $[a, b]$, на котором ведется восстановление зависимости, разбит на $N+1$ частей

$$[a_0, a_1),\ [a_1, a_2),\ \ldots,\ [a_N, b].$$

Рассмотрим следующий класс функций: на каждом из $N+1$ подынтервалов функция совпадает с полиномом степени m (разными на разных подынтервалах) и непрерывна на всем интервале вместе со своими $m-1$ производными.

Такой класс функций назовем классом сплайнов степени m, сопряженных в N точках $a_1, \ldots, a_N$.

В дальнейшем будем считать, что $m=3$, а точки $a_1, \ldots, a_N$, задающие подынтервалы, получены в результате разбиения отрезка $[a, b]$ на $N+1$ равных частей. Обозначим класс таких сплайнов $V_N^3(x, \alpha)$.

Проблема состоит в том, чтобы найти функцию $N_N^3(x, \alpha_э)$ из $V_N^3(x, \alpha)$, минимизирующую эмпирический риск

$$I_э(\alpha) = \frac{1}{l} \sum_{i=1}^{l} (y_i - V_N^3(x_i, \alpha))^2. \qquad (12.12)$$

Строить сплайны удобно, введя в рассмотрение систему фундаментальных сплайнов. Для кубических сплайнов с N сопряжениями на сетке $(a, a_1, \ldots, a_N, a_{N+1} = b)$ вводятся $N+4$ фундаментальных сплайна

$$\mu_1(x), \ \mu_2(x), \ \mu_3(x), \ \ldots, \ \mu_{N+4}(x). \qquad (12.13)$$

Фундаментальные сплайны (12.13) однозначно определяются условиями

$$\mu_1(a_i) = 0, \quad \mu_1'(a_0) = 1, \quad \mu_1'(a_{N+1}) = 0 \quad (i = 1, 2, \ldots, N+1),$$
$$\mu_2(a_i) = 0, \quad \mu_2'(a_0) = 0, \quad \mu_2'(a_{N+1}) = 1 \quad (i = 1, 2, \ldots, N+1).$$
$$\mu_r(a_k) = \delta_{k, r-3}, \quad \mu_r'(a_0) = \mu_r'(a_{N+1}) = 0$$
$$(r = 3, \ldots, N+4; \ k = 0, 1, \ldots, N+1),$$
$$a_0 = a, \quad a_{N+1} = b.$$

В определении фундаментальных сплайнов δ_{ij} означает символ Кронекера

$$\delta_{ij} = \begin{cases} 1, & \text{если } i = j, \\ 0, & \text{если } i \neq j. \end{cases}$$

Поскольку любой сплайн $V_N^3(x, \alpha)$ полностью определяется $N+2$ значениями в узлах a_i $(i = 0, 1, \ldots, N+1)$ и значениями первой производной на концах отрезка, то имеет место равенство

$$V_N^3(x, \alpha) = \sum_{j=0}^{N+1} V_N^3(a_j, \alpha) \mu_{j+3}(x) +$$
$$+ [V_N^3(x, \alpha)]_a' \mu_1(x) + [V_N^3(x, \alpha)]_b' \mu_2(x).$$

Именно таким представлением мы будем пользоваться дальше при восстановлении регрессии в классе сплайнов $V_N^3(x, \alpha)$. Ниже мы найдем конкретные выражения для системы фундаментальных сплайнов $\mu_1(x), \mu_2(x), \mu_3(x), \ldots, \mu_{N+4}(x)$, и с помощью этой системы представим класс сплайнов степени 3 с N сопряжениями в параметрическом

виде

$$V_r^3(x,\ \alpha) = \sum_{i=1}^{N+4} \alpha_i \mu_i(x).$$

Тем самым мы сведем задачу отыскания сплайна, минимизирующего эмпирический риск (12.11), к определению параметров α, минимизирующих функционал

$$I_9(\alpha) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{N+4} \alpha_j \mu_j(x_i) \right)^2,$$

т. е. приведем решение задачи к тем же схемам линейной алгебры (12.10), к которым приводилась задача восстановления регрессии в классе полиномов.

Итак построим систему кубических фундаментальных сплайнов на равномерной сетке с шагом Δ: $a_{i+1} - a_i = \Delta$.

Пусть m_{i+1}, m_i — значения второй производной сплайна $V_N^3(x,\ \alpha)$ в узлах a_{i+1} и a_i. Так как вторая производная полинома третьей степени — линейная функция, то для $x \in [a_i,\ a_{i+1}]$ справедливо

$$[V_N^3(x,\ \alpha)]'' = m_{i+1} \frac{x - a_i}{\Delta} + m_i \frac{a_{i+1} - x}{\Delta},$$

где

$$m_{i+1} = [V_N^3(x,\ \alpha)]''_{a_{i+1}}, \quad m_i = [V_N^3(x,\ \alpha)]''_{a_i}.$$

Проинтегрировав дважды эту функцию с учетом условия непрерывности сплайна на концах отрезка $[a_i,\ a_{i+1}]$, получим, что кубический сплайн на отрезке $[a_i,\ a_{i+1}]$ описывается формулой

$$V_N^3(x,\ \alpha) = \frac{1}{6\Delta} [m_i (a_{i+1} - x)^3 + m_{i+1} (x - a_i)^3 + (6V_N^3(a_i,\ \alpha) - \Delta^2 m_i)(a_{i+1} - x) + (6V_N^3(a_{i+1},\ \alpha) - \Delta^2 m_{i+1})(x - a_i)].$$

На всем отрезке $[a,\ b]$ полученная функция непрерывна, но ее первая производная может претерпевать разрывы в узлах сопряжений.

Чтобы избежать этого, выберем величины m из условия непрерывности производной сплайна на всем отрезке $[a,\ b]$. Приравняв односторонние производные сплайна

в точках a_{i+1}, получим уравнения

$$[V_N^3(x,\alpha)]'_{a_{i+1}-0}=\frac{\Delta}{3}m_i+\frac{\Delta}{3}m_{i+1}+\frac{V_N^3(a_{i+1},\alpha)-V_N^3(a_i,\alpha)}{\Delta}=$$
$$=-\frac{\Delta}{3}m_{i+1}-\frac{\Delta}{3}m_{i+2}+\frac{V_N^3(a_{i+2},\alpha)-V_N^3(a_{i+1},\alpha)}{\Delta}=$$
$$=[V_N^3(x,\alpha)]'_{a_{i+1}+0}.$$

Таким образом, для определения $N+2$ значений m_i получено N линейных уравнений. Еще два уравнения дают краевые условия

$$[V_N^3(x,a)]'_a=V'_a,\quad [V_N^3(x,a)]'_b=V'_b,$$

откуда получаем

$$2m_1+m_2=\frac{6}{\Delta}\left(\frac{V_N^3(a_1,\alpha)-V_N^3(a_0,\alpha)}{\Delta}-V'_a\right),$$
$$m_{N+1}+2m_{N+2}=\frac{6}{\Delta}\left(V'_b-\frac{V_N^3(a_{N+1},\alpha)-V_N^3(a_N,\alpha)}{\Delta}\right).$$

В матричной записи система имеет вид

$$\mathscr{C}\mathscr{M}^*=\mathscr{D},$$

где

$$\mathscr{C}=\begin{Vmatrix}
2 & 1 & 0 & \dots & & & & & \\
\frac{1}{2} & 2 & \frac{1}{2} & 0 & \dots & & & & \\
0 & \frac{1}{2} & 2 & \frac{1}{2} & 0 & \dots & & & \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
 & & \dots & & 0 & \frac{1}{2} & 2 & 0 & \\
 & & \dots & & & 0 & \frac{1}{2} & 2 & \frac{1}{2} \\
 & & \dots & & & & 0 & 1 & 2
\end{Vmatrix},$$

$$\mathscr{M}^*=\begin{Vmatrix} m_1 \\ \vdots \\ m_{N+2} \end{Vmatrix},\qquad \mathscr{D}=\begin{Vmatrix} d_1 \\ \vdots \\ d_{N+2} \end{Vmatrix},$$

$$d_1=\frac{6}{\Delta}\left(\frac{V_N^3(a_1,\alpha)-V_N^3(a_0,\alpha)}{\Delta}-V'_a\right);$$

$$d_i=\frac{3}{\Delta^2}(V_N^3(a_{i+2},\alpha)+V_N^3(a_i,\alpha)-2V_N^3(a_{i+1},\alpha)),$$
$$i=2,\ \dots,\ N.$$

$$d_{N+2}=\frac{6}{\Delta}\left(V'_b-\frac{V_N^3(a_{N+1},\alpha)-V_N^3(a_N\alpha)}{\Delta}\right).$$

Для построения фундаментальных сплайнов $\mu_1(x)$, ..., $\mu_{N+4}(x)$ удобно представить вектор $\mathscr{D} = (d_1, \ldots, d_{N+2})^{\mathrm{T}}$ как произведение вектора определяющих значений $V_l = ([V_N^3(a_0, \alpha)]', V_N^3(a_0, \alpha), \ldots, V_N^3(a_{N+1}, \alpha), [V_N^3(a_{N+1}, \alpha)]')^{\mathrm{T}}$ на матрицу $\mathscr{B}$, которая имеет вид

$$\mathscr{B} = \begin{Vmatrix} -\frac{6}{\Delta} & -\frac{6}{\Delta^2} & \frac{6}{\Delta^2} & 0 & \ldots & \ldots & \ldots \\ 0 & \frac{3}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{3}{\Delta^2} & 0 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & \frac{3}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{3}{\Delta^2} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \frac{6}{\Delta^2} & -\frac{6}{\Delta^2} & \frac{6}{\Delta} \end{Vmatrix}.$$

Определяющими значениями фундаментальных сплайнов являются векторы V_l^*, у которых $N+3$ координаты равны нулю, а значение одной координаты — единице. Положение единицы в векторе определяется номером фундаментального сплайна. При надлежащем упорядочении фундаментальных сплайнов матрица определяющих значений оказывается единичной. Ниже указано такое упорядочение:

$$\begin{matrix} \mu_1 & \mu_3 & \ldots & \mu_{N+4} & \mu_2 \end{matrix}$$
$$\begin{Vmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{Vmatrix}.$$

Матрица $\mathscr{M}$ значений вторых производных $N+4$ фундаментальных сплайнов

$$\mathscr{M} = \begin{Vmatrix} m_{1,1} & \ldots & m_{1,N+4} \\ \ldots & \ldots & \ldots \\ m_{N+2,1} & \ldots & m_{N+2,N+4} \end{Vmatrix}$$

определяется как решение матричного уравнения

$$\mathscr{C}\mathscr{M} = \mathscr{B}.$$

Зная матрицу $\mathscr{M}$, легко вычислить значения фундаментальных сплайнов. Они вычисляются по формулам: для

$x \in [a_i,\ a_{i+1}]$

$$\mu_1(x) = m_{i1}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,1}\frac{(x-a_i)^3}{6\Delta} - \\ - m_{i+1,1}\frac{x-a_i}{6}\Delta - m_{i,1}\frac{a_{i+1}-x}{6}\Delta,$$

$$\mu_2(x) = m_{i,N+4}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,N+4}\frac{(x-a_i)^3}{6\Delta} - \\ - m_{i+1,N+4}\frac{x-a_i}{6}\Delta - m_{i,N+4}\frac{a_{i+1}-x}{6}\Delta,$$

$$\mu_j(x) = m_{i,j-2}\frac{(a_{i+1}-x)^3}{6\Delta} + m_{i+1,j-2}\frac{(x-a_i)^3}{6\Delta} + \\ + \left(\delta_{i,j-3} - m_{ij-2}\frac{\Delta^2}{6}\right)\frac{a_{i+1}-x}{\Delta} + \left(\delta_{i+1,j-3} - m_{i+1,j-2}\frac{\Delta^2}{6}\right)\frac{x-a_i}{\Delta}.$$

Матрицу $\mathcal{M}$ можно вычислить аналитически:

$$\mathcal{M} = \mathcal{C}^{-1}\mathcal{B}.$$

Для этого достаточно найти матрицу $\mathcal{C}^{-1}$.

Обозначим через D_n определитель порядка n:

$$D_n = \det\begin{Vmatrix} 2 & 1 & 0 & & & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & 0 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & 0 & \frac{1}{2} & 2 & \frac{1}{2} \\ & 0 & & 0 & \frac{1}{2} & 2 \end{Vmatrix}.$$

Раскладывая этот определитель по алгебраическим дополнениям элементов последнего столбца, получим рекуррентную формулу для вычисления определителя

$$D_n = 2D_{n-1} - \frac{1}{2}D_{n-2}, \quad D_1 = 2, \quad D_0 = 1.$$

Вычислим теперь элементы c_{ij}^{-1} матрицы $\mathcal{C}^{-1}$, используя алгебраические дополнения матрицы $\mathcal{C}$, выраженные

с помощью определителей D_n. Получаем:

$$\text{I.}\quad c_{ij}^{-1} = \frac{(-1)^{i+j}}{2^{j-i}} \frac{D_{i-1} D_{N+2-j}}{D_{N+2}} \qquad (1 < i \leqslant j \leqslant N+2).$$

$$\text{II.}\quad c_{ij}^{-1} = \frac{(-1)^{i+j}}{2^{i-j}} \frac{D_{j-1} D_{N+2-i}}{D_{N+2}} \qquad (1 \leqslant j \leqslant i < N+2).$$

$$\text{III.}\quad c_{1,j}^{-1} = \frac{(-1)^{1+j}}{2^{j-2}} \frac{D_{N+2-j}}{D_{N+2}} \qquad (1 < j \leqslant N+2). \tag{12.14}$$

$$\text{IV.}\quad c_{N+2,\,j}^{-1} = \frac{(-1)^{N+j}}{2^{N+1-j}} \frac{D_{j-1}}{D_{N+2}} \qquad (1 \leqslant j < N+2).$$

Схему применимости формул I — IV удобно представить графически (рис. 22).

Итак, для того чтобы найти систему кубических фундаментальных сплайнов с N сопряжениями, надо:

1) вычислить величины $\mathscr{D}_n$;

2) по формулам I — IV получить матрицу $\mathscr{C}^{-1}$ размерности $(N+2)\times(N+2)$;

3) вычислить матрицу $\mathscr{M}$ (размерности $(N+2)\times(N+4)$), умножив матрицу $\mathscr{C}^{-1}$ на $\mathscr{B}$ (матрица $\mathscr{B}$ имеет размерность $(N+2)\times$ $\times(N+4)$);

4) по формулам (12.14) получить фундаментальные сплайны.

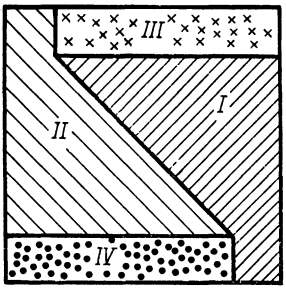

Рис. 22.

Для того чтобы сохранить единство обозначений, используем для системы фундаментальных сплайнов те же обозначения, что и для системы полиномов, т. е. будем считать, что

$$\mu_1(x) = \varphi_1(x), \ \ldots, \ \mu_{N+4}(x) = \varphi_{N+4}(x).$$

В этих обозначениях задача отыскания кубического сплайна, минимизирующего эмпирический риск (12.12), записывается в виде (12.9). А ее решение — определение век-

тора α коэффициентов разложения искомой функции по системе фундаментальных сплайнов, определяется формулой (12.10). Таким образом, после того как фундаментальная система сплайнов построена, вычисление сплайна, минимизирующего эмпирический риск, проводится точно по той же схеме, по которой определяются коэффициенты линейной (по параметрам α) регрессии.

§ 4. Алгоритмы восстановления функции в классе сплайнов

Рассмотрим теперь *алгоритм 12-2* упорядоченной минимизации риска в классе сплайнов. Для этого зададим следующую структуру на множестве кусочно-полиномиальных зависимостей. К классу S_1 отнесем константы, к S_2 — все полиномы степени единица, к S_3 — полиномы степени два, к четвертому классу S_4 отнесем полиномы степени три (назовем их кубическими сплайнами с нулем сопряжений).

Начиная с пятого класса, рассматриваются кусочно-полиномиальные функции. В пятый класс S_5 попадают сплайны с одним сопряжением, в S_6 — с двумя и т. д.

Емкость множества функций, образованного сплайнами с r сопряжениями, равна $h = r + 4$.

Таким образом, проблема состоит в том, чтобы выбрать элемент структуры S_{r+4}, для которого достигается минимум по α и r функционала

$$R(\alpha, r) = \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \sum_{p=1}^{r+4} \alpha_p \varphi_p(x_i) \right)^2}{1 - \sqrt{\frac{(r+4)\left(\ln \frac{l}{r+4} + 1\right) - \ln \eta}{l}}} \right]_{\infty}. \quad (12.15)$$

Особенность задачи восстановления регрессии в классе кусочно-полиномиальных зависимостей заключается в том, что каждый раз при переходе к классу сплайнов с большим числом сопряжений используется своя фундаментальная система (а не добавляется еще одна функция, как это было при восстановлении регрессии в классе полиномов).

Строго говоря, это приводит к тому, что элемент структуры S_{p+1} не содержит S_p. Однако последнее обстоятельство не является здесь принципиальным.

При восстановлении нехарактеристической функциональной зависимости целесообразно провести селекцию обучающей последовательности, т. е. исключить такое количество векторов $t=0, 1, 2, \ldots$, чтобы функционал

$$\hat{R}(\alpha, r)=\left[\frac{\frac{1}{l-t}\sum_{i}^{(t)}\left(y_i-\sum_{p=1}^{r+4}\alpha_p\varphi_p(x_i)\right)^2}{1-\sqrt{\frac{h\left(\ln\frac{l-t}{h}+1\right)-\ln\eta+\ln C_l^t}{l-t}}}\right]_{\infty} \tag{12.16}$$

достиг наиболее глубокого минимума. Функция, на которой (12.16) достигает минимума, принимается за решение задачи минимизации среднего риска $\left(\sum_{i}^{(t)}$ означает, что суммируется лишь $l-t$ членов$\right)$.

Отыскание точного минимума функционала (12.16) требует большого перебора вариантов. Поэтому рационально здесь применить метод последовательного уменьшения функционала. Сначала найти вектор, исключение которого из обучающей последовательности минимизирует функционал для $t=1$. Если эта величина окажется меньше минимальной величины функционала (12.16) при $t=0$ (для всей обучающей выборки), то соответствующий вектор исключается и делается попытка аналогично исключить еще один вектор, т. е. минимизировать (12.16) при $t=2$.

Если же никакое исключение одного вектора не приводит к уменьшению функционала, то исключение векторов прекращается

§ 5. Алгоритмы решения некорректных задач интерпретации измерений

В этом параграфе мы рассмотрим алгоритмы решения некорректных задач интерпретации результатов косвенных экспериментов для случая, когда операторное уравнение

$$Af(t)=F(x) \tag{12.17}$$

есть интегральное уравнение Фредгольма I рода

$$\int_a^b K(t,\ x) f(t)\, dt = F(x). \qquad (12.18)$$

Пусть заданы измерения функции $F(x)$ в l точках x_i:

$$x_1,\ y_1;\ \ldots;\ x_l,\ y_l.$$

Согласно теории, решением уравнения (12.18) является функция $f(t)$, доставляющая минимум функционалу

$$I = \int \left(y - \int_a^b K(t,\ x) f(t)\, dt \right)^2 P(y \mid x)\, dy\, dx. \qquad (12.19)$$

Будем минимизировать средний риск (12.19) методом упорядоченной минимизации в классе кубических сплайнов.

Для этого найдем функцию $V_r^3(t, \alpha_э)$, минимизирующую функционал в классе кубических сплайнов с r сопряжениями

$$R(\alpha,\ r) = \left[\frac{\frac{1}{l} \sum_{i=1}^{l} \left(y_i - \int_a^b K(t,\ x_i)\, V_r^3(t,\ \alpha)\, dt \right)^2}{1 - \sqrt{\frac{(r+4)\left(\ln \frac{l}{r+4} + 1\right) - \ln \eta}{l}}} \right]_{\infty}. \qquad (12.20)$$

Построим *алгоритм 12-3* минимизации функционала (12.20) как некоторую модификацию алгоритмов 12-2. Для этого введем обозначения

$$\int_a^b K(t,\ x_i)\, \mu_p(t)\, dt = \varphi_p(x). \qquad (12.21)$$

Так как, согласно § 3,

$$V_r^3(t,\ \alpha) = \sum_{p=1}^{r+4} \alpha_p \mu_p(t),$$

то в обозначениях (12.21) минимизация функционала (12.20) сводится к минимизации функционала (12.15). Минимум по α и r функционала может быть найден по схеме алгоритма 12-2. Пусть минимум достигается при r^*, $\alpha_э$. Тогда решением интегрального уравнения

объявляется функция

$$f(t) = \sum_{p=1}^{r^*+4} \alpha_p^3 \mu_p(t).$$

При интерпретации результатов косвенных экспериментов целесообразно проводить селекцию измерений. Селекция измерений также проводится по схеме алгоритма 12-2, т. е. сводится к минимизации функционала (12.16) и выбору в качестве решения прообраза функции, доставляющей минимум этому функционалу.

§ 6. Алгоритмы восстановления многомерной регрессии в классе линейных функций

Рассмотрим *алгоритм 12-4* восстановления многомерной линейной регрессии.

Пусть требуется в классе функций

$$F(x, \beta) = \sum_{i=1}^{n} \beta_i \varphi_i(x) = \beta^{\mathrm{T}} \varphi(x) \qquad (\varphi(x) = (\varphi_1(x), \ldots, \varphi_n(x))^{\mathrm{T}}), \quad (12.22)$$

восстановить регрессию. Пусть

$$\zeta_1, \ldots, \zeta_n \qquad (12.23)$$

— система собственных векторов матрицы $(\Phi^{\mathrm{T}}\Phi)$ (матрицы (12.11)), ранжированная в порядке убывания собственных чисел. Представим (12.22) в виде

$$F(x, \alpha) = \sum_{p=1}^{n} \alpha_p \zeta_p^{\mathrm{T}} \varphi(x) = \sum_{p=1}^{n} \alpha_p \chi_p(x), \qquad (12.24)$$

где

$$\chi_p(x) = \zeta_p^{\mathrm{T}} \varphi(x).$$

Зададим на классе функций $F(x, \alpha)$ структуру

$$S_1 \subset \ldots \subset S_n, \qquad (12.25)$$

где S_p содержит лишь те функции, которые разложимы по первым p членам ряда [1]). Тогда наилучшим элементом

[1]) Заметим, что такое задание структуры является априорным лишь в постановке восстановления значений функции, где матрица Φ образована по всем $l+k$ векторам полной выборки. Тем не менее мы используем здесь структуру (12.25).

структуры будет тот, на котором достигается минимум функционала

$$R(\alpha,\ p)=\left[\frac{\frac{1}{l}\sum_{i=1}^{l}\left(y_i-\sum_{s=1}^{p}\alpha_s\chi_s(x_i)\right)^2}{1-\sqrt{\frac{p\left(\ln\frac{l}{p}+1\right)-\ln\eta}{l}}}\right]_{\infty}.$$

Реализация этого алгоритма та же, что и 12-2.

При построении линейной регрессии часто оказывается целесообразным проведение селекции обучающей последовательности. Необходимо исключить такие t элементов ($t=0,\ 1,\ 2,\ \ldots,\ s$), чтобы минимизировать по α, p функционал

$$R(\alpha,\ p)=\left[\frac{I_{\text{э}}^{t}(\alpha)}{1-\sqrt{\frac{p\left(\ln\frac{l-t}{p}+1\right)+\ln C_l^t-\ln\eta}{l-t}}}\right]_{\infty}, \tag{12.26}$$

где $I_{\text{э}}^{t}(\alpha)$ — функционал эмпирического риска, построенный на обучающей выборке, из которой исключены соответствующие элементы.

Минимизацию функционала (12.26), так же как и раньше в аналогичных случаях, следует проводить с помощью эвристической процедуры последовательной минимизации. Перебором следует найти такую обучающую последовательность, состоящую из $l-1$ элементов, для которой оценка (12.26) (при $t=1$) минимальна. Если полученная оценка для $t=1$ меньше, чем при $t=0$, то соответствующий вектор удаляется из обучающей последовательности и делается попытка удалить еще один вектор и т. д. Удаляется такое множество векторов, при котором оценка (12.26) достигает минимума.

Выше при построении линейной регрессии структура (12.25) была задана в соответствии с порядком следования членов ряда (12.23). Однако часто порядок (12.23) определяется не в соответствии с величинами собственных чисел, а устанавливается в ходе построения регрессии. Рассмотрим такой пошаговый алгоритм построения регрессии. Сначала выбирается один фактор (функция $\chi_1(x)$),

с помощью которого достигается наилучшее приближение к эмпирическим данным. Для определения такого фактора n раз (n—число факторов) решается задача построения регрессии по одному фактору и выбирается тот, для которого величина эмпирического риска минимальна.

Этот фактор фиксируется, а затем перебором по оставшимся $n-1$ факторам подбирается второй фактор такой, чтобы построенная на этих двух факторах линейная функция доставляла наименьшую величину эмпирическому риску. Найденный второй фактор фиксируется и подбирается третий и т. д.

При такой схеме упорядочения по факторам в качестве окончательного решения выбирается функция, доставляющая минимум по α и p функционалу

$$R(\alpha,\ p)=\left[\frac{I_{\text{э}}(\alpha)}{1-\sqrt{\dfrac{p\left(\ln\dfrac{l}{p}+1\right)+\ln C_n^p-\ln\eta}{l}}}\right]_{\infty}.$$

§ 7. Алгоритмы восстановления значений произвольной функции в классе линейных по параметрам функций

Алгоритмы 12-5 и 12-6 восстановления по выборке

$$x_1,\ y_1;\ \ldots;\ x_l,\ y_l$$

значений функции в заданных точках

$$x_{l+1},\ \ldots,\ x_{l+k}$$

в классе линейных функций основаны на двух различных идеях задания структуры.

В алгоритме 12-5 на множестве $F(x,\ \alpha)$ задается та же самая структура, что и в алгоритме 12-4:

$$S_1 \subset S_r \subset \ldots \subset S_n,$$

элемент S_p содержит функции, разложимые по p первым членам ряда (12.23). Однако, в отличие от алгоритма 12-4, здесь матрица Φ составлена по всем $l+k$ элементам обучающей и рабочей выборок. (Поэтому задание структуры здесь является априорным.)

Проблема состоит в том, чтобы выбрать элемент структуры S_p, а в нем функцию $F(x, \alpha)$, минимизирующие оценку

$$R(\alpha, p) = \left[\frac{1+\frac{l}{l+k}\varkappa_*}{1-\frac{k}{l+k}\varkappa_*}\right]_\infty I_э(\alpha),$$

где $\varkappa_*$ — наименьшее решение неравенства

$$p\left(\ln\frac{l+k}{p}+1\right)+\ln\Gamma_{l,k}(\varkappa)\leqslant\ln\eta.$$

Реализация алгоритма 12-5 проводится по той же схеме, которая была рассмотрена в предыдущем параграфе.

Здесь целесообразно провести селекцию полной выборки с тем, чтобы добиться меньшей оценки суммарного риска. Селекция также проводится методом последовательного уменьшения оценки.

Пусть исключено t точек выборки ($l_и$ — из обучающей и $k_и$ — из рабочей, $l_и + k_и = t$). Тогда:

1) Строится упорядоченная система собственных векторов матрицы $\Phi^T\Phi$ (элементы матрицы $\Phi^T\Phi$ построены по выборке, из которой исключено t элементов). По построенной упорядоченной системе функций задается структура на классе линейных функций.

2) По этой структуре отыскивается такой элемент S_p, а в нем такая функция, что достигается минимум выражения

$$R(\alpha, p) = \left[\frac{1+\varkappa_*\frac{l-l_и}{l+k-t}}{1-\varkappa_*\frac{k-k_и}{l+k-t}}\right]_\infty I_э(\alpha), \qquad (12.27)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$p\left(\ln\frac{l+k-t}{p}+1\right)+\ln\Gamma_{l-l_и,\,k-k_и}(\varkappa)+\ln C_{l+k}^{t}\leqslant\ln\eta.$$

3) Перебором по числу исключенных точек отыскивается решение, которое минимизирует функционал (12.27).

Алгоритм 12-6 основан на идее упорядочения классов эквивалентности линейных функций. Согласно теории все множество линейных функций делится на конечное число классов эквивалентности, одинаково упорядочивающих

полную выборку $x_1, \ldots, x_{l+k}$ (§ 9 гл. X). Каждый класс эквивалентности характеризуется числом $[\rho^2/D^2]$. Зададим на множестве линейных функций структуру

$$S_1 \subset \ldots \subset S_n,$$

где к элементу S_p относятся все те классы эквивалентности, для которых

$$\left[\frac{\rho^2}{D^2}\right] \geqslant \frac{1}{p-1}, \quad p < n,$$

$$\left[\frac{\rho^2}{D^2}\right] \geqslant 0, \qquad p = n, \quad p \geqslant 2.$$

Для множества S_p этой структуры емкость равна p.

Таким образом, возникает необходимость найти такой элемент S_p структуры и такую в нем функцию, для которых оценка суммарного риска (12.27) минимальна. Решение этой задачи в полном объеме — проблема чрезвычайно трудная. Поэтому здесь мы рассмотрим лишь частное решение.

Пусть

$$y = L^{\text{э}}(x) = \alpha^{\text{т}} x + \alpha_0 \tag{12.28}$$

— линейная функция, минимизирующая эмпирический риск на элементах обучающей выборки.

Предположим, что известна функция

$$y = L^*(x) = \beta^{\text{т}} x + \beta_0, \tag{12.29}$$

принадлежащая элементу S_p с минимальным номером p, для которой выполняется условие $\mathscr{A}$: если для элементов обучающей последовательности x_i, y_i и x_j, y_j окажется $y_i > y_j$, то выполняется неравенство $L^*(x_i) \geqslant L^*(x_j)$. (Ниже будет приведен алгоритм построения такой функции.)

Функция $L^*(x)$ минимизирует первый сомножитель оценки (12.27) при выполнении условия $\mathscr{A}$, а функция $L_{\text{э}}(x)$ — второй сомножитель.

Рассмотрим теперь параметрическое (по b, c_0 и γ) семейство линейных функций

$$y = b\left(\gamma \frac{\alpha}{\|\alpha\|} + (1-\gamma)\frac{\beta}{\|\beta\|}\right)^{\text{т}} x + c_0 = c^{\text{т}} x + c_0; \tag{12.30}$$

при $\gamma = 1$, $b = 1$, $c_0 = \alpha_0$ функция (12.30) совпадает с (12.28), а при $\gamma = 0$, $b = 1$, $c_0 = \beta_0$ функция (12.30) совпадает с (12.29).

Найдем среди семейства (12.30) функцию (значение параметров b, γ и c_0), которая доставляет минимум оценке (12.27). Поиск такой тройки может быть проведен по следующему алгоритму:

для всякого фиксированного γ значение параметров c_0 и b определяется из условия минимизации по c_0, b эмпирического риска.

Для того чтобы оценить качество полученного решения, необходимо определить величину $d = [D^2/\rho^2] + 1$.

Перенумеруем векторы x полной выборки в порядке возрастания величин

$$\left(\gamma \frac{\alpha}{\|\alpha\|} + (1-\gamma)\frac{\beta}{\|\beta\|}\right)^{\mathrm{T}} x_l.$$

Значение

$$\rho = \min_i \sup_{\|\varphi\|=1} \varphi^{\mathrm{T}} \{x_{i+1} - x_i\} \tag{12.31}$$

вычислим с помощью алгоритма 11-2 (модификация 2).

Вектор φ_0, определяющий (12.31), выражается через минимальный по модулю вектор ψ_0, удовлетворяющий неравенству

$$\psi^{\mathrm{T}}(x_{i+1} - x_i) \geqslant 1, \quad i = 1, 2, \ldots, l + k - 1.$$

А именно $\varphi_0 = \psi_0 / \|\psi_0\|$.

В качестве диаметра минимальной сферы, содержащей полную выборку, примем диаметр выборки D и найдем величину d. По найденным величинам d и $I_э$ вычислим оценку (12.27).

Перебором отыскивается такое γ, для которого оценка (12.27) окажется минимальной. С помощью найденной функции вычисляются значения y в точках рабочей выборки.

Итак, для реализации алгоритма 12-6 осталось определить способ построения линейной функции, принадлежащей элементу структуры с наименьшим номером и удовлетворяющей условию $\mathscr{A}$. Построение такой функции будем проводить, строя оптимальную разделяющую гиперплоскость (см. § 6 гл. XI).

Пусть на полной выборке, состоящей из l элементов обучающей и k элементов рабочей выборки, задан фиксированный порядок следования

$$x_{i_1}, \ldots, x_{i_l}, \quad x_{i_{l+1}}, \ldots, x_{i_{l+k}}.$$

Построим минимальный по модулю вектор ψ, для которого выполнится неравенство

$$\psi^{\mathrm{T}}(x_{i_{r+1}} - x_{i_r}) \geqslant 1. \qquad (12.32)$$

Для построения вектора ψ используем алгоритм 11-1 (модификация 2). С помощью этого алгоритма удается построить вектор ψ, который при условии выполнения (12.32) минимизирует функционал $\psi^{\mathrm{T}}\psi$, т. е. определяет направление, которое максимизирует ρ.

Нам осталось определить такой порядок следования векторов, при котором для элементов обучающей выборки будет выполнено условие: x_i предшествует x_j, если $y_i > y_j$, и при этом модуль вектора ψ_0, удовлетворяющего условию (12.32), достигает минимума. Вектор ψ_0 и определяет $L^*(x)$ ($\beta = \psi_0$).

Точное решение этой задачи может быть получено полным перебором по всем порядкам полной выборки, у которых подвыборка элементов обучающей последовательности упорядочена в соответствии с уменьшением значения y.

При реализации же алгоритма мы используем эвристический метод последовательной минимизации.

1. Сначала упорядочим элементы обучающей последовательности в соответствии с величинами y и найдем минимальный по модулю вектор ψ, удовлетворяющий неравенствам (12.32).

2. Затем найдем такой элемент x^* рабочей выборки и так его расположим в ряду

$$x_1, \ldots, x_q, x^*, x_{q+1}, \ldots, x_l, \qquad (12.33)$$

чтобы минимизировать модуль вектора ψ при условии выполнения неравенств

$$\psi^{\mathrm{T}}(x_{l+1} - x_l) \geqslant 1,$$
$$\psi^{\mathrm{T}}(x^* - x_{q+l}) \geqslant 1,$$
$$\psi^{\mathrm{T}}(x_q - x^*) \geqslant 1.$$

Перенумеруем последовательность (12.33) от 1 до $l+1$.

3. Найдем еще один элемент рабочей выборки, который при соответствующем расположении в ряду приводит к минимизации нормы вектора ψ, для которого выполнены условия (12.32) и т. д.

Таким образом, будет найден такой порядок расположения векторов x и такой вектор ψ, при которых будут выполнены условия (12.32), а модуль ψ достигнет малой величины.

Итак, последовательность действий алгоритма 12-6 следующая:

1. Отыскивается функция

$$L^*(x) = \beta^T x + \beta_0.$$

2. Отыскивается функция

$$y = \alpha^T x + \alpha_0,$$

минимизирующая эмпирический риск.

3. Определяется трехпараметрическое семейство функций

$$y = b\left(\frac{\alpha}{\|\alpha\|}\gamma + \frac{\beta}{\|\beta\|}(1-\gamma)\right)^T x + c_0, \qquad (12.34)$$

в котором находится оптимальное решение по следующему правилу: для каждого фиксированного γ определяются величины $d = [D^2/\rho^2] + 1$ и I_3^*.

С помощью величин d и I_3^* определяется оценка суммарного риска (12.27). Перебором по γ находится такая функция, для которой оценка риска минимальна. С помощью найденной функции определяются величины y для точек рабочей выборки.

Алгоритм может быть усилен за счет селекции выборки.

§ 8. Алгоритмы восстановления регрессии в классе кусочно-линейных функций

Алгоритмы восстановления кусочно-линейных функций, рассматриваемые в этом параграфе, построены по той же схеме, по которой реализованы алгоритмы построения кусочно-линейных решающих правил (см. § 8 гл. XI).

Определяется таксонная структура обучающей последовательности X_l, элемент которой $X_{l_1}, \ldots, X_{l_k}$ задает разделение обучающей выборки на k подмножеств. В соответствии с полученным разбиением выборки на таксоны исходное пространство X делится на k подпространств $\Gamma_1, \ldots, \Gamma_k$ по правилу: вектор x относится к области Γ_p,

если из k чисел

$$\rho(x, X_{l_1}), \ldots, \rho(x, X_{l_k})$$

число $\rho(x, X_{l_p})$ наименьшее. Величина $\rho(x, X_{l_i})$ есть расстояние от точки x до множества X_{l_i}.

Для каждой области строится своя линейная функция, минимизирующая эмпирический риск. Таким образом, в соответствии p-му элементу таксонной структуры ставится кусочно-линейная функция.

С помощью *алгоритма 12-7* определяется такая кусочно-линейная функция (т. е. такой элемент таксонной структуры S_p множества X_l обучающей последовательности), для которой достигает минимума функционал

$$R(\alpha, p) = \left[\frac{I_э(\alpha)}{1 - \sqrt{\frac{np\left(\ln\frac{l}{np} + 1\right) - \ln\eta}{l}}}\right]_\infty. \qquad (12.35)$$

Реализуется алгоритм 12-7 по схеме:

1. С помощью алгоритма таксономии, описанного в § 8 гл. XI, задается таксонная структура множества X обучающей последовательности.

2. Перебором по элементам таксонной структуры строятся кусочно-линейные функции, состоящие из одного, двух, трех и т. д. кусков плоскостей.

3. Выбирается такая кусочно-линейная функция, для которой оценка (12.35) минимальная.

Алгоритм 12-7 может быть усилен за счет селекции векторов обучающей последовательности. В этом случае минимизируется функционал

$$R(\alpha, p) = \left[\frac{I^l_э(\alpha)}{1 - \sqrt{\frac{np\left(\ln\frac{l-t}{np} + 1\right) + \ln C^t_l - \ln\eta}{l-t}}}\right]_\infty.$$

§ 9. Алгоритмы восстановления значений произвольной функции в классе кусочно-линейных функций

Идея построения кусочно-линейных функций путем использования таксонной структуры множества X обучающей последовательности являются эвристической. Такая

идея может быть проведена строго лишь для задачи восстановления значений функции в заданных точках. Реализует ее *алгоритм 12-8.*

На множестве x, состоящем из l элементов обучающей последовательности и из k элементов рабочей выборки, задается таксонная структура. Элемент таксонной структуры S_p определяет разбиение множества векторов X_{l+k}, состоящего из векторов обучающей и рабочей выборок на p подмножеств $X_1, \ldots, X_p$.

Для каждого из подмножеств X_r может быть поставлена задача восстановления значений функции в заданных точках в классе линейных функций. Обозначим через l_r число векторов обучающей последовательности, принадлежащих X_r, а через k_r — число элементов рабочей выборки, принадлежащих X_r.

Тогда, используя алгоритмы § 7, для каждого множества X_r в классе линейных функций восстановим значения функции в k_r точках рабочей выборки (в тех точках, которые принадлежат X_r) по l_r эмпирическим данным (также из X_r).

При этом воспользуемся следующей оценкой величины суммарного риска:

$$I_\Sigma(\alpha_э) < \left[\frac{1+\frac{l_r}{l_r+k_r}\varkappa_*}{1-\frac{k_r}{l_r+k_r}\varkappa_*}\right]_\infty I_э(\alpha_э) = \mathscr{L}_r,$$

где $\varkappa_*$ — наименьшее решение неравенства

$$d_r\left(\ln\frac{l_r+k_r}{d_r}+1\right)+\ln\Gamma_{l_r,\,k_r}(\varkappa)\leqslant\ln\eta,$$

$F(x,\ \alpha_э)$ — функция, минимизирующая на обучающей последовательности из X_r величину эмпирического риска.

Величина суммарного риска по всем подмножествам равна

$$I_\Sigma = \frac{1}{k}\sum_{r=1}^{p}\mathscr{L}_r k_r.$$

Минимизацией по элементам S_p таксонной структуры найдем наилучшее решение.

Наконец, рассмотрим локальный *алгоритм 12-9* восстановления значений функции в заданных точках. Этот

алгоритм с точностью до реализации метода минимизации суммарного риска в классе линейных функций повторяет локальный алгоритм распознавания образов 11-8.

1. Так же как и в алгоритме 11-8, для каждого элемента полной выборки

$$x_1, \ldots, x_{l+k}$$

рассматривается система окрестностей (окрестность X_i^t вектора x_i состоит из тех векторов, для которых $\rho(x_i x) \leqslant \leqslant c_t$).

Иначе говоря, рассматривается $l+k$ систем вложенных множеств

$$1) \quad (x_1)_1 \subset (x_1, x_{i_1})_2 \subset \ldots \subset (x_1, \ldots, x_{l+k})\, q,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$(l+k) \quad (x_{l+k})_1 \subset (x_{l+k}, x_j, x_s)_2 \subset \ldots \subset (x_1, \ldots, x_{l+k})\, q.$$

2. Пусть исследуются окрестности точки x_i, и пусть X_i^t — t-я окрестность точки x_i.

Тогда оценка суммарного риска векторов рабочей выборки, принадлежащих этой окрестности, полученных с помощью линейной функции, минимизирующей эмпирический риск на векторах из X_i^t, оценивается величиной

$$R_\Sigma = \left[\frac{1 + \dfrac{l_t}{l_t + k_t}\varkappa_*}{1 - \dfrac{k_t}{l_t + k_t}\varkappa_*} \right]_\infty I_э(\alpha_э), \qquad (12.36)$$

где $\varkappa_*$ — наименьшее решение неравенства

$$d\left(\ln \frac{l_t + k_t}{d} + 1\right) + \ln \Gamma_{l_t,\, k_t}(\varkappa) \leqslant \ln \eta.$$

В (12.36) l_t — число элементов обучающей последовательности, k_t — число элементов рабочей выборки с векторами x из X_i^t.

Определим такую окрестность точки x_i, а вместе с ней и такие значения векторов рабочей выборки, для которых достигается минимум выражения (12.36).

3. Таким образом, для каждого вектора x_i, задающего систему окрестностей, могут быть указаны значения y векторов рабочей выборки, принадлежащих экстремальной окрестности, и получена оценка величины суммарного риска классификации.

Составим табл. 1.

Таблица 1

Окрестности точек	Значения точек					Оценка величины суммарного риска
	x_{l+1},	. . .	x_{l+j}	. . .	x_{l+k}	
x_1	—	. . .	y_{l+j}^1	. . .	—	$m_1 = R_\Sigma k_1$
⋮	. . .	. . .	. . .	. . .	. . .	. . .
x_{l+1}	y_{l+1}^{l+1}	. . .	—	. . .	. . .	$m_{l+1} = R_\Sigma k_{l+1}$
⋮	. . .	. . .	. . .	. . .	. . .	. . .
x_{l+k}	—	. . .	y_{l+j}^{l+k}	. . .	y_{l+k}^{l+k}	$m_{l+k} = R_\Sigma k_{l+k}$

Прочерк в таблице означает, что соответствующий вектор рабочей выборки не принадлежит экстремальной окрестности. x_i; k_i — число элементов рабочей выборки, попавших в экстремальную окрестность точки x_i.

4. Обозначим через $y_{l+1}^*, \ldots, y_{l+k}^*$ искомые значения функции в точках $x_{l+1}, \ldots, x_{l+k}$. Поставим в соответствие каждой строке таблицы неравенство и рассмотрим систему

$$\begin{aligned} \sum_i{}' (y_i^* - y_i^1)^2 &< m_1, \\ \sum{}' (y_i^* - y^{l+k})^2 &< m_{l+k}. \end{aligned} \tag{12.37}$$

В неравенствах (12.37) штрих у суммы означает, что суммирование производится лишь по значениям функции в тех точках x, которые принадлежат экстремальной окрестности.

В тех случаях, когда неравенства (12.37) совместны, существует допустимое множество $\{Y\}$ решений $Y = (y_{l+1}, \ldots, y_{l+k})^T$. В качестве окончательного ответа выберем такой вектор Y^*, который находится на наименьшем расстоянии от наиболее далекой точки допустимого множества, т. е. найдем вектор Y^* такой, который является минимаксным множества $\{Y\}$

$$Y^* = \arg \min_{\hat{Y}} \max_{\overline{Y}} \| \hat{Y} - \overline{Y} \|.$$

Отыскание такого вектора является задачей выпуклого программирования. С помощью соответствующих алгоритмов вектор Y может быть найден.

Однако в целях упрощения вычислений мы будем искать минимаксный вектор не множества $\{Y\}$, а минимаксный вектор более широкого множества $\{Y\}'$. Так же как и допустимое множество $\{Y\}$, множество $\{Y\}'$ задано пересечением $l+k$ цилиндров. Однако вместо цилиндров со сферическими основаниями

$$\sum_i{}' (y_i^* - y_i^s)^2 = m_s$$

рассмотрим цилиндры с прямоугольными основаниями

$$y_i^s - \sqrt{m_s} \leqslant y_i^* \leqslant y_i^s + \sqrt{m_s}.$$

Очевидно, что допустимое множество $\{Y\}$ содержит допустимое множество $\{Y\}'$. Минимаксная же точка множества $\{Y\}'$ отыскивается по следующему простому правилу: для каждого вектора x_{l+i} рабочей выборки найдем с помощью $i+1$-го и последнего столбцов таблицы два числа

$$a_i = \min_s (y_i^s - \sqrt{m_s}), \quad b_i = \max_s (y_i^s + \sqrt{m_s}).$$

Минимаксный вектор множества $\{Y\}'$ есть вектор с координатами

$$y_i = \frac{a_i + b_i}{2}, \quad i = l+1, \ldots, l+k.$$

ПОСЛЕСЛОВИЕ

В этой книге задача восстановления зависимостей по эмпирическим данным рассматривалась с позиций приближения функций.

В ней были реализованы две новые идеи:

— задание структуры на классе функций, в котором ведется восстановление и минимизация риска по элементам структуры (метод упорядоченной минимизации риска);

— выделение классов эквивалентности и задание на них структуры (восстановление значений функции в заданных точках).

Было показано, что развитие этих идей приводит к созданию более совершенных методов восстановления, чем традиционные.

Однако все конкретные структуры, рассмотренные в книге, возникли скорее из соображений здравого смысла, чем в результате анализа. Между тем принятое определение структуры удовлетворяет аксиоматике алгебраических структур. Поэтому можно ожидать, что средствами анализа удастся найти структуры более содержательные, чем те, которые использовались.

Исследований в этом направлении нет.

КОММЕНТАРИИ

К главе I

Проблема минимизации среднего риска по эмпирическим данным является одной из основных проблем прикладного анализа. Она изучалась многими авторами: *Л. Ле-Камом* [95, 96], *П. Хубером* [87, 90], *Я. З. Цыпкиным* [67, 68], *В. Н. Вапником* [7—14], *А. Я. Червоненкисом* [11—14] и др.

В этой книге рассматривается специальный класс задач минимизации среднего риска — задачи восстановления зависимостей, к которым относятся задачи: обучения распознаванию образов, восстановления регрессии, интерпретации результатов косвенных экспериментов.

Теория распознавания образов появилась в конце 50-х годов. В 60—70 годах ей были посвящены монографии *М. А. Айзермана, Э. М. Бравермана, Л. И. Розоноэра* [2], *Я. З. Цыпкина* [66, 67], *В. Н. Вапника, А. Я. Червоненкиса* [12], *Н. Г. Загоруйко* [19], *В. А. Ковалевского* [25], *К. Фу* [64], *Н. Нильсона* [43], *В. Н. Фомина* [62], *Ю. И. Журавлева* [18] и др.

Проблему восстановления регрессии изучали еще со времен Гаусса. Ей посвящена многочисленная литература и, в частности, такие классические работы, как монографии *С. Рао* [49], *Ю. В. Линника* [34], *М. Кендалла, А. Стьюарта* [24].

Наконец, проблема интерпретации результатов косвенных экспериментов приводится к решению операторных уравнений, образующих некорректно поставленные задачи.

Теории некорректно поставленных задач в 50-х—70-х годах было посвящено много работ (см. библиографию в [56]). Среди этих работ мы отметим монографию *А. Н. Тихонова, В. Я. Арсенина* [56] и работы *В. К. Иванова* [20, 21], по материалам которых написано приложение к главе I.

В книге выделен специальный класс стохастических некорректных задач — задача интерпретации результатов косвенных экспериментов.

К главе II

Применение методов стохастической аппроксимации для решения задач минимизации среднего риска на больших выборках связано с работами *Я. З. Цыпкина* [66, 67] и *М. А. Айзермана, Э. М. Бравермана, Л. И. Розоноэра* [2]. В этих работах наряду с условиями сходимости процедур типа стохастической аппроксимации рассмотрены конкретные применения к задачам распознавания образов и восстановления регрессии. Математические вопросы теории стохастической

аппроксимации рассмотрены в монографии *М. Б. Невельсона Р. З. Хасьминского* [42].

При минимизации функционала среднего риска по ограниченному множеству эмпирических данных различаются два направления исследования: классическое направление, основанное на методах параметрической статистики, и направление, основанное на минимизации эмпирического риска.

Методы параметрической статистики были разработаны в 20—40-х годах и связаны с именами таких замечательных статистиков как *Р. Фишер, К. Пирсон, Г. Крамер*. Сейчас методы параметрической статистики являются рабочим инструментом в решении многих задач. Они излагаются во всех руководствах по статистике. См. например, *С. Уилкс* [58], *М. Кендалл, А. Стьюарт* [24].

Вопрос о применимости методов минимизации эмпирического риска для поиска минимума среднего риска начали изучать позже.

В 1954 г. появился результат *Л. Ле-Кама* [95], согласно которому для определенных классов функций потерь метод минимизации эмпирического риска с ростом объема выборки определяет функцию, минимизирующую средний риск. В этой работе Ле-Кам впервые связал проблему минимизации риска с условиями равномерной сходимости средних к математическим ожиданиям и нашел условия равномерной сходимости для определенных видов функции потерь. В 1968 г. *П. Хубер* [87] показал, что метод минимизации эмпирического риска применим и для функций потерь более общего вида. Однако как работы Ле-Кама, так и работы Хубера исследуют асимптотические возможности метода.

В 1971 г. в работе *В. Н. Вапника, А. Я. Червоненкиса* [11] были найдены необходимые и достаточные условия равномерной сходимости частот появления событий к их вероятностям и получены оценки скорости такой сходимости. На базе этих оценок удалось обосновать применимость метода минимизации эмпирического риска для решения задачи обучения распознаванию образов на выборках ограниченного объема. Позже в 1974 г. этот результат был распространен и на задачи восстановления зависимостей более общей природы (*В. Н. Вапник, А. Я. Червоненкис* [13]).

К главе III

Проблеме восстановления плотности вероятностей заданной с точностью до конечного числа параметров посвящены многочисленные работы [49, 34, 24].

Однако оказалось, что до недавнего времени почти все работы в этом направлении сводились к оцениванию неизвестных параметров плотности, а не восстановлению функции плотности. Лишь в 1965 г. *Д. Кин* [93] получил байесову оценку плотности нормального закона (она приведена в § 7), которая оказалась не принадлежащей классу нормальных.

В 1969 г. *П. Я. Лумельский* и *П. Н. Сапожников* получили наилучшую несмещенную оценку плотности многомерного нормального закона [36]. (Этот результат приведен в § 10.) Наилучшую несмещенную оценку плотности одномерного закона ранее получил *А. Н Колмогоров* [27].

Что же касается задачи оценивания параметров, то здесь основные результаты получены еще *Р. Фишером* [82]. Эти результаты и составляют основу методов параметрического анализа.

Проблемы дискриминантного анализа в основном концентрируются вокруг построения линейной дискриминантной функции. Постановка этой задачи впервые была дана *Р. Фишером* [82], который для ее решения предложил минимизировать функционал, приведенный в § 2. В 1966 г. задача построения линейной дискриминантной функции для нормальных распределений была решена *Т. В. Андерсеном* и *Р. Р. Бахадуром* [71].

Другие исследования здесь связаны с попыткой выписать функционал, минимизация которого приводила бы к построению линейной дискриминантной функции не только для нормальных распределений.

Сначала в качестве такого функционала использовался функционал Р. Фишера, а затем рассматривались и другие функционалы. Подробный обзор литературы по дискриминантному анализу приведен в [60].

Случай независимо распределенных дискретных признаков также рассматривался в дискриминантном анализе.

В 1952 г. *А. М. Аптли* построил дискриминантный автомат, алгоритм которого, по существу, мало отличается от современных дискриминантных автоматов, построенных в соответствии с гипотезой о независимости дискретных признаков [105].

К главе IV

Идея построения устойчивого в заданном классе плотностей метода оценивания параметра сдвига принадлежит *П. Хуберу*. В 1967 г. он получил устойчивый метод оценивания параметра сдвига в классе плотностей, заданных смесью [88] (результат Хубера и приведен в § 8).

Затем другими авторами были получены устойчивые методы оценивания параметра сдвига в разных классах функций. В частности, устойчивые методы оценивания были получены в классе плотностей, сосредоточенных в основном на отрезке, классе плотностей с функциями распределения, близкими к нормальным, и т. д. Подробно обзор имеющихся методов устойчивого оценивания дан в работе *Б. Т. Поляка* и *Я. З. Цыпкина* [46].

Применение методов устойчивого оценивания параметра сдвига к оцениванию пораметров регрессии также связано с работами *Б. Т. Поляка* и *Я. З. Цыпкина* [46]. На различных модельных примерах они показали преимущество устойчивого метода оценивания параметров регрессии в условиях ограниченного объема выборки.

К главе V

Оценивание параметров является традиционным методом решения задачи восстановления регрессии. Центральное место в теории оценивания параметров регрессии по выборке ограниченного объема занимают исследования метода наименьших квадратов, устанавливающие его экстремальность (теорема о нормальной регрессии, теорема Гаусса — Маркова).

Эти теоремы устанавливают оптимальность метода наименьших квадратов среди некоторого заданного множества методов. При этом предполагается, что метод наименьших квадратов является наилучшим методом оценивания параметров не только в заданном узком классе методов, но и хорошим вообще (в достаточно широком классе методов).

В 1956 г. *К. Стейн* [103] неожиданно привел пример, показывающий, что наилучшая оценка среднего многомерного нормального закона с известной ковариационной матрицей $\sigma^2 I$ (σ^2 — известное число, I — единичная матрица) отлична от вектора реализаций (т. е. не приводится к методу наименьших квадратов).

В 1961 г. *В. Джеймс* и *К. Стейн* [91] нашли метод оценки среднего для многомерного нормального закона с неизвестной величиной σ^2 ковариационной матрицы $\sigma^2 I$ равномерно лучший, чем оценка с помощью реализации. Наконец, в 1970 г. *А. Я. Баранчик* [73] построил класс оценок, равномерно лучших оценки с помощью реализации. Этот класс оценок и приведен в книге для получения оценок параметров нормальной регрессии равномерно лучших, чем оценки метода наименьших квадратов. Метод построения оценок параметров регрессии, использующий оценки Джеймса — Стейна — Баранчика, приведенный в § 3, получен с помощью теоремы *П. К. Бхаттачария* [75].

Пример, приведенный К. Стейном, показал необоснованность гипотезы о том, что несмещенные методы оценивания всегда содержат «хорошие». (Ведь уже в самой простой ситуации строятся методы оценивания, равномерно лучшие, чем классические.)

Приведенная в главе теория построения наилучшего линейного метода оценивания принадлежит *В. А. Кощееву* [31]. Эта теория дает возможность, используя априорную информацию, получить линейные оценки лучшие, чем те, которые следуют из метода наименьших квадратов.

Однако вопрос о том, существует ли метод оценивания параметров регрессии лучший, чем метод наименьших квадратов в случае, когда не используется дополнительная априорная информация, остается открытым и связан с построением метода оценивания среднего равномерно лучшего, чем эмпирическое среднее для случайных векторов, которые являются реализацией не обязательно нормального закона.

Иначе говоря, проблема сводится к получению оценок стейновского типа, инвариантных по отношению к законам плотности вероятностей. Такие оценки возможны. (См., например, работу *Дж. Бергера* [74].)

К главам VI и VII

Проблема равномерной сходимости частот появления событий к их вероятностям впервые была рассмотрена в работах *В. И. Гливенко* [85] и *Ф. П. Кантелли* [92]. В 1933 г. они показали, что имеет место равномерная сходимость эмпирических кривых распределения к функции распределения (равномерная сходимость частот к вероятностям по специальному классу событий). В том же году *А. Н. Колмогоров* [94] нашел асимптотическую оценку скорости сходимости, которая позже была уточнена *Н. В. Смирновым* [53].

Обоснование применимости метода минимизации эмпирического риска для решения задач обучения распознавания образов связано с установлением условий равномерной сходимости частот к вероятностям для произвольных классов событий.

В 1971 г. *В. Н. Вапник* и *А. Я. Червоненкис* [11] нашли необходимые и достаточные условия равномерной сходимости частот появления событий к их вероятностям для произвольной системы событий и получили оценки скорости такой сходимости.

В этой книге используются лишь достаточные условия. Подробно необходимые и достаточные условия изложены в монографии *В. Н. Вапника, А. Я. Червоненкиса* [12].

Содержание главы VII является прямым обобщением результатов, полученных при оценке скорости равномерного относительного уклонения частот от вероятностей на оценку скорости равномерного относительного уклонения средних от математических ожиданий. Они получены *В. Н. Вапником* и *А. Я. Червоненкисом* в 1974 г. [13].

Оценки скорости сходимости равномерного относительного уклонения, выраженные через ε-энтропию множества функции, приводятся здесь впервые.

К главе VIII

Метод упорядоченной минимизации риска был сформулирован для решения задачи обучения распознаванию образов в монографии *В. Н. Вапника, А. Я. Червоненкиса* [12].

Однако, по существу, при построении алгоритмов минимизации риска к нему обращаются каждый раз, когда метод минимизации эмпирического риска приводит к абсурдным результатам. (Например, при восстановлении полиномиальной регрессии.)

Двухуровневая процедура выбора (элемента структуры и наилучшей функции, принадлежащей данному элементу структуры) содержится во всех эвристических алгоритмах, цель которых получить решение лучшее, чем то, которое следует из стандартной методики минимизации эмпирического риска (см., например, работы *И. Ш. Пинскера* [44] и *А. Г. Ивахненко* [22]).

В этой книге в качестве критерия выбора элемента структуры используются две идеи: оценка процедуры «скользящий контроль» и равномерная оценка величины среднего риска по величинам эмпирического.

Оценка среднего риска следует из теории равномерной сходимости. Что же касается процедуры «скользящий контроль», то, видимо, впервые она была предложена *М. Н. Вайнцвайгом* в 1968 г. В 1969 г. *А. Л. Лунц* и *В. Л. Брайловский* показали несмещенность оценки [37]. В главе VIII дано эквивалентное представление оценки «скользящий контроль» для регрессии, позволяющее существенно сократить объем вычислений.

Селекция обучающей выборки рассмотрена здесь впервые.

К главе IX

Идея применения метода упорядоченной минимизации риска для решения некорректных задач измерений была реализована в 1974 г. в работе *В. Н. Вапника* и *А. И. Михальского* [8]. Однако и ранее

использовались различные (эвристические) приемы, позволяющие выбрать подходящий вид приближения (см., например, работу *Л. А. Вайнштейна* [6] и работу *Л. П. Грабарь* [17]).

В 1975 г. в работе [14] *В. Н. Вапник* и *А. Я. Червоненкис* установили факт сходимости с ростом объема эмпирических данных последовательности решений, получаемых методом упорядоченной минимизации риска к искомому при условии, что решение ищется в виде разложения по специальной системе функции, если же решение искать в виде разложения по полиномам, то такой сходимости может и не быть.

В 1974 г. *А И Михальский* показал, что для некоторых классов операторных уравнений существует сходимость с ростом объема выборки решений, определяемых с помощью метода упорядоченной минимизации риска, к искомой функции, если решение искать в классе сплайнов. Им же была развита техника построения сплайнов с заданным числом сопряжений, минимизирующих эмпирический риск [38].

Идея восстановления плотности распределения вероятностей, как решение некорректной задачи численного дифференцирования была реализована в работе *В. Н. Вапника, А. Р. Стефанюка* [10]. В этой же работе было получено обобщение теорем А. Н. Тихонова на стохастический случай, приведенное в приложении к главе IX.

К главе X

Впервые задача восстановления значений функции в заданных точках была рассмотрена в монографии *В. Н. Вапника* и *А. Я. Червоненкиса* [12] для восстановления значений характеристических функций. В работе *В. Н. Вапника, А. М. Стерина* [9] были рассмотрены различные структуры на классах эквивалентности характеристических функций.

Методы восстановления значений произвольной функции в заданных точках рассмотрены здесь впервые. Также впервые здесь исследуется селекция полной выборки.

К главам XI—XII

Библиотека программ метода обобщенного портрета была разработана *Т. Г. Глазковой* и *А. А. Журавель*. Алгоритмы восстановления значений характеристической функции в заданных точках реализовал и исследовал *А. М. Стерин*.

Алгоритмы восстановления регрессии были созданы *Т. Г. Глазковой, В. А. Кощеевым, А. И. Михальским*.

Алгоритмы интерпретации некорректных задач измерений созданы *А. И. Михальским*.

ЛИТЕРАТУРА

1. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. — М.: Статистика, 1974.
2. Айзерман М. А., Браверман Э. М., Розоноэр Л. И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970.
3. Алгоритмы обучения распознаванию образов. Сб. под ред. В. Н. Вапника. — М.: Советское радио, 1973.
4. Альберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения. — М.: Мир, 1972.
5. Андерсон Т. Введение в многомерный статистический анализ. — М.: Физматгиз, 1963.
6. Вайнштейн Л. А. О численном решении интегральных уравнений первого рода с использованием априорных сведений о восстанавливаемой функции. — ДАН СССР, 1972, т. 204, № 6.
7. Вапник В. Н. Машины, обучающиеся распознаванию образов. — Алгоритмы обучения распознаванию образов. Сб. под ред. В. Н. Вапника, М.: Советское радио, 1973.
8. Вапник В. Н,, Михальский А. И. О поиске зависимостей методом упорядоченной минимизации риска. — Автоматика и телемеханика, 1974, № 10.
9. Вапник В. Н., Стерин А. М. Об упорядоченной минимизации суммарного риска в задаче распознавания образов. — Автоматика и телемеханика, 1977, № 10.
10. Вапник В. Н., Стефанюк А. Р. Непараметрические методы восстановления плотности вероятностей. — Автоматика и телемеханика, 1978, № 8.
11. Вапник В. Н., Червоненкис А. Я. О равномерной сходимости частот появления событий к их вероятностям. — Теория вероятностей и ее применения, 1971, т. XVI, № 2.
12. Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974.
13. Вапник В. Н., Червоненкис А. Я. О методе упорядоченной минимизации риска. — Автоматика и телемеханика, 1974, № 8, 9.
14. Вапник В. Н., Червоненкис А. Я. Об асимптотических свойствах метода упорядоченной минимизации риска. — Автоматика и телемеханика, 1975, № 12.
15. Витушкин А. Г. Оценка сложности задачи табулирования. — М.: Физматгиз, 1959.
16. Гнеденко Б. В. Курс теории вероятностей. — М.: Физматгиз, 1961.
17. Грабарь Л. П. Применение полиномов Чебышева, ортонормированных на системе равноотстоящих точек для решения интегральных уравнений первого рода. — ДАН СССР, 1967, т. 172, № 4.

18. Журавлев Ю. И. Математические модели в задачах распознавания и классификации. — М.: Наука, 1978.
19. Загоруйко Н. Г. Методы распознавания и их применения. — М.: Советское радио, 1972.
20. Иванов В. К. О линейных некорректных задачах. — ДАН СССР, 1962, т. 145, № 2.
21. Иванов В. К. О некорректно поставленных задачах. — Математический сборник, 1963, т. 61, № 2.
22. Ивахненко А. Г., Зайченко Ю. П., Дмитров В. Д. Принятие решений на основе самоорганизации. — М.: Советское радио, 1976.
23 Каган А. М., Шалаевский О. В. Допустимость оценок наименьших квадратов — исключительное свойство нормального закона. — Математические заметки, 1969, т. 6, № 1.
24. Кендалл М., Стьюарт А. Статистические выводы и связи. — М.: Наука, 1973.
25. Ковалевский В. А. Методы оптимальных решений в распознавании изображений. — М.: Наука, 1976.
26. Колмогоров А. Н. Несмещенные оценки. — Изв. АН СССР, сер. математическая, 1950, № 4.
27. Колмогоров А. Н., Тихомиров В. М. ε-энтропия и ε-емкость в функциональных пространствах. — Успехи математических наук, 1959, № 2.
28. Колмогоров А. Н., Фомин С. В. Элементы теории функции и функционального анализа. — М.: Физматгиз, 1961.
29. Корбут А. А., Финкельштейн Ю. Ю. Дискретное программирование. — М.: Наука, 1969.
30. Коровкин П. П. Линейные операторы и теории приближений. — М.: Физматгиз, 1959.
31. Кощеев В. А. Метод учета априорной информации в линейном оценивании параметров. — Статистические методы теории управления. М.: Наука, 1978.
32. Лаврентьев М. М. О некоторых некорректных задачах математической физики. — Изд-во СО АН СССР, 1962.
33. Ле-Кам Л. О некоторых асимптотических свойствах оценок максимума правдоподобия. — Математика, 1960, № 1.
34. Линник Ю. В. Метод наименьших квадратов. — М.: Физматгиз, 1962.
35. Линник Ю. В. Статистические задачи с мешающими параметрами. — М.: Наука, 1966.
36. Лумельский П. Я., Сапожников П. Н. Несмещенные методы для плотностей распределений. — Теория вероятностей и ее применения, 1969, № 2.
37. Лунц А. Л., Браиловский В. Л. Об оценке признаков, полученных в статистических процедурах распознавания. — Изв. АН СССР, сер. Техническая кибернетика, 1969, № 3.
38. Михальский А И. Метод осредненных сплайнов в задаче приближения зависимостей по эмпирическим данным. — Автоматика и телемеханика, 1974, № 3.
39 Морозов В. А. О принципе невязки при решении несовместных уравнений методом регуляризации А. Н. Тихонова. — Журн. выч. мат. и мат. физ., 1973, т. 13, № 5.

40. Морозов В. А. О вычислении нижних граней функционалов по приближенной информации. — Журн. выч. мат. и мат. физ., 1973, т. 13, № 4.
41. Надарая Э. А. О непараметрических оценках плотности вероятностей и регрессии. — Теория вероятностей и ее применения, 1965, № 1.
42. Невельсон М. Б., Хасьминский Р. З. Стохастическая аппроксимация и рекуррентное оценивание. — М.: Наука, 1972.
43. Нильсон Н. Обучающиеся машины. — М.: Мир, 1967.
44. Пинскер И. Ш. Выбор структуры и вычисление параметров решающего правила при ограниченной выборке. — сб. Моделирование и автоматический анализ электрокардиограмм, М.: Наука, 1973.
45. Поляк Б. Т., Цыпкин Я. З. Псевдоградиентные алгоритмы адаптации и обучения. — Автоматика и телемеханика, 1973, № 3.
46. Поляк Б. Т., Цыпкин Я. З. Помехоустойчивая идентификация. — Труды IV симпозиума ИФАК по идентификации, часть I. — Тбилиси: Мецниереба, 1976.
47. Пугачев В. С. Статистическая теория обучающихся автоматических систем. — Техническая кибернетика, 1967, № 6.
48. Пугачев В. С. Статистические проблемы теории распознавания образов. — Труды III Всесоюзного совещания по автоматическому управлению, М.: Наука, 1967.
49. Рао С. Р. Линейные статистические методы и их применения. — М.: Наука, 1968.
50. Растригин Л. А., Эренштейн Р. Х. Принятие решений коллективом решающих правил в задачах распознавания образов. — Автоматика и телемеханика, 1975, № 9.
51. Раудис Ш. Ю. Ограниченность выбора в задачах классификации. — Статистические проблемы управления, вып. 18, Вильнюс: Институт физики и математики АН Лит. ССР, 1977.
52. Рыжик И. М., Градштейн И. С. Таблицы интегралов, сумм, рядов и произведений. — М.: Гостехиздат, 1956.
53. Смирнов Н. В. Теория вероятностей и математическая статистика (Избранные труды). — М.: Наука, 1970.
54. Тихонов А. Н. О решении некорректно поставленных задач. — ДАН СССР, 1963, т. 151, № 3.
55. Тихонов А. Н. О регуляризации некорректно поставленных задач. — ДАН СССР, 1963, т. 153, № 1.
56. Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. — М.: Наука, 1974.
57. Турчин В. Ф., Козлов В. Н., Малкевич М. С. Использование методов математической статистики для решения некорректных задач. — Успехи физических наук, 1970, т. 102, № 3.
58. Уилкс С. Математическая статистика. — М.: Наука, 1967.
59. Уилкинсон Дж., Алгебраическая проблема собственных значений. — М.: Наука, 1970.
60. Урбах В. Ю. Дискриминантный анализ: основные идеи и приложения. — сб. Статистические методы классификации, вып. 1, Изд-во МГУ, 1969.
61. Феллер В. Введение в теорию вероятностей и ее приложения. — М.: Мир, 1964.

62. Фомин В. Н. Математическая теория обучаемых опознающи› систем. — Изд-во ЛГУ, 1976.
63. Форсайт Дж., Моллер К. Численное решение систем линей ных алгебраических уравнений. — М.: Мир, 1968.
64. Фу К. Последовательные методы в распознавании образов и обучении машин. — М.: Наука, 1971.
65. Хедли Дж. Нелинейное и динамическое программирование. — М.: Мир, 1967.
66. Цыпкин Я. З. Адаптация и обучение в автоматических системах. — М.: Наука, 1968.
67. Цыпкин Я. З. Основы теории обучающихся систем. — М.: Наука, 1970.
68. Ченцов Н. Н. Оценка неизвестной плотности по наблюдениям. — ДАН СССР, 1962, т. 147, № 1.
69. Якубович В. А. Некоторые общие теоретические принципы построения обучаемых опознающих систем. — сб. Вычислительная техника и вопросы программирования, Изд-во ЛГУ, 1965.
70. Якубович В. А. Рекуррентные конечно-сходящиеся алгоритмы решения системы неравенств. — ДАН СССР, 1966, т. 166, № 6.
71. Anderson T. W., Bahadur R. R. Classification into two multivariate normal distributions with different covariance metrices. — The Annals of Mathematical Statistics, June, v. 133, No. 2, 1966.
72. Andrews H. Introduction to mathematical techniques in pattern recognition. — Wiley, N. Y., 1972.
73. Baranchik A. J. A Family of minimax estimators of the mean of a multivariate normal distribution. — The Annals of Mathematical Statistics, v. 41, No. 2, 1970.
74. Berger J. O. Minimax estimation of location vectors of a wide class of densities. — The Annals of Statistics, v. 3, No. 6, 1975.
75. Bhattacharya P. K. Estimating the mean of a multivariate normal population with general quadratic loss function. — The Annals of Mathematical Statistics, v. 37, No. 18, 1966.
76. Cacoullos T. Estimation of a multivariate density. — Inst. Stat. Math. Tokyo v. 18, No. 2, 1966.
77. Chen C. H. Statistical pattern recognition. — Hayden, N. Y., 1973.
78. Devroye L. P., Wagner T. J. A distribution-free performance bound in error estimation. — IEEE Transaction on Information Theory, v. IT-22, No. 5, 1976.
79. Dvoretsky A. On stochastic approximation. — Proceedings of the III Berkeley Symposium on Mathematical Statistics and Probability, v. 1, 1956.
80. Dancel J. W. The conjugate gradient method for linear and nonlinear operator equation. — SIAM Numbr. Anal, v. 4, No. 1, 1967.
81. Fix I. R., Hodges J. L. Discriminatory analysis; nonparametric discrimination: consistensy properties. — Report 4 of the USAF School of Aviation Medicine, Raudolph Field, Texas, 1952.

82. Fisher R. A. Contributions to Mathematical Statistics. — N. Y., 1952.
83. Fraser D. A. S. Nonparametric Method in Statistics. — Wiley, N. Y., 1957.
84. Fukunaga K. Introduction to statistical pattern recognition. — Academic, N. Y., 1972.
85. Glivenko V. I. Sulla determinatione empirics di probabilits.— Giornale dell' Institute Italiano degli Attuari, 4, 1933.
86. Hostenes M. R., Stiefel E. Method of conjugate gradient's for solving linear systems. — J. Res. Mat, Bur. Standards, 49, No. 6, 1952.
87. Huber P. The behaviour of maximum likelyhood estimates under nonstandard condition. — Proc. Fifth Berkley Symp. on Math Statistics, Prob., v. 1, 1967.
88. Huber P. Robust estimation of the a location parameter. — The Annals of Mathematical Statistics, v. 35, No. 1, 1964.
89. Huber P. Robust statistics: a review. — The Annals of Mathematical Statistics, v. 43, No. 4, 1972.
90. Huber P. Robust regression: asymptotics, conjectures and Monte Carlo. — The Annals Statistics, v. 1, No. 5, 1973.
91. James W., Stein C. Estimation with quadratic loss. — Proc. Fourth Berkeley Symposium on Math. Statist. Prob, v. 1 Univ. of California Press, 1961.
92. Kantelly F. P. Sulla determinazione empirica della leggi di probabilita. — Giornale dell' Institute Italiano degli Attuari, 1933.
93. Keehn D. G. Note on learning for Gaussian properties. — IEEE. Transaction on Information Theory, v. — IT-11, No. 1, 1965.
94. Kolmogoroff A. N. Sulla determinatione empirica di una legge di distributione. — Giornale dell' Institute Italiano degli Attuari, N. 4, 1933.
95. Le Kam L. On some asymptotic properties of maximum likehood estimates and related Bayes' estimates. — Calif. Public. Statist. v. 11, 1953.
96. Le Kam L. On asymptotic theory of estimation and testing hypotheses. — Proc. III Berkeley Symposium on Math. Statist. and Probabilyty, v. 1, 1956.
97. Lawler E. L., Waod D. E. Branch-and-bound method. — A survey. Oper. Res. v. 14, No. 4, 1966.
98. Meisel W. S. Computer-oriented approaches to pattern recognition. — Academic, N. Y., 1972.
99. Parsen E. On the estimation of a probability function and mode. — The Annals of Mathematical Statistics, v. 33, No. 3, 1962.
100 Reiss R. D. Consistensy of certain class of empirical densiti finction. — Metrika, v. 22, No. 4, 1975.
101. Saks J., Ylvisaker D. A note on Hubers Robust estimation of a location parameter. — The Annals of Mathematical Statistics, v. 43, No. 4, 1972.
102. Sclove S. L. Improved estimators for coefficient in Linear Regression. — Journal of the American Statistical Association, v. 63, No. 322, June, 1968.

103. Stein C. Inadmissibility of usual estimator for mean of a multivariate normal distribution — Proc. Third Berkeley Symp. Math. Stat. Prob. v. 1, Univ. California Press, 1956.
104. Wald A. Note on consistency of the most likehood estimate. — The Annals of Mathematicil Statistics, № 20, 1949.
105. Uttley A. M. A theory of the mechanism of learning on the computation of conditional probabilities. — Proc. 1 st., Congress Int. Cybernetics, Namur, 1956.

Владимир Наумович Вапник

ВОССТАНОВЛЕНИЕ ЗАВИСИМОСТЕЙ
ПО ЭМПИРИЧЕСКИМ ДАННЫМ

М., 1979 г., 448 стр. с илл.

Редактор *В. А. Кощеев*
Техн. редактор *И. Ш. Аксельрод*
Корректоры *О. А. Сигал, М. Л. Медведская*

ИБ № 2250

Сдано в набор 07.07.78. Подписано к печати 02.02.79. Т-05314. Бумага 84 × 108$^{1}/_{32}$, тип. № 1. Литературная гарнитура. Высокая печать. Условн. печ. л. 23,52. Уч.-изд. л. 23,14. Тираж 5500 экз. Заказ № 68. Цена книги 1 р. 70 к.

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

Отпечатано в ордена Трудового Красного Знамени Ленинградской типографии № 2 имени Евгении Соколовой «Союзполиграфпрома» при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 198052, Ленинград, Л-52, Измайловский проспект, 29 с матриц ордена Октябрьской Революции, ордена Трудового Красного Знамени Ленинградского производственно-технического объединения «Печатный Двор» имени А. М. Горького «Союзполиграфпрома» при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 197136, Ленинград, П-136, Гатчинская, 26.