Байесовская классификация. Непараметрические методы

К. В. Воронцов

23 февраля 2010 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 Байесовский подход к классификации

Байесовский подход является классическим в теории распознавания образов и лежит в основе многих методов. Он опирается на теорему о том, что если плотности распределения классов известны, то алгоритм классификации, имеющий минимальную вероятность ошибок, можно выписать в явном виде. Для оценивания плотностей классов по выборке применяются различные подходы. В этом курсе рассматривается три: параметрический, непараметрический и оценивание смесей распределений.

§1.1 Вероятностная постановка задачи классификации

Пусть X — множество объектов, Y — конечное множество имён классов, множество $X \times Y$ является вероятностным пространством с плотностью распределения $p(x,y) = \mathsf{P}(y)p(x|y)$. Вероятности появления объектов каждого из классов $P_y = \mathsf{P}(y)$ называются *априорными вероятностями* классов. Плотности распределения $p_y(x) = p(x|y)$ называются *функциями правдоподобия* классов[1]. Вероятностная постановка задачи классификации разделяется на две независимые подзадачи.

Задача 1.1. *Имеется простая выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ из неизвестного распределения $p(x,y) = P_y p_y(x)$. Требуется построить эмпирические оценки[2] априорных вероятностей $\hat{P}_y$ и функций правдоподобия $\hat{p}_y(x)$ для каждого из классов $y \in Y$.*

Задача 1.2. *По известным плотностям распределения $p_y(x)$ и априорным вероятностям P_y всех классов $y \in Y$ построить алгоритм $a(x)$, минимизирующий вероятность ошибочной классификации.*

Вторая задача решается относительно легко, и мы сразу это сделаем. Первая задача имеет множество решений, поскольку многие распределения $p(x,y)$ могли бы дать одну и ту же выборку X^ℓ. Приходится привлекать различные предположения о плотностях, что и приводит к большому разнообразию байесовских методов.

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 $A_y = \{x \in X \mid a(x) = y\}$, $y \in Y$. Вероятность того, что появится объект класса y и алгоритм a отнесёт его к классу s, равна $P_y\,\mathsf{P}(A_s|y)$. 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, как правило, известны заранее.

[1] Большой буквой P будем обозначать вероятности, а строчной p — плотности распределения.

[2] Символами с «крышечкой» принято обозначать *выборочные* (*эмпирические*) оценки вероятностей, функций распределения или случайных величин, вычисляемые по выборке.

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

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. Для произвольного $t \in Y$ запишем функционал среднего риска:

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

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

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

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

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

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. В этом случае формула оптимального алгоритма упрощается.

Теорема 1.2. *Если P_y и $p_y(x)$ известны, $\lambda_{yy} = 0$ 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, $\sum\limits_{y \in Y} (\lambda_{ys} - \lambda_{yt}) P_y\, p_y(x) = \lambda_t P_t p_t(x) - \lambda_s P_s p_s(x) = \tilde{g}_t(x) - \tilde{g}_s(x)$, 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 s и t — 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 s, t, что не повлияет на средний риск $R(a)$.

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. 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*.

Минимальное значение среднего риска $R(a)$, достигаемое байесовским решающим правилом, называется *байесовским риском* или *байесовским уровнем ошибки*.

Если классы равнозначны ($\lambda_y \equiv 1$), то байесовское правило называется также *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.

1.1.3 Задача восстановления плотности распределения

Перейдём к Задаче 1.1. 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 $X^\ell_y = \big\{(x_i,y_i)_{i=1}^\ell \;\big|\; y_i = y\big\}$.

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 других содержательных соображений.

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.*

«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 проще, чем одну n-мерную плотность. Однако оно крайне редко выполняется на практике, поэтому алгоритмы классификации, использующие (1.4), называются *наивными байесовскими* (naïve Bayes).

Подставим эмпирические оценки одномерных плотностей $\hat{p}_{yj}(\xi)$ в (1.4) и затем в (1.2). Получим алгоритм

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

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

Основной его недостаток — низкое качество классификации. Он используется либо как эталон при экспериментальном сравнении алгоритмов, либо как элементарный «строительный блок» в алгоритмических композициях, **??**.

§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).

1.2.1 Непараметрические оценки плотности

Локальное оценивание опирается на само определение плотности. Рассмотрим сначала одномерные оценки. Несмотря на простоту, они уже могут быть использованы для конструирования «наивных» байесовских классификаторов (1.5).

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, что приводит к появлению широких *зон неуверенности*, в которых максимум (1.2) достигается одновременно для нескольких классов $y \in Y$. Проблема решается с помощью *локальной непараметрической оценки* Парзена-Розенблатта [16, 15]:

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

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. 1, стр. 8.

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).

Обоснованием оценки (1.7) служит теорема, утверждается, что $\hat{p}_h(x)$ поточечно сходится к истинной плотности $p(x)$ для широкого класса ядер при увеличении длины выборки m и одновременном уменьшении ширины окна h [15, 16, 8].

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, 6]:

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

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. Сходимость оценки (1.10) доказана при некоторых дополнительных ограничениях на ядро K и метрику ρ, причём скорость сходимости не сильно отличается от одномерного случая [7].

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 y, то в байесовском классификаторе (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):

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

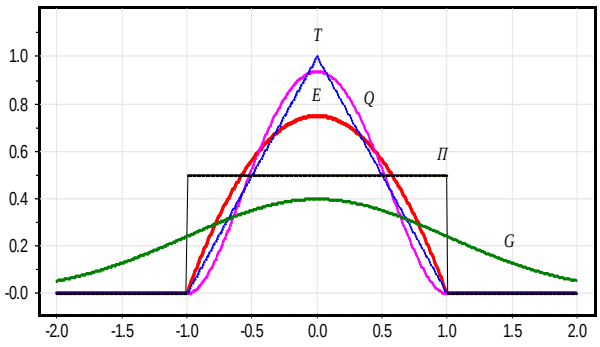

Рис. 1. Часто используемые ядра:
E — Епанечникова;
Q — квартическое;
T — треугольное;
G — гауссовское;
$П$ — прямоугольное.

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

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

Ширина окна h существенно влияет на качество восстановления плотности. При $h \to 0$ плотность концентрируется вблизи обучающих объектов, и функция $\hat{p}_h(x)$ претерпевает резкие скачки. При $h \to \infty$ плотность чрезмерно сглаживается и вырождается в константу. Следовательно, должно существовать компромиссное значение ширины окна h^*. На практике его находят по скользящему контролю:

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

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

Переменная ширина окна $h(x)$. Если распределение объектов в пространстве X сильно неравномерно, то возникает *проблема локальных сгущений*. Одно и то же значение ширины окна h приводит к чрезмерному сглаживанию плотности в одних областях пространства X, и недостаточному сглаживанию в других. Проблему решает *переменная ширина окна*, определяемая в каждой точке $x \in X$ как расстояние до $(k+1)$-го соседа $h(x) = \rho\big(x, x^{(k+1)}\big)$, если считать, что обучающие объекты ранжированы по возрастанию расстояний до x.

Нормирующий множитель $V(h)$ не должен зависеть от y, поэтому в числе соседей должны учитываться объекты всех классов, хотя плотности $\hat{p}_{y,h(x)}(x)$ оцениваются по подвыборкам X_y^ℓ для каждого класса $y \in Y$ в отдельности. Оптимальное значение k^* определяется по критерию скользящего контроля, аналогично h^*.

Функция ядра K практически не влияет на качество восстановления плотности и на качество классификации. В то же время, она определяет степень гладкости функции $\hat{p}_h(x)$. 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, рис. 1), и для них достаточно взять только те точки выборки, которые попадают в окрестность точки x радиуса h.

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* предполагается, что плотность распределения выборки $X^m = \{x_1, \ldots, x_m\}$ известна с точностью до параметра, $p(x) = \varphi(x; \theta)$, где φ — фиксированная функция. Вектор параметров θ оценивается по выборке X^m с помощью *принципа максимума правдоподобия* (maximum likelihood).

Нормальный дискриминантный анализ — это специальный случай байесовской классификации, когда предполагается, что плотности всех классов $p_y(x)$, $y \in Y$ являются многомерными нормальными. Этот случай интересен и удобен тем, что задача оценивания параметров распределения по выборке решается аналитически.

1.3.1 Многомерное нормальное распределение

Пусть $X = \mathbb{R}^n$, то есть объекты описываются n числовыми признаками.

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 математическим ожиданием (центром) $\mu \in \mathbb{R}^n$ и ковариационной матрицей $\Sigma \in \mathbb{R}^{n \times n}$. Предполагается, что матрица Σ симметричная, невырожденная, положительно определённая.

Интегрируя по $\mathbb{R}^n$, можно убедиться в том, что это действительно распределение, а параметры μ и Σ оправдывают своё название:

$$\int \mathcal{N}(x; \mu, \Sigma)\, dx = 1;$$

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

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

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.

1.3.2 Квадратичный дискриминант

Рассмотрим задачу классификации с произвольным числом классов.

Теорема 1.3. *Если классы имеют n-мерные нормальные плотности распределения*

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

то байесовский классификатор задаёт квадратичную разделяющую поверхность. Она вырождается в линейную, если ковариационные матрицы классов равны.

Доказательство. Поверхность, разделяющая классы s и t, описывается уравнением $\lambda_s P_s p_s(x) = \lambda_t P_t p_t(x)$, которое после логарифмирования принимает вид

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

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. Рассмотрим простейший случай, когда классы s, t равновероятны и равнозначны ($\lambda_s P_s = \lambda_t P_t$), ковариационные матрицы равны, признаки некоррелированы и имеют одинаковые дисперсии ($\Sigma_s = \Sigma_t = \sigma I_n$). 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 свойством

оптимальности — это та прямая, в одномерной проекции на которую классы разделяются наилучшим образом, то есть с наименьшим значением байесовского риска $R(a)$.

Если признаки коррелированы ($\Sigma_s = \Sigma_t \neq \sigma I_n$), то ортогональность исчезает, однако разделяющая гиперплоскость по-прежнему проходит посередине между классами, касательно к линиям уровня обоих распределений.

Если классы не равновероятны или не равнозначны ($\lambda_s P_s \neq \lambda_t P_t$), то разделяющая гиперплоскость отодвигается дальше от более значимого класса.

Если ковариационные матрицы не диагональны и не равны, то 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 парадоксальным ситуациям. Например, может возникнуть область, в которой объекты будут относиться к менее плотному классу, хотя объекты более плотного класса находятся ближе.

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 евклидовой метрикой. В этом случае оптимальным (байесовским) решающим правилом является *классификатор ближайшего среднего* (nearest mean classifier), относящий объект к классу с ближайшим центром.

Принцип максимума правдоподобия составляет основу параметрического подхода. Пусть задано множество объектов $X^m = \{x_1, \ldots, x_m\}$, выбранных независимо друг от друга из вероятностного распределения с плотностью $\varphi(x;\theta)$. *Функцией правдоподобия* называется совместная плотность распределения всех объектов выборки:

$$p(X^m;\theta) = p(x_1,\ldots,x_m;\theta) = \prod_{i=1}^{m} \varphi(x_i;\theta).$$

Значение параметра θ, при котором выборка максимально правдоподобна, то есть функция $p(X^m;\theta)$ принимает максимальное значение, называется *оценкой максимума правдоподобия*. Как известно из математической статистики, эта оценка обладает рядом оптимальных свойств [4, 5].

Вместо максимизации правдоподобия удобнее максимизировать его логарифм:

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

Для решения этой задачи можно использовать стандартные методы оптимизации. В некоторых случаях решение выписывается в явном виде, исходя из необходимого условия оптимума (если функция $\varphi(x;\theta)$ достаточно гладкая по параметру θ):

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

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

Теорема 1.4. *Пусть задана независимая выборка объектов $X^m = (x_1, \ldots, x_m)$. Тогда оценки параметров гауссовской плотности $\varphi(x;\theta) \equiv \mathcal{N}(x;\mu,\Sigma)$, доставляющие максимум функционалу правдоподобия (1.13), 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 $\mathsf{E}_{X^m}\hat{\theta}(X^m) = \theta_0$.*

Оценка $\hat{\mu}$ является несмещённой, $\mathsf{E}\hat{\mu} = \mu$. Однако оценка $\hat{\Sigma}$ уже смещена: $\mathsf{E}\hat{\Sigma} = \frac{m}{m-1}\Sigma$. Это связано с тем, что при вычислении $\hat{\Sigma}$ вместо неизвестного точного значения матожидания μ приходится подставлять его выборочную оценку $\hat{\mu}$. Для учёта этого небольшого смещения вводится *поправка на смещение*:

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

Подстановочный алгоритм. Оценим параметры функций правдоподобия $\hat{\mu}_y$ и $\hat{\Sigma}_y$ по частям обучающей выборки $X_y^\ell = \{x_i \in X^\ell \mid y_i = y\}$ для каждого класса $y \in Y$. Затем эти выборочные оценки подставим в формулу (1.2). Получим *байесовский нормальный классификатор*, который называется также *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 оценки, сделанные по коротким выборкам, могут быть не достаточно точными.

Недостатки подстановочного алгоритма вытекают из нескольких чрезмерно сильных базовых предположений, которые на практике часто не выполняются.

- Функции правдоподобия классов могут существенно отличаться от гауссовских. В частности, когда имеются признаки, принимающие дискретные значения, или когда классы распадаются на изолированные сгустки.

- Если длина выборки меньше размерности пространства, $\ell_y < n$, или среди признаков есть линейно зависимые, то матрица $\hat{\Sigma}_y$ становится вырожденной. В этом случае обратная матрица не существует и метод вообще неприменим.

- На практике встречаются задачи, в которых признаки «почти линейно зависимы». Тогда матрица $\hat{\Sigma}_y$ является *плохо обусловленной*, то есть близкой к некоторой вырожденной матрице. Это так называемая *проблема мультиколлинеарности*, которая влечёт неустойчивость как самой обратной матрицы $\hat{\Sigma}_y^{-1}$, так и вектора $\hat{\Sigma}_y^{-1}(x - \mu_{st})$. Они могут непредсказуемо и сильно изменяться при незначительных вариациях исходных данных, например, связанных с погрешностями измерений. Неустойчивость снижает качество классификации.

- Выборочные оценки чувствительны к нарушениям нормальности распределений, в частности, к редким большим выбросам.

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 до крайности прост и сводится к вычислению параметров $\hat{\mu}_{yj}$ и $\hat{\sigma}_{yj}$ для всех $y \in Y$ и всех признаков $j = 1, \ldots, n$.

1.3.3 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 [13]. Предположим, что ковариационные матрицы классов одинаковы и равны Σ. Оценим $\hat{\Sigma}$ по всем ℓ обучающим объектам. С учётом поправки на смещённость,

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

Согласно теореме 1.3, разделяющая поверхность линейна, если классов два, и кусочно-линейна, если классов больше. Запишем подстановочный алгоритм:

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

Этот алгоритм называется *линейным дискриминантом Фишера* (ЛДФ). Он неплохо работает, когда формы классов действительно 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.

Регуляризация ковариационной матрицы. Эвристика Фишера не является радикальным решением проблемы мультиколлинеарности: общая ковариационная матрица классов $\hat{\Sigma}$ также может оказаться плохо обусловленной (близкой к вырожденной). Признаком плохой обусловленности является наличие у матрицы собственных значений, близких к нулю. Поэтому один из способов решения проблемы — модифицировать матрицу $\hat{\Sigma}$ так, чтобы все её собственные значения λ увеличились на заданное число τ, а все собственные векторы v сохранились. Для этого достаточно прибавить к матрице единичную, умноженную на τ:

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

Известны и другие способы решения проблемы плохой обусловленности. Можно пропорционально уменьшать недиагональные элементы — вместо $\hat{\Sigma}$ брать матрицу $(1-\tau)\hat{\Sigma} + \tau \operatorname{diag}\hat{\Sigma}$ [11]. Можно занулять недиагональные элементы матрицы, соответствующие тем парам признаков, ковариации которых незначимо отличаются от нуля [2]; при этом матрица становится разреженной, и для её обращения могут применяться специальные, более эффективные, алгоритмы. Можно разбивать множество признаков на группы и полагать, что признаки из разных групп не коррелированы. Тогда матрица $\hat{\Sigma}$ приобретает блочно-диагональный вид. Для таких матриц также существуют специальные эффективные алгоритмы обращения.

Отбор и преобразование признаков. Другой способ устранения мультиколлинеарности заключается в том, чтобы отбросить наименее значимые признаки. Различные методы *отбора признаков* (features selection) рассматриваются в разделе **??**. Обратим внимание на кажущийся парадокс: информация отбрасывается, но качество классификации повышается.

Существует простой «жадный» метод отбора признаков, в котором исходная n-мерная задача сводится к серии двумерных задач. Это метод *редукции размерности* А. М. Шурыгина [11]. Сначала находятся два признака, в подпространстве которых линейный дискриминант Фишера наилучшим образом разделяет классы (т. е. байесовский риск принимает наименьшее значение). Обозначим его параметры через $\left\{\alpha_y^{(2)}, \beta_y^{(2)} \colon y \in Y\right\}$, где в векторе $\alpha_y^{(2)} \in \mathbb{R}^n$ лишь два коэффициента не равны нулю. Функция проекции на нормаль разделяющей гиперплоскости принимается за новый признак: $\psi(x) = x^{\intercal}\alpha_y^{(2)}$. Из оставшихся $n-2$ признаков выбирается тот, который в паре с $\psi(x)$ наилучшим образом разделяет классы. Снова строится двумерный линейный дискриминант и определяются параметры $\left\{\alpha_y^{(3)}, \beta_y^{(3)} \colon y \in Y\right\}$, где в векторе $\alpha_y^{(3)} \in \mathbb{R}^n$ уже три коэффициента не равны нулю. Так продолжается до тех пор, пока присоединение новых признаков улучшает качество классификации (понижает байесовский риск). Достоинства метода редукции — возможность отбросить неинформативные признаки и обойтись без обращения ковариационных матриц размера более 2×2. В некоторых прикладных задачах он превосходит другие методы классификации [11]. Недостатком является отсутствие строгого теоретического обоснования. Как и всякая «жадная» стратегия, он находит неоптимальный набор

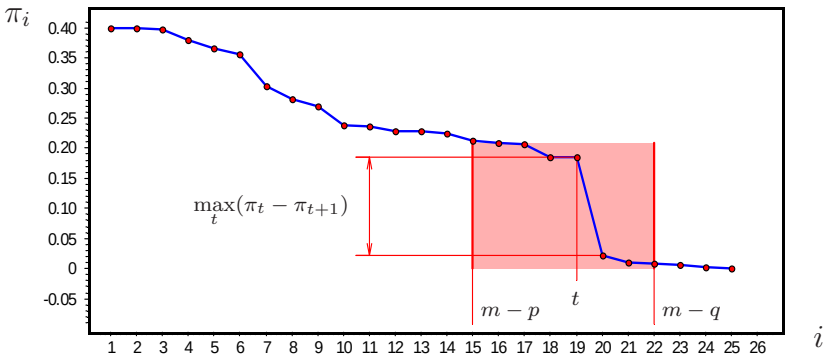

Рис. 2. Эксперт предположил, что в выборке длины $m = 25$ находится от $q = 3$ до $p = 10$ выбросов. Более точное число выбросов, равное 6, удалось определить по критерию «крутого склона».

признаков. По всей видимости, этот метод подходит для тех задач, в которых имеется небольшое число признаков, существенно более информативных, чем остальные.

Ещё один способ сокращения размерности заключается в том, чтобы из имеющегося набора признаков построить новый набор, состоящий из наименьшего числа максимально информативных признаков. Оптимальное *линейное* преобразование пространства признаков строится с помощью *метода главных компонент* (principal component analysis), который будет рассмотрен в разделе **??**.

Робастные методы оценивания. Оценки, устойчивые относительно редких больших выбросов или других несоответствий реальных данных модельному распределению $\varphi(x;\theta)$, называются *робастными* (robust — здравый).

Простейший метод робастного оценивания параметра θ по выборке X^m основан на фильтрации выбросов путём двукратного решения задачи. Сначала оценка максимума правдоподобия для параметра $\hat{\theta}$ вычисляется по всей выборке X^m. Для каждого объекта $x_i \in X^m$ вычисляется значение правдоподобия $\pi_i = \varphi(x_i;\hat{\theta})$ и объекты сортируются по убыванию правдоподобий: $\pi_1 \geqslant \ldots \geqslant \pi_m$. Объекты, оказавшиеся в конце этого ряда, считаются нетипичными (выбросами) и удаляются из выборки. Обычно для этого применяется критерий «крутого склона»: задаются два параметра p и q, и находится значение $t \in \{m-p, \ldots, m-q-1\}$, для которого скачок правдоподобия $\pi_t - \pi_{t+1}$ максимален. Затем последние $(m-t)$ объектов удаляются из выборки, см. рис. 2. После этого оценка $\hat{\theta}$ вычисляется вторично по отфильтрованной выборке. В некоторых случаях приходится проводить несколько таких фильтраций подряд.

Заметим, что удаление объекта может не требовать полного пересчёта оценки $\hat{\theta}$. В частности, оценки нормального распределения $\hat{\mu}$, $\hat{\Sigma}$ аддитивны по объектам, и для них достаточно вычесть слагаемое, соответствующее удаляемому объекту.

§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 1.2. *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 1.1. Общая идея EM-алгоритма

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

Этот алгоритм был предложен и исследован М. И. Шлезингером как инструмент для *самопроизвольной классификации образов* [10]. Двенадцать лет спустя он был открыт заново в [12] под названием *EM-алгоритма*. Область его применения чрезвычайно широка — дискриминантный анализ, кластеризация, восстановление пропусков в данных, обработка сигналов и изображений [9]. 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

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

Введём обозначение $g_{ij} \equiv \mathsf{P}(\theta_j \mid x_i)$. 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

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

Таким образом, М-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 [12, 17, 14].

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 1.2. На Е-шаге вычисляется матрица скрытых переменных G по формуле (1.17). На М-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 М-шаге алгоритма. Достаточно лишь сместиться в направлении максимума, сделав одну или несколько итераций, и затем выполнить Е-шаг. Этот алгоритм также обладает неплохой сходимостью и называется *обобщённым EM-алгоритмом* (generalized EM-algorithm, GEM) [12].

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

Алгоритм 1.2. 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;

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

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.

Алгоритм 1.3. ЕМ-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: $\texttt{EM}\,(X^m, k, \Theta, \delta)$;

ЕМ-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 ЕМ-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 1.3.

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

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

Стохастический EM-алгоритм. Максимизируемый функционал $Q(\Theta)$ в общем случае может иметь большое количество локальных экстремумов. Поэтому EM-алгоритму присущи обычные недостатки любого детерминированного процесса многоэкстремальной оптимизации: застревание в локальных максимумах, 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 1.2 в том, что на M-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

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

где выборки X_j генерируются из X^m путём стохастического моделирования. Для каждого объекта $x_i \in X^m$ разыгрывается случайное значение $j(i)$ из дискретного распределения вероятностей $\{g_{ij} \colon j = 1, \ldots, k\}$, и объект $x_i \in X^m$ включается только в выборку $X_{j(i)}$.

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, $|X_j| \leqslant m_0$, 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 максимумов:

- SEM работает быстрее обычного детерминированного EM, и его результаты меньше зависят от начального приближения.
- Как правило, SEM находит экстремум $Q(\Theta)$, более 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 1.3. *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 1.5. *Если справедлива Гипотеза 1.3, 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 **??** и оценки (1.19).

Таким образом, М-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 Е-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 1.3 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, $\Sigma_j = \sigma_j^2 I_n$. 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 1.4. *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, $j = 1, \ldots, k$:*

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

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

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

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*.

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 1.5. *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 3.

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.

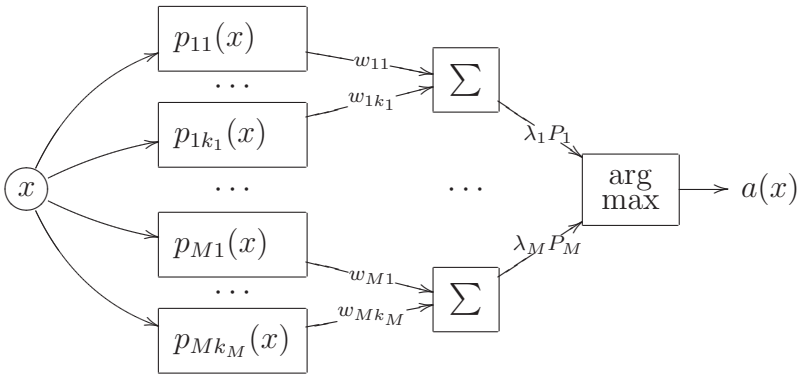

Рис. 3. Сеть радиальных базисных функций представляет собой трёхуровневую суперпозицию.

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 плотностей классов $p_y(x)$ с помощью EM-алгоритма. Результатом обучения являются центры μ_{yj} и дисперсии Σ_{yj} компонент $j = 1, \ldots, k_y$. Оценивая дисперсии, мы фактически подбираем веса признаков в метриках $\rho_{yj}(x, \mu_{yj})$ для каждого центра μ_{yj}. При использовании Алгоритма 1.3 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 **??**.

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] *Закс Ш.* Теория статистических выводов. — М.: Мир, 1975.

[5] *Лагутин М. Б.* Наглядная математическая статистика. — М.: П-центр, 2003.

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

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

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

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

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

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

[12] *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.

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

[14] *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.

[15] *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.

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

[17] *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.