1. Проблема обучения распознаванию образов.

1.1 Прикладные задачи, приводящие к проблемам малонаполненных выборок и необходимости использования метрического пространства признаков.

В теории распознавания образов часто встречаются задачи, оперирующие с большим числом признаковых переменных. Это, во-первых, приложения, в которых объединены данные от большого числа датчиков, во-вторых, задачи, объединяющие множество многомерных моделей, параметры которых моделей могут быть использованы для классификации, и, в-третьих, приложения, основанные на получении (извлечении) скрытых зависимостей между признаками.

Яркими примерами задач первого типа могут служить задачи распознавания речевых команд.

Рассмотрим речевую команду. Подобно тому, как сигнал с некоторым неизменным характером колебаний характеризуется его спектром, представляющим собой функцию частоты $x(f)$, адекватной характеристикой нестационарного сигнала будет последовательность его спектров $x_t(f)$, каждый из которых отражает локальный характер колебаний в некоторой окрестности текущей точки. На практике в каждый момент времени достаточно вычислить конечное число спектральных составляющих для некоторых фиксированных частот $f^{(1)},...,f^{(n)}$. Таким образом, результат спектрально-временного анализа сигнала $X=(x_t, t\in T)$ представляет собой последовательность его мгновенных спектров $x_t=(x_t^{(1)},...,x_t^{(n)})$, принимающих значения из множества X действительных векторов некоторой фиксированной размерности. В большинстве современных систем распознавания речи анализируется именно последовательность мгновенных спектров сигнала. Естественно, что как число спектров в последовательности, так и число частот, образующих каждый из них, должны быть достаточно велики, чтобы отразить особенности каждого звука в составе произнесенной команды. Как результат, произнесенная команда оказывается представленной вектором с весьма большой размерности, содержащим, по крайней мере, сотни компонент.

На рис. 1.1 показан сигнал речи, зарегистрированный при произнесении слова “один”, и его представление в виде последовательности мгновенных интенсивностей спектральных составляющих в семи полосах частот, охватывающих диапазон от 10 до 3000 герц. Спектральные кривые упорядочены в соответствии с увеличением частоты. Интенсивности спектральных составляющих показаны в условных индивидуальных масштабах, что позволяет визуально сравнить динамику их изменения во времени.

К задачам подобного типа относится также задача распознавания рукописных символов при их вводе в ЭВМ непосредственно в процессе написания с помощью специального пера. При таком способе ввода каждый символ первоначально оказывается представленным сигналом, состоящим из двух компонент, а именно, текущих координат пера по вертикали и горизонтали, однако, может оказаться целесообразным использовать и дополнительные локальные характеристики процесса написания, например, угловой азимут мгновенного направления движения пера, его скорость, силу прижатия к бумаге, временные отрывы от нее, наклон и т.п. В качестве аргумента сигнала может выступать либо время, либо длина пути, пройденного пером от точки первого касания бумаги. На рис.1.2 представлен трехкомпонентный сигнал в виде функций координат и азимута движения пера от пройденного пути вдоль его траектории, полученный при написании рукописной буквы “d”.

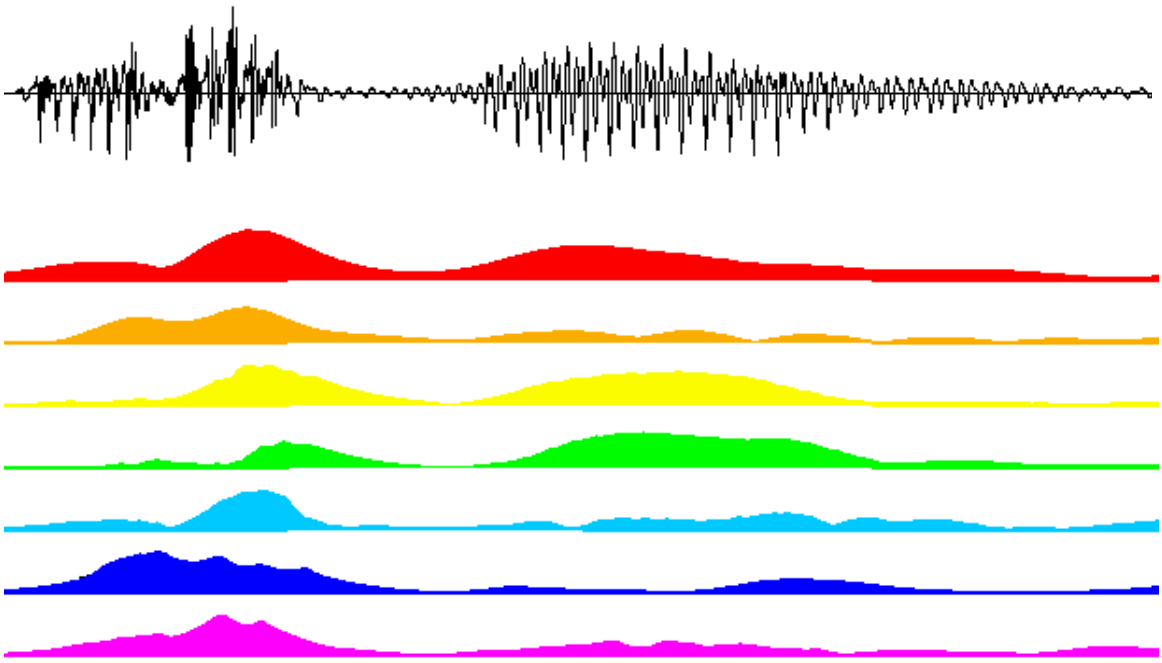


Рисунок. 1.1 Сигнал речи, зарегистрированный при произнесении слова “один”, и его представление в виде последовательности мгновенных интенсивностей спектральных составляющих в семи полосах частот, охватывающих диапазон от 10 до 3000 герц.

Примером задачи второго типа может служить задача классификации использования земельных ресурсов с использованием SAR (синтетического апертурного радара) изображений (рис. 1.2). Например, Солберг и Джэйн [49] использовали текстурные характеристики, рассчитанные на SAR изображениях для классификации каждого пикселя. Все 18 признаков для каждого образа (пикселя) были рассчитаны на основе четырех текстурных моделей: локальные статистики (5 признаков), матрицы уровня серого (6 признаков), фрактальные признаки (2 признака) и логнормальная модель случайного поля (5 признаков).

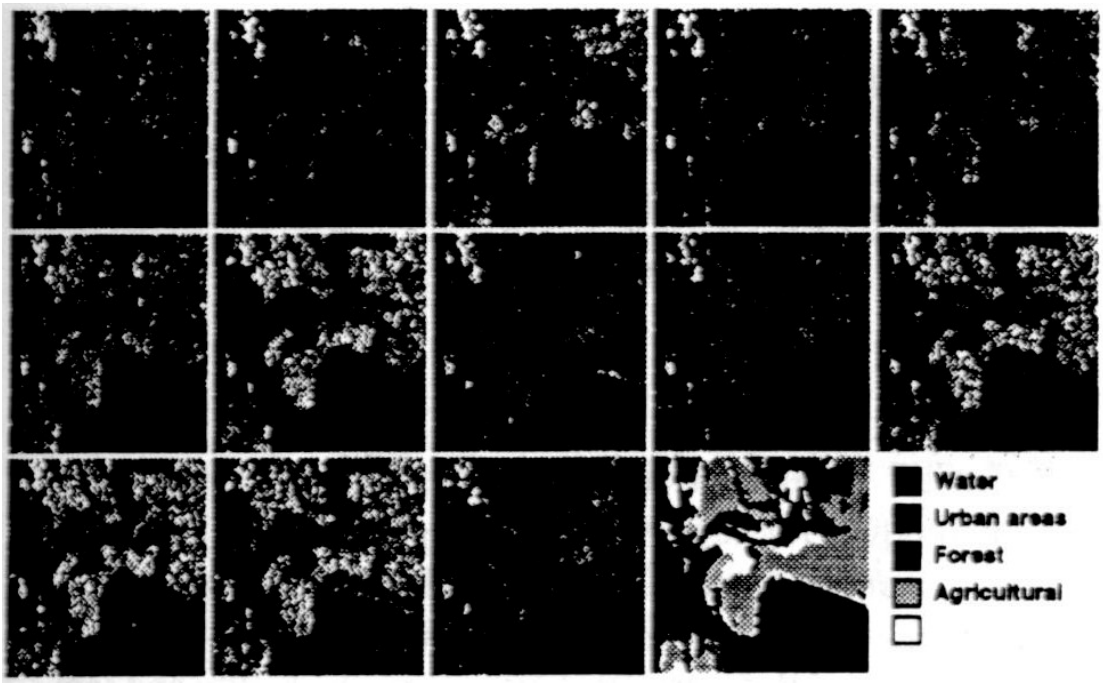


Рисунок. 1.2. Пример SAR изображения.

В качестве примера задач третьего типа [24] может выступить описание физико-химических объектов, которые могут быть представлены в виде совокупности (сочетаний) из n упорядоченных элементов (a, b, c, d, e…) по m [20]. Чтобы отчетливо представлять о каких элементах идет речь, перечислим их.

Во-первых, под (a, b, c, d, e) можно понимать химические элементы H, F, Cl, Br, и J. Тогда, например, метан запишется как CHHHH, а дифторхлорбромметан - CHFFCl. Образуя различные сочетания из пяти элементов по четыре с четырьмя повторениями, получают весь набор галогенметанов (70 молекул).

Во-вторых, молекулярной комбинаторикой могут быть охвачены смеси газов и жидкостей. Принимая за a, b, c, d 25-процентный шаг в концентрациях соответствующего чистого компонента (aaaa), (bbbb), (cccc) или (dddd), путем комбинации элементов получим соответствующую смесь. Например,

100% H_2O	75% H_2O	50% H_2O	25% H_2O	
	25%	50%	75%	100%
	CH_3OH	CH_3OH	CH_3OH	CH_3OH

Таким образом, поиск закономерностей формирования класса молекул в структуры приводят к большой размерности вектора признаков.

Следует отметить, что приведенные задачи являются наиболее яркими примерами класса приложений, оперирующих с многомерными данными, а не исключениями из правил. Круг подобных задач чрезвычайно широк. Даже если разработчику и удается

провести в той или иной мере эффективное снижение размерности или определить наиболее важные факторы, на первых этапах исследования он все же старается запастись как можно большей информацией, чтобы наиболее полно описать изучаемое явление.

Следует заметить, что выбор признаков, образующих удобное для распознавания пространство, представляет собой отдельную весьма сложную проблему. В то же время существует широкий класс прикладных задач распознавания образов, в которых легко удается непосредственно вычислить степень «непохожести» любых двух объектов, но трудно указать набор осмысленных характеристик объектов, которые могли бы служить координатными осями пространства признаков.

Наглядным примером задачи распознавания образов в метрическом пространстве является задача распознавания рукописных символов, например, букв, при их вводе в ЭВМ непосредственно в процессе написания с помощью специального пера. При таком способе ввода каждый символ первоначально оказывается представленным сигналом, состоящим из двух компонент, а именно, текущих координат пера по вертикали и горизонтали, однако, может оказаться целесообразным использовать и дополнительные локальные характеристики процесса написания, например, угловой азимут мгновенного направления движения пера, его скорость, силу прижатия к бумаге, временные отрывы от нее, наклон и т.п. В качестве аргумента сигнала может выступать либо время, либо длина пути, пройденного пером от точки первого касания бумаги. На рис.1.3 представлен трехкомпонентный сигнал в виде функций координат и азимута движения пера от пройденного пути вдоль его траектории, полученный при написании рукописной буквы “d”.

В этой задаче трудно указать заранее фиксированное число признаков сигнала, которые могли бы сформировать пространство, удовлетворяющее гипотезе компактности. Нельзя использовать в качестве признаков и отсчеты сигнала, взятые с некоторым шагом вдоль оси аргумента, поскольку сигналы, полученные от разных написаний даже одного и того же символа неизбежно будут иметь разную длину, и, следовательно, не существует единого линейного пространства, в котором могли бы быть представлены написания распознаваемых символов [43].

Заметим, что разные варианты написания одного и того же символа естественно представить как результат некоторого нелинейного преобразования оси аргумента, приводящего к ее «короблению». Эти различия между разными написаниями, несущественные с точки зрения распознавания символов, легко компенсировать с помощью процедуры так называемого парного выравнивания (рис. 1.3), тогда остающееся несовпадение сигналов будет нести информацию об «истинной»

непохожести сигналов, которую естественно принять в качестве рабочей метрики при построении процедуры обучения распознаванию символов.

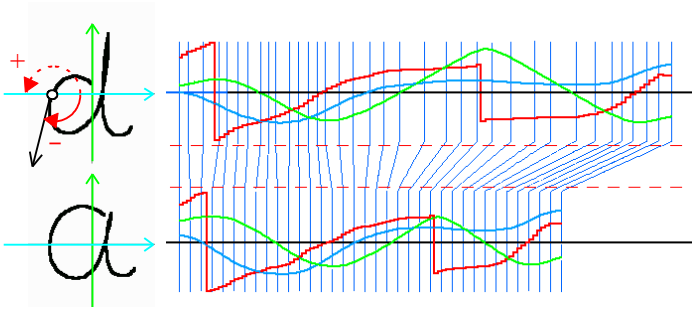


Рисунок. 1.3 Два сигнала в виде функций координат и азимута движения пера от пройденного пути вдоль его траектории, полученные при вводе рукописных символов в компьютер непосредственно в процессе написания. Совмещение произведено по значениям азимута.

Другим примером задач, требующих введения метрического пространства распознавания является задача классификации пространственной структуры белков опираясь на знание лишь первичной структуры (последовательности аминокислот). Информация о пространственной организации белка (третичная структура) является очень важной для понимания механизмов работы макромолекул и их функций. Третичную структуру макромолекул определяют экспериментальными методами (рентгеноструктурный анализ, ядерный магнитный резонанс). Эти методы являются чрезвычайно трудоемкими и требуют больших затрат времени, но позволяют получить достоверные сведения о пространственной организации молекул. Характерно, что для больших групп эволюционно родственных белков, подчас очень значительно отличающихся по первичной структуре и, значит, по распределению всех атомов в пространстве, способ укладки полипептидной цепи остается в главных чертах неизменным. С другой стороны, при всем разнообразии пространственных структур белков удается выделить относительно небольшое число типов укладки полипептидной цепи. Налицо задача классификации – выделение групп белков, достаточно близких друг к другу по пространственной структуре. Один из фундаментальных принципов молекулярной биологии говорит о том, что последовательность аминокислотных остатков полипептидной цепи белка несет в себе всю информацию, необходимую и достаточную для формирования однозначной пространственной структуры. Учитывая это положение, в настоящее время большие усилия прилагаются для разработки методов предсказания третичной структуры молекул на основе известной первичной структуры. Разумеется, абсолютно точно произвести такой прогноз невозможно, остается надежда на то, чтобы правильно «угадать» группу к которой относится исследуемый белок.

В качестве примера, на рис. 1.4 представлено схематичное трехмерное изображение белка Cytochrome C4 и его первичная структура. На сегодняшний момент биологи выделяют определенные типы (семейства, фолды) известных пространственных структур [44]. Положительным результатом процедуры распознавания считается достоверное отнесение белка к одному из таких классов. Налицо задача распознавания образов.

Пространственная структура

Первичная структура:

```
1    AGDAEAGQGK
11   VAVCGACHGV
21 DGNSPAPNFP
31   KLAGQGERYL
41   LKQLQDIKAG
51 STPGAPEGVG
61   RKVLEMTGML
71 DPLSDQDLED
81   IAAYFSSQKG
91   SVGYADPALA
101  KQGEKLFRGG
111  KLDQGMPACT
121  GCHAPNGVGN
131  DLAGFPKLGG
141  QHAAYTAKQL
151  TDFREGNRTN
161  DGDTMIMRGV
171  AAKLSNKDIE
181  ALSSYIQGLH
```

Рисунок. 1.4. 3D представление и первичная структура белка Cytochrome C4

Имея в наличии информацию только о первичной структуре белков входящих в изучаемые данные, первым шагом представляется попытка получить некоторые количественные характеристики, которые бы отражали существо пространственной классификации. В настоящее время открыто более четырехсот признаков аминокислот (наиболее важные - гидрофобность, степень поляризации, размер и др.), однако, прямое использование этих признаков затруднено тем, что различные протеины имеют разную длину, и, следовательно, непосредственное представление первичных структур как векторных "сигналов" их свойств потребует учета специфики работы с задачами такого типа. Другой простейший подход получения количественных признаков из аминокислотной последовательности заключается в подсчете относительного числа остатков каждой из аминокислот к общей длине последовательности. В таком случае каждый протеин будет представлен точкой в

двадцатимерном пространстве. Однако наиболее разумной представляется схема, использующая знания о взаимной близости аминокислотных последовательностей. Существуют априорные объективные данные о близости в химико-биологическом смысле всех пар аминокислот (210 – пар, включая близость «самой с собой»), которые обычно выражаются в виде матрицы соответствия 20×20. Для двух аминокислотных последовательностей пытаются найти такое их взаимное соответствие, чтобы величина «невязки» близостей аминокислот была по возможности минимальной. При этом, для белков различной длины более короткий приходится искусственно «вытягивать» за счет введения в определенные позиции делеций, т.е. разрывать исходную структуру. Результат такой процедуры обычно выражается величиной несходства выравниваемых последовательностей. Опираясь на какую либо процедуру выравнивания последовательностей, например Fasta3 (ftp://ftp.virginia.edu/pub/fasta), строится матрица всех взаимных расстояний между первичными структурами. Такая матрица и рассматривается как метрическое пространство.

1.2 Современные методы обучения распознаванию образов в пространствах действительнозначных признаков.

1.2.1 Постановка задачи обучения распознаванию образов.

Центральными понятиями формальной постановки задачи обучения распознаванию образов [2, 5, 10, 15, 24, 27, 29, 45] являются:

1. Гипотетическое множество (генеральная совокупность) Ω объектов распознавания. Существенно, что это множество не подлежит восприятию.

2. Индикаторная функция $g(\omega):\Omega \to \mathsf{M}$, $\mathsf{M} = \{1,...,m\}$, ставящая в соответствие каждому элементу из Ω индекс из конечного множества M, который называется классом. Эта функция разбивает множество Ω на m непересекающихся классов $\Omega^1,...,\Omega^m$. Она также не известна наблюдателю.

3. Некоторое пространство наблюдений X , $\mathbf{x} \in \mathsf{X}$ в пределах которого некоторая функция $\mathbf{x}(\omega) : \Omega \to \mathsf{X}$, также неизвестная наблюдателю, ставит в соответствие каждому объекту $\omega \in \Omega$ его образ $\mathbf{x}(\omega) \in \mathsf{X}$, непосредственно воспринимаемый наблюдателем.

Конечной целью распознавания образов является способность наблюдателя угадать класс объекта $\omega \in \Omega$ по его видимому образу $\mathbf{x} \in \mathsf{X}$, т.е. определить функцию $\hat{g}(\mathbf{x}) : \mathsf{X} \to \mathsf{M}$ - конечное решающее правило, которое позволило бы угадать класс скрытого объекта и при этом и “не слишком часто” ошибаться.

Пункты 1-3 представляют собой первичную модель источника данных. Если бы мы эту модель знали в виде функций $g(\omega)$ и $\mathbf{x}(\omega)$, то задача распознавания , т.е. построение функции $\hat{g}(x)$, свелась бы к инвертированию первичной модели.

Но у наблюдателя нет модели. При этом доступная наблюдателю информация о функциях $g(\omega)$ и $\mathbf{x}(\omega)$, составляющих вместе с множествами Ω, M и X первичную модель источника данных, ограничивается результатами измерений над конечным числом объектов ω_j, $j = 1,\ldots,N$, составляющих обучающую совокупность. В зависимости от того, какие измерения могут быть произведены на объектах обучающей совокупности, различают задачи обучения распознаванию образов с учителем, без учителя, а также промежуточный вариант.

Задача обучения с учителем предполагает, что каждый объект ω_j в обучающей совокупности представлен номером своего класса $g_j = g(\omega_j)$ и образом в пространстве наблюдений $\mathbf{x}_j = \mathbf{x}(\omega_j)$, то есть обучающая совокупность в целом есть конечное множество пар $(g_j, \mathbf{x}_j)$, $j = 1,\ldots,N$. Такую задачу называют также задачей обучения по классифицированной обучающей совокупности.

В случае задачи обучения без учителя в обучающей совокупности отсутствуют данные о принадлежности объектов к классам, а обучающая совокупность в целом представляет собой просто конечное множество образов объектов в пространстве наблюдений $x_j, j = 1,\ldots,N$. При таком понимании задачи обучения говорят об обучении по неклассифицированной обучающей совокупности.

Если в обучающей совокупности принадлежность объектов к классам известна для части объектов и неизвестна для остальных, то обучающую совокупность называют частично классифицированной.

В принципе, пространство наблюдений X может иметь любую природу, но, как правило, под наблюдением $\mathbf{x}(\omega)$ понимают вектор с некоторым фиксированным числом компонент $\mathbf{x} = (x_1,\ldots,x_n)^T$, которые называются признаками объекта. Обычно полагают, что признаки принимают действительные $x_i \in \mathsf{R}$ либо дискретные значения, в последнем случае обычно $x_i \in \{0,1\}$. Мы остановимся на первом варианте, полагая в дальнейшем, что пространство наблюдений X является *n*-мерным евклидовым пространством R^n, или пространством признаков, или признаковым пространством.

Качество решающего правила распознавания содержательно интерпретируется в простейшем случае как “частота” правильных решений о классе объекта, но легко придумать ситуацию, когда не все правильные ответы равносильны друг другу и поэтому обычно вводят понятие функции потерь

$$\lambda(g,\hat{g}) \geq 0 \; \textit{если} \; g \neq \hat{g}$$
$$\lambda(g,\hat{g}) = 0 \; \textit{если} \; g = \hat{g}$$

Понятие частоты ошибки обычно формализуют, рассматривая множество Ω как вероятностное пространство $\langle \Omega, \mathsf{F}, P \rangle$, наделяя его некоторой σ-алгеброй подмножеств F и вероятностной мерой P. В этом случае для функции $\hat{g}(x)$ также необходимо потребовать измеримость.

Пусть $\hat{g}(x)$ - конкретное решающее правило. Рассмотрим в множестве Ω подмножество Ω^-, для которых наше решение о классе объекта не совпадает с истинным

$$\Omega^- = \{\omega \in \Omega | \; g(\omega) \neq \hat{g}[x(\omega)]\}.$$

Потребуем, чтобы функции $\hat{g}(x)$, $g(x)$ и $x(\omega)$ были таким, чтобы подмножество Ω^- было измеримо в пространстве F. Тогда для него существует вероятностная мера $P(\Omega^-) \geq 0$, которая характеризует частоту неправильного решения, т.е. оказывается определенным функционал на множестве всех решающих правил

$$J[g(\cdot)] = P\{\Omega^-[g(\cdot)]\}.$$

Естественно выбирать решающее правило таким образом, чтобы $J[\,]$ был минимальным.

Тогда решающее правило следует выбирать из условия минимума вероятности ошибки $P(\hat{g}[\mathbf{x}(\omega)] \neq g(\omega))$, где ω понимается как случайная переменная, принимающая значения из множества Ω.

Остается дать количественное выражение понятию качества решающего правила распознавания. Пусть зафиксирована функция потерь

$$\lambda(g, \hat{g}) \geq 0 \; \textit{если} \; g \neq \hat{g}$$
$$\lambda(g, \hat{g}) = 0 \; \textit{если} \; g = \hat{g},$$

зафиксируем также решающее правило, тогда для каждого значения объекта $\omega \in \Omega$ определено значение потерь при распознавании его класса с помощью решающего правила.

$$\omega \in \Omega : [\lambda\{g(\omega), \hat{g}[x(\omega)]\}]$$

Поскольку множество Ω наделено структурой вероятностного пространства то потери от неверного распознавания представляют собой случайную величину. Ее математическое ожидание будем понимать как степень некачественности данного решающего правила. Математическое ожидание потерь при распознавании принято называть средним риском ошибки распознавания:

$$J[\hat{g}(\cdot)] = M[\lambda\{g(\omega), \hat{g}[x(\omega)]\}]$$

1.2.2 Структура оптимального решающего правила

Предположим, что модель источника данных известна. Тогда для каждого класса объектов $\omega \in \Omega^k$, $g(\omega) = k$ в пространстве наблюдений X определено некоторое распределение вероятности. Допустим, что это распределение выразимо в виде

плотности вероятности $\varphi^k(x)$, $x \in \mathsf{X}$. Пусть для каждого класса определена вероятность появления объекта этого класса

$$q^k = P[g(\omega) = k], \quad \sum_{k=1}^{m} q^k = 1,\ k = 1,\ldots,m$$

Т.е., в сущности, наблюдатель имеет дело с двухкомпонентной случайной величиной (g, x) $g \in \mathsf{M} = \{1..m\}$, которая принимает значения из декартового произведения $\mathsf{M} \times \mathsf{X}$, где множество M - дискретна. Вероятностная модель источника данных определяет некоторое совместное распределение вероятности на этом множестве для двухкомпонентной случайной величины (g, x). Известно, что полное распределение вероятности для двухкомпонентного случайного объекта можно выразить двумя способами:

$$\psi(k, x) = P[g(\omega) = k]\,\varphi[x(\omega)|g(\omega) = k] = g^k \varphi^k(x) \qquad (1.2.1)$$

и

$$\psi(k, x) = \varphi[x(\omega)]\,P[g(\omega) = k|x(\omega) = x] = f(x)\,\pi^k(x) \qquad (1.2.2)$$

, где $f(x)$ - плотность полного распределения вероятности в пространстве наблюдений X, $\pi^k(x)$ - условная вероятность того, что объект принадлежит к классу k, если известно, что он отобразился в точку x пространства наблюдений X.

Таким образом, с точки зрения наблюдателя полная вероятностная модель источника данных может быть выражена двумя способами.

I. q^k, $k = 1..m$, - априорная вероятность класса,

$$\sum_{k=1}^{m} q^k = 1;$$

$\varphi^k(x)$ -условные плотности вероятности распределений в пространстве наблюдений X для некоторого класса k.

$$\varphi^k(x) \geq 0,\ \int_{\mathsf{X}} \varphi^k(x)dx = 1,\ k = 1..m$$

II. $f(x)$, $x \in \mathsf{X}$ - полная плотность распределения в пространстве наблюдений

$$\mathsf{X}\ \ f(x) \geq 0, \int_{\mathsf{X}} f(x)dx = 1;$$

$\pi^k(x)$, $k = 1..m$ - апостериорная вероятность класса объекта.

$$\sum_{k=1}^{m} \pi^k(x) = 1,\ 0 \leq \pi^k(x) \leq 1 .$$

Функцию $\pi^k(x)$ принято называть функцией степени достоверности. Почему это так будет понятно из дальнейшего изложения.

Предположим, что вероятностная модель источника данных известна полностью. выясним какое решающее правило является оптимальным, т.е. доставляет минимальное значение функционалу среднего риска

$$J[\hat{g}(\cdot)] = M\{\lambda[g,\hat{g}(x)]\} = \int_{M\times X} \lambda[g,\hat{g}(x)]\psi(g,x)\mu(d(g,x)) = \sum_{k=1}^{m}\int_{X} \lambda[k,\hat{g}(x)]\psi(k,x)dx \quad (1.2.3)$$

Подставив в выражение для среднего риска определение для полной вероятности (1.2.2), получим

$$J[\hat{g}(\cdot)] = \int_{X}\left\{\sum_{k=1}^{m}\lambda[k,\hat{g}(x)]\pi^{k}(x)\right\} f(x)dx \quad (1.2.4).$$

Выражение в фигурных скобках естественно называть условным риском при условии, что объект отобразился в точку x пространства наблюдений X. Внутри скобок $\hat{g}(x)$ есть константа при фиксированном x, т.е. это тот класс, в пользу которого решающее правило $\hat{g}(x)$ принимает решение.

Общий средний риск $J[\hat{g}(\cdot)]$ есть ни что иное, интеграл или сумма условных рисков в каждой точке пространства наблюдений.

Кроме того, мы договорились не налагать никаких ограничений на вид решающего правила $\hat{g}(x)$. Это означает, что мы каждой точке пространства наблюдений в праве поставить в соответствие тот индекс класса, который будем приписывать, когда появится объект этого класса. Тогда совершенно очевидно, что для минимизации среднего риска надо выбрать такое решающе правило $\hat{g}(x)$, которое каждой точке пространства наблюдений $x \in X$ ставит в соответствие индекс класса $j = \hat{g}(x)$, доставляющее минимальное значение условному риску ошибки в данной точке, т.е. выражению в фигурных скобках.

Отсюда

$$\hat{g}(x) = \arg\min_{j=1..m}\sum_{k=1}^{m}\lambda[k,j]\pi^{k}(x) \quad (1.2.5).$$

Из этой общей структуры решающего правила распознавания видно, что оно опирается не на всю модель источника данных в целом, а лишь на часть этой модели, а именно на векторную функцию $\pi^{k}(x)$, которая называет функцией степени достоверности.

Полное же распределение вероятности в пространстве наблюдений $f(x)$ оказалось никак не влияющим на структуру оптимального решающего правила.

1.2.3. Восстановление плотностей распределений классов в пространстве признаков.

В выражении (1.2.5) для оптимального решающего правила присутствует функция $\pi^{k}(x)$, выражающая зависимость вероятности класса от точки пространства наблюдений. Эти вероятности легко получить, если известны априорные вероятности классов и условные плотности распределения для каждого класса [2,10,45]

$$\pi^{k}(x) = \frac{\varphi^{k}(x)g^{k}}{f(x)}$$

Обратим внимание, что после подстановки в (1.2.5)общая плотность распределения играет роль лишь общего коэффициента и не влиет на вид решающего правила. Т.о. надо по обучающей выборке оценить априорные вероятности классов q^k и условные плотности распределений φ^k, $k = 1..m$, $x \in \mathsf{X}$.

Сделаем основной акцент на случай того, что $\mathsf{X} = R^n$ и $x = \mathbf{x} = (x_1,...,x_n)^T$

1. q^k -априорные вероятности классов. Оценкой максимального правдоподобия для них являются частоты

$$\hat{q}_N^k = \frac{N^k}{N},$$

где N – объем выборки, N^k - число объектов k-го класса в выборке.

Возникает вопрос - насколько хороша такая оценка. Предположим, что из генеральной совокупности выбираются объекты независимо, случайно, с возвратом. в результате такого бесконечного эксперимента мы получим последовательность $\hat{q}_N^k$. Эта последовательность может либо иметь предел, либо не иметь. Факт наличия у последовательность предела будем рассматривать как случайное событие. Можно доказать, что вероятность этого события равна 1. в математической статистике в этом случае говорят, что событие происходит почти наверное.

В свою очередь, если числовая последовательность $\hat{q}_N^k$ сходится, то она может сходится либо к истинной вероятности класса q^k, либо к какому-то другому числу. Можно доказать, что пределом последовательности $\hat{q}_N^k$ с вероятностью 1 будет q^k. Сходимость почти наверное является наиболее сильным видом сходимости.

2. $\varphi^k(x)$ -условные плотности вероятности распределений в пространстве наблюдений. Ситуация с оцениванием $\varphi^k(x)$ бесконечно сложнее.

Среди методов оценивания плотности распределения различают параметрические [41, 42] и непараметрические.

Обратим внимание, что восстановить плотность распределения – это восстановить функцию. Ограничимся рассмотрением пространства наблюдений в виде $\mathsf{X} = R^n$ $x = \mathbf{x} = (x_1,...,x_n)^T$, $x_1,...x_2$ - признаки объекта распознавания, R^n - признаковое пространство.

Тогда плотность распределения k-го класса в признаковом пространстве есть плотность распределения некоторой n-мерной случайной величины.

Плотность распределения некоторой есть действительная функция векторного аргумента

$$\varphi^k(\mathbf{x}),\ \mathbf{x} \in R^n\ \varphi^k(x) \geq 0,\ \int_{\mathsf{X}} \varphi^k(x)dx = 1.$$

Непараметрическое восстановление плотности.

Пусть учитель предъявил *N* объектов *k*-го класса $\mathbf{x_1}\ldots\mathbf{x_N}$. На рис.1.2..1 показана иллюстрация двумерного случая. Рассмотрим точку $\mathbf{x}\in R^n$. Представляется естественным присвоить этой точке тем большее значение плотности распределения, чем больше объектов обучающей выборки попали в достаточно близкую окрестность этой точки. Один из наиболее распространенных способов реализации такой идеи заключается в выборе так называемой потенциальной функции $\eta(\mathbf{x}',\mathbf{x}'')$. Эта функция принципиально двух аргументов $\mathbf{x}',\mathbf{x}''\in R^n$. Она максимальна, когда $\mathbf{x}'=\mathbf{x}''$, и равна нулю, когда $\|\mathbf{x}'-\mathbf{x}''\|\to\infty$. Пример такой функции для одномерного вектора признаков показана на рис 1.2.2.

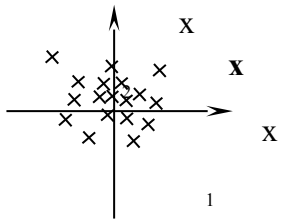


Рисунок 1.2.1

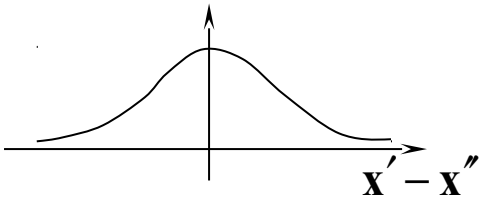


Рисунок 1.2.2. Потенциальная функция $\eta(\mathbf{x}',\mathbf{x}'')$ для одномерного случая

Потребуем от потенциальной функции, чтобы она была неотрицательна

$$\eta(\mathbf{x}',\mathbf{x}'')\ge 0,$$

и, чтобы

$$\int_{R^n}\eta(\mathbf{x}',\mathbf{x}'')d\mathbf{x}'=1,\ \mathbf{x}''\in\mathbf{R^n}.$$

Отсюда следует, что потенциальная функция неизбежно стремится к нулю при $\|\mathbf{x}'-\mathbf{x}''\|\to\infty$. Кроме того потребуем, чтобы потенциальная функция монотонно убывала при увеличении нормы разности $\|\mathbf{x}'-\mathbf{x}''\|\to\infty$.

В соответствии с вышеизложенным будем оценивать плотность распределения следующим образом:

$$\hat{\varphi}^k(\mathbf{x})=\frac{1}{N}\sum_{j=1}^{N}\eta(\mathbf{x},\mathbf{x}_j).$$

Особенности данной оценки:

1. Чтобы вычислить значение функции в точке $\mathbf{X}$ необходимо хранить в памяти всю обучающую выборку
2. Оценка плотности существенным образом зависит от выбора потенциальной функции. Если потенциальную функцию выбрать «острой», то на малых выборках оценка плотности будет неровной, причем эта неровность будет существенно зависеть от данной выборки. Т.е. если $\varphi^*(\mathbf{x})$ - истинная плотность, то ее оценки

$\hat{\varphi}(\mathbf{x})$ при острой форме функции $\eta(\mathbf{x}', \mathbf{x}'')$ будут отличаться очень большой вариабельностью. В тоже время, если выбрать функцию $\eta(\mathbf{x}', \mathbf{x}'')$ очень «размытой», то оценка $\hat{\varphi}(\mathbf{x})$ также окажется очень отличной от истинной. Ясно, что здесь есть золотая середина, и она будет зависеть от размера выборки. Чем меньше выборка, тем «размытее» должна быть потенциальная функция. Чем больше выборка, чем лучше она отражает детали истинной плотности, тем более островершинной должна быть потенциальная функция для отражения этих деталей.

Вопрос о том, как выбирать вид потенциальной функций и степень ее островершинности в зависимости от размера выборки и расположения точек в ней, называется теорией непараметрического оценивания [5,11,45].

Параметрические оценки.

Идея параметрического оценивания заключается в том, что оценка $\hat{\varphi}(\mathbf{x})$ некоторой плотности $\varphi^*(\mathbf{x})$, представленной в виде выборки, ищется в пределах некоторого семейства плотностей распределения, задаваемых некоторой общей формулой $\varphi(\mathbf{x}; \mathbf{a})$, содержащей свободно изменяемый в некоторой области параметр **a**. Естественно, что такое параметрическое семейство должно удовлетворять условиям:

$$\varphi(\mathbf{x}; \mathbf{a}) \geq 0,\ \mathbf{x} \in \mathbf{R}^n,\ \mathbf{a} \in \mathsf{A},$$

$$\int_X \varphi(\mathbf{x}; \mathbf{a}) dx = 1,\ \mathbf{a} \in \mathsf{A}$$

Тогда указать конкретную плотность распределения значит указать конкретное значение параметра **a**. Наиболее распространенным параметрическим семейством является семейство нормальных распределений

$$\varphi(\mathbf{x};\ \mathbf{x_0}, \mathbf{K}) = \frac{1}{|\mathbf{K}|^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left[(\mathbf{x} - \mathbf{x_0})\mathbf{K}^{-1}(\mathbf{x} - \mathbf{x_0})\right],$$

где $\mathbf{a}(\mathbf{x_0}, \mathbf{K})$ - вектор параметров, $\mathbf{x_0}$-вектор математического ожидания, **K** – ковариационная матрица. Но это семейство принципиально унимодально. Существует способ формирования более сложных параметрических семейств. Этот способ заключается в формировании так называемых смесей распределений. Особенно часто используют смеси нормальных распределений.

Договоримся обозначать плотность нормального распределения как $\mathsf{N}(\mathbf{x};\ \mathbf{x_0}, \mathbf{K})$. пусть в построении смеси участвуют *к* нормальных распределений

$$\mathsf{N}^i\left(\mathbf{x};\ \mathbf{x}_0^i, \mathbf{K}^i\right),\ i = 1..k.$$

Поставим в соответствие каждому нормальному распределению его вес $p_i, i = 1..k, \sum p_i = 1, p_i \geq 0$. Общую плотность распределения смеси выберем в виде линейной комбинации

$$\varphi(\mathbf{x};\mathbf{a}) = \sum_{i=1}^{k} p_i \mathrm{N}^i\left(\mathbf{x};\, \mathbf{x}_0^i,\, \mathbf{K}^i\right)$$

Вектор параметров такого распределения

$$\mathbf{a} = \left(\mathbf{x}_\mathbf{0}^i, \mathbf{K}^i, p^i, i = 1..k\right)$$

Смесь распределений имеет очень простую вероятностную интерпретацию. При выборе значения случайной величины **x** сначала разыгрывают номер распределения с вероятностями p_i, а уже потом разыгрывается то распределение, на которое пал случайный выбор.

Проинтерпретируем некоторые особенности оценивания смесей на примере смесей нормальных распределений.

Чем более сложную форму имеет истинная плотность распределения, подлежащая восстановлению, тем большее число элементов смеси понадобится для ее аппроксимации. Однако нетрудно понять, что число элементов в смеси должно быть меньше размера обучающей выборки, чтобы на каждое распределение приходилось по несколько элементов выборки. При увеличении выборки число элементов можно увеличивать.

Рассмотрим один из методов оценивания параметров распределения, называемый методом максимального правдоподобия.

Пусть $\varphi(\mathbf{x};\mathbf{a})$ параметрическое семейство распределений. Предположим, что выборка $\mathbf{x}_j, j = 1..N$ получена в результате n независимых испытаний с одним распределением $\varphi(\mathbf{x};\mathbf{a}^*)$ из этого семейства.

Вся выборка – это вектор из векторов, полученных в каждом эксперименте

$$\mathbf{X} = \left(\mathbf{x}_1,...,\mathbf{x}_N\right)^T.$$

Поскольку компоненты этого вектора независимы по условиям эксперимента, то плотность распределения этого вектора есть произведений плотностей

$$f(\mathbf{X};\mathbf{a}) = \prod_{j=1}^{N} \varphi\left(\mathbf{x}_j;\mathbf{a}\right).$$

При каждом значении параметра **a** в комбинированном пространстве $\mathbf{X} = \left(\mathbf{x}_1,...,\mathbf{x}_N\right)^T$ общая плотность распределения $f(\mathbf{X};\mathbf{a})$ имеет, вообще говоря, свой вид. Выборка же в целом образует одну точку в комбинированном пространстве. Идея оценивания по методу максимального правдоподобия заключается в том, чтобы выбрать параметр **a** т.о., чтобы эта точка приходилась на максимум плотности распределения.

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in A}{\operatorname{argmax}} f(\mathbf{X};\mathbf{a}) .$$

Очевидно что ничего не изменится, если от $f(\mathbf{X};\mathbf{a})$ взять любую монотонную функцию, например логарифм, тогда получим следующее выражения для оценки максимального правдоподобия

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in A}{\operatorname{argmax}} \sum_{i=1}^{N} \log \varphi(\mathbf{x}_j;\mathbf{a})$$

Для оценивания параметра распределения рассмотрим еще один способ, основанный на оценке корня уравнения регрессии.

Пусть в пространстве R^n задано параметрическое семейство $\varphi(\mathbf{x};\mathbf{a})$, $\mathbf{a}\in A$, в рамках которого оценивается истинная плотность распределения $\varphi(\mathbf{x};\mathbf{a}^*)$. Пусть максимальное значение плотности определяется неизвестным наблюдателю значением **x**. Рассмотрим плотность распределения $\varphi(\mathbf{x};\mathbf{a})$ как функцию двух переменных **x** и **a**. Нам будет удобнее рассматривать не саму эту функцию, а ее логарифм. Здесь **x** – случайная величина, имеющая плотность распределения $\varphi(\mathbf{x};\mathbf{a}^*)$, тогда $\log[\varphi(\mathbf{x};\mathbf{a})]$ - случайная функция параметра **a**.

Рассмотрим математическое ожидание этой случайной функции, которая также будет функцией этого варьируемого параметра **a**

$$L(\mathbf{a}) = M_{\mathbf{x}}[\log \varphi(\mathbf{x};\mathbf{a})] = \int_{R^n} \varphi(\mathbf{x};\mathbf{a}^*) \log \varphi(\mathbf{x};\mathbf{a}) d\mathbf{x} .$$

Можно показать, что этот интеграл максимален при $\mathbf{a} = \mathbf{a}^*$.

Обратим внимание, что $\log[\varphi(\mathbf{x};\mathbf{a})]$ есть случайная функция варьируемого параметра **a**. Т.о. мы имеем случайную величину, зависящую от параметра. Допустим, что при любом **a** эта случайная величина имеет математическое ожидание, в нашем случае это $L(\mathbf{a})$. Это условное математическое ожидание принято называть уравнением регрессии. Т.о. мы свели задачу оценивания параметра **a** к задаче оценивания максимума функции регрессии

$$\mathbf{a}^* = \arg\max_{\mathbf{a}\in A} M_{\mathbf{x}}[\log \varphi(\mathbf{x};\mathbf{a})]$$

Если функция регрессии гладкая, то в точке **a** должно выполняться условие:

$$\nabla_{\mathbf{a}} M_{\mathbf{x}}[\log \varphi(\mathbf{x};\mathbf{a})] = 0 \qquad (1.2.6)$$

Если параметрическое семейство $\varphi(\mathbf{x};\mathbf{a})$ удовлетворяет условию регулярности, то операции дифференцирования и взятия математического ожидания можно поменять местами, т.е. справедливо следующее

$$M_{\mathbf{x}}[\nabla_{\mathbf{a}} \log \varphi(\mathbf{x};\mathbf{a})] = 0 \text{ при } \mathbf{a} = \mathbf{a}^*. \qquad (1.2.7)$$

Уравнение такого вида принято называть уравнением регрессии.

Итак для оценивания параметра неизвестного распределения можно воспользоваться либо утверждением (1.2.6), либо утверждением (1.2.7).

Однако в реальной ситуации мы не знаем функцию регрессии, т.е. не знаем условного математического ожидания. Единственно, чем мы обладаем это выборкой $\mathbf{x}_j,\ j=1..N$. Идея оценивания случайной величины заключается в использовании вместо математического ожидания случайной величины его оценки в виде среднего арифметического выборочных значений. Если пользоваться выражением (1.2.5) то получим оценку следующего вида

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}\in\mathsf{A}} \frac{1}{N}\sum_{j=1}^{N} \log\varphi(\mathbf{x}_j;\mathbf{a})$$

Если не обращать внимание на *N*, то это ни что иное как оценка максимального правдоподобия. Если же опираться на утверждение (1.2.5), то получим

$$\frac{1}{N}\sum_{j=1}^{N} \nabla_{\mathbf{a}} \log\varphi(\mathbf{x}_j;\mathbf{a}) = 0 \quad \Rightarrow \hat{\mathbf{a}}_N$$

Уравнение этого типа принять называть уравнениями правдоподобия.

Если выборка бесконечна и у наблюдателя нет запоминающего устройства, способного поместить даже часть этой выборки, используют рекуррентные процедуры оценивания.

Пусть на шаге *j* получена оценка $\hat{\mathbf{a}}_j$, пусть пришло очередное наблюдение $\mathbf{x}_{j+1}$. Новая оценка ищется как некоторая функция

$$\hat{\mathbf{a}}_{j+1} = \eta(\hat{\mathbf{a}}_j, \mathbf{x}_{j+1}).$$

Очень простые оценки для поиска точки максимума функции регрессии дает процедура Киффера-Вольфовица [39].

Аналогичная процедура для оценки корня уравнения регрессии носит название процедуры Робинса-Монро [16,39,51].

Эти процедуры практически эквивалентны друг другу и обеспечивают сходимость почти наверное при очень необременительных предположениях о семействе $\varphi(\mathbf{x};\mathbf{a})$.

1.2.3 Непосредственное восстановление функции степени достоверности.

Напомним, что для поиска оптимального решающего правила нужны были не плотности распределения, а апостериорные вероятности классов в точке наблюдения

$$\pi^k(\mathbf{x}) = P(g(\omega) = k | \mathbf{x}(\omega) = \mathbf{x}),\ \mathbf{x}\in\mathsf{X}$$

Идея восстанавливать плотности появилась из формулы Байеса в которой наличие $f(\mathbf{x})$ мы сочли излишним

$$\pi^k(\mathbf{x}) = \frac{\varphi^k(\mathbf{x})g^k}{f(\mathbf{x})},\ k=1..m$$

Именно этот шаг - удаления из знаменателя общей функции распределения $f(\mathbf{x})$ - является очень неблагоприятным. Нетрудно убедиться, что в большинстве случаев функции апостериорных вероятностей $\pi^k(\mathbf{x})$ много проще, чем исходные функции распределения $\varphi^k(\mathbf{x})$. Даже если функции $\varphi^k(\mathbf{x})$ достаточно вычурны, в формуле Байеса они делятся практически на свою копию $f(\mathbf{x}) = g^1\varphi^1(\mathbf{x}) + g^2\varphi^2(\mathbf{x})$ (случай двух классов *m=2*) и очень интенсивно «выглаживаются» (рис. 1.2.3).

Поэтому всегда целесообразно непосредственно восстанавливать апостериорные вероятности классов, а не исходные плотности распределения.

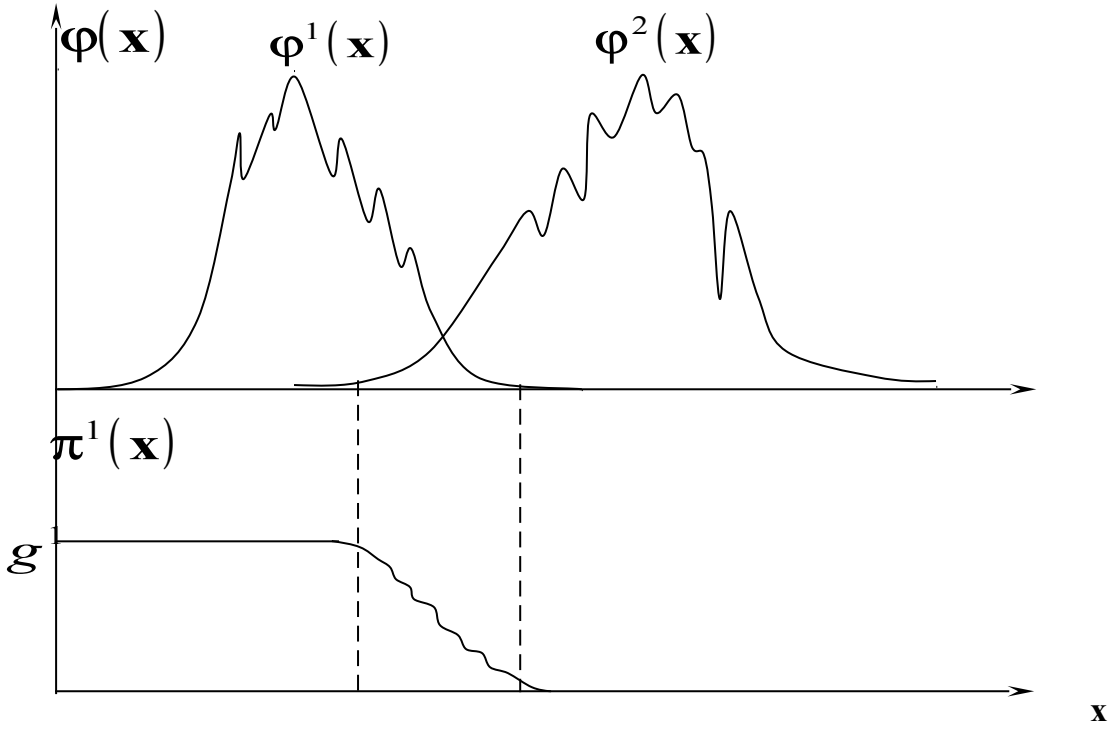


Рисунок 1.2.3 Плотности вероятности классов и апостериорная вероятность одного из классов.

Восстановление функций апостериорных вероятностей классов [17] удобно осуществлять в рамках некоторого параметрического семейства $\pi^k(\mathbf{x};\mathbf{a})$. Такое семейство должно удовлетворять следующим условиям

$$\mathbf{x} \in R^n, \mathbf{a} \in \mathsf{A}$$

$$\pi^k(\mathbf{x};\mathbf{a}) \geq 0 \sum_{k=1}^{m} \pi^k(\mathbf{x};\mathbf{a}) = 1 .$$

В простейшем случае, когда число классов два, достаточно определить параметрическое семейство для вероятности лишь одного класса $0 \leq p(\mathbf{x};\mathbf{a}) \leq 1$ $\mathbf{x} \in R^n, \mathbf{a} \in \mathsf{A}$

$$\pi^1 = p(\mathbf{x};\mathbf{a})$$
$$\pi^2 = 1 - p(\mathbf{x};\mathbf{a})$$

Пусть $p(\mathbf{x};\mathbf{a})$ как раз такое параметрическое семейство. Это семейство необходимо выбирать таким образом, чтобы степень его изменчивости как функции $\mathbf{x}$

не слишком подчинялась параметру **a.** График апостериорной вероятности первого класса удобно представить в пространстве (на плоскости) как совокупность поверхностей равных значений. Надо так ограничить параметрическое семейство функций, чтобы эти поверхности не были слишком сложными. Иногда, после того, как представлена обучающая выборка $(g_j, \mathbf{x}_j),\ \ j = 1,..., N$, и мы потребуем от алгоритма подобрать такое значение параметра **a**, чтобы $p(\mathbf{x};\mathbf{a}) \to 1$ в тех точках, которые помечены индексом первого класса, и, чтобы $p(\mathbf{x};\mathbf{a}) \to 0$ в точках, помеченных индексом второго класса, то у алгоритма появляется соблазнительная возможность сделать это буквально.

Ниже мы это богатство назовем емкостью класса решающих правил.

Если эта емкость слишком велика, то функция $p(\mathbf{x};\mathbf{a})$, идеально аппроксимировав обучающую выборку, будет плохо согласовываться с другими контрольными выборками. При слишком большой емкости класса решающих правил $p(\mathbf{x};\mathbf{a})$ будет аппроксимировать индивидуальные особенности обучающей выборки, ошибочно принимая их за истинные различия между классами.

Именно этот аспект есть центральный момент обучения распознаванию образов.

Итак, пусть параметрическое семейство $p(\mathbf{x};\mathbf{a})$ выбрано. Пусть $(g_j, \mathbf{x}_j),\ \ j = 1,..., N$ обучающая выборка $g_j = 1..m$. Рассмотрим один из путей поиска параметра **a**, при котором $p(\mathbf{x};\mathbf{a})$ наилучшим образом может быть согласована с выборкой.

Рассмотрим случайную величину $z(\omega) = \begin{cases} 1\ g(\omega) = 1 \\ 0\ g(\omega) = 2 \end{cases}$

Если выбранное параметрическое семейство таково что существует значение параметра **a,** при котором в точности воспроизводится апостериорная вероятность первого класса в точке **x**, то в каждой точке **x** условное математическое ожидание случайной величины *z* совпадает с апостериорной вероятностью класса в этой точке $P(g(\omega) = 1 | \mathbf{x}(\omega) = \mathbf{x}),\ \mathbf{x} \in \mathsf{X}$, т.е. $M[z(\omega) - p(\mathbf{x};\mathbf{a}) | \mathbf{x}(\omega) = \mathbf{x}] = 0$ для всех $\mathbf{x} \in \mathsf{X}$.

В свою очередь, рассматривая $\mathbf{x} = \mathbf{x}(\omega)$ как случайный вектор, получим

$$M\{z(\omega) - p[\mathbf{x}(\omega);\mathbf{a}]\} = 0 \qquad (1.2.8).$$

Именно это условие и будем использовать для поиска параметра **a**.

Рассмотрим неотрицательную случайную величину $(z(\omega) - p[\mathbf{x};\mathbf{a}])^2$. Известно, что математическое ожидание случайной величины $z(\omega)$ минимизирует квадрат разности этой случайной величины и наперед заданной неслучайной величины, поэтому поиск корня будем проводить из условия

$$M\left\{\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)^2\right\}\to\min_{\mathbf{a}\in A} \qquad (1.2.9)$$

Если параметрическое семейство $p(\mathbf{x};\mathbf{a})$ дифференцируемо по **a**, то условием минимума будет

$$\nabla_{\mathbf{a}} M\left\{\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)^2\right\}=0$$

Если семейство $p(\mathbf{x};\mathbf{a})$ удовлетворяет условию регулярности, то знаки взятия производной и математического ожидания можно поменять местами

$$M\left\{\nabla_{\mathbf{a}}\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)^2\right\}=0$$

$$\nabla_{\mathbf{a}}\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)^2=-2\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)\cdot\nabla_{\mathbf{a}}p[\mathbf{x};\mathbf{a}].$$

Таким образом, приходим к следующему условию оценивания

$$M\left\{\left(z(\omega)-p[\mathbf{x};\mathbf{a}]\right)\cdot\nabla_{\mathbf{a}}p[\mathbf{x};\mathbf{a}]\right\}=0 \qquad (1.2.10)$$

До сих пор мы исходили из того, что условие (3.3) при некотором **a** выполняется точно. Но в действительности $p(\mathbf{x};\mathbf{a})$ не воспроизводит в точности функцию апостериорной вероятности классов. Однако ясно, что условия (3.4) и (3.5) остаются вполне разумными критериями для выбора подходящего параметра **a** в данном случае.

Однако использовать эти условия напрямую нельзя, так как нам не известны математические ожидания. Но вместо них у нас есть обучающая выборка $(g_j,\mathbf{x}_j),\quad j=1,...,N$. Тогда параметр **a** следует выбирать исходя из следующих условий

$$\frac{1}{N}\sum_{j=1}^{N}\left\{I[g_j=1]-p(\mathbf{x}_j;\mathbf{a})\right\}^2\to\min$$

$$\frac{1}{N}\sum_{j=1}^{N}\left\{I[g_j=1]-p(\mathbf{x}_j;\mathbf{a})\right\}\nabla_{\mathbf{a}}p(\mathbf{x}_j;\mathbf{a})=0.$$

Эти условия и являются рабочими для построения алгоритма обучения.

Здесь

$$I[g_j=1]=\begin{cases}1, g_j=1\\ 0, g_j=0\end{cases}$$

индикаторная функция.

1.2.4. Прямое восстановление решающего правила распознавания.

В тех случаях, когда невозможно принять какие-либо простые предположения о вероятностной модели данных, а следовательно и о априорной вероятности классов $\pi^k(x)$, $k=1..m$, то неуместно ставить задачу обучения как задачу восстановления этой модели. Но ведь модель нужна только для того, чтобы затем, опираясь на нее, построить решающее правило распознавания $\hat{g}(x)$, поэтому наиболее

распространенный подход к распознаванию образов заключается в прямом восстановлении решающего правила распознавания.

Однако, говоря о восстановлении модели источника данных, мы всегда ограничивали сложность этой модели. Естественно, из простой модели получались простые решающие правила. Теперь между обучающей выборкой и решающим правилом промежуточного звена нет, и необходимо ограничивать непосредственно сложность решающего правила. Обычно это означает необходимость использования некоторого достаточно простого параметрического семейства решающих правил.

В случае двух классов $m = 2$ такие параметрические семейства строятся на основании некоторой дискриминантной функции $d(\mathbf{x};\mathbf{c})$, принимающей в каждой точке признакового пространства X действительные значения, $\mathbf{c}$ – некоторый векторный параметр $\mathbf{c} \in \mathsf{C}$.

$$\hat{g}(\mathbf{x}) = \begin{cases} 1, d(\mathbf{x};\mathbf{c}) > 0 \\ 2, d(\mathbf{x};\mathbf{c}) \le 0 \end{cases}$$

В простейшем случае дисктриминантная функция может быть взята линейной

$$d(\mathbf{x};\mathbf{a},b) = \mathbf{a}^T\mathbf{x} + b,\ \mathbf{a} \in R^n,\ b \in R.$$

Заметим, что уравнение $\mathbf{a}^T\mathbf{x} + b = 0$ при $b = 0$ задает в пространстве R^n множество точек $\mathbf{x} \in R^n$, удовлетворяющих этому уравнению и это уравнение принять называть гиперплоскостью.

Вектор $\mathbf{a}$ называют в этом случае направляющим вектором гиперплоскости. Очевидно, что гиперплоскость не изменится если направляющий вектор умножить на любой коэффициент отличный от нуля. Обратим внимание, что гиперплоскость является подпространством пространства R^n и его размерность на единицу меньше размерности исходного пространства.

Существенным свойством подпространства вообще и гиперплоскости в частности является то, что ему всегда принадлежит нулевая точка.

Рассмотрим ситуацию, когда $b \neq 0$. В этом случае точка $\mathbf{x} = \mathbf{0}$ уже не будет принадлежать множеству точек пространства, удовлетворяющего уравнению $\mathbf{a}^T\mathbf{x} + b = 0$. Следовательно, множество таких точек не является подпространством и не является, таким образом, гиперплоскостью. Множества такого типа принято называть аффинными многообразиями (рис. 3.4). Нетрудно убедиться, что аффинное многообразие, а следовательно и

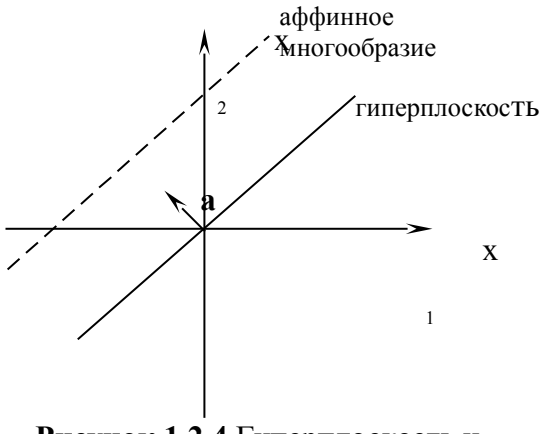


Рисунок 1.2.4 Гиперплоскость и аффинное многообразие

гиперплоскость разбивают все пространство на два непересекающихся подмножества $\mathbf{a}^T\mathbf{x}+b>0$ и $\mathbf{a}^T\mathbf{x}+b\le 0$. В дальнейшем мы буден называть множество $\mathbf{a}^T\mathbf{x}+b=0$ гиперплоскостью и при $b\neq 0$.

В данных терминах линейная дискриминантная функция определяет некоторую гиперплоскость, разбивающую пространство на области принятия решений первого и второго класса соответственно

$$d(\mathbf{x};\mathbf{a},b)=\mathbf{a}^T\mathbf{x}+b\begin{cases}>0\ k=1\\ \le 0\ k=2\end{cases}.$$

Однако, во многих случаях возникает надобность использования в качестве границ между классами более сложные поверхности, чем гиперплоскости. этого добиваются, используя в качестве дискриминантной функции полиномиальную функцию вида

$$d(\mathbf{x};\mathbf{a},n)=\sum_{i=1}^{n}a_i y_i(\mathbf{x}),$$

где $y_0(\mathbf{x}), y_1(\mathbf{x})\ldots y_n(\mathbf{x})$ - подходящее семейство базисных функций.

Нетрудно убедиться, что линейная дискриминантная функция есть частный случай полиномиальной

$$\begin{aligned}y_0(\mathbf{x})&\equiv 1, i=0\\ y_i(\mathbf{x})&=x_i,\ i=1..n\end{aligned}.$$

В теории обучения распознаванию образов достаточно ограничиться рассмотрением линейных дискриминантных функций, поскольку значения базисных функций всегда можно рассматривать как новые признаки объекта, обеспечивающие переход в новое признаковое пространство, в котором дискриминантные функции уже являются линейными.

В силу этого обстоятельства пространство, образованное значениями базисных функций называется спрямляющим базисным пространством.

Поскольку мы задали класс дискриминантных функций параметрически $d(\mathbf{x};\mathbf{a})$, то и класс решающих правил распознавания также оказывается параметрически задан. Ранее мы определяли функционал, численно равный среднему риску ошибки распознавания $J[\hat{g}(\cdot)]$. Поскольку решающее правило зависит от параметра, тои средний риск зависит от параметра $J[\mathbf{a}]$.

Задачу обучения распознаванию образов естественно поставить как задачу определения такого параметра дискриминантной функции, при котором средний риск минимален

$$J[\hat{g}(\cdot)]=M[\lambda\{g(\omega),\hat{g}[\mathbf{x}(\omega);\mathbf{a}]\}]\to\min_{\mathbf{a}\in A}\qquad(1.2.11).$$

Как и всякую функция средний риск можно минимизировать, анализирую на каждом шаге величину градиента $\nabla_{\mathbf{a}} J[\mathbf{a}]$. необходимым условием минимума среднего риска является равенство нулю градиента

$$\nabla_{\mathbf{a}} M[\lambda\{g(\omega), \hat{g}[\mathbf{x}(\omega); \mathbf{a}]\}] = 0 \qquad (1.2.12).$$

Однако для функции потерь $\lambda[g, \hat{g}]$, зависящей только от истинного и экспериментального значения класса объекта перестановка местами операций взятия градиента и математического ожидания

$$M[\nabla_{\mathbf{a}} \lambda\{g(\omega), \hat{g}[\mathbf{x}(\omega); \mathbf{a}]\}] = 0 \qquad (1.2.13)$$

неправомерна. Это происходит в силу того скачкообразной зависимости решающего правила $\hat{g}[\mathbf{x}(\omega); \mathbf{a}]$ от параметра **a** при любом $\mathbf{x} \in \mathsf{X}$.

для преодоления этой трудности переходят к функциям потерь вида $\lambda[\mathbf{x}, \mathbf{a}, g]$, штрафуя не просто факт принадлежности объекта к неверному классу, но и слишком близкой приближение вектора **x** к разделяющей гиперплоскости в области своего класса и тем более расстояния, на которое вектор вторгается в чужую область.

В реальных условиях, так как вероятностные характеристики источника данных нам неизвестны, математические ожидания в формулах (3.6), (3.7), (3.8) не могут быть вычислены, и не может быть вычислено значение среднего риска. Тем не менее, нашей целью является приближение к оптимальному значению

$$\mathbf{a}^* = \arg\min_{\mathbf{a} \in \mathsf{A}} J[\mathbf{a}].$$

Наблюдателю доступна лишь обучающая выборка $(g_j, \mathbf{x}_j), \quad j = 1, \ldots, N$. Компромисс заключается в замене математического ожидания средним арифметическим

$$\hat{J}_N(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^{N} \lambda\left(\hat{g}[\mathbf{x}_j; \mathbf{a}], g_j\right).$$

Такую оценку называют эмпирическим риском, т.е. риском, измеренным на выборке.

Минимизируя эмпирический риск, мы фактически заменим оптимальное значение

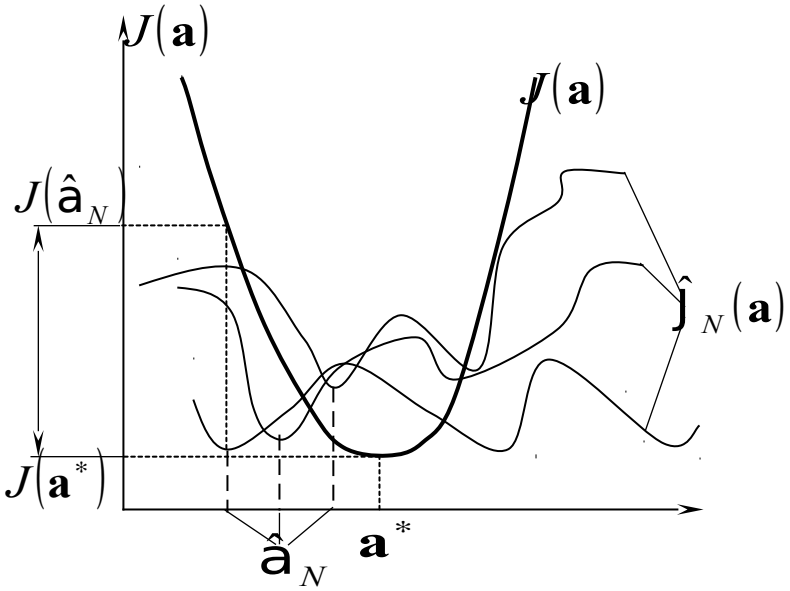


Рисунок 1.2.5 Средний и эмпирический риски ошибки распознавания

параметра $\mathbf{a}^*$ его оценкой $\hat{\mathbf{a}}_N = \arg\min_{\mathbf{a}\in A} \hat{J}_N[\mathbf{a}]$. Очевидно, что эмпирический риск является случайной функцией $\hat{J}_N(\mathbf{a})$ параметра $\mathbf{a}$, определяемой размером N и конкретными значениями ее элементов $(g_j, \mathbf{x}_j)$, $j = 1,..., N$, где $(g_j, \mathbf{x}_j)$ представляет собой двухкомпонентную случайную величину. Как следствие случайным оказывается и значение параметра $\hat{\mathbf{a}}_N = \arg\min_{\mathbf{a}\in A} \hat{J}_N(\mathbf{a})$, а также и значение среднего риска $J(\hat{\mathbf{a}}_N)$ (рис. 3.5). Для того, чтобы контролировать качество наблюдения необходимо контролировать величину случайного отклонения $|J(\hat{\mathbf{a}}_N) - J(\mathbf{a}^*)|$.

Теоретически осуществить такой контроль методами статистики очень трудно. Этот вопрос относится к так называемой статистике экстремальных значений [10,19]. Мы подменим вопрос об отклонении минимума случайной реализации от минимума ее математического ожидания более общим вопросом об отклонении реализации случайной функции от ее математического ожидания.

Говорят, что случайная функция $\hat{J}_N(\mathbf{a})$, зависящая от выборки размера *N,* равномерно сходится к некоторой неслучайной функции , если выполняется следующие условие

$$\lim_{N\to\infty} P\left(\left|\hat{J}_N(\mathbf{a}) - J(\mathbf{a})\right| > \varepsilon \text{ хотя бы для одного } \mathbf{a}\right) = 0 \text{ для всех } \varepsilon > 0.$$

Справедливо утверждение:

если для всех $\mathbf{a}$ $|\hat{J}_N(\mathbf{a}) - J(\mathbf{a})| \le \varepsilon$, то $|J(\hat{\mathbf{a}}_N) - J(\mathbf{a}^*)| \le \varepsilon$, где $\hat{\mathbf{a}}_N = \arg\min_{\mathbf{a}\in A} \hat{J}_N(\mathbf{a})$, а $\mathbf{a}^* = \arg\min_{\mathbf{a}\in A} J(\mathbf{a})$ (рис. 1.2.6).

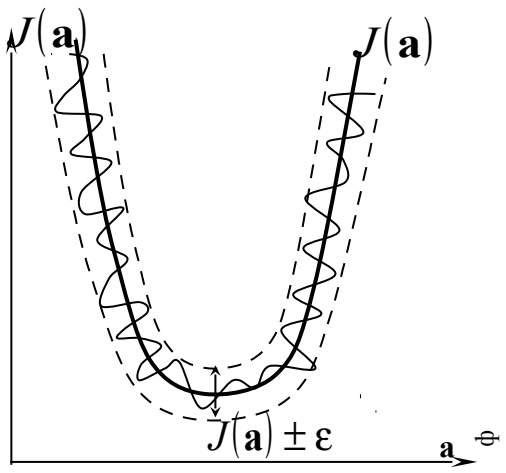


Рисунок 1.2.6 Иллюстрация сходимости случайной функции к ее математическому ожиданию

Т.о. из равномерной сходимости случайной функции $\hat{J}_N(\mathbf{a})$ к ее математическому ожиданию $J(\mathbf{a})$ непосредственно следует, что $\lim_{N\to\infty} P\left(\left|J(\hat{\mathbf{a}}_N) - J(\mathbf{a}^*)\right| > \varepsilon\right) = 0$ для всех ε.

Иначе говоря, при $N \to \infty$ средний риск, обеспечиваемый эмпирически оптимальным решающим правилом неограниченно приближается по вероятности к минимально возможному среднему риску в данном семействе решающих правил.

Мы до сих пор вели разговор о среднем риске ошибки, основанном на произвольной функции потерь $\lambda[g, \hat{g}]$. Однако функция потерь только тогда соответствует своему назначению, когда она удовлетворяет следующим условиям

1) $\lambda[g, \hat{g}] = 0 \; g = \hat{g}$,
2) $0 \le \lambda[g, \hat{g}] \le \lambda_{\max} < \infty \; g \ne \hat{g}$.

В простейшем случае потери для любых видов ошибки распознавания могут быть назначены одинаковыми

$$\lambda[g, \hat{g}] = 0 \; g = \hat{g},$$
$$\lambda[g, \hat{g}] = \lambda_{\max} \; g \ne \hat{g}.$$

Тогда нет никакого необходимости брать $\lambda_{\max}$ отличной от единицы. В этом случае матрица потерь будет называться антидиагональной

$$\lambda[g, \hat{g}] = \begin{matrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{matrix}.$$

Интуитивно понятно, что если равномерная сходимость эмпирического риска к среднему риску имеет место для антидиагональной матрицы потерь, то она имеет место и при более щадящей функции потерь.

Если матрица потерь антидиагональна, то средний риск ошибки распознавания есть ни что иное как вероятность оцененного класса с истинным

$$P(\hat{g}[\mathbf{x}(\omega)] \ne g[\omega]) = J(\mathbf{a})$$

с точностью до $\lambda_{\max}$.

Однако мы на обучающей выборке можем вычислять лишь частоту ошибки распознавания в пределах обучающей выборки

$$\frac{1}{N}\sum_{j=1}^{N} I(\hat{g}[\mathbf{x}_j; \mathbf{a}] \ne g_j).$$

Будем говорить, что частота появления события, зависящего от параметра, сходится по вероятности равномерно по всем значениям параметра к вероятности этого события, если выполняется условие

$$\lim_{N\to\infty} P\left\{\frac{1}{N}\sum_{j=1}^{N} I\left(\hat{g}[\mathbf{x}_j;\mathbf{a}] \neq g_j\right) - P\left(\hat{g}[\mathbf{x}(\omega)] \neq g[\omega]\right) > \varepsilon \text{ хотя бы для одного } \mathbf{a}\in \mathsf{A}\right\} = 0, \text{для } \forall\varepsilon > 0$$

Интуитивно ясно, что сходимость частоты неверного определения класса объекта к вероятности этого события является достаточным условием равномерной сходимости эмпирического риска к среднему риску при любой функции потерь выделенного выше класса.

Этот факт имеет или не имеет места в зависимости, во-первых, от того, какое параметрическое семейство решающих правил выбрано, во-вторых, от того вероятностные характеристики источника данных. Мы сейчас займемся выработкой терминологии, которая позволит нам различать разные классы решающих правил с точки зрения их способности обеспечить равномерную сходимость частоты неверного распознавания к вероятности этого события.

Пусть $\hat{g}[\mathbf{x};\mathbf{a}]$ - некоторое семейство решающих правил распознавания. Пусть $\mathbf{x}_1,\dots\mathbf{x}_N$ - некоторое конечное множество точек в пространстве признаков. При каждом значении параметра **a** решающее правило распознавания делит эти *N* точек на два класса (рис.1.2.7). Причем это деление зависит от выбора параметра. Вообще говоря, если не обращать внимание на класс решающих правил, *N* объектов можно разбить на два класса

$$\sum_{K=1}^{N} C_N^{N-K} = \sum_{K=1}^{N} \frac{N!(N-K)!}{K!}$$

способами. Однако в рамках принятого класса решающих правил не все из возможных способов можно реализовать. Кроме того число способов которыми данное параметрическое семейство решающих правил можно разбить эти точки на два класса зависит еще и от расположения этих точек в признаковом пространстве.

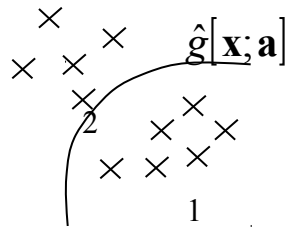


Рисунок 1.2.7 Решающее правило распознавания

Обозначим через $\Delta(\mathbf{x}_1,\dots\mathbf{x}_N)$ число способов, которым данное параметрическое семейство разбивает на два класса данную совокупность точек за счет варьирования параметра **a**. Будем рассматривать точки $\mathbf{x}_1,\dots\mathbf{x}_N$ как случайные точки в составе обучающей выборке.

Плотность распределения такого случайного вектора – это полная функция распределения **x** в признаковом пространстве по всем классам

$$f(\mathbf{x}) = \sum_{k=1}^{m} g^k \varphi^k(\mathbf{x}),$$

тогда $\Delta(\mathbf{x}_1, \dots \mathbf{x}_N)$ -случайная величина.

Математическое ожидание логарифма этой случайной величины

$$H(N) = M\{\log_2 \Delta(\mathbf{x}_1, \dots \mathbf{x}_N)\}$$

называют энтропией данного семейства решающих правил на выборке $\mathbf{x}_1, \dots \mathbf{x}_N$.

Теорема. Для равномерной сходимости частоты появления событий $\hat{g}[\mathbf{x}(\omega)] \neq g[\omega]$ к их вероятности необходимо и достаточно, чтобы вероятностные меры на множестве Ω и семейство решающих правил удовлетворяли условию

$$\lim_{N \to \infty} \frac{H(N)}{N} = 0.$$

Однако на практике удобнее пользоваться значительно более конструктивным, хотя и только достаточным условием равномерной сходимости частот появления ошибок к их вероятностям, опирающимся только на решающее правило $\hat{g}[\mathbf{x}; \mathbf{a}]$ и инвариантное к конкретному виду распределения вероятностей на множестве Ω.

Емкостью семейства решающих правил называется максимальное число ν точек $\mathbf{x}_1, \dots \mathbf{x}_\nu$ в пространстве R^n, которые можно разбить на два класса всеми возможными способами за счет варьирования параметра **a**.

В частности $\nu = n+1$ для семейства линейных решающих правил

$$\hat{g}(\mathbf{x}; \mathbf{a}, b) = \begin{cases} 1; \mathbf{a}^T \mathbf{x} + b > 0 \\ 2; \mathbf{a}^T \mathbf{x} + b \leq 0 \end{cases}$$

Теорема. Справедлива оценка сверху для вероятности отклонения частоты ошибочной классификации от ее вероятности

$$P\left\{ \sup_{\mathbf{a} \in \mathrm{A}} \left| \frac{1}{N} \sum_{j=1}^{N} I\left(\hat{g}[\mathbf{x}_j; \mathbf{a}] \neq g_j\right) - P(\hat{g}[\mathbf{x}(\omega)] \neq g[\omega]) \right| > \varepsilon \right\} < 4{,}5 \frac{(2N)^\nu}{\nu!} e^{-\frac{\varepsilon^2 (N-1)}{4}}$$

$$P(\hat{g}[\mathbf{x}(\omega)] \neq g[\omega])$$

Заметим, что величина $P(\hat{g}[\mathbf{x}(\omega)] \neq g[\omega])$ неслучайна и является характеристикой источника данных, а $\frac{1}{N} \sum_{j=1}^{N} I\left(\hat{g}[\mathbf{x}_j; \mathbf{a}] \neq g_j\right)$ - случайная величина, зависящая от выборки. Из этой теоремы непосредственно следует, что для равномерной сходимости частот к вероятности достаточно, чтобы емкость ν семейства решающих правил $\hat{g}[\mathbf{x}; \mathbf{a}]$ была конечной.

Допустим, что по обучающей выборке $(g_j, \mathbf{x}_j),\ j = 1,..., N$ в рамках параметрического семейства $\hat{g}[\mathbf{x}; \mathbf{a}]$ проведено обучение распознаванию образов по методу минимизации эмпирического риска и получено значение параметра $\hat{\mathbf{a}}_N = \arg\min_{\mathbf{a}\in\mathsf{A}} \hat{J}_N(\mathbf{a})$. фундаментальным вопросом теории распознавания образов является вопрос о величине среднего риска ошибки распознавания $J(\hat{\mathbf{a}}_N)$ для решающего правила $\hat{g}[\mathbf{x}; \hat{\mathbf{a}}_N]$.

Ограничимся рассмотрением случая антидиагональной матрицы потерь. тогда средний риск ошибки распознавания $J(\mathbf{a})$ - это вероятность ошибки распознавания, а эмпирический риск – это доля неверно классифицированных объектов, т.е. в сформулированной выше теореме под знаком модуля сумма $\frac{1}{N}\sum_{j=1}^{N} I\left(\hat{g}[\mathbf{x}_j; \mathbf{a}] \neq g_j\right)$ есть эмпирический риск, а вероятность $P(\hat{g}[\mathbf{x}(\omega)] \neq g[\omega])$ - средний риск. Поэтому теорему можно переписать следующим образом

$$P\left\{\sup_{\mathbf{a}\in\mathsf{A}} \left|\hat{J}_N(\mathbf{a}) - J(\mathbf{a})\right| < \varepsilon\right\} > 1 - \eta_N,$$

где

$$\eta_N = 4{,}5\frac{(2N)^{\nu}}{\nu!} e^{-\frac{\varepsilon^2(N-1)}{4}}.$$

Отсюда

$$P\left\{\left|\hat{J}_N(\mathbf{a}) - J(\mathbf{a})\right| < \varepsilon\right\} > 1 - \eta_N,$$

и следовательно

$$P\left\{J(\hat{\mathbf{a}}_N) < \hat{J}_N(\hat{\mathbf{a}}_N) + \varepsilon\right\} > 1 - \eta_N \qquad (1.2.14)$$

В сущности это то неравенство, которое позволят судить о качестве решающего правила распознавания, полученного по обучающей выборке. Однако это неравенство является чрезвычайно осторожным и очень сильно занижает вероятность выполнения условия $J(\hat{\mathbf{a}}_N) < \hat{J}_N(\hat{\mathbf{a}}_N) + \varepsilon$. Осторожность этой оценки качества решающего правила вытекает по-видимому из того факта, что мы заменили понятие энтропии семейства решающих правил его емкостью. Грубость понятия емкости семейства решающих правил связана с тем, что оно сформулировано без учета вероятностных характеристик источника данных. В результате утверждения на основе этого понятия покрывают «самые плохие» распределения вероятности, даже те, которых никогда не бывает в реальности.

Обратим внимание, что теорема о связи равномерной сходимости частот ошибки к их вероятностям на основе свойств энтропии семейства решающих правил имеет вид

необходимого и достаточного условия, т.е. вовсе не является излишне осторожной гарантией.

В тоже время аналогичное условие в терминах емкости класса решающих правил уже имеет лишь достаточный характер.

1.2.6 Детерминистский вариант задачи обучения распознавания образов.

Оптимальная разделяющая гиперплоскость.

В качестве пространства признаков будем рассматривать n-мерное действительное пространство $\mathbf{R}^n$, понимая его точки как вектор-столбцы $\mathbf{x} = (x_1, \ldots, x_n)^T$. Будем также рассматривать только случай двух классов, причем нам будет удобнее выбрать в качестве индексов классов не их номера не 1 и 2, а индексы $g = 1$ и $g = -1$.

Пусть Ω - гипотетическое множество всех мыслимых объектов распознавания $\omega \in \Omega$, каждый из которых характеризуется фактической принадлежностью к первому либо второму классу, выражаемой неизвестным, вообще говоря, значением индикаторной функции класса $g(\omega)$, соответственно, $g(\omega) = 1$ либо $g(\omega) = -1$, и наблюдаемым значением вектора признаков $\mathbf{x}(\omega)$.

Под задачей распознавания понимается задача построения некоторого решающего правила, которое позволило бы судить о скрытом классе $g(\omega)$ предъявленного объекта на основе анализа вектора его наблюдаемых признаков $\mathbf{x}(\omega) \in \mathbf{R}^n$, т.е правила вида $\hat{g}(\mathbf{x})$, и “не слишком часто” ошибаться. Мы ограничимся здесь рассмотрением только решающих правил, опирающихся на линейные дискриминантные функции

$$\hat{g}(\mathbf{x};\mathbf{a},b) = \begin{cases} 1, \text{ если } \mathbf{a}^T\mathbf{x} + b > 0, \\ -1, \text{ если } \mathbf{a}^T\mathbf{x} + b < 0, \end{cases} \qquad (1.2.15)$$

где вектор $\mathbf{a} \in \mathbf{R}^n$ и скаляр $b \in \mathbf{R}$ являются параметрами, полностью определяющими линейную дискриминантную функцию. Заметим, что уравнение $\mathbf{a}^T\mathbf{x} - b = 0$ определяет линейное многообразие размерности $n-1$, разделяющее в пространстве признаков $\mathbf{R}^n$ области принятия решений в пользу первого и второго класса. Мы, как и ранее, будем называть такое многообразие разделяющей гиперплоскостью, хотя, строго говоря, под гиперплоскостью принято понимать частный случай, когда линейное многообразие размерности на единицу меньшей, чем все линейное пространство, является его подпространством, т.е. содержит нулевую точку, что имеет место только в случае $b = 0$.

Пусть $(g_j, \mathbf{x}_j), \quad j = 1,..., N$ - обучающая выборка, где индексы классов принимают значения ± 1. Исходя из принципа минимизации эмпирического риска, нам хотелось бы выбрать такие параметры линейной дискриминантной функции, которые бы обеспечивали бы наименьшее количество ошибок распознавания в пределах обучающей выборке.

Обратим внимание на тот факт, что если некоторая линейная дискриминантная функция с параметрами (**a,** b) обеспечивает некоторое вполне определенное значение доли ошибок, то такую же долю ошибок будет обеспечивать целое множество дискриминантных функций с близкими параметрами, что для случая размерности признакового пространства *n=2* показано на рис. 1.2.8. Тем не менее, очевидно, что значения среднего риска, обеспечиваемые этими, казалось бы эквивалентными с точки зрения обучающей выборки дискриминантными функциями, будут практически всегда различными по отношению к генеральной совокупности. Поскольку задача распознавания всегда решается на конечной выборке, то разговор о эмпирического риска к среднему при увеличении ее размера является малоутешительным. Вопрос о выборе одной дискриминантной функции из множества эквивалентных является очень актуальным.

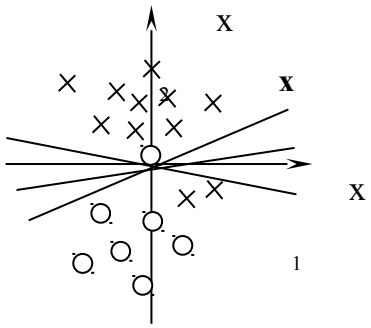
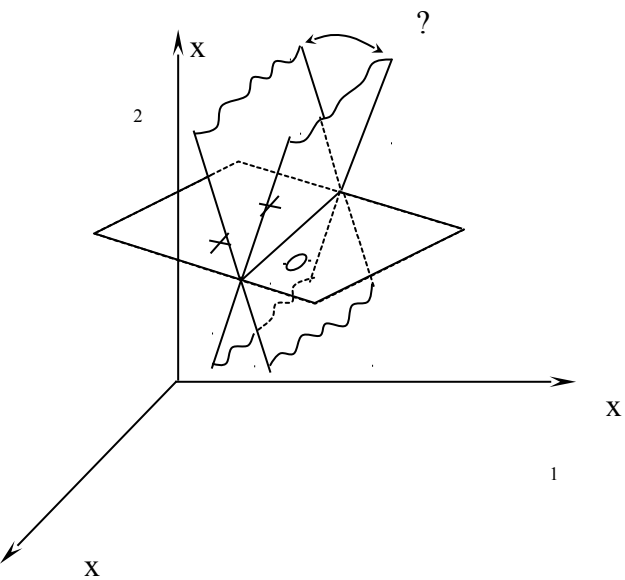


Рисунок 1.2.8 Иллюстрация неоднозначности решения задачи обучения распознаванию образов по обучающей выборке**.**

Этот вопрос становится принципиальным, если размерности признакового пространства велика по сравнению с размером выборки.

Очень часто возникают ситуации, когда число элементов выборки меньше размерности пространства признаков. Тогда элементы выборки образуют в пространстве признаков подпространство (аффинное многообразие). Этот случай проиллюстрирован на рис 1.2.9.



Исходное предположение заключается в том, что в пространстве признаков существуют области не слишком сложной формы, пересекающиеся или непересекающиеся, в которых полностью сосредоточены распределения вероятности, связанные с первым и вторым классом объектов.

Если обучающая выборка имеет число элементов меньшее, чем размерность признакового пространства, то такая выборка несет информацию лишь о форме сечения областей обоих классов в

n **Рисунок 1.2.9** Аффинное многообразие

подпространстве, образованном выборкой. Поэтому, строго говоря, такая выбора не позволяет нам принять обоснованное решение о линейной дискриминантной функции во всем признаковом пространстве. Мы можем лишь выбрать форму следа такой разделяющей гиперплоскости в подпространстве выборки. Выбрать же “пространственный наклон” разделяющей гиперплоскости мы не в состоянии, поскольку выборка не содержит никакой информации, на которую можно было бы опереться.

Для того, чтобы выбрать одну разделяющую гиперплоскость из целого пучка возможных, необходимо привлечь некоторую априорную информацию, содержащуюся в выборке.

В нашем случае, выбрав след разделяющей гиперплоскости в подпространстве выборки, мы фактически оценили форму сечений областей классов с тех сторон, которыми они обращены друг к другу. Что же касается формы этих областей вне пространства выборки, то поскольку нет никакой фактической информации, естественно принять, что эти формы такие же. В геометрических терминах такое предположение означает, сто разделяющую гиперплоскость надо выбрать так, чтобы она была ортогональна подпространству выборки.

В метрических пространствах это означает, что мы выбираем разделяющую гиперплоскость так, чтобы точки первого и второго класса обучающей выборки были как можно дальше от нее, каждые со своей стороны.

Заметим, что такая идея представляется вполне разумной и для больших выборок. Ведь точки обучающей выборки несут лишь приблизительную информацию о формах областей классов. Нет гарантии, что область данного класса “заканчивается” прямо сразу за крайней точкой обучающей выборки. Поэтому представляется целесообразным на всякий случай отодвинуть разделяющую гиперплоскость от точек как первого, ток и второго класса.

Для того, чтобы реализовать эту идею надо выбрать способ измерения удаленности гиперплоскости от выборки объектов определенного класса, т.е. способ измерения того, насколько вся выборка находится по нужную сторону от разделяющей гиперплоскости.

В данном курсе в качестве меры удаленности мы примем удаленность от гиперплоскости “самой плохой” точки выборки.

Сразу же заметим, что такой выбор представляется естественным далеко не всегда, поскольку исходит из того, что выборка не содержит “диких точек”, что на самом деле встречается довольно часто.

Изложенный принцип построения линейной дискриминантной функции реализуется понятием оптимальной разделяющей гиперплоскости для точек двух классов.

1.3. Особенности обучения в условиях малого относительного размера обучающей выборки по сравнению с размерностью пространства признаков

1.3.1. Селекция признаков (сокращение признакового пространства)

Одной из классических задач теории распознавания образов является понижение размерности вектора признаков $\mathbf{x}$. Следует отметить тот факт, что интерес к этой процедуре сохраняется и в последнее время т.к. появление нового поколения быстродействующих ЭВМ, и как следствие, относительная независимость исследователей, разработчиков в области анализа данных, и распознавания образов в частности, от вычислительной сложности не явилась эволюционным решением этой проблемы. Дело в том, что помимо (а) сокращения объема вычислений, отбор признаков, в большей или меньшей степени направлен и на (б) сжатие объема данных, (в) сокращение стоимости сбора данных, (г) улучшение классификации, (д) возможность визуализации многомерных данных [1,13]. Приложения, требующие применения методов отбора признаков встречаются в следующих задачах: (1) Приложения, в которых объединены данные от большого числа датчиков. (2) Объединение многомерных моделей, когда все параметры различных моделей могут быть использованы для классификации; и (3) приложения по основанные на получении (извлечении) данных, где целью является определение скрытых зависимостей между признаками.

Таким образом, очевидно, что выбор признаков играет в распознавании важную роль. Выбор адекватного множества признаков, учитывающий трудности, которые связаны с реализацией процессов выделения или выбора признаков, и обеспечивающий в то же время необходимое качество классификации, представляет собой одну из наиболее трудных задач построения распознающих систем. Для того чтобы облегчить анализ этой задачи в [29] предлагается разделить признаки на три категории: (1) "физические", (2) структурные и (3) математические.

Физические и структурные признаки обычно используются людьми при распознавании образов, поскольку такие признаки легко обнаружить на ощупь, визуально, с помощью других органов чувств. Поскольку органы чувств обучены распознаванию физических и структурных признаков, человек, естественно

пользуется в основном такими признаками при классификации и распознавании. В случае же построения вычислительной системы распознавания образов эффективность таких признаков с точки зрения организации процесса распознавания может существенно снижаться, так как, вообще говоря, в большинстве практических ситуаций довольно сложно имитировать возможности органов чувств человека. С другой стороны, можно создать систему, обеспечивающую выделение математических признаков образов, что может оказаться затруднительным для человека при отсутствии "механической" помощи. Примерами признаков этого типа являются статистические средние, коэффициенты корреляций, характеристические числа и собственные векторы ковариационных матриц, и прочие инвариантные свойства объектов.

Предварительная обработка образов обычно включает решение двух основных задач: преобразование кластеризации и выбор признаков. Основной задачей распознавания образов является построение решающих функций, представляющих некоторые классы. Эти функции должны обеспечивать разделение пространства измерений на области, каждая из которых содержит точки, представляющие образы только одного из рассматриваемых классов. Данное положение приводит к идее преобразований кластеризации, реализуемого в пространстве измерений, для того чтобы обеспечить группировку точек, представляющих выборочные образы одного класса. В результате такого преобразования максимизируется расстояние между множествами и минимизируются внутримножественные расстояния.

Выбор наиболее эффективных признаков позволяет снизить размерность вектора измерений. Выбор признаков можно осуществлять вне связи с качеством схемы классификации. Оптимальный выбор признаков при этом определяется максимизацией или минимизацией некоторого критерия. Такой подход принято считать выбором признаков без учета ограничений. Другой подход связывает выбор признаков с качеством классификации, причем обычно эта связь выражается в терминах вероятности правильного распознавания.

Преобразование кластеризации и упорядочение признаков

Практически всегда измерения характеристик образа, соответствующие отдельным признакам x_j , $j = 1,\ldots,L$, не важны в одинаковой степени для задачи классификации. Преобразование кластеризации признакам с меньшей значимостью принято приписывать меньшие веса. Назначение весов признаков можно осуществить посредством линейного преобразования, которое обеспечивает более благоприятную группировку точек в новом, т.н. вторичном пространстве.

Рассмотрим векторы образов $\mathbf{a}$ и $\mathbf{b}$, который после применения к ним преобразования $\mathbf{W}$ перешли в векторы $\mathbf{a}^*$ и $\mathbf{b}^*$. Тогда справедливо $\mathbf{a}^* = \mathbf{W}\mathbf{a}$ и $\mathbf{b}^* = \mathbf{W}\mathbf{b}$, где

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1L} \\ w_{21} & w_{22} & \dots & w_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ w_{L1} & w_{L2} & \dots & w_{LL} \end{pmatrix},$$

и w_{jk} - весовые коэффициенты.

Таким образом каждый элемент преобразованного вектора образа представляет собой линейную комбинацию элементов исходного вектора. В новом пространстве евклидово расстояние между векторами $\mathbf{a}^*$ и $\mathbf{b}^*$ определяется как

$$D(\mathbf{a}^*, \mathbf{b}^*) = \sqrt{\sum_{j=1}^{L} (a_j^* - b_j^*)^2} = \sqrt{\sum_{j=1}^{L} \left[\sum_{i=1}^{L} w_{ji}(a_i - b_i) \right]^2}. \qquad (1.3.1)$$

В тех случаях, когда линейное преобразование сводится к изменению масштабных коэффициентов координатных осей, матрица $\mathbf{W}$ является диагональной, и выражение для евклидова расстояния сводится к

$$D(\mathbf{a}^*, \mathbf{b}^*) = \sqrt{\sum_{j=1}^{L} w_{jj}^2 (a_j^* - b_j^*)^2}, \qquad (1.3.2)$$

где элементы w_{jj} представляют собой весовые коэффициенты при признаках. Задача преобразования кластеризации заключается в том чтобы определить весовые коэффициенты признаков w_{jj}, минимизирущие внутримножественные расстояния внутри классов с учетом определенных ограничений, наложенных на коэффициенты w_{jj}. В качестве подобных ограничений обычно рассматриваются ограничения вида $\sum_{j=1}^{L} w_{jj} = 1$ и $\prod_{j=1}^{L} w_{jj} = 1$. С учетом того факта, что в новом пространстве внутреннее расстояние для множества точек, представляющих образы определяется как

$$\overline{D^2} = 2\sum_{j=1}^{L} (w_{jj}\sigma_j)^2, \qquad (1.3.3)$$

где σ_j^2 - несмещенная оценка выборочной дисперсии компонент, соответствующих координат x_j, минимизация $\overline{D^2}$ с учетом первого ограничения дает значения весовых коэффициентов

$$w_{jj} = \frac{1}{\sigma_j^2 \sum_{j=1}^{L} (1/\sigma_j^2)}, \qquad (1.3.4)$$

и с учетом второго ограничения

$$w_{jj} = \frac{1}{\sigma_j}\left(\prod_{j=1}^{L}\sigma_j\right)^{1/L} \qquad (1.3.5)$$

Формулы (1.3.4) и (1.3.5) определяют матрицу преобразования $\mathbf{W}$ с учетом введенных выше ограничений. Если векторы образов переводятся из пространства X в пространство X^* с помощью преобразования

$$\mathbf{x}^* = \mathbf{W}\mathbf{x}, \qquad (1.3.6)$$

то внутреннее расстояние множества в пространстве X^* минимизируется. После этого требуется повести второе преобразование

$$\mathbf{x}^{**} = \mathbf{A}\mathbf{x}^*$$

с целью выделения компонент, имеющих заданную дисперсию, и обеспечения возможности провести упорядочение и выбор признаков. Это преобразование превращает ковариационную матрицу точек, представляющих образы в пространстве X^{**} в диагональную. Кроме того, для обеспечения неизменности расстояний необходимо наложить условие ортонормированности на матрицу $\mathbf{A}$.

Пусть $\mathbf{e}_1$, $\mathbf{e}_2,\ldots,\mathbf{e}_L$ - собственные векторы ковариационной матрицы $\mathbf{C}$ и λ_1, $\lambda_2,\ldots,\lambda_L$ - соответствующие характеристические числа. Элементы ортогональной матрицы преобразования $\mathbf{A}$ выбираются так, что в преобразованном пространстве ковариационная матрица становится диагональной. Этого можно достичь используя l транспонированных собственных векторов ковариационной матрицы $\mathbf{C}$ в качестве строк ортогональной матрицы $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \\ \vdots \\ \mathbf{e}_l^T \end{pmatrix}.$$

Таким образом, размерность исходного признакового пространства понижена до $l < L$. После этого можно определить матрицу весов $\mathbf{W}$ так, чтобы расстояние $\overline{D^2}$ принимало при выполнении заданных ограничений экстремальное значение. Для ограничения вида $\prod_{j=1}^{L} w_{jj} = 1$ матрица $\mathbf{W}$ определяется следующим образом:

$$\mathbf{W} = \left(\prod_{j=1}^{l}\lambda_j\right)^{1/2m}\begin{pmatrix} \lambda_1^{-1/2} & 0 & \cdots & 0 \\ 0 & \lambda_2^{-1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_l^{-1/2} \end{pmatrix}. \qquad (1.3.7)$$

Следовательно, если мы хотим минимизировать внутреннее расстояние множества, то в качестве векторов признаков следует выбирать собственные векторы,

соответствующие наименьшим характеристическим числам ковариационной матрицы $\mathbf{C}$.

При ограничении $\sum_{j=1}^{L} w_{jj} = 1$ матрица $\mathbf{W}$ определяется как

$$\mathbf{W} = \left(\sum_{j=1}^{l} \frac{1}{\lambda_j} \right)^{-1} \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \vdots & \frac{1}{\lambda_l} \end{pmatrix} \qquad (1.3.8)$$

Таким образом, значение внутреннего расстояния множества достигнет глобального минимума, если в качестве характеристических чисел λ_j выбраны l наименьших из L характеристических чисел ковариационной матрицы и матрица преобразования $\mathbf{A}$ составлена из l соответствующих собственных векторов.

Выбор признаков при помощи минимизации энтропии

В теории статистики статистическую неопределенность принято выражать таким понятием как энтропия. Энтропия представляет собой статистическую неопределенность. Хорошей мерой внутреннего разнообразия для заданного семейства объектов распознавания служит энтропия совокупности, определяемая как

$$H = -E\{\ln p\}, \qquad (1.3.9)$$

где p - плотность вероятности совокупности образов, а E - оператор математического ожидания плотности p. Понятие энтропии удобно использовать в качестве критерия при организации оптимального выбора признаков. Признаки, уменьшающие неопределенность заданной ситуации, считаются более информативными, чем те, которые приводят к противоположному результату. Таким образом, если считать энтропию мерой неопределенности, то разумным правилом является выбор признаков, обеспечивающих минимизацию энтропии рассматриваемых классов. Поскольку это правило эквивалентно минимизации дисперсии в различных совокупностях образов, то вполне можно ожидать, что соответствующая процедура будет обладать кластеризационными свойствами.

Пусть существует m классов, характеризующихся плотностями распределения $p(\mathbf{x} \mid \omega_1)$, $p(\mathbf{x} \mid \omega_2)$,..., $p(\mathbf{x} \mid \omega_m)$. В соответствии с (1.3.9) энтропия i-ой совокупности образов определяется как

$$H_i = -\int_{\mathbf{x}} p(\mathbf{x} \mid \omega_i) \ln p(x \mid \omega_i) d\mathbf{x}. \qquad (1.3.10)$$

Очевидно, что при $p(\mathbf{x} \mid \omega_i) = 1$, т.е. при отсутствии неопределенности, имеем $H_i = 0$.

Данный метод предлагает, что каждая из M совокупностей образов характеризуется плотностью нормального распределения с математическим ожиданиями $\mathbf{m}_i$ и ковариационными матрицами $\mathbf{C}_i$, соответственно для i-ой совокупности образов. Кроме того, предполагается, что ковариационные матрицы, описывающие статистические характеристики всех m классов идентичны.

Основная идея заключается в определении матрицы линейного преобразования $\mathbf{A}$, переводящей заданные векторы образов в новые векторы меньшей размерности. Это преобразование можно представить как

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \tag{1.3.11}$$

причем матрица преобразования отыскивается при помощи минимизации энтропий совокупности образов, входящих в рассматриваемые классы. Предполагается, что вектор $\mathbf{X}$ - размерности L, $\mathbf{У}$ - отображенный вектор, размерности l, $l < L$ и $\mathbf{A}$ - матрица размерности $l \times L$. Строками матрицы $\mathbf{A}$ служат l выбранных векторов $\mathbf{a}_1^T$, $\mathbf{a}_2^T, \ldots, \mathbf{a}_l^T$, представляющих собой вектор-строки. Таким образом, матрица $\mathbf{A}$ имеет вид

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_l^T \end{pmatrix} \tag{1.3.12}$$

Задача состоит в определении такого способа выбора l векторов признаков, чтобы вектор $\mathbf{X}$ преобразовывался в изображение $\mathbf{У}$ и одновременно минимизировалась величина энтропии, определяемая (1.3.10).

Многомерное нормальное распределение полностью определяется вектором математического ожидания и ковариационной матрицей, которая в свою очередь характеристическими числами и собственными векторами. Последние можно рассматривать как векторы, представляющие свойства рассматриваемых образов. Часть из этих векторов свойств содержит меньше информации, ценной для распознавания, чем другие векторы, и поэтому ими можно пренебречь. Это явление приводит к процедуре выбора признаков, предусматривающей использование наиболее важных свойств в качестве векторов-признаков. Такие векторы-признаки можно затем использовать для формирования матрицы преобразования $\mathbf{A}$. В [27, 57] показано, что функция энтропии H_i^* принимает минимальное значение, если матрица преобразования $\mathbf{A}$ составлена из l нормированных собственных векторов, соответствующих наименьшим характеристическим числам ковариационной матрицы

$\mathbf{C}$. Применяя этот результат, надо иметь в виду, что число векторов, используемых для формирования матрицы $\mathbf{A}$, должно быть достаточно большим, чтобы изображения несли достаточное количество различительной информации.

Преобразование Карунена-Лоэва

Основанием применения дискретного разложения Карунена-Лоэва [27] в качестве средства выбора признаков является наличие у него следующих оптимальных свойств.

Во-первых, оно минимизирует среднеквадратичную ошибку при использовании лишь конечного числа базисных функций в разложении

$$\mathbf{x}_i = \sum_{j=1}^{L} c_{ij}\boldsymbol{\phi}_j \,, \tag{1.3.13}$$

где

$$x_i = \begin{pmatrix} x_i(t_1) \\ x_i(t_2) \\ \vdots \\ x_i(t_L) \end{pmatrix}, \qquad \boldsymbol{\phi}_j = \begin{pmatrix} \phi_j(t_1) \\ \phi_j(t_2) \\ \vdots \\ \phi_j(t_L) \end{pmatrix}, \tag{1.3.14}$$

L - количество наблюдений для функции $x_i(t)$, осуществленных в интервале $[T_1, T_2]$, $\phi_j(t)$ - базисные функции, в качестве которых используется множество детерминированных ортонормированных функций, заданных на интервале $[T_1, T_2]$. Причем относительно коэффициентов предполагается, что они удовлетворяют условию $E\{c_{ij}\} = 0$.

Во-вторых, данное преобразование минимизирует функцию энтропии, выраженную через дисперсии коэффициентов разложения.

Принцип минимизации среднеквадратичной ошибки предполагает, что разложение Карунена-Лоэва минимизирует ошибку аппроксимации при использовании в приведенном разложении числа базисных векторов, меньшего L. Принцип минимизации энтропии обеспечивает искомые эффекты кластеризации, описанные выше.

Применение дискретного разложения Карунена-Лоэва при выборе признаков можно рассматривать как линейное преобразование. Если

$$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1\boldsymbol{\phi}_2 \ldots \boldsymbol{\phi}_l)\,, \quad l < L \tag{1.3.15}$$

- матрица преобразования, то преобразованные образы являются коэффициентами разложения Карунена-Лоэва, т.е. для любого образа $\mathbf{x}_i$,принадлежащего классу $\boldsymbol{\omega}_i$, выполняется

$$\mathbf{c}_i = \mathbf{\Phi}^T \mathbf{x}_i . \qquad (1.3.16)$$

Поскольку $\mathbf{\Phi}^T$ - матрица размера $l \times L$ и $\mathbf{X}$ - L-мерный вектор, то $\mathbf{c}_i$ при $l < L$ представляют собой изображения, имеющие размерность, меньшую чем L.

Условия оптимальности разложения Карунена-Лоэва выполняются, если в качестве столбцов матрицы преобразования $\mathbf{\Phi}$ выбраны l нормированных собственных векторов, соответствующих наибольшим характеристическим числам корреляционной матрицы $\mathbf{R}$. В таком случае для любого вектора $\mathbf{X}$ его изображения меньшей размерности определяются как

$$\mathbf{y} = \mathbf{A}\mathbf{x},$$

где $\mathbf{A}$ - матрица преобразования, строками которой служат нормированные собственные векторы, соответствующие наибольшим характеристическим числам корреляционной матрицы $\mathbf{R}$.

Для того, чтобы применение разложения Карунена-Лоэва приводило к получению оптимальных результатов, необходимо выполнение условия $E\{\mathbf{x}_i\} = 0$, которое выполняется автоматически, если отдельные классы характеризуются нулевыми математическими ожиданиями. Однако надо иметь в виду, что за исключением непосредственно этапа обучения, вообще говоря, отсутствуют сведения о принадлежности образа к определенному классу.

Хотя предположение об идентичности математических ожиданий всех совокупностей образов ограничивает возможности применения разложения Карунена-Лоэва, не следует считать, что этот подход к выбору признаков не имеет достоинств. Допущения такого рода характерны для большинства статистических методов анализа.

1.3.2. Стабилизация классификаторов

В дискриминантном анализе часто встречаются задачи с малыми размерами обучающих выборок. В практических задачах ситуация складывается так, что должно быть рассмотрено большое число различных измерений при обучении на объектах реального мира. Как результат, во многих областях применения массивы данных могут иметь большое число признаков, например 100-200 и более. Статистическая зависимость между отдельными компонентами в таких данных часто имеет нелинейный характер. Следовательно, необходимо использовать дополнительные полиномиальные признаки для четкого оценивания упомянутых нелинейных зависимостей [52]. В этих случаях размерность признакового пространства становится крайне большой. Если число объектов, по которым необходимо провести

обучение, все же ограничено, может возникнуть большие трудности при построении дискриминантной функции [38,47,48].

Стандартные классические статистические методы требуют обращения ковариационной матрицы, например, линейная дискриминантная функция Фишера (ЛДФ) [28,35]:

$$g_F(\mathbf{x}) = [x - \tfrac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]S^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \qquad (1.3.17)$$

в которой S представляет собой оценку максимума правдоподобия ковариационной матрицы $\mathbf{C}$ размером $n \times n$, $\mathbf{x}$ представляет собой p-размерный вектор, который необходимо классифицировать, $\bar{\mathbf{x}}^{(i)}$ - вектор среднего по выборке i-го класса. Необходимо определить такое решающее правило $(\mathbf{a}, a_0)$, чтобы выполнялось

$$g_F(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + a_0 = d, \qquad (1.3.18)$$

где $\mathbf{x} \in \mathsf{X}$ и d принимает положительные значения для объектов первого класса и отрицательные для второго.

Прямые вычисления невозможны в случае, когда число признаков n превышает число объектов N [46]. Для больших размеров признаков при уменьшении n ожидаемая вероятность ошибок классификаций сильно возрастает [47]. Для преодоления этих проблем существует несколько различных методов.

Один заключается в снижении числа признаков до $n < N$, используя знания нескольких экспертов или с помощью методов отбора и экстракции наиболее информативных признаков, например, как это было описано в предыдущем разделе. Второй подход заключается в модификации стандартной ЛДФ каким либо способом. Целью настоящего раздела и является обзор некоторых таких возможностей. Главные из них – это применение функции псевдо-линейного разделения Фишера, которая рассматривает классификаторы в подпространстве, определенном некоторыми доступными объектами [34] и классификатор ближайших средних, который игнорирует ковариации.

Третий подход состоит способе стабилизации, использующем различные методы регуляризации, такие как ридж-оценка ковариационной матрицы и бэггинг (бутстрэп, совмещенный с агрегированием [37]). Часто оптимальная ридж-оценка ковариационной матрицы или бэггинг могут улучшить работу линейных классификаторов. Иногда подобные методы стабилизации не только не помогают, но и более того даже увеличивают ошибку классификации. Это четко связано со стабильностью классификаторов на специфических данных.

Одной из простейших модификаций процедуры ЛДФ является классификатор ближайших средних

$$g_{NM}(\mathbf{x}) = [x - \tfrac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^T(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \qquad (1.3.19)$$

Следовательно, классификатор ближайших средних генерирует перпендикулярную биссектрису между средними по классам, и таким образом, строит оптимальный линейный классификатор для классов с одинаковыми нормальными сферическими распределениями. Преимущество этого классификатора заключается в относительной нечувствительности к размеру выборки [47]. Однако такой классификатор не учитывает разницы между дисперсиями и ковариациями.

Другая модификация линейной дискриминантной функции Фишера (1.3.17), позволяет преодолеть обращение вырожденной ковариационной матрицы для выборок малого размера $n < N$, это так называемый псевдо линейный дискриминатор Фишера (ПЛДФ) [34]. Здесь непосредственное решение для (1.3.18) получается как (используя расширенный вектор):

$$g_{PF}(\mathbf{x}) = (\mathbf{a}, a_0)^T (\mathbf{x}, 1) = (\mathbf{x}, 1)(\mathbf{X}, \mathbf{I})^{-1} d \qquad (1.3.20)$$

где, $(\mathbf{x}, 1)$ - расширенный вектор, который необходимо классифицировать и $(\mathbf{X}, \mathbf{I})$ - расширенная матрица данных. Процедура обращения матрицы $(\mathbf{X}, \mathbf{I})^{-1}$ представляет собой псевдо обращение Мур-Пенроса, которое дает минимальное нормированное решение. Перед обращением данные необходимо сдвинуть таким образом, чтобы средние значения по признакам были равны нулю. Этот метод близок к методу декомпозиции вырожденных значений.

Для значений $N \geq n$ ПЛДФ, максимизируя расстояния между выборками разных классов, идентичен ЛДФ (1.3.17). Однако для значений $N < n$, ПЛДФ находит линейное подпространство, где располагаются все данные, и в этом подпространстве оцениваются средние значения и ковариационные матрицы и строит линейное разделяющее правило, которое ортогонально этому подпространству во всех других направлениях, в которых нет заданных объектов.

Постоянное расстояние для всех обучающих выборок, найденное ПЛДФ может быть увеличено коррекцией обучающей выборки, позволяющей некоторым выборкам иметь большее взаимное расстояние. Такая процедура, названная классификатором малых выборок предложена в [34].

Ридж-оценка ковариационной матрицы

Хорошо известный метод обращения вырожденной ковариационной матрицы, использованный при построении стандартного ЛДФ (1.3.17), заключается в добавлении некоторых постоянных значений к диагональным элементам, оцениваемой ковариационной матрицы

$$\mathbf{C}_R = \mathbf{C} + \lambda \mathbf{I}, \qquad (1.3.21)$$

и $\mathbf{I}$ - есть единичная матрица размером $n \times n$, и λ - параметр регуляризации.

Новая оценка $\mathbf{C}_R$ называется "реберной" оценкой ковариационной матрицы, термин, который был заимствован из регрессионного анализа [1]. Этот подход называется регуляризованным дискриминантным анализом. Модификация (1.3.21) дает нам "реберную" или регуляризованную дискриминантную функцию

$$g_F(\mathbf{x}) = [x - \tfrac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})](\mathbf{S} + \lambda\mathbf{I})^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \qquad (1.3.22)$$

Очевидно, что при $\lambda \to \infty$, теряются значения дисперсий. В этом случае классификатор (1.3.22) является приближением к процедуре ближайших средних (1.3.19). Одновременно обобщенная ошибка может быть существенным образом уменьшена. Малые значения параметра регуляризации λ могут быть полезны для стабилизации решения. Очень малые значения λ могут быть не достаточно эффективны. Обычно для выбора приемлемого параметра регуляризации используется два метода: cross validation [40] и бутстрэпа. Они оба требуют большого объема вычислений, и оба не достаточно точны при малом размере обучающей выборки. Смотри, например, Фридман [36], Раудис и Скурихина [50].

Бэггинг

Другим методом стабилизации является бэггинг, основанный на бутстрэпе, и соединяющий концепции предложенные Бриманом. Предлагается строить бэггинг-классификатор посредством усреднения параметров линейного классификатора, построенного на нескольких повторениях процедуры бутстрэпа. Случайный выбор с замещением N образцов из набора из N штук называется дублированием бутстрэпа. Из каждой такой копии и строится "бутстрэпная" версия классификатора. Усреднение этих версий как раз и дает бэггинг-классификатор. Бримэн показал что бэггинг может уменьшить ошибку линейной регрессии и классификации. Отмечается, что такой метод полезен только для нестабильных процедур. Для стабильных методов он может даже ухудшить результат работы классификатора.

Таким образом, вопрос стабильности и нестабильности классификатора на специфических данных очень важен. Для классификатора можно предсказать меру нестабильности на разных данных, необходимость использования бэггинг или ридж-оценки ковариационной матрицы.

1.4.1 Основные задачи исследования

Для реализации предлагаемого принципа необходимости учета априорной информации об исследуемых данных при построении решающих правил распознавания образов в данной работе ставятся следующие основные задачи:

1. Разработать эффективный алгоритм обучения в признаковом пространстве большой размерности по сравнению с объемом выборки.
2. Разработать эффективный алгоритм обучения для классов задач, в которых легко удается непосредственно вычислить степень «непохожести» любых двух объектов, но трудно указать набор осмысленных характеристик объектов, которые могли бы служить координатными осями пространства признаков.
3. Разработать комплекс вспомогательных процедур, направленных на отображение многомерных данных и решающего правила распознавания.
4. Исследовать работоспособность предложенных алгоритмов на модельных и реальных данных.
5. Создать программно-алгоритмический комплекс, реализующий разработанные алгоритмы и обеспечивающий наглядное представление данных и результатов обучения и распознавания.

   Конструируемые процедуры должны реализовывать схему обучения распознаванию образов с учителем.