Восстановление плотности и байесовская теория классификации

К. В. Воронцов
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 25 апреля 2024

Содержание

Восстановление плотности — задача обучения без учителя

Дано: простая (i.i.d.) выборка $X^\ell = \{x_1, \ldots, x_\ell\} \sim p(x)$.

Найти параметрическую модель плотности распределения:

$$p(x) = \varphi(x; \theta),$$

где θ — параметр, φ — фиксированная функция.

Критерий — максимум (логарифма) правдоподобия выборки:

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

Необходимое условие оптимума:

$$\frac{\partial}{\partial \theta} L(\theta; X^\ell) = \sum_{i=1}^{\ell} \frac{\partial}{\partial \theta} \ln \varphi(x_i; \theta) = 0,$$

где функция $\varphi(x; \theta)$ достаточно гладкая по параметру θ.

Восстановление многомерной гауссовской плотности

Пусть объекты x описываются n признаками $f_j(x) \in \mathbb{R}$
и выборка порождена n-мерной гауссовской плотностью:

$$p(x) = \mathcal{N}(x; \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^n \det \Sigma}}$$

$\mu \in \mathbb{R}^n$ — вектор математического ожидания, $\mu = \mathsf{E}x$
$\Sigma \in \mathbb{R}^{n\times n}$ — ковариационная матрица, $\Sigma = \mathsf{E}(x-\mu)(x-\mu)^\mathsf{T}$
(симметричная, невырожденная, положительно определённая)

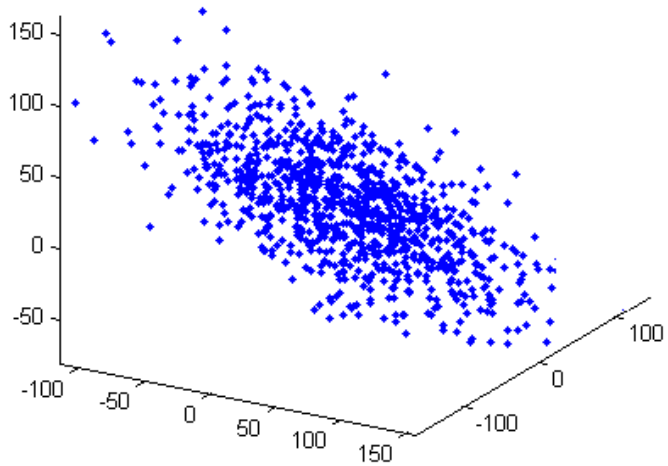
Выборочные оценки максимального правдоподобия:

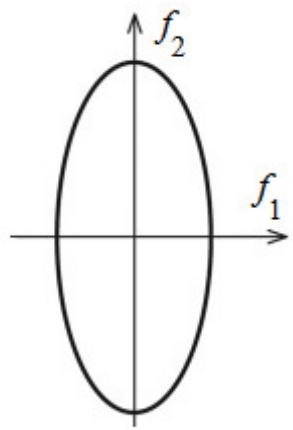
$$\frac{\partial}{\partial \mu} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i$$

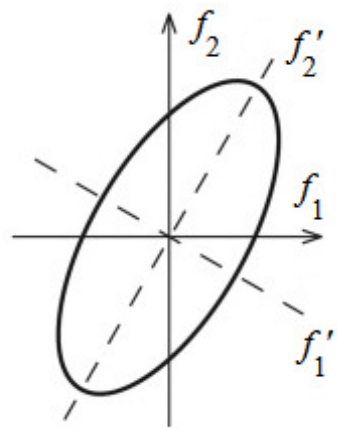
$$\frac{\partial}{\partial \Sigma} \ln L(\mu, \Sigma; X^\ell) = 0 \quad \Rightarrow \quad \hat{\Sigma} = \frac{1}{\ell}\sum_{i=1}^{\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\mathsf{T}$$

Геометрический смысл многомерной нормальной плотности

Эллипсоид рассеяния — облако точек эллиптической формы:







При $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ оси эллипсоида параллельны ортам.
В общем случае: $\Sigma = VSV^\mathsf{T}$ — спектральное разложение,
$V = (v_1, \ldots, v_n)$ — ортогональные собственные векторы,
$S = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ — собственные значения матрицы Σ

$$(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu) = (x-\mu)^\mathsf{T}VS^{-1}V^\mathsf{T}(x-\mu) = (x'-\mu')^\mathsf{T}S^{-1}(x'-\mu').$$

$x' = V^\mathsf{T}x$ — *декоррелирующее* ортогональное преобразование

Проблема мультиколлинеарности

Проблема: при $\ell < n$ матрица $\hat{\Sigma}$ вырождена, но даже при $\ell \geqslant n$ она может оказаться плохо обусловленной.

Регуляризация ковариационной матрицы $\hat{\Sigma} + \tau I_n$ увеличивает собственные значения на τ, сохраняя собственные векторы (параметр τ можно подбирать по скользящему контролю)

Диагонализация ковариационной матрицы — оценивание n одномерных плотностей признаков $f_j(x)$, $j = 1, \ldots, n$:

$$\hat{p}_j(\xi) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} \exp\left(-\frac{(\xi-\hat{\mu}_j)^2}{2\hat{\sigma}_j^2}\right), \quad j = 1, \ldots, n$$

где $\hat{\mu}_j$ и $\hat{\sigma}_j^2$ — оценки среднего и дисперсии признака j:

$$\hat{\mu}_j = \frac{1}{\ell}\sum_{i=1}^{\ell} f_j(x_i)$$
$$\hat{\sigma}_j^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} (f_j(x_i) - \hat{\mu}_j)^2$$

Задача разделения смеси распределений

Порождающая модель смеси k распределений:

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

$\varphi(x, \theta_j) = p(x|j)$ — *функция правдоподобия* j-й компоненты;
$w_j = P(j)$ — *априорная вероятность* j-й компоненты.

Задача максимизации логарифма правдоподобия:

$$L(w, \theta) = \ln \prod_{i=1}^{\ell} p(x_i) = \sum_{i=1}^{\ell} \ln \sum_{j=1}^{k} w_j \varphi(x_i, \theta_j) \to \max_{w,\theta}$$

при ограничениях $\sum_{j=1}^{k} w_j = 1; \;\; w_j \geqslant 0.$

EM-алгоритм для разделения смеси распределений

Теорема (необходимые условия экстремума)

Точка $(w_j, \theta_j)_{j=1}^k$ локального экстремума $L(w, \theta)$ удовлетворяет системе уравнений относительно w_j, θ_j и g_{ij}:

$$\text{E-шаг:} \quad g_{ij} = \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)}, \quad i = 1, \ldots, \ell, \quad j = 1, \ldots, k;$$

$$\text{M-шаг:} \quad \theta_j = \arg\max_{\theta} \sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i, \theta), \quad j = 1, \ldots, k;$$

$$w_j = \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij}, \quad j = 1, \ldots, k.$$

EM-алгоритм — метод простых итераций для решения системы

Вероятностная интерпретация

E-шаг — это формула Байеса:

$$g_{ij} = \mathrm{P}(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{\sum_{s=1}^{k} w_s\varphi(x_i,\theta_s)}.$$

Очевидно, выполнено условие нормировки: $\sum_{j=1}^{k} g_{ij} = 1$.

M-шаг — это максимизация взвешенного правдоподобия, с весами объектов g_{ij} для j-й компоненты смеси:

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

$$w_j = \frac{1}{\ell}\sum_{i=1}^{\ell} g_{ij}.$$

Доказательство. Условия Каруша–Куна–Таккера

Лагранжиан оптимизационной задачи $L(w,\theta) \to \max$:

$$\mathscr{L}(w,\theta) = \sum_{i=1}^{\ell} \ln \underbrace{\left(\sum_{j=1}^{k} w_j \varphi(x_i,\theta_j) \right)}_{p(x_i)} - \lambda \left(\sum_{j=1}^{k} w_j - 1 \right)$$

Приравниваем нулю производные:

$$\frac{\partial \mathscr{L}}{\partial w_j} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\ell} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} = \lambda w_j; \quad \lambda = \ell; \quad w_j = \frac{1}{\ell} \sum_{i=1}^{\ell} g_{ij}$$

$$\frac{\partial \mathscr{L}}{\partial \theta_j} = \sum_{i=1}^{\ell} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} \frac{\frac{\partial}{\partial \theta_j} \varphi(x_i,\theta_j)}{\varphi(x_i,\theta_j)} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{\ell} g_{ij} \ln \varphi(x_i,\theta_j) = 0$$

■

EM-алгоритм для разделения смеси распределений

вход: $X^\ell = \{x_1, \ldots, x_\ell\}, \quad k;$

выход: $(w_j, \theta_j)_{j=1}^k$ — параметры смеси распределений;

инициализировать $(\theta_j)_{j=1}^k, \quad w_j := \frac{1}{k};$

повторять

E-шаг (expectation): для всех $i = 1, \ldots, \ell, \quad j = 1, \ldots, k$

$$g_{ij} := \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)};$$

M-шаг (maximization): для всех $j = 1, \ldots, k$

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

$$w_j := \frac{1}{\ell} \sum_{i=1}^\ell g_{ij};$$

пока w_j, θ_j и/или g_{ij} не сошлись;

Разделение смеси гауссиан (Gaussian Mixture Model, GMM)

вход: $X^\ell = \{x_1, \ldots, x_\ell\}$, k;

выход: $(w_j, \mu_j, \Sigma_j)_{j=1}^k$ — параметры смеси гауссиан;

инициализировать $(\mu_j, \Sigma_j)_{j=1}^k$, $w_j := \frac{1}{k}$;

повторять

E-шаг (expectation): для всех $i = 1, \ldots, \ell$, $j = 1, \ldots, k$

$$g_{ij} := \frac{w_j \mathcal{N}(x_i; \mu_j, \Sigma_j)}{\sum_{s=1}^k w_s \mathcal{N}(x_i; \mu_s, \Sigma_s)};$$

M-шаг (maximization): для всех $j = 1, \ldots, k$

$$\mu_j := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} x_i;$$

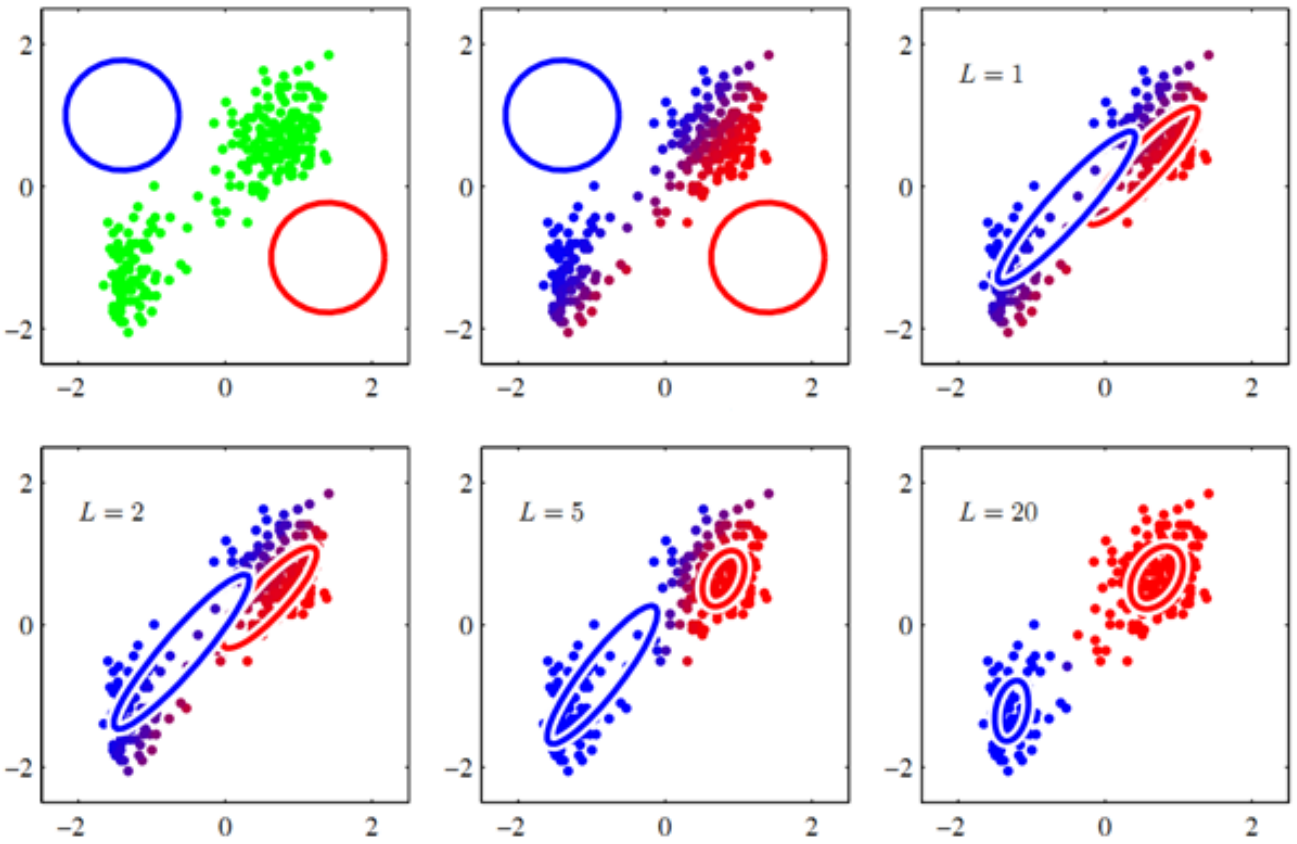
$$\Sigma_j := \frac{1}{\ell w_j} \sum_{i=1}^\ell g_{ij} (x_i - \mu_j)(x_i - \mu_j)^\mathsf{T};$$

$$w_j := \frac{1}{\ell} \sum_{i=1}^\ell g_{ij};$$

пока (w_j, μ_j, Σ_j) и/или g_{ij} не сошлись;

Пример

Две гауссовские компоненты $k = 2$ в пространстве $X = \mathbb{R}^2$.
Расположение компонент в зависимости от номера итерации L:



Задача непараметрического восстановления плотности

Задача: по выборке $X^\ell = (x_i)_{i=1}^\ell$ оценить плотность $\hat{p}(x)$, без введения параметрической модели плотности

Дискретный случай: $x_i \in D$, $|D| \ll \ell$. Гистограмма частот:

$$\hat{p}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} [x_i = x]$$

Одномерный непрерывный случай: $x_i \in \mathbb{R}$. По определению плотности, если $P[a, b]$ — вероятностная мера отрезка $[a, b]$:

$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h]$$

Эмпирическая оценка плотности по окну ширины h (заменяем вероятность на долю объектов выборки):

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

Локальная непараметрическая оценка Парзена-Розенблатта

Эмпирическая оценка плотности по окну ширины h:

$$\hat{p}_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} \frac{1}{2} \left[\frac{|x - x_i|}{h} < 1 \right].$$

Обобщение: *оценка Парзена-Розенблатта* по окну ширины h (другое название — Kernel Density Estimate, KDE):

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

где $K(r)$ — *ядро*, удовлетворяющее требованиям:
— чётная функция;
— нормированная функция: $\int K(r)\, dr = 1$;
— невозрастающая при $r > 0$, неотрицательная функция.

В частности, при $K(r) = \frac{1}{2}\big[|r| < 1\big]$ имеем эмпирическую оценку.

Два варианта обобщения на многомерный случай

1. Если объекты описываются n признаками $f_j : X \to \mathbb{R}$:

$$\hat{p}_{h_1 \ldots h_n}(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{f_j(x) - f_j(x_i)}{h_j} \right)$$

2. Если на X задана функция расстояния $\rho(x, x')$:

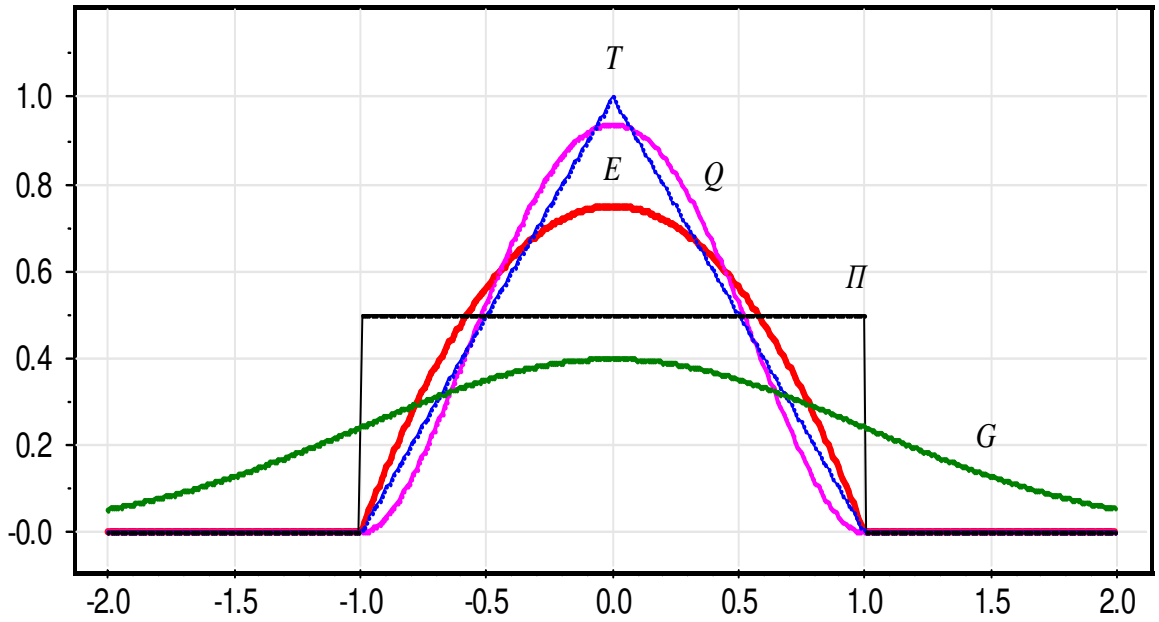
$$\hat{p}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h} \right)$$

где $V(h) = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ — нормировочный множитель

Сферическое гауссовское ядро — частный случай обоих:

$$\hat{p}_h(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi} h} \exp\left(-\frac{(f_j(x) - f_j(x_i))^2}{2h^2} \right)$$

Выбор ядра



$E(r) = \frac{3}{4}(1-r^2)\big[|r| \leqslant 1\big]$ — оптимальное (Епанечникова);

$Q(r) = \frac{15}{16}(1-r^2)^2\big[|r| \leqslant 1\big]$ — квартическое;

$T(r) = \big(1-|r|\big)\big[|r| \leqslant 1\big]$ — треугольное;

$G(r) = (2\pi)^{-1/2}\exp(-\frac{1}{2}r^2)$ — гауссовское;

$\Pi(r) = \frac{1}{2}\big[|r| \leqslant 1\big]$ — прямоугольное.

Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

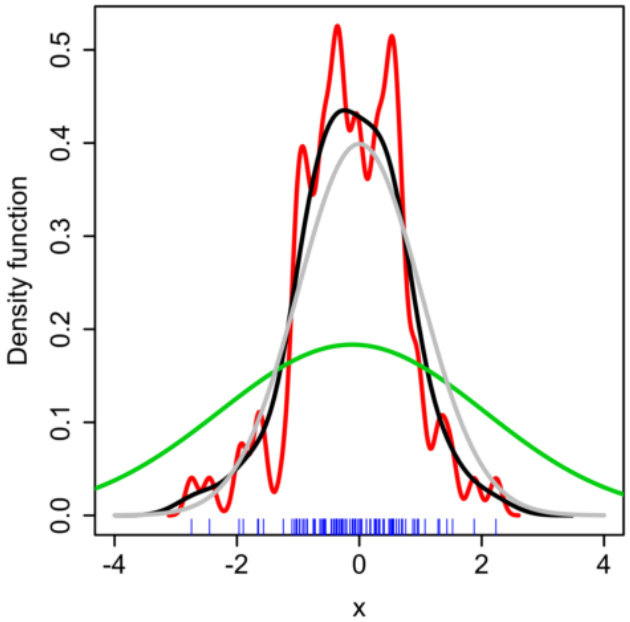
$$J(K) = \int_{-\infty}^{+\infty} \mathsf{E}\big(\hat{p}_h(x) - p(x)\big)^2\, dx.$$

Асимптотические значения отношения $J(K^*)/J(K)$ при $\ell \to \infty$ не зависят от вида распределения $p(x)$.

ядро $K(r)$	степень гладкости	$J(K^*)/J(K)$
Епанечникова $K^*(r)$	$\hat{p}'_h$ разрывна	1.000
Квартическое	$\hat{p}''_h$ разрывна	0.995
Треугольное	$\hat{p}'_h$ разрывна	0.989
Гауссовское	∞ дифференцируема	0.961
Прямоугольное	$\hat{p}_h$ разрывна	0.943

Зависимость оценки плотности от ширины окна

Оценка $\hat{p}_h(x)$ при различных значениях ширины окна h:



истинная плотность
(стандартная гауссовская)

$h = 0.05$ — переобучение
$h = 0.337$ — оптимальная
$h = 2.0$ — недообучение

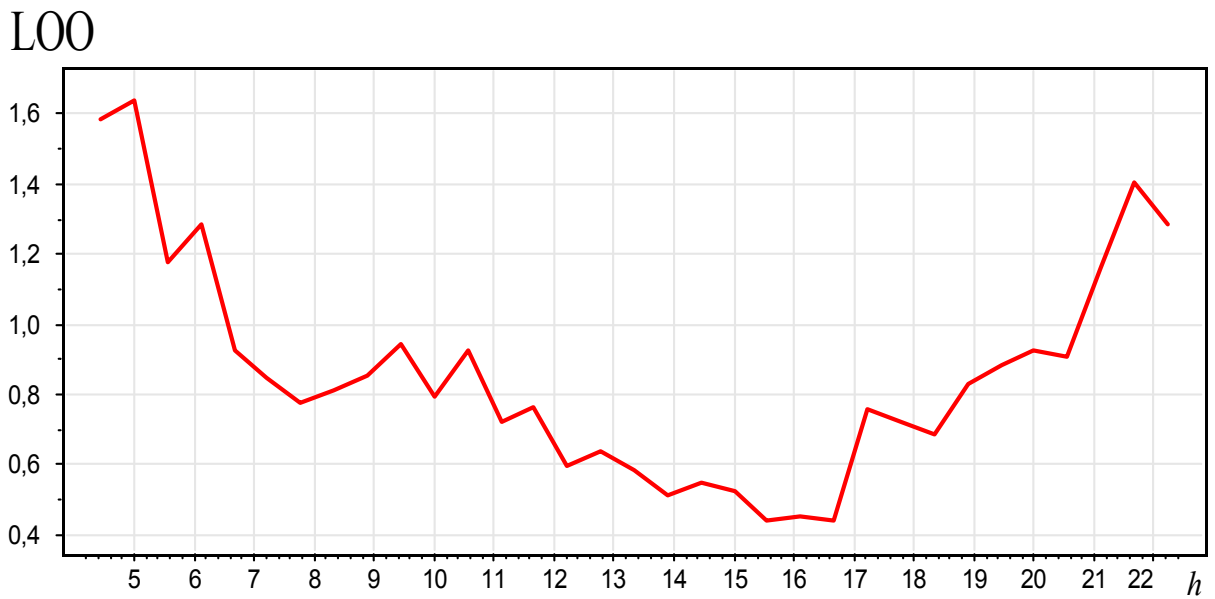
- Качество восстановления плотности существенно зависит от ширины окна h, но слабо зависит от вида ядра K
- При неоднородности локальных сгущений плотности можно задавать $h_k(x) = \rho\left(x, x^{(k+1)}\right)$, где k — число соседей

Выбор ширины окна

Скользящий контроль *Leave One Out* для оценки плотности:

$$\mathrm{LOO}(h) = -\sum_{i=1}^{\ell} \ln \hat{p}_h(x_i; X^\ell \setminus x_i) \to \min_h,$$

Типичный вид зависимости $\mathrm{LOO}(h)$ или $\mathrm{LOO}(k)$:



Ретроспектива: (непара)метрические методы анализа данных

Восстановление плотности. Метод Парзена–Розенблатта:

$$\hat{p}_h(x; X^\ell) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

Классификация. Метод парзеновского окна:

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

Регрессия. Метод *ядерного сглаживания* Надарая–Ватсона:

$$a_h(x; X^\ell, Y^\ell) = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)}$$

Вероятностная постановка задачи классификации

X — объекты, Y — классы, $X \times Y$ — в.п. с плотностью $p(x, y)$

Дано: $X^\ell = (x_i, y_i)_{i=1}^\ell \sim p(x, y)$ — простая выборка (i.i.d.)

Найти: $a\colon X \to Y$ с минимальной вероятностью ошибки

Пусть известна совместная плотность

$$p(x, y) = p(x)\,\mathrm{P}(y|x) = \mathrm{P}(y)p(x|y)$$

$\mathrm{P}(y)$ — *априорная вероятность* класса y

$p(x|y)$ — *функция правдоподобия* класса y

$\mathrm{P}(y|x)$ — *апостериорная вероятность* класса y

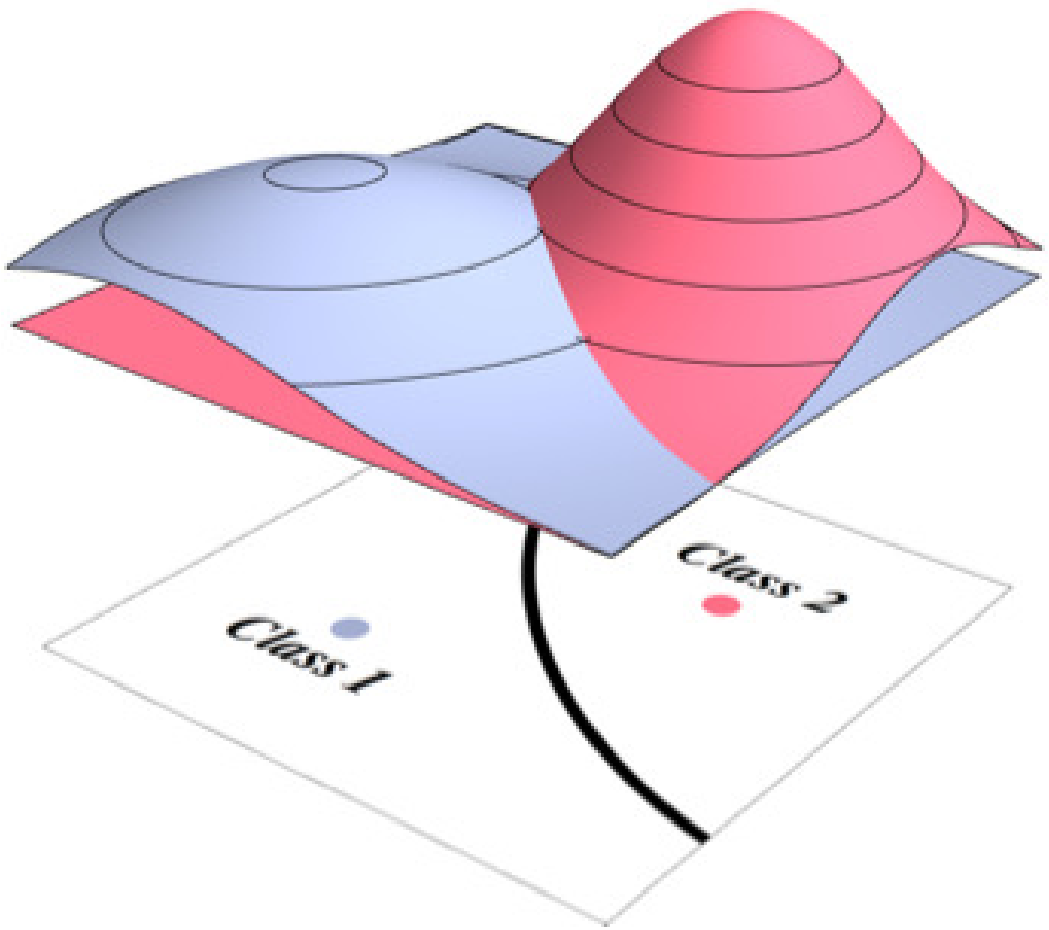
По формуле Байеса: $\mathrm{P}(y|x) = \dfrac{P(y)p(x|y)}{p(x)}$

Байесовский классификатор:

$$a(x) = \arg\max_{y \in Y} \mathrm{P}(y|x) = \arg\max_{y \in Y} P(y)p(x|y)$$

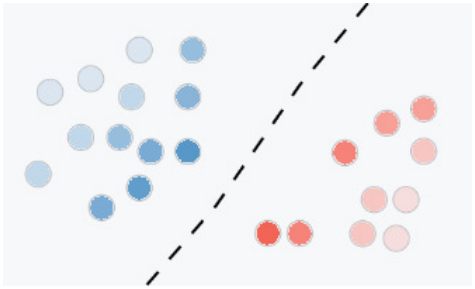
Классификация по максимуму функции правдоподобия

Частный случай: $a(x) = \arg\max_{y \in Y} p(x|y)$ при равных $P(y)$

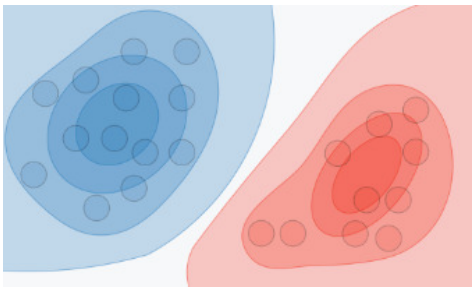


Два подхода к обучению классификации

1. **Дискриминативный** (discriminative):
   x — неслучайные векторы
   $P(y|x,w)$ — модель классификации
   Примеры: LR, GLM, SVM, RBF



2. **Генеративный** (generative):
   $x \sim p(x|y)$ — случайные векторы
   $p(x|y,\theta)$ — модель генерации данных
   Примеры: NB, PW, FLD, RBF



Байесовские модели классификации — генеративные:

- моделируют форму классов не только вдоль границы, но и на всём пространстве, что избыточно для классификации
- требуют больше данных для обучения
- более устойчивы к шумовым выбросам

Оптимальный байесовский классификатор

Теорема

Пусть $P(y)$ и $p(x|y)$ известны, $\lambda_y \geqslant 0$ — *потеря от ошибки* на объекте класса $y \in Y$. Тогда минимум среднего риска

$$R(a) = \sum_{y \in Y} \lambda_y \int [a(x) \neq y]\, p(x, y) dx$$

достигается *оптимальным байесовским классификатором*

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

Замечание 1: после подстановки эмпирических оценок $\hat{P}(y)$ и $\hat{p}(x|y)$ байесовский классификатор уже не оптимален

Замечание 2: задача оценивания плотности распределения — более сложная, чем задача классификации

Наивный байесовский классификатор (Naïve Bayes)

Наивное предположение:
признаки $f_j : X \to D_j$ — независимые случайные величины с плотностями распределения, $p_j(\xi|y)$, $y \in Y$, $j = 1, \ldots, n$

Тогда функции правдоподобия классов представимы в виде произведения одномерных плотностей по признакам, $x^j \equiv f_j(x)$:

$$p(x|y) = p_1(x^1|y) \cdots p_n(x^n|y), \quad x = (x^1, \ldots, x^n), \quad y \in Y$$

Прологарифмировав под argmax, получим классификатор

$$a(x) = \arg\max_{y \in Y} \left(\ln \lambda_y \hat{P}(y) + \sum_{j=1}^{n} \ln \hat{p}_j(x^j|y) \right)$$

Восстановление n одномерных плотностей
— намного более простая задача, чем одной n-мерной

Признаки с плотностями экспоненциального вида

Предположение: одномерные плотности экспоненциальны:

$$p(x^j|y;\theta_{yj},\varphi_{yj}) = \exp\left(\frac{x^j\theta_{yj} - c(\theta_{yj})}{\varphi_{yj}} + h(x^j,\varphi_{yj})\right)$$

где θ_{yj}, φ_{yj} — параметры, $c(\theta)$, $h(x,\varphi)$ — параметры-функции.

Задача максимизации log-правдоподобия

$$L(\theta,\varphi) = \sum_{j=1}^{n}\sum_{y\in Y}\left(\sum_{x_i\in X_y} \ln p(x_i^j|y;\theta_{yj},\varphi_{yj})\right) \to \max_{\theta,\varphi}$$

распадается на независимые подзадачи для каждого (y,j):

$$\sum_{x_i\in X_y}\left(\frac{x^j\theta_{yj} - c(\theta_{yj})}{\varphi_{yj}} + h(x^j,\varphi_{yj})\right) \to \max_{\theta_{yj},\varphi_{yj}}$$

По θ_{yj} задача решается аналитически, по φ_{yj} не всегда

Линейный наивный байесовский классификатор

Решение θ_{yj} через среднее значение признака j в классе y:

$$\frac{\partial L}{\partial \theta_{yj}} = 0 \quad \Rightarrow \quad c'(\theta_{yj}) = \sum_{x_i \in X_y} \frac{x_i^j}{|X_y|} \equiv \bar{x}_{yj} \quad \Rightarrow \quad \theta_{yj} = [c']^{-1}(\bar{x}_{yj})$$

Решение φ_{yj} не всегда выражается из уравнения $\frac{\partial L}{\partial \varphi_{yj}} = 0$, но для распределений Пуассона, Бернулли, биномиального $\varphi_{yj} = 1$; для гауссовского распределения (и если φ_{yj} не зависит от y):

$$\frac{\partial L}{\partial \varphi_{yj}} = 0 \quad \Rightarrow \quad \varphi_{yj} = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(x_i^j - \bar{x}_{y_i j}\right)^2$$

В итоге Naïve Bayes оказывается линейным классификатором:

$$a(x) = \arg\max_{y \in Y} \left(\sum_{j=1}^{n} x^j \underbrace{\frac{\theta_{yj}}{\varphi_{yj}}}_{w_{yj}} + \underbrace{\ln(\lambda_y P(y)) - \sum_{j=1}^{n} \frac{c(\theta_{yj})}{\varphi_{yj}}}_{b_y} + \underbrace{\cancel{h(x^j, \varphi_{yj})}}_{\text{если от } y \text{ не зависит}} \right)$$

Напоминание. Примеры экспоненциальных распределений

μ — параметр матожидания, $\theta = g(\mu)$ — функции связи:

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

$$\mu^x(1-\mu)^{1-x} = \exp\left(x\ln\frac{\mu}{1-\mu} + \ln(1-\mu)\right)$$

$$C_k^x\left(\frac{\mu}{k}\right)^x\left(1-\frac{\mu}{k}\right)^{k-x} = \exp\left(x\ln\frac{\mu}{k-\mu} + k\ln(k-\mu) + \ln C_k^x - k\ln k\right)$$

$$\frac{1}{x!}e^{-\mu}\mu^x = \exp\left(x\ln(\mu) - \mu - \ln x!\right)$$

распределение	значения	$c(\theta)$	$c'(\theta)$	$[c']^{-1}(\mu)$	φ	$h(x,\varphi)$
нормальное	$\mathbb{R}$	$\frac{1}{2}\theta^2$	θ	μ	σ^2	$-\frac{x^2}{2\varphi} - \frac{\ln(2\pi\varphi)}{2}$
Бернулли	$\{0,1\}$	$\ln(1+e^\theta)$	$\frac{1}{1+e^{-\theta}}$	$\ln\frac{\mu}{1-\mu}$	1	0
биномиальное	$\{0,\ldots,k\}$	$k\ln\frac{1+e^\theta}{k}$	$\frac{k}{1+e^{-\theta}}$	$\ln\frac{\mu}{k-\mu}$	1	$\ln C_k^x - k\ln k$
Пуассона	$\{0,1,\ldots\}$	e^θ	e^θ	$\ln\mu$	1	$-\ln x!$

Выводы про наивный байесовский классификатор

Достоинства:

- очень быстрое обучение за $O(\ell n)$ — вычисление $\bar{x}_{yj}$, φ_{yj}
- почти нет переобучения, даже на коротких выборках
- единообразная обработка разнотипных признаков
- хорошее начальное приближение для других методов
- оценка полезности и отбор признаков: $\max_y p(y|j)$
- базовый уровень качества при классификации текстов
- при классификации текстов отбор признаков по полезности удаляет стоп-слова, общую и нерелевантную лексику

Ограничения и недостатки:

- гипотеза о независимости признаков
- низкий уровень качества в большинстве приложений

Напоминание. Метод парзеновского окна (Parzen Window, PW)

Непараметрическая оценка плотности Парзена–Розенблатта с функцией расстояния $\rho(x, x')$, для каждого класса $y \in Y$:

$$\hat{p}_h(x|y) = \frac{1}{\ell_y V_h} \sum_{x_i \in X_y} K\left(\frac{\rho(x, x_i)}{h}\right),$$

Метод окна Парзена — это метрический классификатор:

$$a(x) = \arg\max_{y \in Y} \lambda_y \frac{P(y)}{\ell_y} \sum_{x_i \in X_y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

Замечание 1: нормирующий множитель $V_h = \int_X K\left(\frac{\rho(x,x_i)}{h}\right) dx$ сокращается под argmax, если он не зависит от x_i и y_i.

Замечание 2 (напоминание): имеем проблемы выбора ядра $K(r)$, ширины окна h, функции расстояния $\rho(x, x')$.

Квадратичный дискриминант (Quadratic Discriminant Analysis)

Гипотеза: каждый класс $y \in Y$ имеет n-мерную гауссовскую плотность с центром μ_y и ковариационной матрицей Σ_y:

$$p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y) = \frac{\exp\left(-\frac{1}{2}(x-\mu_y)^\mathsf{T}\Sigma_y^{-1}(x-\mu_y)\right)}{\sqrt{(2\pi)^n \det \Sigma_y}}$$

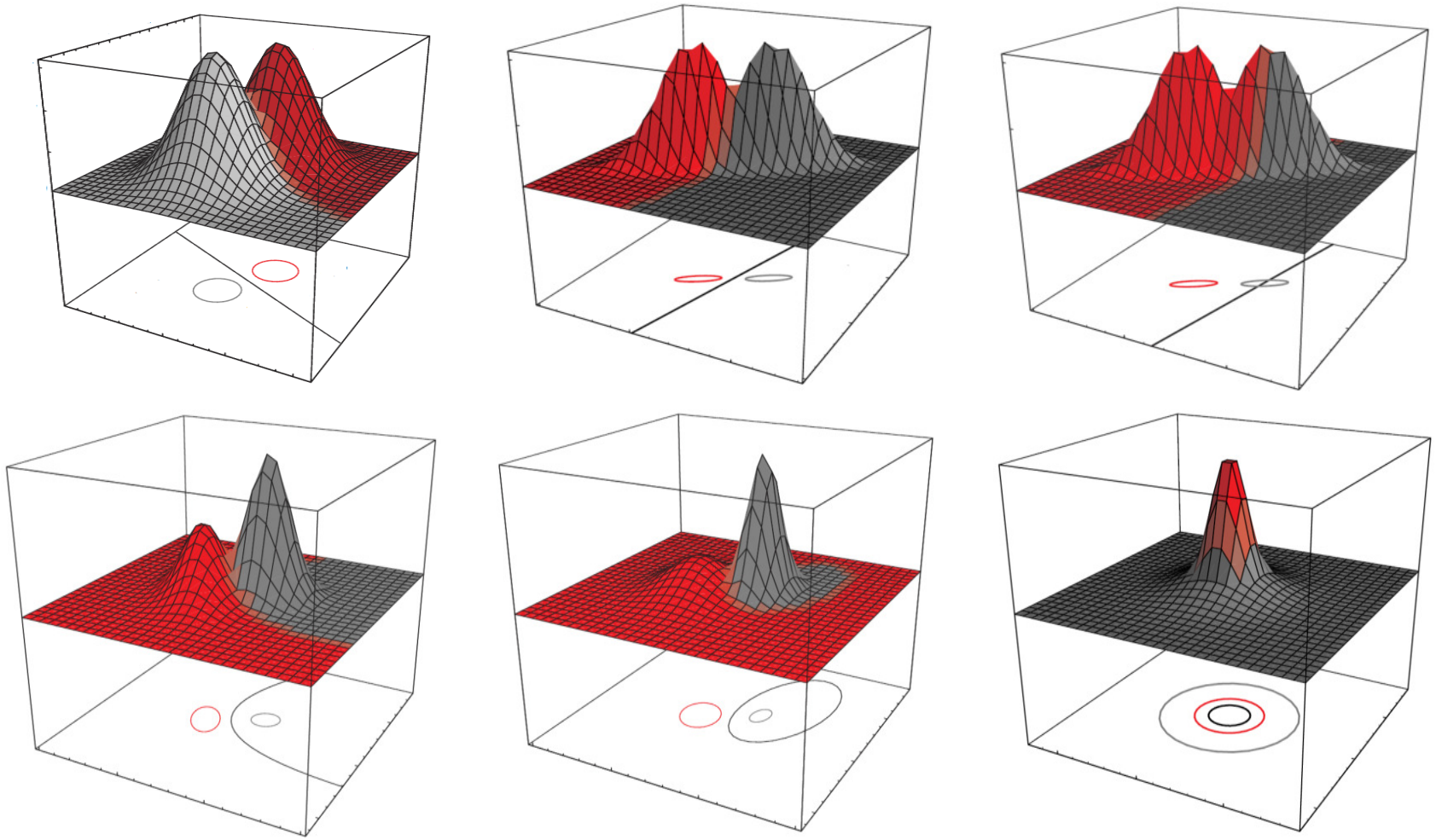
Теорема

1. Разделяющая поверхность, определяемая уравнением $\lambda_y P(y)p(x|y) = \lambda_s P(s)p(x|s)$, квадратична для всех $y, s \in Y$.
2. Если $\Sigma_y = \Sigma_s$, то поверхность вырождается в линейную.

Квадратичный дискриминант — подстановочный алгоритм:

$$a(x) = \arg\max_{y \in Y} \left(\ln \lambda_y P(y) - \frac{1}{2}(x-\hat{\mu}_y)^\mathsf{T}\hat{\Sigma}_y^{-1}(x-\hat{\mu}_y) - \frac{1}{2}\ln\det\hat{\Sigma}_y\right)$$

Геометрический смысл квадратичного дискриминанта



Линейный дискриминант Фишера (Fisher Linear Discriminant)

Проблема: для малочисленных классов возможно $\det \hat{\Sigma}_y = 0$.

Пусть ковариационные матрицы классов равны: $\Sigma_y = \Sigma$, $y \in Y$.

Оценка максимума правдоподобия для Σ:

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

Линейный дискриминант — подстановочный алгоритм:

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

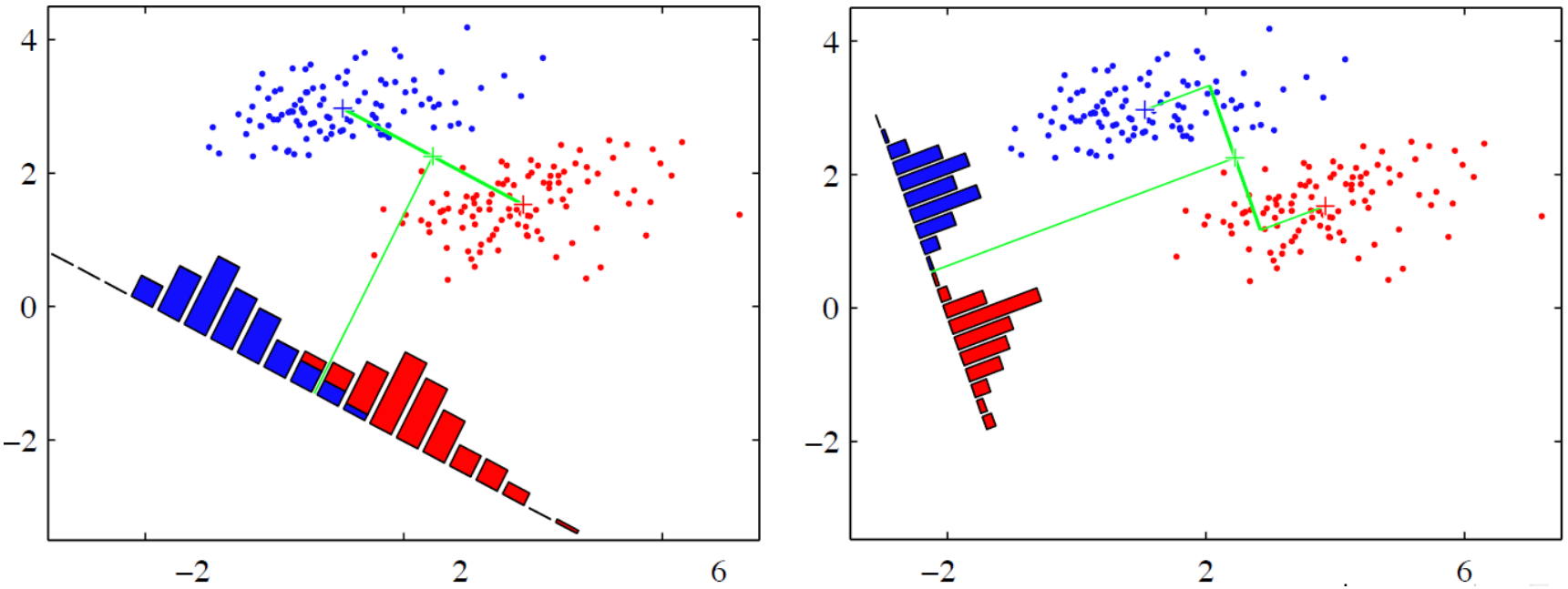
$$= \arg\max_{y \in Y} \Big(\underbrace{\ln(\lambda_y \hat{P}(y)) - \tfrac{1}{2} \hat{\mu}_y^{\mathsf{T}} \hat{\Sigma}^{-1} \hat{\mu}_y}_{\beta_y} + x^{\mathsf{T}} \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y} \Big);$$

$$a(x) = \arg\max_{y \in Y} \left(x^{\mathsf{T}} \alpha_y + \beta_y \right).$$

В случае мультиколлинеарности — обращать матрицу $\hat{\Sigma} + \tau I_n$.

Геометрическая интерпретация линейного дискриминанта

В одномерной проекции на направляющий вектор разделяющей гиперплоскости классы разделяются наилучшим образом, то есть с минимальной вероятностью ошибки:



Ось проекции перпендикулярна общей касательной эллипсоидов рассеяния

Fisher R. A. The use of multiple measurements in taxonomic problems. 1936.

Гауссовская смесь с диагональными матрицами ковариации

Гауссовская смесь GMM — Gaussian Mixture Model

Допущения:

1. Функции правдоподобия классов $p(x|y)$ представимы в виде смесей k_y компонент, для каждого класса $y \in Y$
2. Компоненты $j = 1, \ldots, k_y$ имеют n-мерные гауссовские плотности с некоррелированными признаками: $\mu_{yj} = (\mu_{yj1}, \ldots, \mu_{yjn})$, $\Sigma_{yj} = \operatorname{diag}(\sigma^2_{yj1}, \ldots, \sigma^2_{yjn})$:

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj})$$

$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geqslant 0$$

EM-алгоритм. Эмпирические оценки средних и дисперсий

Числовые признаки: $f_d : X \to \mathbb{R}$, $d = 1, \ldots, n$.

E-шаг: для всех $y \in Y$, $j = 1, \ldots, k_y$, $d = 1, \ldots, n$:

$$g_{yij} = \frac{w_{yj}\mathcal{N}(x_i; \mu_{yj}, \Sigma_{yj})}{p(x_i|y)} \equiv \mathrm{P}(j|x_i, y_i = y)$$

M-шаг: для всех $y \in Y$, $j = 1, \ldots, k_y$, $d = 1, \ldots, n$

$$w_{yj} = \frac{1}{\ell_y} \sum_{i:\, y_i = y} g_{yij}$$

$$\hat{\mu}_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} f_d(x_i)$$

$$\hat{\sigma}^2_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} \big(f_d(x_i) - \hat{\mu}_{yjd}\big)^2$$

Замечание: компоненты «наивны», но смесь не «наивна»

Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

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

$\mathcal{N}_{yj} = (2\pi)^{-\frac{n}{2}}(\sigma_{yj1} \cdots \sigma_{yjn})^{-1}$ — нормировочные множители;
$\rho_{yj}(x, \mu_{yj})$ — *взвешенная евклидова метрика* в $X = \mathbb{R}^n$:

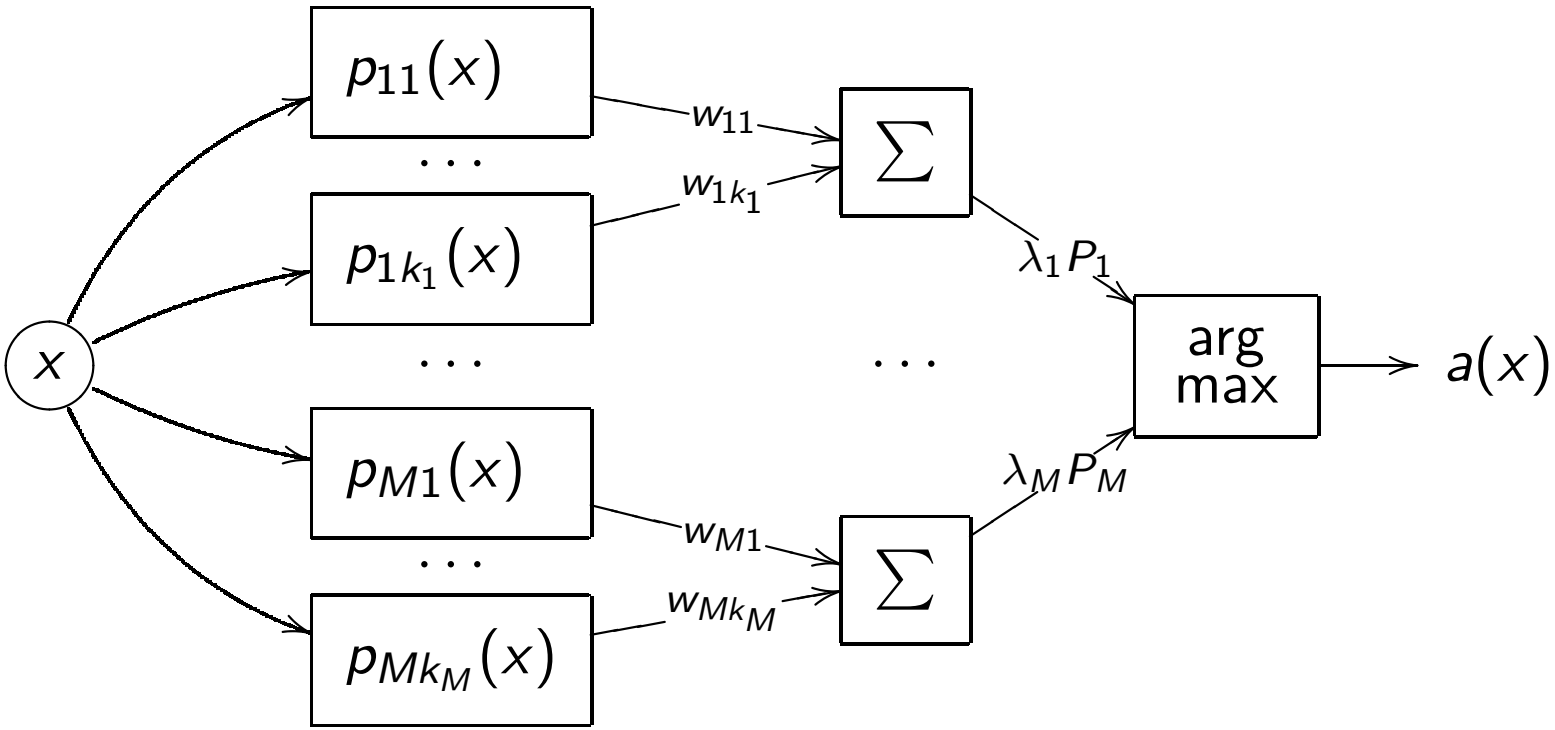
$$\rho^2_{yj}(x, \mu_{yj}) = \sum_{d=1}^{n} \frac{1}{\sigma^2_{yjd}} \big(f_d(x) - \mu_{yjd}\big)^2.$$

Интерпретация: это метрический классификатор, в котором
$p_{yj}(x)$ — близость объекта x к j-й компоненте класса y;
$\Gamma_y(x)$ — близость объекта x к классу y.

Сеть радиальных базисных функций (RBF)

Трёхслойная сеть RBF (Radial Basis Functions):

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} p_{yj}(x)$$



EM-алгоритм как метод обучения радиальных сетей

Отличия генеративного RBF-EM от дискриминативного RBF-SVM:

- опорные векторы μ_{yj} — это не пограничные объекты выборки, а центры локальных сгущений классов
- автоматически строится *структурное описание* каждого класса в виде совокупности компонент — *кластеров*

Преимущества EM-алгоритма:

- EM-алгоритм легко сделать устойчивым к шуму
- как правило, EM-алгоритм довольно быстро сходится

Недостатки EM-алгоритма:

- EM-алгоритм чувствителен к начальному приближению
- Определение числа компонент — трудная задача (простые эвристики могут плохо работать)

Резюме

Три подхода к восстановлению плотности по выборке:

- *Параметрический:* $\hat{p}(x) = \varphi(x, \theta)$
- *Разделение смеси:* $\hat{p}(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \quad k \ll \ell$
- *Непараметрический:* $\hat{p}(x) = \sum_{i=1}^{\ell} \frac{1}{\ell V_h} K\left(\frac{\rho(x, x_i)}{h}\right)$

Байесовская теория классификации:

- Основная формула: $a(x) = \arg\max_{y \in Y} \lambda_y P(y) \hat{p}(x|y)$
- Байесовские генеративные модели классификации
  — моделируют форму классов на всём пространстве,
  — требуют большего объёма данных для обучения,
  — менее чувствительны к шумовым выбросам
- Наивный байесовский классификатор
  — основан на предположении о независимости признаков,
  — неплохо работает в задачах категоризации текстов