Лекции по статистическим (байесовским) алгоритмам классификации

К. В. Воронцов

16 января 2009 г.

Материал находится в стадии разработки, может содержать ошибки и неточности. Автор будет благодарен за любые замечания и предложения, направленные по адресу `vokov@forecsys.ru`, либо высказанные в обсуждении страницы «Машинное обучение (курс лекций, К.В.Воронцов)» вики-ресурса `www.MachineLearning.ru`.

Содержание

1 Статистические алгоритмы классификации

Статистический байесовский подход является одним из старейших в теории классификации и лежит в основе многих методов обучения. Он опирается на предположение, что плотности распределения каждого из классов известны. В этом случае удаётся в явном виде выписать алгоритм классификации, имеющий минимальную вероятность ошибок. На практике плотности классов, конечно же, неизвестны. Их приходится оценивать (восстанавливать) по обучающей выборке, что невозможно сделать с абсолютной точностью. В результате байесовский алгоритм перестаёт быть оптимальным. Чем короче выборка, тем выше шансы «подогнать» распределение под конкретные данные и столкнуться с эффектом переобучения.

Наиболее распространены три подхода к восстановлению плотностей: параметрический, непараметрический и разделение смеси распределений. Первый подход, при дополнительном предположении, что классы имеют гауссовские плотности, приводит к *линейному* или *квадратичному дискриминанту*. Второй подход основан введении функции расстояния (метрики) между объектами и приводит к *методу парзеновского окна*. Третий подход занимает промежуточное положение между первыми двумя. Если классы описываются смесью гауссовских плотностей, он приводит к *методу радиальных базисных функций*.

Существует ещё *«наивный» байесовский классификатор* основан на дополнительном предположении о статистической независимости признаков. Этот подход может комбинироваться с любым из трёх способов восстановления плотности. Как правило, он приводит к простым и устойчивым методам обучения, обладающим сравнительно низким качеством классификации.

§1.1 Вероятностная постановка задачи классификации

Рассмотрим вероятностную постановку задачи классификации, разделив её на две независимые подзадачи.

Задача 1.1. *Имеется множество объектов X и конечное множество имён классов Y. Множество прецедентов $X \times Y$ является вероятностным пространством с известной плотностью распределения $p(x,y) = \mathsf{P}(y)p(x|y)$. Вероятности появления объектов каждого из классов $P_y = \mathsf{P}(y)$ известны и называются априорными вероятностями классов. Плотности распределения классов $p_y(x) = p(x|y)$ также известны и называются функциями правдоподобия классов. Требуется построить алгоритм $a(x)$, минимизирующий вероятность ошибочной классификации.*

Задача 1.2. *Имеется множество прецедентов $X^\ell = (x_i, y_i)_{i=1}^\ell$, выбранных случайно и независимо из неизвестного распределения $p(x,y) = P_y p_y(x)$. Требуется построить эмпирические оценки*[1] *априорных вероятностей $\hat{P}_y$ и функций правдоподобия $\hat{p}_y(x)$ для каждого из классов $y \in Y$, которые приближали бы, соответственно, вероятности P_y и функции $p_y(x)$ на всём множестве X.*

[1] Здесь и далее символами с «крышечкой» обозначаются оценки вероятностей, функций распределения или случайных величин, вычисляемые по обучающей выборке. Такие оценки принято называть *выборочными* или *эмпирическими*.

Первая задача решается относительно легко, и мы сразу это сделаем. Вторая задача не имеет единственного решения, поскольку многие распределения $p(x, y)$ могли бы дать одну и ту же выборку X^ℓ. Для обеспечения единственности привлекаются дополнительные предположения о плотностях классов. Сделать это можно по-разному, что и приводит к большому разнообразию байесовских алгоритмов.

1.1.1 Функционал среднего риска

Знание функций правдоподобия позволяет находить вероятности событий вида «$x \in \Omega$ при условии, что x принадлежит классу y»:

$$\mathsf{P}(\Omega|y) = \int_\Omega p_y(x)\,dx, \quad \Omega \subset X.$$

Рассмотрим произвольный алгоритм $a\colon X \to Y$. Он разбивает множество X на непересекающиеся области:

$$A_y = \{x \in X \mid a(x) = y\}, \quad y \in Y.$$

Вероятность появления объекта класса y, который будет отнесён алгоритмом a к классу s, равна $P_y\,\mathsf{P}(A_s|y)$. Если $y = s$, то это вероятность правильной классификации. Если $y \neq s$, то это вероятность ошибочной классификации. В зависимости от конкретной задачи потери от ошибок разного рода могут быть различны. Каждой паре $(y, s) \in Y \times Y$ поставим в соответствие *величину потери* λ_{ys} при отнесении объекта класса y к классу s. Обычно полагают $\lambda_{yy} = 0$, и $\lambda_{ys} > 0$ при $y \neq s$. Соотношения потерь на разных классах зависят от конкретной задачи.

Пример 1.1. В задачах радиолокационной разведки класс $y = 1$ — самолёты противника, класс $y = 0$ — ложные цели, например, стая птиц. Наибольшая потеря возникает в том случае, когда объект класса 1 принимается за объект класса 0. Это называется *ошибкой I-го рода* или «пропуском цели». Когда объект класса 0 принимается за объект класса 1, говорят об *ошибке II-го рода* или «ложной тревоге». В данном случае $\lambda_{01} < \lambda_{10}$.

Пример 1.2. В задаче обнаружения спама класс $y = 1$ — нежелательные сообщения, класс $y = 0$ — обычные сообщения. Здесь, наоборот, пропуск спама является менее существенной потерей, чем «ложная тревога», поэтому $\lambda_{01} > \lambda_{10}$.

Опр. 1.1. *Функционалом среднего риска называется ожидаемая величина потери при классификации объектов алгоритмом a:*

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P_y\,\mathsf{P}(A_s|y).$$

Если величина потерь одинакова для ошибок любого рода, $\lambda_{ys} = [y \neq s]$, то средний риск $R(a)$ совпадает с вероятностью ошибки алгоритма a.

1.1.2 Оптимальное байесовское решающее правило

Докажем, что знание функций правдоподобия позволяет выписать в явном виде алгоритм a, минимизирующий средний риск $R(a)$.

Теорема 1.1. *Если известны априорные вероятности P_y и функции правдоподобия $p_y(x)$, то минимум среднего риска $R(a)$ достигается алгоритмом*

$$a(x) = \arg\min_{s\in Y} \sum_{y\in Y} \lambda_{ys} P_y p_y(x).$$

Доказательство.

Выделив произвольный $t \in Y$, распишем функционал полного риска:

$$R(a) = \sum_{y\in Y}\sum_{s\in Y} \lambda_{ys} P_y \,\mathsf{P}(A_s|y) =$$
$$= \sum_{y\in Y} \lambda_{yt} P_y \,\mathsf{P}(A_t|y) + \sum_{s\in Y\setminus\{t\}}\sum_{y\in Y} \lambda_{ys} P_y \,\mathsf{P}(A_s|y).$$

Применив формулу полной вероятности, $\mathsf{P}(A_t|y) = 1 - \sum\limits_{s\in Y\setminus\{t\}} \mathsf{P}(A_s|y)$, получим:

$$R(a) = \underbrace{\sum_{y\in Y} \lambda_{yt} P_y}_{\text{const}(a)} + \sum_{s\in Y\setminus\{t\}}\sum_{y\in Y} (\lambda_{ys} - \lambda_{yt}) P_y \,\mathsf{P}(A_s|y) =$$
$$= \text{const}(a) + \sum_{s\in Y\setminus\{t\}} \int_{A_s} \sum_{y\in Y} (\lambda_{ys} - \lambda_{yt}) P_y p_y(x)\, dx. \tag{1.1}$$

Введём для сокращения записи обозначение $g_s(x) = \sum\limits_{y\in Y} \lambda_{ys} P_y p_y(x)$, тогда

$$R(a) = \text{const}(a) + \sum_{s\in Y\setminus\{t\}} \int_{A_s} \big(g_s(x) - g_t(x)\big)\, dx.$$

В выражении (1.1) неизвестны только области A_s. Функционал $R(a)$ есть сумма $|Y|-1$ слагаемых $I(A_s) = \int_{A_s} \big(g_s(x) - g_t(x)\big)\, dx$, каждое из которых зависит только от одной области A_s. Минимум $I(A_s)$ достигается, когда A_s совпадает с областью неположительности подынтегрального выражения. В силу произвольности t

$$A_s = \big\{x \in X \;\big|\; g_s(x) \leqslant g_t(x),\ \forall t \in Y,\ t \neq s\big\}.$$

С другой стороны, $A_s = \big\{x \in X \;\big|\; a(x) = s\big\}$. Значит, $a(x) = s$ тогда и только тогда, когда $s = \arg\min\limits_{t\in Y} g_t(x)$. Если минимум $g_t(x)$ достигается при нескольких значениях t, то можно взять любое из них, что не повлияет на риск $R(a)$, так как подынтегральное выражение в этом случае равно нулю.

Теорема доказана. ■

Часто можно полагать, что величина потери зависит только от истинной классификации объекта, но не от того, к какому классу он был ошибочно отнесён: $\lambda_{ys} \equiv \lambda_y$ для всех $y, s \in Y$. В этом случае формула оптимального алгоритма упрощается.

Теорема 1.2. *Если известны априорные вероятности P_y и функции правдоподобия $p_y(x)$, и, кроме того, $\lambda_{yy} = 0$ и $\lambda_{ys} \equiv \lambda_y$ для всех $y, s \in Y$, то минимум среднего риска достигается алгоритмом*

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x). \tag{1.2}$$

Доказательство.

Рассмотрим выражение (1.1) из доказательства Теоремы 1.1. Поскольку λ_{ys} не зависит от второго индекса, то для любых $s, t \in Y$

$$\lambda_{ys} - \lambda_{yt} = \begin{cases} \lambda_t, & y = t; \\ -\lambda_s, & y = s; \\ 0, & \text{иначе.} \end{cases}$$

Следовательно, $\sum_{y \in Y} (\lambda_{ys} - \lambda_{yt}) P_y\, p_y(x) = \lambda_t P_t p_t(x) - \lambda_s P_s p_s(x) = \tilde{g}_t(x) - \tilde{g}_s(x)$, где $\tilde{g}_y(x) = \lambda_y P_y p_y(x)$ для всех $y \in Y$. Аналогично доказательству Теоремы 1.1 отсюда вытекает, что $a(x) = s$ при тех x, для которых $\tilde{g}_s(x)$ максимально по $s \in Y$. ■

Разделяющая поверхность между классами t и s — это геометрическое место точек $x \in X$ таких, что максимум в (1.2) достигается одновременно при $y = s$ и $y = t$:

$$\lambda_t P_t p_t(x) = \lambda_s P_s p_s(x).$$

Объекты x, удовлетворяющие этому уравнению, можно относить к любому из двух классов, что не повлияет на средний риск $R(a)$.

Отказ от классификации — это «особый ответ» $\varnothing \notin Y$, который в некоторых задачах разрешается выдавать вместо метки класса $y \in Y$. Отказ от классификации выгоден с точки зрения минимизации риска $R(a)$ в тех случаях, когда величина потери при отказе $\lambda_{y\varnothing}$ меньше величины потери от неверной классификации. При этом вместо разделяющей поверхности между классами возникает *разделяющая полоса* ненулевой ширины, см. Упражнение 1.1.

Апостериорная вероятность класса y для объекта x — это условная вероятность $\mathsf{P}(y|x)$. Она может быть вычислена по формуле Байеса, если известны $p_y(x)$ и P_y:

$$\mathsf{P}(y|x) = \frac{p(x,y)}{p(x)} = \frac{p_y(x) P_y}{\sum_{s \in Y} p_s(x) P_s}.$$

Во многих приложениях важно не только классифицировать объект x, но и сказать, с какой вероятностью $\mathsf{P}(y|x)$ он принадлежит каждому из классов $y \in Y$. Одно дело, когда объект x уверенно относится к одному из классов, и совсем другое — когда он находится на границе и может быть отнесён к нескольким классам. Через апостериорные вероятности выражается величина ожидаемых потерь на объекте x:

$$R(x) = \sum_{y \in Y} \lambda_y \mathsf{P}(y|x).$$

Принцип максимума апостериорной вероятности. Оптимальный алгоритм классификации (1.2) можно переписать через апостериорные вероятности:

$$a(x) = \arg\max_{y \in Y} \lambda_y \mathsf{P}(y|x).$$

Поэтому выражение (1.2) называют *байесовским решающим правилом*.

Если классы равнозначны ($\lambda_y \equiv 1$), то данное правило классификации называется *принципом максимума апостериорной вероятности*. Если классы ещё и равновероятны ($P_y \equiv \frac{1}{|Y|}$), то объект x просто относится к классу y с наибольшим значением плотности распределения $p_y(x)$ в точке x.

Можно было бы с самого начала принять принцип максимума апостериорной вероятности в качестве исходного постулата. Мы исходили из принципа минимума среднего риска, что позволило доказать оптимальность байесовского алгоритма и обобщить его на случай произвольной матрицы потерь $(\lambda_{ys})_{|Y|\times|Y|}$.

О тестировании методов обучения на модельных данных. Благодаря свойству оптимальности байесовское решающее правило удобно использовать в качестве эталона при тестировании методов обучения на модельных данных. Любой другой метод не может быть лучше байесовского; вопрос лишь в том, насколько он хуже.

Методика тестирования заключается в следующем. Задаются функции правдоподобия $p_y(x)$ и априорные вероятности P_y. Часто берут двумерные гауссовские плотности $p_y(x)$, чтобы отобразить и выборку, и разделяющую линию на плоском графике. Согласно распределению $P_y p_y(x)$ генерируются две выборки: обучающая $X^\ell = (x_i, y_i)_{i=1}^\ell$ и контрольная $X^k = (x'_i, y'_i)_{i=1}^k$. По обучающей выборке X^ℓ настраивается тестируемый алгоритм $a(x)$. По контрольной выборке вычисляется эмпирическая оценка среднего риска:

$$\hat{R}(a, X^k) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} \frac{1}{k} \sum_{i=1}^{k} [a(x'_i) = s][y'_i = y] = \frac{1}{k} \sum_{i=1}^{k} \lambda_{y'_i, a(x'_i)}.$$

Эта оценка является несмещённой, $\mathsf{E}_{X^k} \hat{R}(a, X^k) = R(a)$, поэтому длину контроля k можно подобрать так, чтобы $\hat{R}(a, X^k) \approx R(a)$ с любой заданной точностью.

Затем в явном виде выписывается оптимальный байесовский классификатор a^*, уравнение разделяющей поверхности (для отображения на графике), значение среднего риска $R(a^*)$. Если интегрирование функций $p_y(x)$ по областям A_s не удаётся выполнить аналитически, то вычисляется эмпирическая оценка $\hat{R}(a^*, X^k)$. Фактически, это вычисление тех же интегралов методом Монте-Карло. Тестируемый алгоритм считается пригодным, если эмпирическая оценка $\hat{R}(a, X^k)$ оказывается не сильно хуже байесовского среднего риска $R(a^*)$ или его эмпирической оценки $\hat{R}(a^*, X^k)$.

1.1.3 Задача восстановления плотности распределения

Перейдём к Задаче 1.2. Требуется оценить, какой могла бы быть плотность вероятностного распределения $p(x, y) = P_y p_y(x)$, сгенерировавшего выборку X^ℓ.

Обозначим подвыборку прецедентов класса y через $X_y^\ell = \big\{(x_i, y_i)_{i=1}^\ell \;\big|\; y_i = y\big\}$.

Проще всего оценить априорные вероятности классов P_y. Согласно закону больших чисел, частота появления объектов каждого из классов

$$\hat{P}_y = \frac{\ell_y}{\ell}, \quad \ell_y = |X^\ell_y|, \quad y \in Y, \tag{1.3}$$

сходится по вероятности к P_y при $\ell_y \to \infty$. Чем больше длина выборки, тем точнее выборочная оценка $\hat{P}_y$.

Оценка (1.3) является несмещённой лишь в том случае, если все без исключения наблюдавшиеся объекты заносились в обучающую выборку. На практике применяются и другие принципы формирования данных. Например, в задачах с *несбалансированными классами* (unbalanced classes) один из классов может встречаться в тысячи раз реже остальных; это может затруднять построение алгоритмов, поэтому выборку формируют неслучайным образом, чтобы объекты всех классов были представлены поровну. Возможна также ситуация, когда обучающая выборка формируется в ходе планируемого эксперимента, а применять построенный алгоритм классификации предполагается в реальной среде с другими априорными вероятностями классов. Во всех подобных ситуациях оценка $\hat{P}_y$ должна делаться не по доле обучающих объектов (1.3), а из каких-то других соображений.

Гораздо труднее оценить (восстановить) функции правдоподобия $\hat{p}_y(x)$ по выборкам X^ℓ_y, для каждого $y \in Y$. Задача восстановления плотности имеет самостоятельное значение, поэтому мы сформулируем её в более общем виде, обозначая выборку через X^m вместо X^ℓ_y, что позволит несколько упростить обозначения.

Задача 1.3. *Задано множество объектов $X^m = \{x_1, \ldots, x_m\}$, выбранных случайно и независимо согласно неизвестному распределению $p(x)$. Требуется построить эмпирическую оценку плотности — функцию $\hat{p}(x)$, приближающую $p(x)$ на всём X.*

Далее будут рассмотрены три подхода к восстановлению плотности, и, соответственно, три типа байесовских классификаторов: параметрический, непараметрический и основанный на разделении смеси распределений.

«Наивный» байесовский классификатор. Допустим, что объекты $x \in X$ описываются n числовыми признаками $f_j\colon X \to \mathbb{R}$, $j = 1, \ldots, n$. Обозначим через $x = (\xi_1, \ldots, \xi_n)$ произвольный элемент пространства объектов $X = \mathbb{R}^n$, где $\xi_j = f_j(x)$.

Гипотеза 1.1. *Признаки $f_1(x), \ldots, f_n(x)$ являются независимыми случайными величинами. Следовательно, функции правдоподобия классов представимы в виде*

$$p_y(x) = p_{y1}(\xi_1) \cdots p_{yn}(\xi_n), \qquad y \in Y, \tag{1.4}$$

где $p_{yj}(\xi_j)$ — плотность распределения значений j-го признака для класса y.

Предположение о независимости существенно упрощает задачу, так как оценить n одномерных плотностей гораздо легче, чем одну n-мерную плотность. К сожалению, оно крайне редко выполняется на практике. Поэтому алгоритмы, основанные на (1.4), называются *наивными байесовскими классификаторами* (naïve Bayes).

Пусть $\hat{p}_{yj}(\xi)$ — эмпирическая оценка плотности распределения признака f_j, вычисленная по подвыборке X^ℓ_y. Подставим эти оценки в (1.4) вместо истинных плотностей $p_{yj}(\xi)$, затем полученную эмпирическую плотность $\hat{p}_y(x)$ подставим в (1.2)

вместо истинной функции правдоподобия $p_y(x)$. Априорную вероятность каждого из классов P_y оценим как долю объектов класса y в выборке, $\hat{P}_y = \ell_y/\ell$.

В итоге получим алгоритм

$$a(x) = \arg\max_{y \in Y} \left(\ln \frac{\lambda_y \ell_y}{\ell} + \sum_{j=1}^{n} \ln \hat{p}_{yj}(\xi_j) \right). \tag{1.5}$$

Наивный байесовский классификатор может быть как параметрическим, так и непараметрическим, в зависимости от того, каким методом восстанавливаются одномерные плотности.

Основные его преимущества — простота реализации и низкие вычислительные затраты при обучении и классификации. В тех редких случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален.

Основной его недостаток — относительно низкое качество классификации в большинстве реальных задач. Чаще всего он используется либо как «примитивный» эталон для сравнения различных моделей алгоритмов, либо как элементарный «строительный блок» в алгоритмических композициях, см. главу **??**.

Преимущества байесовского подхода.

- Байесовское решающее правило оптимально, имеет простую формулу, легко реализуется программно. На его основе строятся многие методы классификации.
- При классификации объекта заодно оцениваются априорные вероятности его принадлежности каждому из классов. Эта информация необходима во многих приложениях для оценки рисков.
- Байесовское решающее правило удобно использовать в качестве эталона при тестировании алгоритмов классификации на модельных данных.

Недостатки байесовского подхода.

- На практике функции правдоподобия классов приходится восстанавливать по конечным выборкам данных. После подстановки восстановленной плотности в формулу (1.2) байесовский классификатор перестаёт быть оптимальным.
- Методов восстановления плотности известно довольно много. Однако ни один из них не является безусловно лучшим. На практике метод либо назначается априори, либо подбирается экспериментальным путём.

§1.2 Непараметрическая классификация

Непараметрические методы классификации основаны на локальном оценивании плотностей распределения классов $p_y(x)$ в окрестности классифицируемого объекта $x \in X$. Для классификации объекта x применяется основная формула (1.2).

Хотя такой подход не требует знания функционального вида плотностей, априорная информация всё равно привлекается. Например, в методе парзеновского окна предполагается, что в пространстве X задана метрика $\rho(x, x')$, адекватно оценивающая степень сходства объектов.

1.2.1 Непараметрические оценки плотности

Локальная аппроксимация опирается, фактически, только на само определение плотности распределения. Рассмотрим несколько случаев.

Дискретный случай. Пусть X — конечное множество, причём $|X| \ll m$. Оценкой плотности служит гистограмма значений x_i, встретившихся в выборке $X^m = (x_i)_{i=1}^m$:

$$\hat{p}(x) = \frac{1}{m} \sum_{i=1}^{m} [x_i = x]. \tag{1.6}$$

Эта оценка не применима, если $|X| \gg m$, и, тем более, в непрерывном случае, так как её значение почти всегда будет равно нулю.

Одномерный непрерывный случай. Пусть $X = \mathbb{R}$. Согласно определению плотности, $p(x) = \lim\limits_{h \to 0} \frac{1}{2h} P[x-h, x+h]$, где $P[a,b]$ — вероятностная мера отрезка $[a,b]$. Соответственно, эмпирическая оценка плотности определяется как доля точек выборки, лежащих внутри отрезка $[x-h, x+h]$, где h — неотрицательный параметр, называемый *шириной окна*:

$$\hat{p}_h(x) = \frac{1}{2mh} \sum_{i=1}^{m} \big[|x - x_i| < h\big]. \tag{1.7}$$

Функция $\hat{p}_h(x)$ является кусочно-постоянной. Это может приводить к возникновению довольно широких *зон неуверенности*, в которых максимум (1.2) достигается одновременно для нескольких классов y из Y. Проблема решается путём обобщения определения плотности. *Локальная непараметрическая оценка* Парзена-Розенблатта [15, 14] даёт сколь угодно гладкие оценки плотности:

$$\hat{p}_h(x) = \frac{1}{mh} \sum_{i=1}^{m} K\left(\frac{x - x_i}{h}\right), \tag{1.8}$$

где $K(z)$ — произвольная чётная функция, называемая *ядром*. Функция $\hat{p}_h(x)$ обладает той же степенью гладкости, что и ядро $K(z)$. Ядро $K(z)$ должно удовлетворять условию нормировки $\int K(z)\,dz = 1$. Тогда $\int \hat{p}_h(x)\,dx = 1$ при любом h, то есть функцию $\hat{p}_h(x)$ действительно можно интерпретировать как плотность вероятности.

На практике часто используются ядра, показанные на Рис. 5, стр. 14.

Прямоугольное ядро $K(z) = \frac{1}{2}\big[|z| < 1\big]$ соответствует простейшей оценке (1.7).

Точечное ядро $K(z) = [z = 0]$ при $h = 1$ соответствует дискретному случаю (1.6).

Следующая теорема даёт обоснование оценке Парзена-Розенблатта. Утверждается, что $\hat{p}_h(x)$ сходится к истинной плотности $p(x)$ для широкого класса ядер при увеличении длины выборки m и одновременном уменьшении ширины окна h.

Теорема 1.3 ([14, 15, 7]). *Пусть выполнены следующие условия:*

1) выборка X^m простая, получена из плотности распределения $p(x)$;

2) ядро $K(z)$ непрерывно, его квадрат ограничен: $\int_X K^2(z)\,dz < \infty$;

3) последовательность h_m такова, что $\lim\limits_{m\to\infty} h_m = 0$ и $\lim\limits_{m\to\infty} mh_m = \infty$.

Тогда $\hat{p}_{h_m}(x)$ сходится к $p(x)$ при $m \to \infty$ для почти всех $x \in X$, причём скорость сходимости имеет порядок $O(m^{-2/5})$.

Многомерный непрерывный случай. Пусть объекты описываются n числовыми признаками $f_j\colon X \to \mathbb{R}$, $j = 1, \ldots, n$. Тогда непараметрическая оценка плотности в точке $x \in X$ записывается в следующем виде [3, 5]:

$$\hat{p}_h(x) = \frac{1}{m} \sum_{i=1}^{m} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j}\right). \tag{1.9}$$

Таким образом, в каждой точке x_i многомерная плотность представляется в виде произведения одномерных плотностей. Заметим, что это никак не связано с «наивным» байесовским предположением о независимости признаков. При «наивном» подходе плотность представлялась бы как произведение одномерных парзеновских оценок (1.8), то есть как произведение сумм, а не как сумма произведений.

Произвольное метрическое пространство. Пусть на X задана функция расстояния $\rho(x, x')$, вообще говоря, не обязательно метрика. Одномерная оценка Парзена-Розенблатта (1.8) легко обобщается и на этот случай:

$$\hat{p}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^{m} K\left(\frac{\rho(x, x_i)}{h}\right), \tag{1.10}$$

где $V(h)$ — нормирующий множитель, гарантирующий, что $\hat{p}_h(x)$ действительно является плотностью:

$$V(h) = \int_X K\left(\frac{\rho(x, x_i)}{h}\right) dx.$$

Сходимость оценки (1.10) доказана при некоторых дополнительных ограничениях на ядро K и метрику ρ, и даже известны оценки скорости сходимости, не сильно отличающиеся от заявленных в Теореме 1.3 для одномерного случая [6].

Замечание 1.1. Пусть объекты описываются n числовыми признаками $f_j\colon X \to \mathbb{R}$, $j = 1, \ldots, n$. На практике в качестве функции расстояния чаще всего используется *взвешенная метрика Минковского*:

$$\rho(x, x') = \left(\sum_{j=1}^{n} w_j \left|f_j(x) - f_j(x')\right|^p\right)^{\frac{1}{p}},$$

где w_j — неотрицательные веса признаков, степень p неотрицательна. В частности, если все веса одинаковы и $p = 2$, то имеем обычную евклидову метрику. Веса служат двум целям. Во-первых, они нужны для нормировки признаков. Если у одного признака типичные значения — десятки тысяч, а у всех остальных — сотые доли, то сравнение объектов будет основываться фактически только на значениях первого признака, и многомерная информация об объектах будет утрачена. Во-вторых, веса можно использовать для выделения более информативных признаков. Чем больше w_j, тем сильнее влияние j-го признака на функцию расстояния. Оптимизация весов признаков является нетривиальной задачей. К сожалению, на практике часто ограничиваются предварительной нормировкой признаков, полагая затем $w_j \equiv 1$. Это дилетантский подход; он быстро приводит к решению задачи, но это решение в некоторых случаях оказывается очень плохим, особенно когда среди признаков имеется много неинформативных (шумовых).

Замечание 1.2. Чтобы определение нормирующего множителя $V(h)$ было корректно, значение интеграла не должно зависеть от x_i. Фактически, это требование однородности пространства X. Действительно, интеграл $V(h)$ есть объём шара с центром в точке x_i и радиусом h, «размытого» с помощью ядра $K\big(\frac{\rho(x,x_i)}{h}\big)$. Этот объём не должен зависеть от того, в какую точку x_i пространства X помещён центр шара. Данное требование не является обременительным, поскольку меру на множестве X можно ввести как угодно (речь идёт не о вероятностной мере, а о некоторой «естественной мере», изначально присущей пространству X; вероятностная мера связана с ней через функцию плотности распределения). В частности, числовое пространство $\mathbb{R}^n$ удовлетворяет требованию однородности.

1.2.2 Метод парзеновского окна

Запишем парзеновскую оценку плотности (1.10) для каждого класса $y \in Y$:

$$\hat{p}_{y,h}(x) = \frac{1}{\ell_y V(h)} \sum_{i=1}^{\ell} [y_i = y]\, K\left(\frac{\rho(x, x_i)}{h}\right), \tag{1.11}$$

где K — ядро, h — ширина окна. Если нормирующий множитель $V(h)$ не зависит от x_i и y, то в байесовском решающем правиле (1.2) его можно убрать из-под знака $\arg\max$ и вообще не вычислять. Подставим оценку плотности (1.11) и оценку априорной вероятности классов $\hat{P}_y = \ell_y/\ell$ в формулу (1.2):

$$a(x; X^\ell, h) = \arg\max_{y \in Y} \lambda_y \sum_{i=1}^{\ell} [y_i = y]\, K\left(\frac{\rho(x, x_i)}{h}\right). \tag{1.12}$$

Выборка X^ℓ сохраняется «как есть» и играет роль параметра алгоритма.

Это очень простой алгоритм. Его можно было бы придумать из чисто эвристических соображений, не прибегая к вероятностной модели данных.

Сначала вводится функция $B(x, x_i) = K\big(\rho(x, x_i)/h\big)$, оценивающая *близость* произвольного объекта x к обучающему объекту x_i. Чем больше расстояние, тем меньше близость, поэтому функция K должна быть невозрастающей. Параметр h регулирует скорость убывания близости с ростом расстояния.

Затем вводится функция $\Gamma_y(x) = \sum\limits_{x_i \in X_y^\ell} B(x, x_i)$, оценивающая суммарную близость объекта x к классу y.

Наконец, классификатор $a(x)$ относит произвольный объект $x \in X$ к классу с наибольшей оценкой суммарной близости: $a(x) = \arg\max_y \Gamma_y(x)$.

Заметим, что именно так вводятся *метрические алгоритмы* классификации (см. **??**) и алгоритм вычисления оценок [4] (см. **??**). При этом функция близости B не обязана быть нормированной плотностью и вообще иметь какую-либо вероятностную интерпретацию.

Если метрика ρ фиксирована, то обучение парзеновского классификатора (1.12) сводится к подбору ширины окна h и вида ядра K.

Ширина окна h решающим образом влияет на качество восстановления плотности. При слишком узком окне ($h \to 0$) плотность концентрируется вблизи обучающих

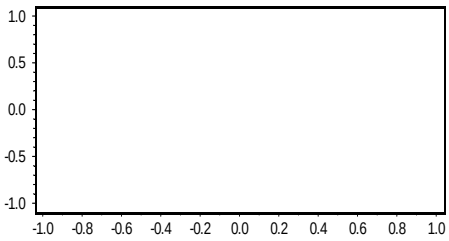

Рис. 1. Локальные оценки плотности при заниженном, завышенном и оптимальном h.

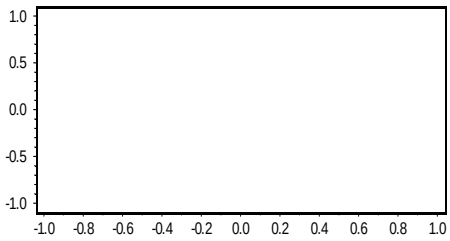

Рис. 2. Оптимизация ширины окна по максимуму LOO-правдоподобия.

объектов, и функция $\hat{p}_h(x)$ претерпевает резкие скачки. При слишком широком окне плотность чрезмерно сглаживается и в пределе $h \to \infty$ вырождается в константу, рис. **??**. Таким образом, должно существовать оптимальное значение ширины окна h^*, при котором достигается компромисс между точностью описания выборки и гладкостью эмпирической плотности $\hat{p}_h(x)$.

Чтобы найти оптимальное h^*, воспользуемся принципом максимума правдоподобия *с исключением объектов по одному* (leave-one-out, LOO):

$$h^* = \arg\max_h \sum_{i=1}^{\ell} \log \hat{p}_h\big(x_i; X^m \setminus x_i\big).$$

Здесь запись $\hat{p}_h\big(x_i; X^m \setminus x_i\big)$ означает, что локальная оценка плотности $\hat{p}_h(x)$ в точке x_i строится по всей выборке X^m за исключением самой точки x_i. Если саму точку не исключать, то максимум правдоподобия будет достигаться при $h \to 0$, то есть при чрезмерно точной подгонке плотности под выборку. Скользящий контроль позволяет выбрать то значение h, при котором локальная оценка плотности $\hat{p}_h(x)$ наилучшим образом описывает новые данные.

Отметим, что в алгоритме классификации (1.12) ширина окна h не должна зависеть от класса. Поэтому в качестве X^m надо брать всю выборку X^ℓ, а не подвыборки отдельных классов X^ℓ_y.

Возможен и другой способ настройки ширины окна — когда принцип скользящего контроля применяется непосредственно к алгоритму классификации:

$$\text{LOO}(h, X^\ell) = \sum_{i=1}^{\ell} \big[a\big(x_i; X^\ell \setminus x_i, h\big) \neq y_i\big] \to \min_h,$$

где $a(x; X^\ell \setminus x_i, h)$ — алгоритм классификации, построенный по обучающей выборке X^ℓ без объекта x_i. Обычно зависимость LOO от h имеет характерный минимум, соответствующий оптимальной ширине окна h^*, рис. 2.

Проблема локальных сгущений возникает в тех случаях, когда распределение объектов в пространстве X сильно неравномерно, и одно и то же значение ширины окна h приводит к чрезмерному сглаживанию плотности в одних местах, и недостаточному сглаживанию в других, рис. 3. Проблему решают *окна переменной ширины*. Ширина окна определяется в каждой точке $x \in X$ как расстояние до $(k+1)$-го соседа $h(x) = \rho\big(x, x^{(k+1)}\big)$, где $x^{(i)}$ — i-й сосед объекта x, если считать, что обучающие объекты ранжированы в порядке возрастания расстояний до x.

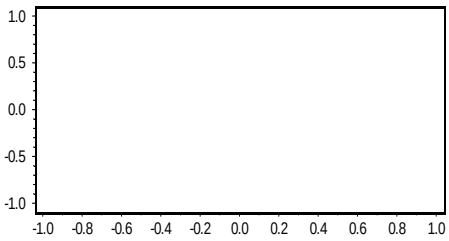

Рис. 3. Локальные оценки плотности с случае локальных сгущений: при заниженном, завышенном и оптимальном числе соседей k.

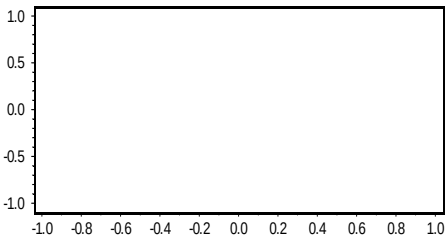

Рис. 4. Оптимизация переменной ширины окна (числа соседей) по функционалу скользящего контроля.

Если ядро $K(r)$ имеет ограниченный носитель $[-1,+1]$, то оценка плотности $\hat{p}_{h(x)}$ для любого x будет зависеть только от k его ближайших соседей. Чем выше локальная плотность объектов в окрестности x, тем меньшей будет ширина окна. В случае прямоугольного ядра этот алгоритм относит объект x к тому классу, которому принадлежит большинство из его k соседей. Это правило классификации называется методом k *ближайших соседей* (k nearest neighbors, kNN).

Целочисленный параметр k определяет компромисс между точностью описания данных и гладкостью функции плотности $\hat{p}_{h(x)}$. Оптимальное значение k^*, аналогично h^*, определяется по критерию скользящего контроля, рис. 4.

Замечание 1.3. Когда ширина окна $h = h(x)$ зависит от классифицируемого объекта x, нормирующий множитель $V(h)$ также становится функцией x. Из требования независимости $V(h(x))$ от y вытекает, что в каждой точке x для всех классов должна использоваться одна и та же ширина окна $h(x)$. Поэтому при вычислении $h(x)$ должны учитываться все объекты выборки, независимо от их классовой принадлежности. В то же время, плотности $\hat{p}_{y,h(x)}(x)$ оцениваются по подвыборкам X_y^ℓ, для каждого класса $y \in Y$ в отдельности.

Функция ядра K практически не влияет на точность восстановления плотности и на качество классификации. Часто используемые ядра показаны на Рис. 5 и в Таблице 1. В последней колонке приведены (для одномерного случая) численные оценки функционала качества восстановления плотности

$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

Минимальное значение $J(K)$, равное J^*, достигается для ядра Епанечникова $E(r)$, которое является оптимальным. Другие ядра доставляют функционалу $J(K)$ значения, лишь немного худшие J^*. Это и позволяет утверждать, что форма ядра практически не влияет на качество восстановления плотности.

В то же время, вид ядра определяющим образом влияет на степень гладкости функции $\hat{p}_h(x)$, см. третью колонку Таблицы 1.

Вид ядра может также влиять на эффективность вычислений. Гауссовское ядро G требует просмотра всей выборки для вычисления значения $\hat{p}_h(x)$ в произвольной точке x. Ядра $E, Q, T, П$ являются финитными (имеют ограниченный носитель), и для них достаточно взять только те точки выборки, которые попадают в окрестность точки x радиуса h.

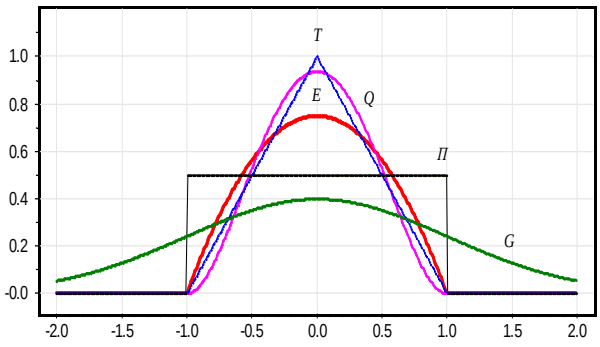

Рис. 5. Часто используемые ядра:
E — Епанечникова;
Q — Квартическое;
T — Треугольное;
G — Гауссовское;
$П$ — прямоугольное.

ядро $K(r)$	формула	степень гладкости	$J^*/J(K)$
Епанечникова	$E(r) = \frac{3}{4}(1-r^2)\big[\lvert r\rvert \leqslant 1\big]$	$\hat{p}'_h$ разрывна	1.000
Квартическое	$Q(r) = \frac{15}{16}(1-r^2)^2\big[\lvert r\rvert \leqslant 1\big]$	$\hat{p}''_h$ разрывна	0.995
Треугольное	$T(r) = \big(1-\lvert r\rvert\big)\big[\lvert r\rvert \leqslant 1\big]$	$\hat{p}'_h$ разрывна	0.989
Гауссовское	$G(r) = (2\pi)^{-1/2}\exp(-\frac{1}{2}r^2)$	∞ дифференцируема	0.961
Прямоугольное	$П(r) = \frac{1}{2}\big[\lvert r\rvert \leqslant 1\big]$	$\hat{p}_h$ разрывна	0.943

Таблица 1. Гладкость и качество восстановления плотности для часто используемых ядер.

Проблема «проклятия размерности». Если используемая метрика $\rho(x, x')$ основана на суммировании различий по всем признакам, а число признаков очень велико, то все точки выборки могут оказаться практически одинаково далеки друг от друга. Тогда парзеновские оценки плотности становятся неадекватны. Это явление называют *проклятием размерности* (curse of dimensionality). Выход заключается в понижении размерности с помощью преобразования пространства признаков (см. раздел **??**), либо путём отбора информативных признаков (см. раздел **??**). Можно строить несколько альтернативных метрик в подпространствах меньшей размерности, и полученные по ним алгоритмы классификации объединять в композицию. На этой идее основаны алгоритмы вычисления оценок, подробно описанные в **??**.

§1.3 Нормальный дискриминантный анализ

В *параметрическом подходе* к восстановлению плотности $p(x)$ по выборке X^m предполагается, что плотность известна с точностью до параметра, $p(x) = \varphi(x;\theta)$, где φ — фиксированная функция. Тогда оптимальное значение вектора параметров θ можно оценить по выборке, исходя из *принципа максимума правдоподобия* (**??**). В общем случае используется функционал *взвешенного правдоподобия*, когда для каждого объекта x_i задаётся неотрицательный *вес* или *степень важности* g_i:

$$L(X^m, G^m; \theta) = \sum_{i=1}^{m} g_i \ln \varphi(x_i;\theta) \to \max_{\theta}, \tag{1.13}$$

где $G^m = (g_1, \ldots, g_m)$ — вектор весов объектов. Для решения этой задачи можно использовать стандартные методы оптимизации. В некоторых случаях удаётся вы-

писать решение в явном виде, исходя из необходимого условия оптимума:

$$\frac{\partial}{\partial\theta}L(X^m, G^m; \theta) = \sum_{i=1}^{m} g_i \frac{\partial}{\partial\theta} \ln \varphi(x_i; \theta) = 0, \tag{1.14}$$

где функция $\varphi(x;\theta)$ достаточно гладкая по параметру θ. В частности, задача решается аналитически, когда $\varphi(x;\theta)$ — многомерное нормальное распределение. Рассмотрим этот классический случай подробно.

Опр. 1.2. *Вероятностное распределение с плотностью*

$$\mathcal{N}(x;\mu,\Sigma) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}} \exp\bigl(-\tfrac{1}{2}(x-\mu)^\intercal\Sigma^{-1}(x-\mu)\bigr), \quad x \in \mathbb{R}^n,$$

называется n-мерным нормальным (гауссовским) распределением с вектором математического ожидания (центром) $\mu \in \mathbb{R}^n$ и ковариационной матрицей $\Sigma \in \mathbb{R}^{n\times n}$. Предполагается, что матрица Σ симметричная, невырожденная и положительно определённая.

Интегрируя по $\mathbb{R}^n$, нетрудно убедиться в том, что параметры распределения μ и Σ оправдывают своё название:

$$\mathsf{E}x = \int x\mathcal{N}(x;\mu,\Sigma)\,dx = \mu;$$
$$\mathsf{E}(x-\mu)(x-\mu)^\intercal = \int (x-\mu)(x-\mu)^\intercal\mathcal{N}(x;\mu,\Sigma)\,dx = \Sigma.$$

Геометрическая интерпретация нормальной плотности. Если признаки некоррелированы, $\Sigma = \operatorname{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, то линии уровня плотности распределения имеют форму эллипсоидов с центром μ и осями, параллельными линиям координат. Если признаки имеют одинаковые дисперсии, $\Sigma = \sigma^2 I_n$, то эллипсоиды являются сферами.

Если признаки коррелированы, то матрица Σ не диагональна и линии уровня имеют форму эллипсоидов, оси которых повёрнуты относительно исходной системы координат. Действительно, как всякая симметричная матрица, Σ имеет спектральное разложение $\Sigma = VSV^\intercal$, где $V = (v_1,\ldots,v_n)$ — ортогональные собственные векторы матрицы Σ, соответствующие собственным значениям $\lambda_1,\ldots,\lambda_n$, матрица S диагональна, $S = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$. Тогда $\Sigma^{-1} = VS^{-1}V^\intercal$, следовательно,

$$(x-\mu)^\intercal\Sigma^{-1}(x-\mu) = (x-\mu)^\intercal VS^{-1}V^\intercal(x-\mu) = (x'-\mu')^\intercal S^{-1}(x'-\mu').$$

Это означает, что в результате ортогонального преобразования координат $x' = V^\intercal x$ оси эллипсоидов становятся параллельны линиям координат. В новых координатах ковариационная матрица S является диагональной. Поэтому линейное преобразование V называется *декоррелирующим*. В исходных координатах оси эллипсоидов направлены вдоль собственных векторов матрицы Σ.

Линейные и квадратичные разделяющие поверхности. Рассмотрим задачу классификации, в которой объекты описываются n вещественными признаками $f_j\colon X \to \mathbb{R}$, $j = 1,\ldots,n$, следовательно, задаются n-мерными векторами, $X = \mathbb{R}^n$. Число классов $|Y|$ произвольно (два или более).

Гипотеза 1.2. *Классы имеют n-мерные нормальные плотности распределения*

$$p_y(x) = \mathcal{N}(x; \mu_y, \Sigma_y), \;\; y \in Y.$$

Теорема 1.4. *Если классы имеют нормальные функции правдоподобия, то байесовское решающее правило имеет квадратичную разделяющую поверхность. Квадратичная поверхность вырождается в линейную тогда и только тогда, когда ковариационные матрицы классов равны.*

Доказательство.

Поверхность, разделяющая классы s и t, описывается уравнением $\lambda_s P_s p_s(x) = \lambda_t P_t p_t(x)$, или, после логарифмирования

$$\ln p_s(x) - \ln p_t(x) = C_{st},$$

где $C_{st} = \ln(\lambda_t P_t / \lambda_s P_s)$ — константа, не зависящая от x. Разделяющая поверхность в общем случае квадратична, поскольку $\ln p_y(x)$ является квадратичной формой по x:

$$\ln p_y(x) = -\tfrac{n}{2} \ln 2\pi - \tfrac{1}{2} \ln |\Sigma_y| - \tfrac{1}{2} (x - \mu_y)^\intercal \Sigma_y^{-1} (x - \mu_y).$$

Если $\Sigma_s = \Sigma_t \equiv \Sigma$, то квадратичные члены сокращаются и уравнение поверхности вырождается в линейную форму:

$$x^\intercal \Sigma^{-1}(\mu_s - \mu_t) - \tfrac{1}{2} \mu_s^\intercal \Sigma^{-1} \mu_s + \tfrac{1}{2} \mu_t^\intercal \Sigma^{-1} \mu_t = C_{st};$$
$$(x - \mu_{st})^\intercal \Sigma^{-1}(\mu_s - \mu_t) = C_{st};$$

где $\mu_{st} = \frac{1}{2}(\mu_s + \mu_t)$ — точка посередине между центрами классов. ■

Геометрия разделяющих поверхностей. Простейший случай: классы равновероятны и равнозначны, ковариационные матрицы равны, признаки некоррелированы и имеют одинаковые дисперсии. Это означает, что классы имеют одинаковую сферическую форму. В этом случае разделяющая гиперплоскость проходит посередине между классами, ортогонально линии, соединяющей центры классов. Нормаль гиперплоскости обладает оптимальным свойством: в одномерной проекции на нормаль классы разделяются наилучшим образом.

Усложнение 1: признаки коррелированы. Тогда ортогональность исчезает, однако разделяющая гиперплоскость по-прежнему проходит посередине между классами, касательно к линиям уровня обоих распределений.

Усложнение 2: классы не равновероятны или не равнозначны. Тогда разделяющая гиперплоскость отодвигается дальше от более значимого класса.

Усложнение 3: ковариационные матрицы общего вида (не диагональны) и не равны. Тогда разделяющая поверхность становится квадратичной и «прогибается» так, чтобы менее плотный класс охватывал более плотный.

В некоторых случаях более плотный класс «разрезает» менее плотный на две несвязные области. Это может приводить к парадоксальной ситуации: возникает область, в которой не было ни одного обучающего прецедента, тем не менее, попадающие в неё объекты относятся к более далёкому классу.

Усложнение 4: Если число классов превышает 2, то разделяющая поверхность является кусочно-квадратичной, а при равных ковариационных матрицах — кусочно-линейной.

Расстояние Махаланобиса. Если классы равновероятны и равнозначны, ковариационные матрицы равны, то уравнение разделяющей поверхности принимает вид

$$(x-\mu_s)^\intercal\Sigma^{-1}(x-\mu_s) = (x-\mu_t)^\intercal\Sigma^{-1}(x-\mu_t);$$
$$\|x-\mu_s\|_\Sigma = \|x-\mu_t\|_\Sigma;$$

где $\|u-v\|_\Sigma \equiv \sqrt{(u-v)^\intercal\Sigma^{-1}(u-v)}$ — метрика в $\mathbb{R}^n$, называемая *расстоянием Махаланобиса*. Разделяющая поверхность является геометрическим местом точек, равноудалённых от центров классов в смысле расстояния Махаланобиса.

Если признаки независимы и имеют одинаковые дисперсии, то расстояние Махаланобиса совпадает с обычной евклидовой метрикой. В этом случае оптимальным (байесовским) решающим правилом является «относить объект к классу с ближайшим центром». Это алгоритм называют *классификатором ближайшего среднего* (nearest mean classifier).

Выборочные оценки параметров нормального распределения. В случае гауссовской плотности с параметрами $\theta \equiv (\mu, \Sigma)$ задача максимизации правдоподобия имеет аналитическое решение, основанное на соотношениях (1.14).

Теорема 1.5. *Пусть задана случайная, независимая, одинаково распределённая выборка наблюдений $X^m = (x_1, \ldots, x_m)$ и вектор весов объектов $G^m = (g_1, \ldots, g_m)$ при условии нормировки $\sum_{i=1}^m g_i = 1$. Тогда оценки параметров гауссовской плотности $\varphi(x;\theta) \equiv \mathcal{N}(x;\mu,\Sigma)$, доставляющие максимум взвешенному функционалу правдоподобия (1.13), имеют вид*

$$\hat{\mu} = \sum_{i=1}^m g_i x_i; \qquad \hat{\Sigma} = \sum_{i=1}^m g_i (x_i-\hat{\mu})(x_i-\hat{\mu})^\intercal.$$

Доказательство вынесено в Упражнение 1.6.

Следствие 1. В условиях предыдущей теоремы оценки параметров гауссовской плотности $\varphi(x;\theta) \equiv \mathcal{N}(x;\mu,\Sigma)$, доставляющие максимум (не взвешенному) функционалу правдоподобия (**??**), имеют вид

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^m x_i; \qquad \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^m (x_i-\hat{\mu})(x_i-\hat{\mu})^\intercal.$$

Поправка на смещение. Естественным требованием к оценке параметра распределения является её несмещённость.

Опр. 1.3. *Пусть X^m есть выборка случайных независимых наблюдений, полученная согласно распределению $\varphi(x;\theta)$ при фиксированном $\theta = \theta_0$. Оценка $\hat{\theta}(X^m)$ параметра θ, вычисленная по выборке X^m, называется несмещённой, если $\mathsf{E}_{X^m}\hat{\theta}(X^m) = \theta_0$.*

Легко убедиться в том, что $\hat{\mu}$ является несмещённой оценкой матожидания μ:

$$\mathsf{E}\hat{\mu} = \mathsf{E}\frac{1}{m}\sum_{i=1}^m x_i = \frac{1}{m}\sum_{i=1}^m \mathsf{E}x_i = \mathsf{E}x = \mu.$$

Аналогично можно показать, что

$$\mathsf{E}\frac{1}{m}\sum_{x=1}^{m}(x_i-\mu)(x_i-\mu)^\intercal = \mathsf{E}xx^\intercal - \mu\mu^\intercal = \Sigma.$$

Однако эта величина не равна $\mathsf{E}\hat{\Sigma}$, ведь при вычислении $\hat{\Sigma}$ вместо неизвестного точного значения матожидания μ подставляется его выборочная оценка $\hat{\mu}$. Аккуратный расчёт показывает, что $\hat{\Sigma}$ является смещённой (несколько заниженной) оценкой Σ.

Теорема 1.6. *Несмещённая оценка ковариационной матрицы имеет вид*

$$\hat{\Sigma} = \frac{1}{m-1}\sum_{x=1}^{m}(x_i-\hat{\mu})(x_i-\hat{\mu})^\intercal. \tag{1.15}$$

Доказательство вынесено в Упражнение 1.8.

1.3.1 Квадратичный дискриминант

В задачах классификации с гауссовскими классами (Гипотеза 1.2) параметры функций правдоподобия $\hat{\mu}_y$ и $\hat{\Sigma}_y$ можно оценить по частям обучающей выборки

$$X_y^\ell = \{x_i \in X^\ell \mid y_i = y\}, \quad y \in Y,$$

для каждого класса y отдельно. Априорные вероятности классов P_y оцениваются согласно (1.3). Полученные выборочные оценки непосредственно подставляются в формулу (1.2). В результате получается алгоритм классификации, который так и называется — *подстановочным* (plug-in).

В асимптотике $\ell_y \to \infty$ оценки $\hat{\mu}_y$ и $\hat{\Sigma}_y$ обладают рядом оптимальных свойств: они не смещены, состоятельны и эффективны. Однако в условиях конечных, зачастую слишком коротких, выборок асимптотические свойства не гарантируют точного восстановления функций правдоподобия, и, следовательно, высокого качества классификации. Приходится изобретать различные эвристические «подпорки», чтобы довести алгоритм до состояния практической пригодности.

Недостатки подстановочного алгоритма.

- Если длина выборки меньше размерности пространства, $\ell_y < n$, то матрица $\hat{\Sigma}_y$ становится вырожденной, поскольку её ранг не может превышать ℓ_y. В этом случае обратная матрица не существует и метод вообще неприменим.

- Даже если длина выборки больше размерности пространства, $\ell_y > n$, матрица $\hat{\Sigma}_y$ всё равно может оказаться вырожденной. Это происходит, когда среди признаков есть линейно зависимые. Обнаружить их в общем случае не так просто; для этого нужны специальные численные методы. Это так называемая *проблема мультиколлинеарности*.

- Даже если среди признаков нет линейно зависимых, матрица $\hat{\Sigma}_y$ может оказаться *плохо обусловленной*, то есть близкой к некоторой вырожденной матрице. Обратная к такой матрице неустойчива, что влечёт несколько неприятностей.

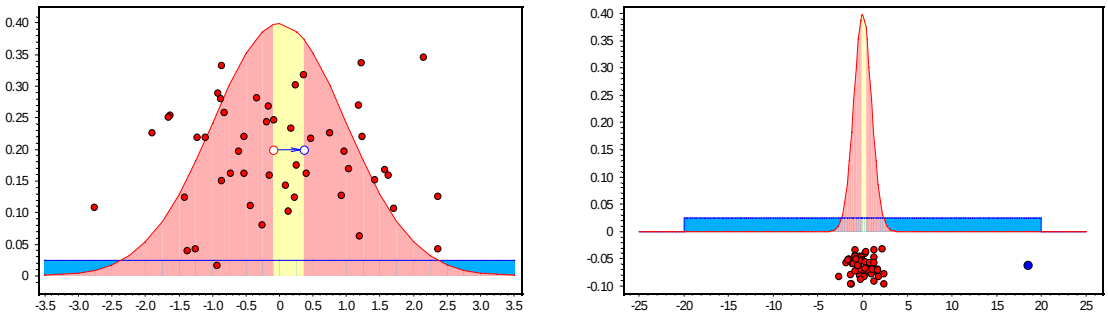

Рис. 6. Одномерная нормальная плотность $\mathcal{N}(0,1)$, загрязнённая равномерным на $[-20,+20]$ распределением. Единственный «выброс» на выборке длины $\ell = 50$ (синяя точка на графике слева) приводит к смещению эмпирического среднего в точку 0.359, значимо отличающуюся от истинного среднего, равного нулю (что хорошо видно на графике справа).

Во-первых, матрица $\hat{\Sigma}_y^{-1}$ может практически непредсказуемо изменяться при незначительных вариациях обучающей выборки. Во-вторых, вектор $\hat{\Sigma}_y^{-1}(x-\mu_{st})$ может практически непредсказуемо изменяться при незначительных вариациях классифицируемого объекта x. Если не предпринимать специальных мер против плохой обусловленности, классификатор будет допускать слишком много ошибок.

- Выборочные оценки чувствительны к нарушениям нормальности распределений, в частности, к редким большим выбросам. В 1960 году Дж. Тьюки показал, что классическая оценка матожидания нормального распределения неустойчива относительно сколь угодно малого ε-загрязнения плотности даже в одномерном случае [10] (загрязнённая плотность имеет вид $(1-\varepsilon)N(x)+\varepsilon\delta(x)$, где $N(x)$ — плотность нормального распределения, $\delta(x)$ — плотность загрязнения). Загрязнения с «тяжёлым хвостом» приводят к появлению редких больших выбросов и значительному смещению оценки матожидания, Рис. 6. При увеличении размерности влияние загрязнений только усиливается.

- Если функции правдоподобия классов существенно отличаются от гауссовских, то методы нормального дискриминантного анализа могут приводить к алгоритмам низкого качества. В частности, когда имеются номинальные признаки, принимающие дискретные значения, или когда классы распадаются на изолированные сгустки.

Прежде, чем обсуждать способы устранения перечисленных недостатков, рассмотрим один важный частный случай.

Наивный байесовский классификатор. Предположим, что все признаки $f_j(x)$ независимы и нормально распределены с матожиданием μ_{yj} и дисперсией σ_{yj}, вообще говоря, отличающимися для разных классов:

$$p_{yj}(\xi) = \frac{1}{\sqrt{2\pi}\sigma_{yj}} \exp\left(-\frac{(\xi-\mu_{yj})^2}{2\sigma_{yj}^2}\right), \quad y \in Y, \quad j = 1,\ldots,n.$$

Тогда, как нетрудно убедиться, ковариационные матрицы Σ_y и их выборочные оценки $\hat{\Sigma}_y$ будут диагональны. В этом случае проблемы вырожденности и мультиколлинеарности не возникают. Метод обучения приобретает до крайности простой вид

и сводится к вычислению параметров $\hat{\mu}_{yj}$ и $\hat{\sigma}_{yj}$ для каждого класса $y \in Y$ и каждого признака $j = 1, \ldots, n$. Доказательство вынесено в Упражнение 1.9.

1.3.2 Линейный дискриминант Фишера

В 1936 г. Р. Фишер предложил простую эвристику, позволяющую увеличить число объектов, по которым оценивается ковариационная матрица, повысить её устойчивость и заодно упростить алгоритм обучения [12]. Эвристика заключается в том, чтобы считать ковариационные матрицы классов равными, даже если они на самом деле не равны. В таком случае достаточно оценить только одну ковариационную матрицу $\hat{\Sigma}$, задействовав для этого все ℓ обучающих объектов. При этом разделяющая поверхность является линейной, если классов два, и кусочно-линейной, если классов больше. Линейные коэффициенты получаются непосредственно из (1.2):

$$\begin{aligned} a(x) &= \arg\max_{y \in Y} \big(\lambda_y P_y p_y(x)\big) = \\ &= \arg\max_{y \in Y} \big(\underbrace{\ln(\lambda_y P_y) - \tfrac{1}{2}\hat{\mu}_y^{\intercal}\hat{\Sigma}^{-1}\hat{\mu}_y}_{\beta_y} + x^{\intercal}\underbrace{\hat{\Sigma}^{-1}\hat{\mu}_y}_{\alpha_y}\big) = \\ &= \arg\max_{y \in Y} \big(x^{\intercal}\alpha_y + \beta_y\big). \end{aligned} \tag{1.16}$$

Обучение сводится к оцениванию матожиданий $\hat{\mu}_y$ для всех $y \in Y$, вычислению общей ковариационной матрицы $\hat{\Sigma}$ и её обращению, см. Алгоритм 1.1. После обучения классификация новых объектов производится по формуле (1.16). Этот алгоритм называется *линейным дискриминантом Фишера* (ЛДФ). Эвристика Фишера неплохо работает, когда формы классов близки к нормальным и не слишком сильно различаются. В этом случае линейное решающее правило близко к оптимальному байесовскому, но существенно более устойчиво, чем квадратичное, и часто обладает лучшей обобщающей способностью.

Замечание 1.4. Формула шага 2 Алгоритма 1.1 отличается от оценки (1.15) поправкой на смещённость, которая вычитается в знаменателе. Можно доказать, что эта поправка равна числу параметров $\hat{\mu}_y$, $y \in Y$, которые оцениваются по выборке и используются при вычислении оценки $\hat{\Sigma}$, см. Упражнение 1.8 в конце главы.

Регуляризация ковариационной матрицы. Общая ковариационная матрица классов $\hat{\Sigma}$ может оказаться плохо обусловленной (близкой к вырожденной), если длина выборки невелика по сравнению с числом признаков, или если среди признаков есть почти линейно зависимые. В этом случае некоторые собственные значения матрицы $\hat{\Sigma}$ будут близки к нулю, обратная матрица и разделяющая поверхность станут неустойчивыми.

Вспомним, что линии уровня гауссовской плотности имеют форму эллипсоидов. Собственные векторы матрицы $\hat{\Sigma}$ задают направления осей эллипсоида. Собственные значения определяют «толщину» эллипсоида вдоль его осей. Существует простой способ увеличить все собственные значения матрицы $\hat{\Sigma}$ на одну и ту же величину τ, оставив неизменными собственные векторы. При этом «форма» распределения немного искажается, зато матрица становится хорошо обусловленной.

Алгоритм 1.1. Обучение линейного дискриминанта Фишера

Вход:
выборка X^ℓ, предполагается $\ell > |Y|$;
величины потерь λ_y, $y \in Y$;

Выход:
коэффициенты линейных разделяющих поверхностей $\alpha_y \in \mathbb{R}^n$, $\beta_y \in \mathbb{R}$, $y \in Y$;

1: $\ell_y := \sum_{i=1}^{\ell} [y_i = y], \quad \hat{P}_y := \ell_y/\ell, \quad \hat{\mu}_y := \frac{1}{\ell_y} \sum_{i=1}^{\ell} [y_i = y] x_i, \quad$ для всех $y \in Y$;

2: $\hat{\Sigma} := \frac{1}{\ell - |Y|} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^\intercal$;

3: $\alpha_y := \hat{\Sigma}^{-1} \hat{\mu}_y, \quad \beta_y := \ln(\lambda_y \hat{P}_y) - \frac{\hat{\mu}_y^\intercal \alpha_y}{2}, \quad$ для всех $y \in Y$;

Пусть v — собственный вектор матрицы $\hat{\Sigma}$, соответствующий собственному значению λ, $\hat{\Sigma} v = \lambda v$. Тогда v является также собственным вектором матрицы $\hat{\Sigma} + \tau I_n$ с собственным значением $\lambda + \tau$, где I_n — единичная матрица размера $n \times n$:

$$(\hat{\Sigma} + \tau I_n) v = \lambda v + \tau v = (\lambda + \tau) v.$$

Таким образом, проблема плохой обусловленности может быть решена путём обращения матрицы $\hat{\Sigma} + \tau I_n$ вместо $\hat{\Sigma}$.

Известны и другие рекомендации.

Можно пропорционально уменьшать недиагональные элементы — вместо $\hat{\Sigma}$ брать матрицу $(1 - \tau)\hat{\Sigma} + \tau \operatorname{diag} \hat{\Sigma}$ [10].

Можно занулять недиагональные элементы матрицы, соответствующие тем парам признаков, ковариации которых незначимо отличаются от нуля [2]. Матрица становится разреженной, и для её обращения могут применяться специальные, более эффективные, алгоритмы.

Для проверки на равенство нулю элементов σ_{ij} ковариационной матрицы $\hat{\Sigma}$ применяется критерий Стьюдента. Для всех $i, j = 1, \ldots, n$, $i < j$, вычисляется коэффициент корреляции $r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$, затем статистика $T = \frac{r_{ij}\sqrt{n-2}}{\sqrt{1-r_{ij}^2}}$. Эта статистика имеет t-распределение Стьюдента с $n - 2$ степенями свободы (оно симметрично и в пределе $n \to \infty$ стремится к нормальному распределению с нулевым ожиданием и единичной дисперсией). Если при заданном уровне значимости α выполняется условие $|T| \leqslant t_{1-\frac{\alpha}{2}}$, где $t_{1-\frac{\alpha}{2}}$ — квантиль распределения Стьюдента, то считается, что элемент ковариационной матрицы σ_{ij} незначимо отличается от нуля и потому *полагается равным* нулю. На практике обычно берут $\alpha = 0.95$.

Можно разбивать множество признаков на группы и полагать, что признаки из разных групп не коррелированы. Тогда матрица $\hat{\Sigma}$ приобретает блочно-диагональный вид. Существуют эффективные алгоритмы обращения таких матриц.

Преобразование пространства признаков. Другой способ решения проблемы мультиколлинеарности заключается в том, чтобы отбросить некоторое количество

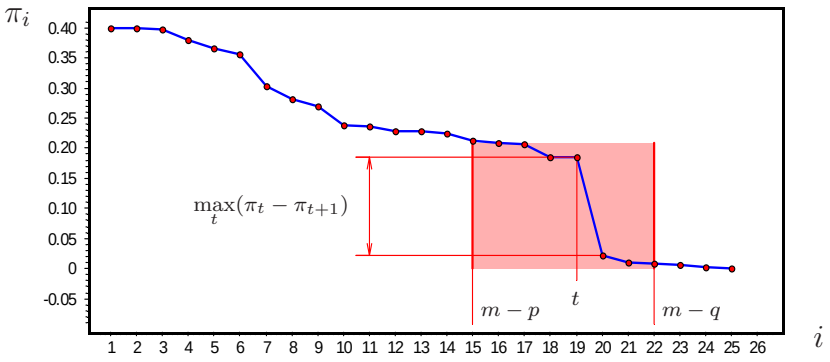

Рис. 7. Эксперт предположил, что в выборке длины $m = 25$ находится от $q = 3$ до $p = 10$ выбросов. Более точное число выбросов, равное 6, удалось определить по критерию «крутого склона».

наименее значимых признаков. Как правило, ими оказываются признаки, почти линейно зависимые от других признаков. Различные методы *отбора признаков* (features selection) рассматриваются в разделе **??**. Обратим внимание на кажущийся парадокс: информация отбрасывается, но решение получается более высокого качества.

Ещё один способ сокращения размерности заключается в том, чтобы из имеющихся признаков построить меньшее количество более информативных признаков. Например, в классе линейных преобразований признаков таким свойством обладают собственные векторы ковариационной матрицы, соответствующие максимальным собственным значениям. Этот факт используется в *методе главных компонент*, см. раздел **??**. Методы *синтеза признаков* (features extraction) подробно рассматриваются в **??**.

Робастные методы оценивания. Оценки, устойчивые относительно редких больших выбросов, связанных с малыми загрязнениями плотности, называются *робастными*[2] (robust — здравый). Простейший метод робастного оценивания параметра θ плотности $\varphi(x;\theta)$ по заданной выборке X^m основан на фильтрации выбросов и состоит из следующих шагов.

1. Оценка параметра $\hat{\theta}$ вычисляется по всей выборке X^m, исходя из принципа максимума правдоподобия.
2. Для каждого объекта $x_i \in X^m$ вычисляется правдоподобие $\pi_i = \varphi(x_i;\hat{\theta})$.
3. Выборка сортируется по убыванию значений правдоподобия: $\pi_1 \geqslant \ldots \geqslant \pi_m$.
4. Объекты, оказавшиеся в конце этого ряда, считаются нетипичными (выбросами) и удаляются из выборки. Здесь возможны варианты. Можно задавать число или долю удаляемых объектов. Можно удалять объекты с низким правдоподобием, $\pi_i < P_0$; в этом случае сортировать выборку не обязательно, но не понятно, из каких соображений выбирать P_0. Ещё лучше применять критерий «крутого склона»: задаются два параметра p и q, и находится значение

[2]Вообще говоря, робастными называют оценки, устойчивые к любым несоответствиям модели $\varphi(x;\theta)$ истинному распределению. Проблема выбросов наиболее часто встречается на практике. Более серьёзные несоответствия свидетельствуют, скорее, о необходимости заменить модель или рассмотреть более общую модель в виде смеси распределений, см. §1.4.

$t \in \{m-p, \ldots, m-q-1\}$, для которого скачок правдоподобия $\pi_t - \pi_{t+1}$ максимален. Затем последние $(m-t)$ объектов удаляются из выборки, Рис. 7.

5. Оценка параметра $\hat{\theta}$ вычисляется вторично по сокращённой выборке.

В некоторых случаях удаление объектов не требует полного пересчёта оценки $\hat{\theta}$. Например, оценки нормального распределения $\hat{\mu}_y$, $\hat{\Sigma}$ аддитивны по объектам выборки, поэтому достаточно вычесть из них слагаемые, соответствующие удаляемым объектам.

Шаги 2–5 можно повторять итерационно, так как после уточнения оценки $\hat{\theta}$ некоторые объекты могут перейти в разряд нетипичных. В большинстве случаев итерационный процесс сходится очень быстро, 1–2 итераций бывает достаточно.

Метод редукции. В дискриминанте Фишера для получения $(n+1)|Y|$ коэффициентов линейного решающего правила (1.16) приходится оценивать $\frac{1}{2}n(n+1)+n|Y|$ параметров. Фактически, одна задача сводится к другой, более сложной. Ещё один подход к уменьшению размерности предложен Шурыгиным и заключается в том, чтобы свести n-мерную задачу к последовательности двумерных [10]. Достоинства метода редукции — простота реализации, отсутствие необходимости оценивать и обращать ковариационную матрицу, возможность отбросить неинформативные признаки. В некоторых прикладных задачах он превосходит другие методы классификации [10]. Недостатком является отсутствие строгого теоретического обоснования. По всей видимости, этот метод хорошо работает на тех задачах, в которых признаки неравноценны и чётко ранжируются по своей «полезности».

§1.4 Разделение смеси распределений

В тех случаях, когда «форму» класса не удаётся описать каким-либо одним распределением, можно попробовать описать её смесью распределений.

Гипотеза 1.3. *Плотность распределения на X имеет вид смеси k распределений:*

$$p(x) = \sum_{j=1}^{k} w_j p_j(x), \qquad \sum_{j=1}^{k} w_j = 1, \qquad w_j \geqslant 0,$$

где $p_j(x)$ — функция правдоподобия j-й компоненты смеси, w_j — её априорная вероятность. Функции правдоподобия принадлежат параметрическому семейству распределений $\varphi(x;\theta)$ и отличаются только значениями параметра, $p_j(x) = \varphi(x;\theta_j)$.

Иными словами, «выбрать объект x из смеси $p(x)$» означает сначала выбрать j-ю компоненту смеси из дискретного распределения $\{w_1, \ldots, w_k\}$, затем выбрать объект x согласно плотности $p_j(x)$.

Задача разделения смеси заключается в том, чтобы, имея выборку X^m случайных и независимых наблюдений из смеси $p(x)$, зная число k и функцию φ, оценить вектор параметров $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$.

1.4.1 EM-алгоритм

К сожалению, попытка разделить смесь, используя принцип максимума правдоподобия «в лоб», приводит к слишком громоздкой оптимизационной задаче. Обойти эту трудность позволяет алгоритм EM (expectation-maximization). Идея алгоритма заключается в следующем. Искусственно вводится вспомогательный вектор *скрытых* (hidden) переменных G, обладающий двумя замечательными свойствами. С одной стороны, он может быть вычислен, если известны значения вектора параметров Θ. С другой стороны, поиск максимума правдоподобия сильно упрощается, если известны значения скрытых переменных.

EM-алгоритм состоит из итерационного повторения двух шагов. На E-шаге вычисляется ожидаемое значение (expectation) вектора скрытых переменных G по текущему приближению вектора параметров Θ. На M-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора Θ по текущим значениям векторов G и Θ.

Алгоритм 1.2. Общая идея EM-алгоритма

1: Вычислить начальное приближение вектора параметров Θ;
2: **повторять**
3: $\quad G := \mathtt{EStep}(\Theta)$;
4: $\quad \Theta := \mathtt{MStep}(\Theta, G)$;
5: **пока** Θ и G не стабилизируются.

Этот алгоритм был предложен и исследован М. И. Шлезингером как инструмент для *самопроизвольной классификации образов* [9]. Двенадцать лет спустя он был открыт заново в [11] под названием *EM-алгоритма*. Область его применения чрезвычайно широка — дискриминантный анализ, кластеризация, восстановление пропусков в данных, обработка сигналов и изображений [8]. Здесь мы рассматриваем его как инструмент разделения смеси распределений.

E-шаг (expectation). Обозначим через $p(x, \theta_j)$ плотность вероятности того, что объект x получен из j-й компоненты смеси. По формуле условной вероятности

$$p(x, \theta_j) = p(x)\,\mathsf{P}(\theta_j|x) = w_j p_j(x).$$

Введём обозначение $g_{ij} \equiv \mathsf{P}(\theta_j|x_i)$. Это неизвестная апостериорная вероятность того, что обучающий объект x_i получен из j-й компоненты смеси. Возьмём эти величины в качестве скрытых переменных. Обозначим $G = (g_{ij})_{m\times k} = (g_1, \ldots, g_j)$, где g_j — j-й столбец матрицы G. Каждый объект обязательно принадлежит какой-то компоненте, поэтому справедлива формула полной вероятности:

$$\sum_{j=1}^{k} g_{ij} = 1 \quad \text{для всех } i = 1, \ldots, \ell.$$

Зная параметры компонент w_j, θ_j, легко вычислить g_{ij} по формуле Байеса:

$$g_{ij} = \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} \quad \text{для всех } i,\ j. \tag{1.17}$$

В этом и заключается E-шаг алгоритма EM.

M-шаг (maximization). Покажем, что знание значений скрытых переменных g_{ij} и принцип максимума правдоподобия приводят к оптимизационной задаче, допускающей эффективное численное (или даже аналитическое) решение. Будем максимизировать логарифм правдоподобия

$$Q(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i) \to \max_{\Theta}.$$

при ограничении $\sum_{j=1}^{k} w_j = 1$. Запишем лагранжиан этой оптимизационной задачи:

$$L(\Theta; X^m) = \sum_{i=1}^{m} \ln\Bigg(\sum_{j=1}^{k} w_j p_j(x_i) \Bigg) - \lambda \Bigg(\sum_{j=1}^{k} w_j - 1 \Bigg).$$

Приравняем нулю производную лагранжиана по w_j:

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{m} \frac{p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} - \lambda = 0, \quad j = 1, \ldots, k. \tag{1.18}$$

Умножим левую и правую части на w_j, просуммируем все k этих равенств, и поменяем местами знаки суммирования по j и по i:

$$\sum_{i=1}^{m} \underbrace{\sum_{j=1}^{k} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)}}_{=1} = \lambda \underbrace{\sum_{j=1}^{k} w_j}_{=1},$$

откуда следует $\lambda = m$.

Теперь снова умножим левую и правую части (1.18) на w_j, подставим $\lambda = m$, и, замечая сходство с формулой (1.17), получим выражение весов компонент через скрытые переменные:

$$w_j = \frac{1}{m} \sum_{i=1}^{m} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \quad j = 1, \ldots, k. \tag{1.19}$$

Легко проверить, что ограничения-неравенства $w_j \geqslant 0$ будут выполнены на каждой итерации, если они выполнены для начального приближения.

Приравняем нулю производную лагранжиана по θ_j, помня, что $p_j(x) \equiv \varphi(x; \theta_j)$:

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{m} \frac{w_j}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} p_j(x_i) = \sum_{i=1}^{m} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) =$$
$$= \sum_{i=1}^{m} g_{ij} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} g_{ij} \ln p_j(x_i) = 0, \quad j = 1, \ldots, k.$$

Полученное условие совпадает с необходимым условием максимума в задаче максимизации взвешенного правдоподобия

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i; \theta), \quad j = 1, \ldots, k. \tag{1.20}$$

Алгоритм 1.3. EM-алгоритм с фиксированным числом компонент

Вход:
выборка $X^m = \{x_1, \ldots, x_m\}$;
k — число компонент смеси;
$\Theta = (w_j, \theta_j)_{j=1}^k$ — начальное приближение параметров смеси;
δ — параметр критерия останова;

Выход:
$\Theta = (w_j, \theta_j)_{j=1}^k$ — оптимизированный вектор параметров смеси;

ПРОЦЕДУРА EM (X^m, k, Θ, δ);
повторять
  E-шаг (expectation):
  для всех $i = 1, \ldots, m,\ \ j = 1, \ldots, k$
    $g_{ij}^0 := g_{ij}; \quad g_{ij} := \dfrac{w_j \varphi(x_i; \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i; \theta_s)}$;
  M-шаг (maximization):
  для всех $j = 1, \ldots, k$
    $\theta_j := \arg\max\limits_{\theta} \sum\limits_{i=1}^m g_{ij} \ln \varphi(x_i; \theta); \quad w_j := \dfrac{1}{m} \sum\limits_{i=1}^m g_{ij}$;
пока $\max\limits_{i,j} |g_{ij} - g_{ij}^0| > \delta$;
вернуть $(w_j, \theta_j)_{j=1}^k$;

Таким образом, M-шаг сводится к вычислению весов компонент w_j как средних арифметических (1.19) и оцениванию параметров компонент θ_j путём решения k независимых оптимизационных задач (1.20). Отметим, что разделение переменных оказалось возможным благодаря удачному введению скрытых переменных.

Условия сходимости алгоритма EM рассматриваются в работах [11, 16, 13].

Критерий останова. Итерации останавливаются, когда значения функционала $Q(\Theta)$ или скрытых переменных G перестают существенно изменяться. Удобнее контролировать скрытые переменные, так как они имеют смысл вероятностей и принимают значения из отрезка $[0, 1]$.

Реализация итерационного процесса показана в Алгоритме 1.3. На E-шаге вычисляется матрица скрытых переменных G по формуле (1.17). На M-шаге решается серия из k задач максимизации взвешенного правдоподобия (1.20), каждая из них — по полной выборке X^m с вектором весов g_j.

Обобщённый EM-алгоритм. Не обязательно добиваться высокой точности решения оптимизационной задачи (1.20) на каждом M-шаге алгоритма. Достаточно лишь сместиться в направлении максимума, сделав одну или несколько итераций, и затем выполнить E-шаг. Этот алгоритм также обладает неплохой сходимостью и называется *обобщённым EM-алгоритмом* (generalized EM-algorithm, GEM) [11].

Проблема выбора начального приближения. Хотя алгоритм EM сходится при достаточно общих предположениях, скорость сходимости может существенно зави-

сеть от «удачности» начального приближения. Сходимость ухудшается в тех случаях, когда делается попытка поместить несколько компонент в один фактический сгусток распределения, либо разместить компоненту посередине между сгустками.

Стандартная (но далеко не самая лучшая) эвристика заключается в том, чтобы выбрать параметры компонент случайным образом. Более разумная идея — найти в выборке k объектов, максимально удалённых друг от друга, и именно в этих точках разместить компоненты.

Проблема выбора числа компонент k. До сих пор предполагалось, что число компонент k известно заранее. На практике это, как правило, не так.

Иногда число компонент удаётся оценить визуально, спроецировав выборку на плоскость каким-либо способом и определив число сгустков точек на полученном графике. С этой целью можно применить метод главных компонент из **??**, многомерное шкалирование из **??** или метод целенаправленного проецирования (Projection Pursuit). Однако визуальный подход обладает очевидными недостатками: проецирование искажает структуру выборки, а необходимость обращаться к эксперту исключает возможность автоматического анализа данных.

Существует ещё один приём — решить задачу несколько раз при последовательных значениях k, построить график зависимости правдоподобия выборки $Q(\Theta)$ от k, и выбрать наименьшее k, при котором график претерпевает резкий скачок правдоподобия. Это называется критерием «крутого склона». К сожалению, он также не лишён недостатков. Во-первых, существенно увеличиваются затраты времени. Во-вторых, если данные плохо описываются моделью компонент $\varphi(x;\theta)$, то «крутой склон» может не наблюдаться. Наличие крутого склона свидетельствует о том, что модель компонент была выбрана удачно.

EM-алгоритм с последовательным добавлением компонент позволяет решить две проблемы сразу — проблему выбора числа компонент и проблему выбора начального приближения. Идея заключается в следующем. Имея некоторый набор компонент, можно выделить объекты x_i, которые хуже всего описываются смесью — это объекты с наименьшими значениями правдоподобия $p(x_i)$. По этим объектам строится ещё одна компонента. Затем она добавляется в смесь и запускаются EM-итерации, чтобы новая компонента и старые «притёрлись друг к другу». Так продолжается до тех пор, пока все объекты не окажутся покрыты компонентами. Реализация этой идеи представлена в Алгоритме 1.4.

На шаге 1 строится первая компонента и полагается $k = 1$. Затем в цикле последовательно добавляется по одной компоненте. Если значение правдоподобия $p(x_i)$ в R раз меньше максимального значения правдоподобия, значит объект x_i плохо описывается смесью. Заметим, что это лишь эвристика; совсем не обязательно сравнивать $p(x_i)$ именно с максимальным правдоподобием; можно брать среднее правдоподобие или фиксированное пороговое значение P_0. На шаге 3 формируется подвыборка U из объектов, которые не подходят ни к одной из компонент. Если длина этой подвыборки меньше порога m_0, то процесс добавления компонент на этом заканчивается, и оставшиеся объекты считаются выбросами. На шаге 6 снова применяется метод максимума правдоподобия для формирования новой компоненты, но теперь уже не по всей выборке, а только по подвыборке U. Веса компонент пересчитываются таким образом, чтобы их сумма по-прежнему оставалась равной единице. На шаге 7

Алгоритм 1.4. EM-алгоритм с последовательным добавлением компонент

Вход:
выборка $X^m = \{x_1, \ldots, x_m\}$;
R — максимальный допустимый разброс правдоподобия объектов;
m_0 — минимальная длина выборки, по которой можно восстановить плотность;
δ — параметр критерия останова;

Выход:
k — число компонент смеси;
$\Theta = (w_j, \theta_j)_{j=1}^k$ — веса и параметры компонент;

1: начальное приближение — одна компонента:
$\theta_1 := \arg\max_\theta \sum_{i=1}^m \ln \varphi(x_i; \theta)$; $\quad w_1 := 1$; $\quad k := 1$;
2: **для всех** $k := 2, 3, \ldots$
3: выделить объекты с низким правдоподобием:
$U := \{x_i \in X^m : p(x_i) < \max_j p(x_j)/R\}$;
4: **если** $|U| < m_0$ **то**
5: **выход** из цикла по k;
6: начальное приближение для k-й компоненты:
$\theta_k := \arg\max_\theta \sum_{x_i \in U} \ln \varphi(x_i; \theta)$; $\quad w_k := \frac{1}{m}|U|$;
$w_j := w_j(1 - w_k)$, $\; j = 1, \ldots, k-1$;
7: $\texttt{EM}\,(X^m, k, \Theta, \delta)$;

все предыдущие компоненты вместе с новой компонентой проходят через цикл итераций EM-алгоритма.

Стохастический EM-алгоритм. Минимизируемый функционал $Q(\Theta)$ в общем случае не является выпуклым и может иметь большое количество локальных экстремумов. Поэтому EM-алгоритму присущи обычные недостатки любого детерминированного процесса многоэкстремальной оптимизации: застревание в локальных минимумах, зависимость решения от начального приближения, медленная сходимость при неудачном выборе начального приближения. Обычно такого рода недостатки преодолеваются методами адаптивной стохастической оптимизации.

Описание одного из вариантов *стохастического EM-алгоритма* (stochastic EM-algorithm, SEM) можно найти в [1, стр. 207]. Основное отличие от Алгоритма 1.3 в том, что на М-шаге (шаг 4) вместо максимизации взвешенного правдоподобия

$$\theta_j := \arg\max_\theta \sum_{i=1}^m g_{ij} \ln \varphi(x_i; \theta)$$

решается задача максимизации обычного, невзвешенного, правдоподобия

$$\theta_j := \arg\max_\theta \sum_{x_i \in X^{(g_j)}} \ln \varphi(x_i; \theta),$$

где выборка $X^{(g_j)}$ генерируется из X^m путём стохастического моделирования: каждый объект $x_i \in X^m$ включается в выборку $X^{(g_j)}$ с вероятностью g_{ij} (например, при условии $r_i < g_{ij}$, где r_i — случайное число из равномерного распределения на $[0,1]$).

Ещё одно отличие алгоритма SEM, описанного в [1], состоит в том, что он последовательно уменьшает число компонент k, начиная с некоторого заведомо избыточного числа $k_{\max}$. Если в результате стохастического моделирования какая-то компонента оказывается слишком малочисленной, $|X^{(g_j)}| \leqslant m_0$, то она вовсе удаляется. Это отличие не принципиально: оба алгоритма, детерминированный и стохастический, могут использовать любую стратегию: и наращивание, и исключение. Возможно также совмещение обеих стратегий. После добавления k-й компоненты на шаге 6 и выполнения основного цикла итераций EM-алгоритма на шаге 7 может оказаться, что некоторая j-я компонента имеет слишком низкое «суммарное правдоподобие» $\sum_{i=1}^m g_{ij}$. В таком случае её следует удалить; и если это та же компонента, которая была только что добавлена, алгоритм прекращает работу.

Преимущества SEM вытекают, главным образом, из того факта, что рандомизация «выбивает» оптимизационный процесс из локальных минимумов:

- SEM работает относительно быстро, и его результаты практически не зависят от начального приближения.
- Как правило, SEM находит экстремум $Q(\Theta)$, близкий к глобальному.

1.4.2 Смеси многомерных нормальных распределений

Рассмотрим решение задачи M-шага в частном случае, когда компоненты имеют нормальные (гауссовские) плотности. В этом случае функционал (1.20) является квадратичным и положительно определенным, поэтому решение выписывается в явном аналитическом виде.

Гауссовские смеси общего вида.

Гипотеза 1.4. *Компоненты смеси имеют n-мерные нормальные распределения $\varphi(x;\theta_j) = \mathcal{N}(x;\mu_j,\Sigma_j)$ с параметрами $\theta_j = (\mu_j,\Sigma_j)$, где $\mu_j \in \mathbb{R}^n$ — вектор математического ожидания, $\Sigma_j \in \mathbb{R}^{n\times n}$ — ковариационная матрица, $j = 1,\ldots,k$.*

Теорема 1.7. *Если справедлива Гипотеза 1.4, то стационарная точка оптимизационной задачи (1.20) имеет вид*

$$\hat{\mu}_j = \frac{1}{mw_j}\sum_{i=1}^m g_{ij}x_i, \quad j = 1,\ldots,k;$$

$$\hat{\Sigma}_j = \frac{1}{mw_j}\sum_{i=1}^m g_{ij}(x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^\intercal, \quad j = 1,\ldots,k.$$

Данное утверждение непосредственно вытекает из Теоремы 1.5 и оценки (1.19).

Таким образом, M-шаг сводится к вычислению выборочного среднего и выборочной ковариационной матрицы для каждой компоненты смеси. При этом для

каждой компоненты используется своё распределение весов объектов. Вес i-го объекта для j-й компоненты равен g_{ij} — оценке принадлежности данного объекта данной компоненте, вычисленной на E-шаге.

Смеси многомерных нормальных распределений позволяют приближать любые непрерывные плотности вероятности. Они являются универсальными аппроксиматорами плотностей, подобно тому, как полиномы являются универсальными аппроксиматорами непрерывных функций. В практических задачах это позволяет восстанавливать функции правдоподобия классов даже в тех случаях, когда на первый взгляд для выполнения Гипотезы 1.4 нет содержательных оснований.

Недостатком гауссовских смесей является необходимость обращать ковариационные матрицы. Это трудоёмкая операция. Кроме того, ковариационные матрицы нередко оказываются вырожденными или плохо обусловленными. Тогда возникает проблема неустойчивости выборочных оценок плотности и самого классификатора. Стандартные приёмы (регуляризация, метод главных компонент) позволяют справиться с этой проблемой. Но есть и другой выход — использовать для описания компонент более простые распределения, например, сферические.

Гауссовские смеси с диагональными матрицами ковариации. Трудоёмкого обращения матриц можно избежать, если принять гипотезу, что в каждой компоненте смеси признаки некоррелированы. В этом случае гауссианы упрощаются, оставаясь, тем не менее, универсальными аппроксиматорами плотности.

Можно было бы предположить, что компоненты имеют сферические плотности, $\Sigma_j = \sigma_j^2 I_n$. Этот случай вынесен в качестве Упражнения 1.10. Однако такое предположение имеет очевидный недостаток: если признаки существенно различаются по порядку величины, то компоненты будут иметь сильно вытянутые формы, которые придётся аппроксимировать большим количеством сферических гауссианов. Предположение о неравных дисперсиях признаков приводит к алгоритму классификации, не чувствительному к различиям в масштабах измерения признаков.

Гипотеза 1.5. *Компоненты смеси имеют n-мерные нормальные распределения с параметрами (μ_j, Σ_j), где $\mu_j = (\mu_{j1}, \ldots, \mu_{jn})$, $\Sigma_j = \operatorname{diag}(\sigma_{j1}^2, \ldots, \sigma_{jn}^2)$ — диагональная матрица, $j = 1, \ldots, k$:*

$$\varphi(x;\theta_j) = \mathcal{N}(x;\mu_j,\Sigma_j) = \prod_{d=1}^{n} \frac{1}{\sigma_{jd}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\xi_d - \mu_{jd}}{\sigma_{jd}}\right)^2\right), \quad x = (\xi_1, \ldots, \xi_n).$$

Отметим, что многомерная нормальная плотность с диагональной матрицей ковариации представима в виде произведения одномерных плотностей. Это означает, что предположение некоррелированности в гауссовском случае равносильно «наивно-байесовскому» предположению о независимости признаков. Отметим, что это предположение делается только относительно компонент; для смеси независимость признаков, вообще говоря, уже не имеет места.

Теорема 1.8. *Если справедлива Гипотеза 1.5, то стационарная точка оптимизационной задачи (1.20) имеет вид*

$$\hat{\mu}_{jd} = \frac{1}{mw_j} \sum_{i=1}^{m} g_{ij} x_{id}, \quad d = 1, \ldots, n;$$
$$\hat{\sigma}^2_{jd} = \frac{1}{mw_j} \sum_{i=1}^{m} g_{ij} (x_{id} - \hat{\mu}_{jd})^2, \quad d = 1, \ldots, n;$$

где $x_i = (x_{i1}, \ldots, x_{in})$ *— объекты выборки* X^m.

Доказательство.

Запишем производные логарифма нормальной плотности $\mathcal{N}(x; \mu_j, \Sigma_j)$ по параметрам μ_{jd}, σ_{jd} в точке $x_i = (x_{i1}, \ldots, x_{in})$:

$$\frac{\partial}{\partial \mu_{jd}} \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) = \sigma_{jd}^{-2} (x_{id} - \mu_{jd});$$
$$\frac{\partial}{\partial \sigma_{jd}} \ln \mathcal{N}(x_i; \mu_j, \Sigma_j) = -\sigma_{jd}^{-1} + \sigma_{jd}^{-3} (x_{id} - \mu_{jd})^2.$$

Приравняем нулю производные взвешенного функционала правдоподобия по параметрам μ_{jd}, σ_{jd}:

$$-\sigma_{jd}^{-2} \sum_{i=1}^{m} g_{ij} (x_{id} - \mu_{jd}) = 0;$$
$$\sigma_{jd}^{-3} \sum_{i=1}^{m} g_{ij} \left(\sigma_{jd}^2 - (x_{id} - \mu_{jd})^2 \right) = 0.$$

Отсюда, вынося параметры μ_{jd}, σ_{jd} за знак суммирования по i, и применяя соотношение (1.19), получаем требуемое. ■

Радиальные функции. Гауссиан $p_j(x) = \mathcal{N}(x; \mu_j, \Sigma_j)$ с диагональной матрицей Σ_j можно записать в виде

$$p_j(x) = \mathcal{N}_j \exp\left(-\tfrac{1}{2} \rho_j^2(x, \mu_j)\right),$$

где $\mathcal{N}_j = (2\pi)^{-\frac{n}{2}} (\sigma_{j1} \cdots \sigma_{jn})^{-1}$ — нормировочный множитель, $\rho_j(x, x')$ — *взвешенная евклидова метрика* в n-мерном пространстве X:

$$\rho_j^2(x, x') = \sum_{d=1}^{n} \sigma_{jd}^{-2} |\xi_d - \xi'_d|^2, \quad x = (\xi_1, \ldots, \xi_n), \quad x' = (\xi'_1, \ldots, \xi'_n).$$

Чем меньше расстояние $\rho_j(x, \mu_j)$, тем выше значение плотности в точке x. Поэтому плотность $p_j(x)$ можно рассматривать как *функцию близости* вектора x к фиксированному центру μ_j.

Функции $f(x)$, зависящие только от расстояния между x и фиксированной точкой пространства X, принято называть *радиальными*.

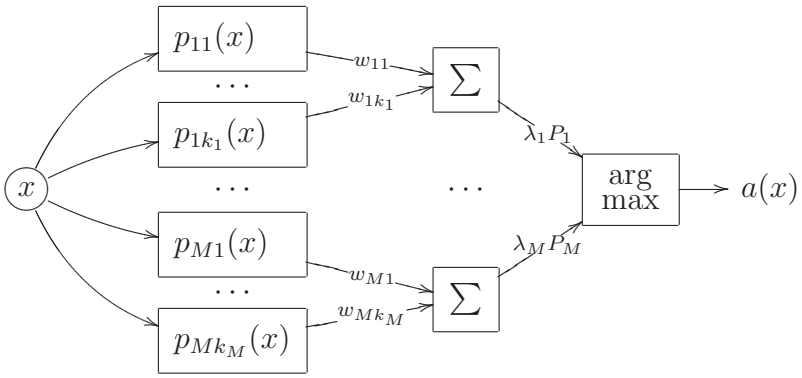

Рис. 8. Сеть радиальных базисных функций представляет собой трёхуровневую суперпозицию.

1.4.3 Сеть радиальных базисных функций

Выше мы рассматривали задачу разделения смеси распределений, забыв на время об исходной задаче классификации.

Пусть теперь $Y = \{1, \ldots, M\}$, каждый класс $y \in Y$ имеет свою плотность распределения $p_y(x)$ и представлен частью выборки $X_y^\ell = \{(x_i, y_i) \in X^\ell \mid y_i = y\}$.

Гипотеза 1.6. *Функции правдоподобия классов $p_y(x)$, $y \in Y$, представимы в виде смесей k_y компонент. Каждая компонента имеет n-мерную гауссовскую плотность с параметрами $\mu_{yj} = (\mu_{yj1}, \ldots, \mu_{yjn})$, $\Sigma_{yj} = \operatorname{diag}(\sigma_{yj1}^2, \ldots, \sigma_{yjn}^2)$, $j = 1, \ldots, k_y$:*

$$p_y(x) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj}), \quad \sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geqslant 0;$$

Алгоритм классификации. Запишем байесовское решающее правило (1.2), выразив плотность каждой компоненты $p_{yj}(x)$ через взвешенное евклидово расстояние от объекта x до центра компоненты μ_{yj}:

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp\left(-\tfrac{1}{2}\rho_{yj}^2(x, \mu_{yj})\right)}_{p_{yj}(x)}, \tag{1.21}$$

где $\mathcal{N}_{yj} = (2\pi)^{-\frac{n}{2}} (\sigma_{yj1} \cdots \sigma_{yjn})^{-1}$ — нормировочные множители. Алгоритм имеет вид суперпозиции, состоящей из трёх уровней или *слоёв*, Рис 8.

Первый слой образован $k_1 + \cdots + k_M$ гауссианами $p_{yj}(x)$, $y \in Y$, $j = 1, \ldots, k_y$. На входе они принимают описание объекта x, на выходе выдают оценки близости объекта x к центрам μ_{yj}, равные значениям плотностей компонент в точке x.

Второй слой состоит из M сумматоров, вычисляющих взвешенные средние этих оценок с весами w_{yj}. На выходе второго слоя появляются оценки принадлежности объекта x каждому из классов, равные значениям плотностей классов $p_{yj}(x)$.

Третий слой образуется единственным блоком arg max, принимающим окончательное решение об отнесении объекта x к одному из классов.

Таким образом, при классификации объекта x оценивается его близость к каждому из центров μ_{yj} по метрике $\rho_{yj}(x, \mu_{yj})$, $j = 1, \ldots, k_y$. Объект относится к тому классу, к чьим центрам он располагается ближе.

Описанный трёхуровневый алгоритм классификации называется *сетью с радиальными базисными функциями* или *RBF-сетью* (radial basis function network). Это одна из разновидностей *нейронных сетей*.

Обучение RBF-сети сводится к восстановлению плотности каждого из классов $p_y(x)$ с помощью EM-алгоритма. Результатом обучения являются центры μ_{yj} и дисперсии Σ_{yj} компонент $j = 1, \ldots, k_y$. Интересно отметить, что, оценивая дисперсии, мы фактически подбираем метрики ρ_{yj}, с помощью которых будут вычисляться расстояния до центров μ_{yj}. При использовании Алгоритма 1.4 для каждого класса определяется оптимальное число компонент смеси.

Преимущества EM-алгоритма. По сравнению с широко известными градиентными методами настройки нейронных сетей (см. главу **??**), EM-алгоритм более устойчив к шуму и быстрее сходится. Кроме того, он описывает каждый класс как совокупность компонент или *кластеров*, что позволяет лучше понимать внутреннюю структуру данных. В частности, центры сферических гауссовских компонент μ_{yj} можно интерпретировать как виртуальные эталонные объекты, с которыми сравниваются классифицируемые объекты. Во многих прикладных задачах виртуальным эталонам удаётся подыскать содержательную интерпретацию. Например, в медицинской дифференциальной диагностике это может быть определённая (j-я) форма данного (y-го) заболевания. Информация о том, что классифицируемый объект близок именно к этому эталону, может быть полезной при принятии решений.

EM-алгоритм может также использоваться для решения задач *кластеризации*, о чём пойдёт речь в главе **??**.

Резюме

1. *Оптимальный байесовский классификатор*, минимизирующий средний риск (вероятность потерь), ещё неоднократно встретится в этом курсе:

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y p_y(x),$$

где $p_y(x)$ — плотность распределения (функция правдоподобия) класса $y \in Y$, P_y — априорная вероятность класса $y \in Y$, λ_y — величина потери от ошибки на объекте класса $y \in Y$.

2. *«Наивный» байесовский классификатор* опирается на дополнительное предположение о статистической независимости признаков, которое крайне редко выполняется на практике. Однако благодаря простоте реализации и высокой устойчивости он всё же используется, чаще всего — как эталон при сравнении алгоритмов, либо как элементарный блок в более сложных моделях.

3. Рассмотрены *три основных подхода* к восстановлению плотностей $p_y(x)$ по подвыборкам X_y^ℓ: параметрический, непараметрический и разделение смеси. Сопоставление формул (1.12) и (1.21) показывает, что непараметрические оценки плотности можно рассматривать как предельный частный случай смеси распределений, в которой каждому обучающему объекту x_i соответствует ровно

одна компонента с априорной вероятностью $w_j = \frac{1}{m}$ и сферической плотностью с центром в точке x_i. С другой стороны, параметрический подход также является крайним случаем смеси — когда берётся только одна компонента. Таким образом, все три подхода отличаются, в первую очередь, количеством аддитивных компонент в модели распределения: $1 \ll k \ll m$. Это приводит к качественным различиям в методах обучения. Требования к форме компонент ослабляются по мере увеличения их числа. Восстановление смеси из произвольного числа компонент k является, по всей видимости, наиболее общим подходом в байесовской классификации.

4. *Непараметрический подход* основан на локальной оценке плотности по Парзену-Розенблатту и приводит к *методу парзеновского окна*. Выбор ширины окна h существенно влияет на качество классификации. При сильно неравномерном распределении объектов рекомендуется использовать окно переменной ширины. Выбор сглаживающего ядра K влияет на гладкость разделяющей поверхности, но почти не влияет на качество классификации.

5. В *параметрическом подходе* предполагается, что плотности известны с точностью до параметра: $p_y(x) = \varphi(x;\theta_y)$, где функция φ фиксирована. Эта гипотеза является довольно сильным априорным предположением. Фактически, утверждается, что «форма классов» может быть приближённо описана (смоделирована) одним из элементов заданного семейства плотностей: $p_y(x) = \varphi(x;\theta_y)$, где θ_y — вектор параметров, свой для каждого класса $y \in Y$. Существует опасность, что модель φ окажется не адекватной; тогда байесовский классификатор будет далёк от оптимального, то есть будет совершать слишком много ошибок. Если плотности гауссовские, то разделяющая поверхность квадратична; если ковариационные матрицы классов равны (классы имеют «одинаковую форму»), то она вырождается в линейную. Параметры нормального распределения — центр μ и ковариационная матрица Σ — легко оцениваются по выборке.

6. *Проблема мультиколлинеарности* затрудняет обращение матрицы Σ. Если Σ вырождена, в частности, если признаков больше, чем объектов, то построить классификатор вообще невозможно. Если Σ плохо обусловлена (близка к вырожденной), то её обращение численно неустойчиво, что приводит к ухудшению качества классификации. Для решения этой проблемы применяется *регуляризация* — вместо матрицы Σ обращается матрица $(\Sigma + \tau I)$. Параметр регуляризации τ либо назначают априори, либо выбирают по скользящему контролю.

7. *Линейный дискриминант Фишера* опирается на предположение, что ковариационные матрицы классов равны. В этом случае Σ оценивается по обучающим объектам всех классов, что повышает устойчивость оценки, особенно в случае малочисленных классов. Линейный дискриминант часто оказывается предпочтительнее квадратичного даже в тех случаях, когда гипотеза равных ковариационных матриц не верна.

8. *Смеси параметрических распределений* $p_y(x) = \sum_{j=1}^{k_y} w_{yj}\varphi(x;\theta_{yj})$ позволяют описывать классы сложной формы. Задача разделения смеси заключается в том, чтобы по выборке оценить веса w_{yj} и параметры θ_{yj} компонент смеси. Эта задача решается EM-алгоритмом. Его специальный вариант позволя-

ет последовательно добавлять компоненты, автоматически определяя число k. Ещё одна небольшая модификация — стохастический EM-алгоритм (SEM) — обладает улучшенными характеристиками сходимости.

9. *Сеть радиальных базисных функций (РБФ)* можно рассматривать как смесь гауссовских компонент с диагональными матрицами ковариации. Построение РБФ с помощью EM-алгоритма — это один из наиболее успешных современных методов классификации.

10. *Наличие выбросов* в данных приводит к смещённым оценкам плотностей классов и ухудшению качества классификации. Во всех трёх подходах возможно применять *фильтрацию выбросов*. После первого пробного решения отбрасываются обучающие объекты, имеющие слишком низкие значения правдоподобия, и задача решается снова, теперь уже по сокращённой выборке.

Упражнения

Упр. 1.1. Сформулировать и доказать теоремы, аналогичные 1.2 и 1.1 для случая, когда вводится величина штрафов $\lambda_{y\varnothing}$ за отказ от классификации объекта класса $y \in Y$. Что представляет из себя оптимальная область отказов?

Упр. 1.2. Пусть $X = \mathbb{R}$, $Y = \{0,1\}$, $\lambda_0 P_0 = C$, $\lambda_1 P_1 = 1$, функции правдоподобия классов имеют вид $p_y(x) = \pi^{-1/2}\exp\bigl(-(x-y)^2\bigr)$. Выписать байесовский алгоритм классификации. Что собой представляет разделяющая поверхность при $C = 1$, при $C = e$?

Упр. 1.3. Пусть $X = \mathbb{R}^2$, $Y = \{0,1\}$, $\ln \lambda_i P_i = C_i$, функции правдоподобия классов гауссовские, $\mu_0 = \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right)$, $\mu_1 = \left(\begin{smallmatrix} -a \\ -b \end{smallmatrix}\right)$, с одинаковыми матрицами ковариации $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & S \end{smallmatrix}\right)$. Выписать байесовский алгоритм классификации и уравнение разделяющей поверхности. Доказать, что разделяющая поверхность касается линий уровня плотностей обоих классов.

Упр. 1.4. Доказать, что наивный байесовский классификатор (1.5) в случае бинарных признаков является линейным разделителем: $a(\xi_1,\ldots,\xi_n) = [\alpha_0 + \alpha_1\xi_1 + \cdots + \alpha_n\xi_n > 0]$. Выписать формулы для вычисления коэффициентов α_j, $j = 0,\ldots,n$ по обучающей выборке.

Упр. 1.5. Производная скалярной функции $f(A)$ по матрице $A = (a_{ij})$ определяется как матрица частных производных $\frac{\partial}{\partial A} f(A) = \bigl(\frac{\partial}{\partial a_{ij}} f(A)\bigr)$. Через $\operatorname{diag} A$ обозначается матрица, диагональные элементы которой совпадают с соответствующими диагональными элементами матрицы A, остальные элементы равны нулю. Доказать, что если A — квадратная $n\times n$-матрица, u — вектор размерности n, то справедливы соотношения:

если A произвольного вида:	если A симметричная:
$\frac{\partial}{\partial u} u^\intercal A u = A^\intercal u + Au$;	$\frac{\partial}{\partial u} u^\intercal A u = 2Au$;
$\frac{\partial}{\partial A} \ln\lvert A\rvert = A^{-1\intercal}$;	$\frac{\partial}{\partial A} \ln\lvert A\rvert = 2A^{-1} - \operatorname{diag} A^{-1}$;
$\frac{\partial}{\partial A} u^\intercal A u = uu^\intercal$;	$\frac{\partial}{\partial A} u^\intercal A u = 2uu^\intercal - \operatorname{diag} uu^\intercal$;

Упр. 1.6. Пользуясь результатами Упражнения 1.5, доказать Теорему 1.5 об оценивании параметров многомерного нормального распределения по максимуму взвешенного правдоподобия.

Упр. 1.7. Доказать Теорему 1.6 о несмещённой оценке ковариационной матрицы многомерного нормального распределения.

Упр. 1.8. Вывести несмещённую оценку общей ковариационной матрицы классов $\hat{\Sigma}$, при условии, что матожидания $\hat{\mu}_y$, $y \in Y$ оцениваются по той же выборке, см. Замечание 1.4.

Упр. 1.9. Выписать алгоритм обучения линейного дискриминанта Фишера при «наивном» предположении о независимости признаков.

Упр. 1.10. Найти стационарную точку оптимизационной задачи (1.20) для случая, когда компоненты смеси имеют n-мерные сферические нормальные распределения с параметрами $\theta_j = (\mu_j, \sigma_j)$, где μ_j — n-мерный вектор, σ_j — скаляр:

$$p_j(x) = (\sigma_j\sqrt{2\pi})^{-n} \exp\left(-\tfrac{1}{2}\sigma_j^{-2}\|x-\mu_j\|^2\right), \quad j = 1,\ldots,k.$$

Решения

Решение 1.5.

1. Распишем производную по вектору u покомпонентно:

$$\frac{\partial}{\partial u_i} u^\intercal A u = \frac{\partial}{\partial u_i} \sum_{s=1}^n \sum_{t=1}^n a_{st} u_s u_t = \sum_{s=1}^n \sum_{t=1}^n a_{st}(\delta_{is} u_t + \delta_{it} u_s) = \sum_{t=1}^n a_{it} u_t + \sum_{s=1}^n a_{si} u_s,$$

где $\delta_{ab} = [a = b]$ — символ Кронекера. Тогда в векторной записи

$$\frac{\partial}{\partial u} u^\intercal A u = A^\intercal u + A u.$$

2. Если матрица A симметричная ($A^\intercal = A$), то

$$\frac{\partial}{\partial u} u^\intercal A u = 2Au.$$

3. Теперь распишем производные по матрице A. Рассмотрим сначала случай, когда A — произвольная матрица, на элементы которой не наложено никаких дополнительных ограничений. Выпишем разложение определителя A по i-й строке:

$$|A| = \sum_{s=1}^n a_{is} A_{is},$$

где A_{is} — алгебраическое дополнение элемента a_{is}. Нам понадобятся два свойства алгебраических дополнений. Во-первых, A_{is} не зависит от элементов i-й строки и s-го столбца матрицы A. Во-вторых, A_{is} связано с элементами обратной матрицы $(b_{ij})_{n\times n} = B = A^{-1}$ соотношением $b_{ij} = A_{ji}/|A|$. Отсюда следует, что

$$\frac{\partial}{\partial a_{ij}} |A| = A_{ij} = b_{ji}|A|,$$

или в векторной записи

$$\frac{\partial}{\partial A} |A| = |A| A^{-1\intercal}.$$

Теперь легко найти и производную от логарифма определителя:

$$\frac{\partial}{\partial A} \ln|A| = \frac{1}{|A|}|A|A^{-1\intercal} = A^{-1\intercal}.$$

4. Наконец, производная квадратичной формы $u^\intercal A u$ по A есть

$$\frac{\partial}{\partial a_{ij}} u^\intercal A u = \frac{\partial}{\partial a_{ij}} \sum_{s=1}^n \sum_{t=1}^n a_{st} u_s u_t = u_i u_j,$$

или в векторной записи

$$\frac{\partial}{\partial A} u^\intercal A u = u u^\intercal.$$

5. Если матрица A — симметричная, всё немного усложняется, так как элементы матрицы теперь связаны дополнительными ограничениями $a_{ij} = a_{ji}$. Теперь любая функция от матрицы A имеет не n^2 аргументов, а только $n(n+1)/2$ аргументов a_{ij}, $i \leqslant j$. Выпишем разложение определителя симметричной матрицы A по i-й строке, пользуясь тем, что $a_{ij} = a_{ji}$ и $A_{ij} = A_{ji}$:

$$|A| = \sum_{s<i} a_{is}A_{is} + \sum_{s>i} a_{is}A_{is} + a_{ii}A_{ii} = 2\sum_{s<i} a_{is}A_{is} + a_{ii}A_{ii},$$

где A_{is} — алгебраическое дополнение элемента a_{is}. Отсюда следует

$$\frac{\partial}{\partial a_{ij}}|A| = \begin{cases} 2A_{ij} & i < j; \\ A_{ij} & i = j; \end{cases}$$

или в векторной записи

$$\frac{\partial}{\partial A}|A| = |A|(2A^{-1} - \operatorname{diag} A^{-1}); \qquad \frac{\partial}{\partial A}\ln|A| = 2A^{-1} - \operatorname{diag} A^{-1}.$$

6. Наконец, для симметричной матрицы A

$$u^{\intercal}Au = \sum_{s=1}^{n}\sum_{t=1}^{n}[s<t]a_{st}u_su_t + \sum_{s=1}^{n}\sum_{t=1}^{n}[s>t]a_{st}u_su_t + \sum_{s=1}^{n}a_{ss}u_su_s =$$
$$= 2\sum_{s=1}^{n}\sum_{t=1}^{n}[s\leqslant t]a_{st}u_su_t - \sum_{s=1}^{n}a_{ss}u_s^2.$$

Производная квадратичной формы $u^{\intercal}Au$ по A есть

$$\frac{\partial}{\partial a_{ij}}u^{\intercal}Au = 2u_iu_j - \delta_{ij}u_i^2; \quad i \leqslant j;$$
$$\frac{\partial}{\partial A}u^{\intercal}Au = 2uu^{\intercal} - \operatorname{diag} uu^{\intercal}.$$

Решение 1.6. Доказательство Теоремы 1.5.

Запишем логарифм плотности нормального распределения. Воспользовавшись тождеством $|\Sigma^{-1}| = |\Sigma|^{-1}$, представим $\mathcal{N}$ как функцию от Σ^{-1}, а не от самой ковариационной матрицы Σ:

$$\ln\mathcal{N}(x;\mu,\Sigma^{-1}) = \operatorname{const}(\mu,\Sigma^{-1}) + \tfrac{1}{2}\ln|\Sigma^{-1}| - \tfrac{1}{2}(x-\mu)^{\intercal}\Sigma^{-1}(x-\mu).$$

Возьмём производные от $\ln\mathcal{N}$ по вектору матожидания μ и матрице Σ^{-1}:

$$\frac{\partial}{\partial\mu}\ln\mathcal{N}(x_i;\mu,\Sigma^{-1}) = -\Sigma^{-1}(x_i-\mu);$$
$$\frac{\partial}{\partial\Sigma^{-1}}\ln\mathcal{N}(x_i;\mu,\Sigma^{-1}) = \frac{1}{2}\big(2\Sigma - \operatorname{diag}\Sigma\big) - \frac{1}{2}\big(2(x_i-\mu)(x_i-\mu)^{\intercal} - \operatorname{diag}(x_i-\mu)(x_i-\mu)^{\intercal}\big).$$

Необходимые условия минимума функционала $L(X^m, G^m;\theta)$, см. (1.13):

$$\frac{\partial L}{\partial\mu} = \sum_{i=1}^{m} g_i\frac{\partial}{\partial\mu}\ln\mathcal{N}(x_i;\mu,\Sigma) = 0;$$
$$\frac{\partial L}{\partial\Sigma^{-1}} = \sum_{i=1}^{m} g_i\frac{\partial}{\partial\Sigma^{-1}}\ln\mathcal{N}(x_i;\mu,\Sigma) = 0;$$

Подставляя сюда производную $\ln\mathcal{N}$ по вектору μ, получим:

$$\sum_{i=1}^{m} g_i\Sigma^{-1}(x_i-\mu) = 0.$$

Умножим это равенство слева на Σ, вынесем μ за знак суммирования, и, с учётом нормировки $\sum_{i=1}^{m} g_i = 1$, получим первое соотношение, утверждаемое теоремой.

Введём обозначения

$$S(x_i) = \Sigma - (x_i - \mu)(x_i - \mu)^\intercal, \qquad S = \sum_{i=1}^{m} g_i S(x_i).$$

В этих обозначениях производная L по матрице Σ^{-1} примет вид

$$\frac{\partial L}{\partial \Sigma^{-1}} = \frac{1}{2}\sum_{i=1}^{m} g_i\big(2S(x_i) - \operatorname{diag} S(x_i)\big) = S - \frac{1}{2}\operatorname{diag} S = 0.$$

Последнее равенство выполняется тогда и только тогда, когда $S = 0$. Следовательно,

$$\Sigma \sum_{i=1}^{m} g_i = \sum_{i=1}^{m} g_i (x_i - \mu)(x_i - \mu)^\intercal.$$

откуда, с учётом той же нормировки, получаем второе соотношение.

Решение 1.8. Чтобы получить несмещённую оценку ковариационной матрицы, найдём матожидание оценки максимума правдоподобия $\hat{\Sigma}$:

$$\begin{aligned}
\mathsf{E}\hat{\Sigma} &= \mathsf{E}\frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^\intercal = \\
&= \mathsf{E}\left(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^\intercal - 2x_i\hat{\mu}^\intercal + \hat{\mu}\hat{\mu}^\intercal\right) = \mathsf{E}\left(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^\intercal - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} x_i x_j^\intercal\right) = \\
&= \mathsf{E}\left(\frac{1}{m}\sum_{i=1}^{m} x_i x_i^\intercal - \frac{1}{m^2}\sum_{i=1}^{m} x_i x_i^\intercal - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m} x_i x_j^\intercal\right) = \\
&= \mathsf{E}xx^\intercal - \frac{1}{m}\mathsf{E}xx^\intercal - \frac{1}{m^2}m(m-1)\mathsf{E}x\mathsf{E}x^\intercal = \\
&= \frac{m-1}{m}\big(\mathsf{E}xx^\intercal - \mu\mu^\intercal\big) = \frac{m-1}{m}\Sigma.
\end{aligned}$$

Следовательно, несмещённой оценкой является $\hat{\Sigma} = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^\intercal$.

Решение 1.9. Выборочные оценки матожидания и дисперсии для y-го класса и j-го признака:

$$\hat{\mu}_{yj} = \frac{1}{\ell_y}\sum_{i=1}^{\ell}[y_i = y] f_j(x_i), \quad y \in Y, \quad j = 1, \ldots, n;$$

$$\hat{\sigma}^2_{yj} = \frac{1}{\ell_y - 1}\sum_{i=1}^{\ell}[y_i = y]\big(f_j(x_i) - \hat{\mu}_{yj}\big)^2, \quad y \in Y, \quad j = 1, \ldots, n.$$

Алгоритм классификации:

$$a(x) = \arg\max_{y\in Y}\left(2\ln\lambda_y\ell_y - \sum_{j=1}^{n}\ln\hat{\sigma}^2_{yj} - \sum_{j=1}^{n}\frac{(f_j(x) - \hat{\mu}_{yj})^2}{\hat{\sigma}^2_{yj}}\right).$$

Список литературы

[1] *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.

[2] *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985.

[3] *Епанечников В. А.* Непараметрическая оценка многомерной плотности вероятности // *Теория вероятностей и её применения.* — 1969. — Т. 14, № 1. — С. 156–161.

[4] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // *Проблемы кибернетики.* — 1978. — Т. 33. — С. 5–68. http://www.ccas.ru/frc/papers/zhuravlev78prob33.pdf.

[5] *Лапко А. В., Ченцов С. В., Крохов С. И., Фельдман Л. А.* Обучающиеся системы обработки информации и принятия решений. Непараметрический подход. — Новосибирск: Наука, 1996.

[6] *Орлов А. И.* Нечисловая статистика. — М.: МЗ-Пресс, 2004.

[7] *Хардле В.* Прикладная непараметрическая регрессия. — М.: Мир, 1993.

[8] *Шлезингер М., Главач В.* Десять лекций по статистическому и структурному распознаванию. — Киев: Наукова думка, 2004.

[9] *Шлезингер М. И.* О самопроизвольном различении образов // Читающие автоматы. — Киев, Наукова думка, 1965. — Pp. 38–45.

[10] *Шурыгин А. М.* Прикладная стохастика: робастность, оценивание, прогноз. — М.: Финансы и статистика, 2000.

[11] *Dempster A. P., Laird N. M., Rubin D. B.* Maximum likelihood from incomplete data via the EM algorithm // *J. of the Royal Statistical Society, Series B.* — 1977. — no. 34. — Pp. 1–38.

[12] *Fisher R. A.* The use of multiple measurements in taxonomic problem // *Ann. Eugen.* — 1936. — no. 7. — Pp. 179–188.

[13] *Jordan M. I., Xu L.* Convergence results for the EM algorithm to mixtures of experts architectures: Tech. Rep. A.I. Memo No. 1458: MIT, Cambridge, MA, 1993.

[14] *Parzen E.* On the estimation of a probability density function and mode // *Annals of Mathematical Statistics.* — 1962. — Vol. 33. — Pp. 1065–1076. http://citeseer.ist.psu.edu/parzen62estimation.html.

[15] *Rosenblatt M.* Remarks on some nonparametric estimates of a density function // *Annals of Mathematical Statistics.* — 1956. — Vol. 27, no. 3. — Pp. 832–837.

[16] *Wu C. F. G.* On the convergence properties of the EM algorithm // *The Annals of Statistics.* — 1983. — no. 11. — Pp. 95–103. http://citeseer.ist.psu.edu/78906.html.