

# Мат. модели машинного обучения: оценивание плотности распределения и байесовская теория классификации

Воронцов Константин Вячеславович

`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)»

## 1 Параметрические методы восстановления плотности

- Задача восстановления плотности распределения
- Восстановление многомерной гауссовской плотности
- Задача разделения смеси распределений

## 2 Непараметрическое восстановление плотности

- Восстановление одномерных плотностей
- Восстановление многомерных плотностей
- Выбор ядра и ширины окна

## 3 Байесовская теория классификации

- Оптимальный байесовский классификатор
- Наивный байесовский классификатор
- Обзор байесовских классификаторов

## Задача восстановления плотности (обучение без учителя)

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

**Найти:** параметрическую модель плотности распределения

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

где  $w$  — вектор параметров,  $\varphi$  — фиксированная функция

**Критерий:** метод максимума правдоподобия (ММП) выборки  
(Maximum Likelihood Estimate, MLE, Maximum log-Likelihood)

$$L(w; X^\ell) = \ln \prod_{i=1}^{\ell} \varphi(x_i; w) = \sum_{i=1}^{\ell} \ln \varphi(x_i; w) \rightarrow \max_w$$

Аналитическое решение: необходимое условие экстремума

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

при условии достаточной гладкости функции  $\varphi(x; w)$  по  $w$

## Частный случай №1: оценка дискретного распределения

**Дано:** простая выборка  $x_i \in X$ ,  $|X| < \infty$ , порождаемая дискретным распределением  $(p_x: x \in X)$ ,  $\sum_x p_x = 1$ ,  $p_x \geq 0$

**Найти:** параметры распределения  $(p_x: x \in X)$

**Критерий:** максимум (логарифма) правдоподобия выборки

$$\ln \prod_{i=1}^{\ell} p_{x_i} = \sum_{x \in X} \underbrace{\sum_{i=1}^{\ell} [x_i = x]}_{\ell_x} \ln p_x = \sum_{x \in X} \ell_x \ln p_x \rightarrow \max_{(p_x)}$$

### Выборочная оценка максимального правдоподобия

$\hat{p}_x = \frac{\ell_x}{\ell}$  — частотные оценки вероятностей  $p_x = P(x_i = x)$ ,  
оценка минимума кросс-энтропии, эмпирическая гистограмма

Доказательство из условий ККТ:  $\frac{\partial}{\partial p_x} \left( \sum_{x \in X} \ell_x \ln p_x + \mu \left( 1 - \sum_{x \in X} p_x \right) \right) = 0$

## Частный случай №2: многомерная гауссовская плотность

**Дано:** выборка  $x_i \in \mathbb{R}^n$ , порождаемая гауссовской плотностью:

$$x_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(x; \mu, \Sigma) = \frac{\exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu))}{\sqrt{(2\pi)^n \det \Sigma}}, \quad x \in \mathbb{R}^n$$

**Найти** параметры распределения:

$\mu \in \mathbb{R}^n$  — вектор математического ожидания,  $\mu = \mathbb{E}x$

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

**Критерий:** максимум (логарифма) правдоподобия выборки

**Выборочные оценки максимального правдоподобия:**

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln L(\mu, \Sigma; X^\ell) = 0 &\Rightarrow \hat{\mu} = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i \\ \frac{\partial}{\partial \Sigma} \ln L(\mu, \Sigma; X^\ell) = 0 &\Rightarrow \hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu})(x_i - \hat{\mu})^\top \end{aligned}$$

## Некоторые приёмы матричного дифференцирования

Производная скалярной функции  $f(A)$  по матрице  $A = (a_{ij})$ :

$$\frac{\partial}{\partial A} f(A) = \left( \frac{\partial}{\partial a_{ij}} f(A) \right)$$

$\text{diag } A$  — диагональ матрицы  $A$ , остальные элементы нули

$A$  — квадратная  $n \times n$ -матрица,  $\det A$  — её детерминант

$u$  — вектор размерности  $n$

если  $A$  произвольного вида:

$$\frac{\partial}{\partial u} u^T A u = A^T u + A u$$

$$\frac{\partial}{\partial A} u^T A u = u u^T$$

$$\frac{\partial}{\partial A} \ln \det A = A^{-1T}$$

если  $A$  симметричная,  $A^T = A$ :

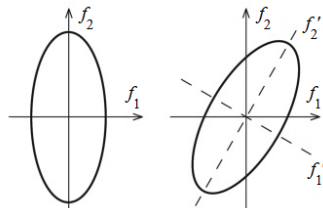
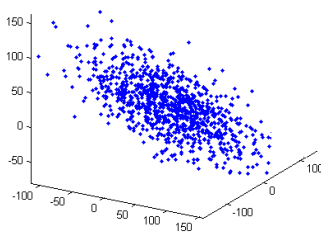
$$\frac{\partial}{\partial u} u^T A u = 2 A u$$

$$\frac{\partial}{\partial A} u^T A u = 2 u u^T - \text{diag } u u^T$$

$$\frac{\partial}{\partial A} \ln \det A = 2 A^{-1} - \text{diag } A^{-1}$$

## Геометрический смысл многомерной нормальной плотности

*Эллипсоид рассеяния* — облако точек эллиптической формы:



При  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$  оси эллипсоида параллельны осям.

В общем случае:  $\Sigma = VSV^T$  — спектральное разложение,

$V = (v_1, \dots, v_n)$  — ортогональные собств. векторы,  $V^T V = I_n$

$S = \text{diag}(\lambda_1, \dots, \lambda_n)$  — собственные значения матрицы  $\Sigma$

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

$x' = V^T x$  — ортогональное преобразование поворот/отражение

## Проблема мультиколлинеарности

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

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

**Диагонализация ковариационной матрицы:**

«наивное» предположение о независимости признаков приводит к оцениванию  $n$  одномерных плотностей признаков:

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

$\hat{\mu}_j = \frac{1}{\ell} \sum_{i=1}^{\ell} f_j(x_i)$  — выборочная оценка среднего признака  $f_j$

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

## Задача разделения смеси распределений

**Дано:** выборка  $\{x_1, \dots, x_\ell\}$ , порождаемая i.i.d. из  $p(x)$

**Найти:** модель вероятностной смеси  $k$  распределений:

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

описывающую двухуровневый процесс порождения данных:

- ①  $j \sim P(j) \equiv w_j$  — дискретное *априорное распределение*
- ②  $x \sim p(x|j) \equiv \varphi(x, \theta_j)$  — плотность  $j$ -й компоненты

**Критерий** максимума log-правдоподобия:

$$\left\{ \begin{aligned} Q(w, \theta) = \ln \prod_{i=1}^{\ell} p(x_i) &= \sum_{i=1}^{\ell} \ln \sum_{j=1}^k w_j \varphi(x_i, \theta_j) \rightarrow \max_{w, \theta} \\ \sum_{j=1}^k w_j &= 1, \quad w_j \geq 0 \end{aligned} \right.$$

## ЕМ-алгоритм для разделения смеси распределений

### Теорема (необходимые условия экстремума)

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

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

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

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

ЕМ-алгоритм — метод простых итераций для решения системы

## Вероятностная интерпретация шагов ЕМ-алгоритма

**Е-шаг** — это формула Байеса:

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

Нормировка условных вероятностей:  $\sum_{j=1}^k g_{ij} = 1$

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

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

вес компоненты определяется как средний вес её объектов:

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

## Доказательство — через условия Каруша–Куна–Таккера

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

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

Приравниваем нулю производные:

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

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



## ЕМ-алгоритм для разделения смеси распределений

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

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

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

**повторять**

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

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

М-шаг (maximization): для всех  $j = 1, \dots, k$

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

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

**пока**  $w_j, \theta_j$  и/или  $g_{ij}$  не сошлись;

## Разделение смеси гауссиан (Gaussian Mixture Model, GMM)

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

**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**

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

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

М-шаг (maximization): для всех  $j = 1, \dots, k$

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

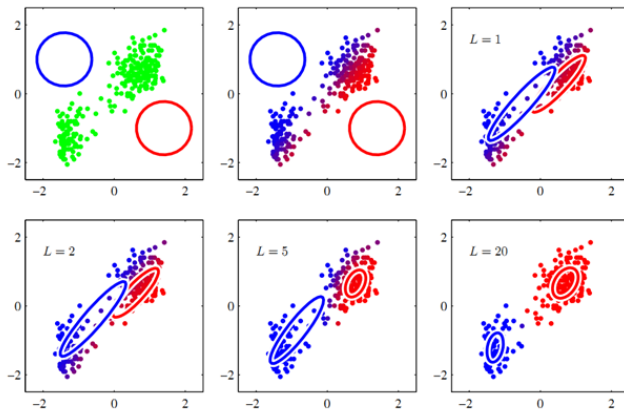
$$\Sigma_j := \frac{1}{\ell w_j} \sum_{i=1}^{\ell} g_{ij} (x_i - \mu_j)(x_i - \mu_j)^T;$$

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

**пока**  $(w_j, \mu_j, \Sigma_j)$  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  $[a, b]$ :

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

Отбросив  $\lim$  и заменив вероятность частотой, получим кусочно-постоянную оценку плотности по окну ширины  $h$ :

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

## Локальная непараметрическая оценка Парзена-Розенблатта

Кусочно-постоянная 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 *ядро*  $K(r)$  — функция, удовлетворяющая требованиям:

- чётная, неотрицательная, невозрастающая при  $r > 0$ ;
- нормированная:  $\int K(r) dr = 1$ ;
- достаточно гладкая, чтобы оценка KDE была гладкой.

При  $K(r) = \frac{1}{2} [ |r| < 1 ]$  это кусочно-постоянная оценка.

## Два варианта обобщения на многомерный случай

- ❶ Если объекты описываются  $n$  признаками  $f_j: X \rightarrow \mathbb{R}$ :

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

- ❷ Если на  $X$  задана функция расстояния  $\rho(x, x')$ :

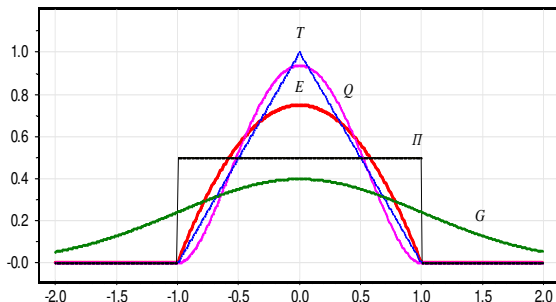
$$\hat{\rho}_h(x) = \frac{1}{\ell V(h)} \sum_{i=1}^{\ell} K\left(\frac{\rho(x, x_i)}{h}\right)$$

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

**Сферическое гауссовское ядро** — частный случай обоих:

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

## Выбор ядра



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

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

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

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

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

## Выбор ядра почти не влияет на качество восстановления

Функционал качества восстановления плотности:

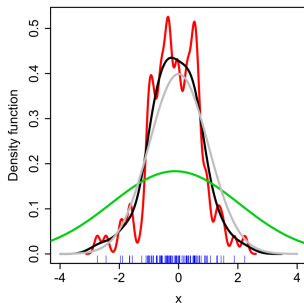
$$J(K) = \int_{-\infty}^{+\infty} E(\hat{p}_h(x) - p(x))^2 dx.$$

Асимптотические значения отношения  $J(K^*)/J(K)$  при  $\ell \rightarrow \infty$  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           | $\infty$ дифференцируема | 0.961         |
| Прямоугольное         | $\hat{p}_h$ разрывна     | 0.943         |

## Зависимость оценки плотности от ширины окна

Оценка  $\hat{\rho}_h(x)$  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  $h_k(x) = \rho(x, x^{(k+1)})$ , где  $k$  — число соседей

## Выбор ширины окна

Скользящий контроль *Leave-One-Out* для оценки плотности по максимуму правдоподобия:

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

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



## Ретроспектива: (непара)метрические методы анализа данных

Восстановление плотности. Метод Парзена–Розенблатта:

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

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:

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

## Вероятностные генеративные модели классификации

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

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

**Найти:** модель классификации  $a: X \rightarrow Y$

**Критерий:** минимизировать вероятность ошибки  $P[a(x) \neq y]$

Пусть известна совместная плотность

$$p(x, y) = p(x) P(y|x) = P(y)p(x|y)$$

$P(y)$  — априорная вероятность класса  $y$

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

$P(y|x)$  — апостериорная вероятность класса  $y$

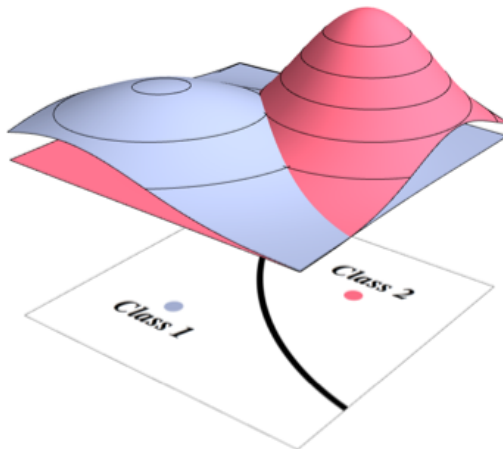
По формуле Байеса:  $P(y|x) = \frac{P(y)p(x|y)}{p(x)}$

**Байесовский классификатор:**

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

## 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

### ❶ Дискриминативный (discriminative):

$x$  — неслучайные векторы

$P(y|x, w)$  — модель классификации

Примеры: LR, GLM, SVM, RBF

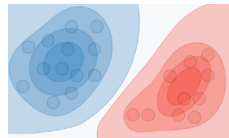


### ❷ Генеративный (generative):

$x \sim p(x|y)$  — случайные векторы

$p(x|y, w)$  — 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
- зато более устойчивы к шумам около границы классов

## Оптимальный байесовский классификатор

### Теорема

Пусть  $P(y)$  и  $p(x|y)$  известны,  $\lambda_y \geq 0$  — 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

## Байесовский классификатор и максимизация правдоподобия

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

**Найти:** параметры  $\{P_y, w_y\}$  модели  $p(x, y) = P_y p(x|y; w_y)$

**Критерий:** метод максимума правдоподобия (ММП)

$$\sum_{i=1}^{\ell} \ln p(x_i, y_i) = \sum_{i=1}^{\ell} \ln P_{y_i} + \sum_{y \in Y} \sum_{x_i \in X_y} \ln p(x_i|y; w_y) \rightarrow \max_{\{P_y, w_y\}}$$

где  $X_y = \{x_i: y_i = y\}$  — все объекты класса  $y$  из выборки  $X^\ell$ .

Декомпозиция на  $|Y| + 1$  независимо решаемых подзадач:

$$\sum_{i=1}^{\ell} \ln P_{y_i} \rightarrow \max: \{\sum_y P_y = 1; P_y \geq 0\} \Rightarrow P_y = \hat{P}(y) = \frac{|X_y|}{\ell}$$

$$\sum_{x_i \in X_y} \ln p(x_i|y; w_y) \rightarrow \max_{w_y} \quad \text{— MLE оценка плотности класса } y$$

## Наивный байесовский классификатор (Naïve Bayes)

**Наивное предположение:**

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

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

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

Прологарифмировав под  $\arg\max$ , 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  $\theta_{yj}$ ,  $\varphi_{yj}$  — параметры,  $c(\theta)$ ,  $h(x, \varphi)$  — параметры-функции.

Задача максимизации log-правдоподобия

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

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

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

По  $\theta_{yj}$  задача решается аналитически, по  $\varphi_{yj}$  не всегда

## Линейный наивный байесовский классификатор

Решение  $\theta_{yj}$  через среднее значение признака  $j$  в классе  $y$ :

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

Решение  $\varphi_{yj}$  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  $\varphi_{yj}$  не зависит от  $y$ ):

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

В итоге Naïve Bayes оказывается линейным классификатором:

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

## Напоминание. Примеры экспоненциальных распределений

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

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= \exp\left(\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)\right) \\ \mu^x (1-\mu)^{1-x} &= \exp\left(x \ln \frac{\mu}{1-\mu} + \ln(1-\mu)\right) \\ C_k^x \left(\frac{\mu}{k}\right)^x \left(1 - \frac{\mu}{k}\right)^{k-x} &= \exp\left(x \ln \frac{\mu}{k-\mu} + k \ln(k-\mu) + \ln C_k^x - k \ln k\right) \\ \frac{1}{x!} e^{-\mu} \mu^x &= \exp\left(x \ln(\mu) - \mu - \ln x!\right) \end{aligned}$$

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

## Выводы про наивный байесовский классификатор

### Достоинства:

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

### 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  $\arg\max$ , 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  $\mu_y$  и ковариационной матрицей  $\Sigma_y$ :

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

### Теорема

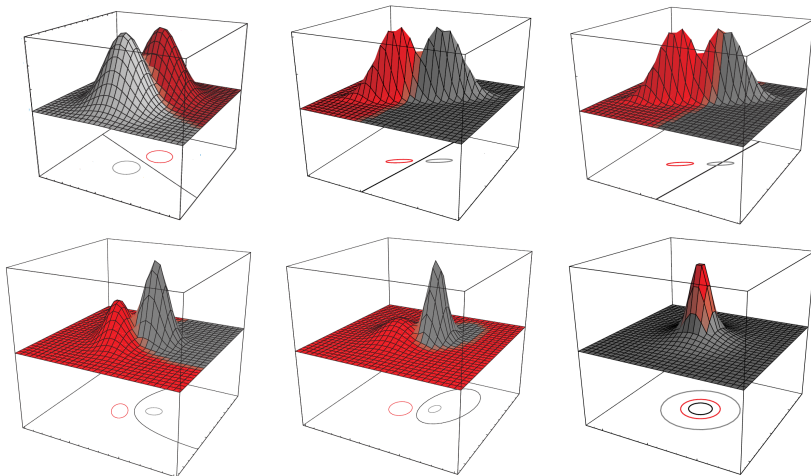
1. Разделяющая поверхность, определяемая уравнением  $\lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)$ , квадратична для всех  $y, s \in Y$ .
2. Если  $\Sigma_y = \Sigma_s$ , то поверхность вырождается в линейную.

**Квадратичный дискриминант** — подстановочный алгоритм:

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

## Геометрический смысл квадратичного дискриминанта

Разделяющая поверхность линейна ( $\Sigma_y = \Sigma_s$ ) или квадратична:



## 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  $\Sigma$ :**

$$\hat{\Sigma} = \frac{1}{\ell} \sum_{i=1}^{\ell} (x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^T, \quad \hat{\mu}_y = \frac{\sum_i [y_i = y] x_i}{\sum_i [y_i = y]}$$

**Линейный дискриминант** — подстановочный алгоритм:

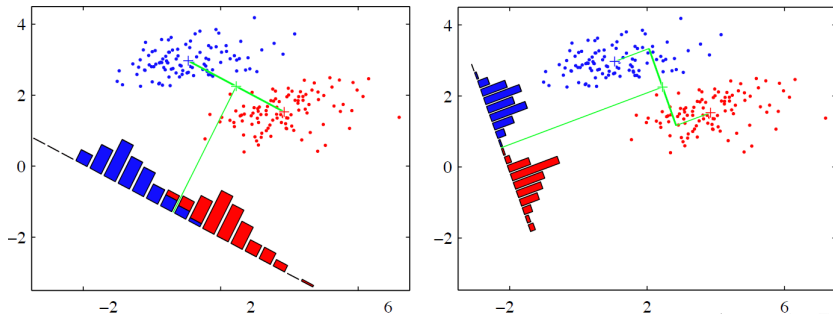
$$\begin{aligned} a(x) &= \arg \max_{y \in Y} \lambda_y \hat{P}(y) \hat{p}(x|y) = \\ &= \arg \max_{y \in Y} \underbrace{\left( \ln(\lambda_y \hat{P}(y)) - \frac{1}{2} \hat{\mu}_y^T \hat{\Sigma}^{-1} \hat{\mu}_y \right)}_{\beta_y} + x^T \underbrace{\hat{\Sigma}^{-1} \hat{\mu}_y}_{\alpha_y}; \end{aligned}$$

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

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  $j = 1, \dots, k_y$  имеют  $n$ -мерные гауссовские плотности с некоррелированными признаками:  
 $\mu_{yj} = (\mu_{yj1}, \dots, \mu_{yjn})$ ,  $\Sigma_{yj} = \text{diag}(\sigma_{yj1}^2, \dots, \sigma_{yjn}^2)$ :

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

## ЕМ-алгоритм. Эмпирические оценки средних и дисперсий

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

**Е-шаг:** для всех  $y \in Y$ ,  $j = 1, \dots, k_y$ ,  $d = 1, \dots, n$ :

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

**М-шаг:** для всех  $y \in Y$ ,  $j = 1, \dots, k_y$ ,  $d = 1, \dots, n$

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

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

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

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

## Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

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

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

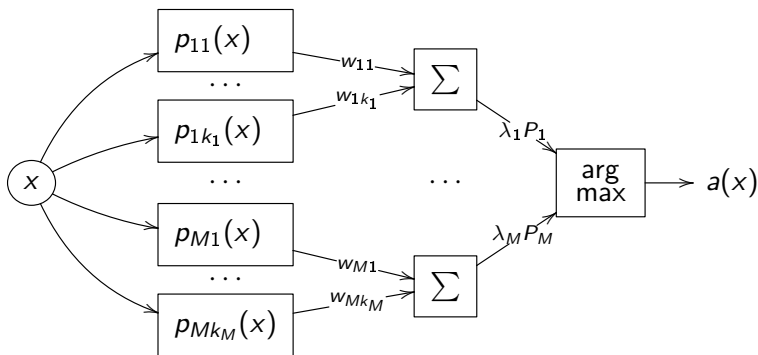
$$\rho_{yj}^2(x, \mu_{yj}) = \sum_{d=1}^n \frac{1}{\sigma_{yjd}^2} (f_d(x) - \mu_{yjd})^2.$$

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

## Сеть радиальных базисных функций (RBF)

В задаче многоклассовой классификации,  $Y = \{1, \dots, M\}$ ,  
 EM-алгоритм обучает трёхслойную сеть RBF:

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



## ЕМ-алгоритм как метод обучения радиальных сетей

Отличия генеративного RBF-ЕМ от дискриминативного RBF-SVM:

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

Преимущества ЕМ-алгоритма:

- ЕМ-алгоритм более устойчив к шумовым выбросам
- как правило, ЕМ-алгоритм довольно быстро сходится

Недостатки ЕМ-алгоритма:

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

## Восстановление плотности по выборке, три подхода:

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

## Вероятностные модели классификации, два подхода, оба основаны на максимизации правдоподобия:

- *дискриминативный*: моделируем  $P(y|x, w)$ ,  $x$  не случайны
- *генеративный*: моделируем  $p(x|y, w)$ , для всех  $y \in Y$ 
  - более требовательный к объёму данных для обучения,
  - менее чувствительный к шумовым выбросам

## Байесовская теория классификации:

- основная формула:  $a(x) = \arg \max_{y \in Y} \lambda_y P(y) \hat{p}(x|y)$
- *наивный байесовский классификатор*
  - основан на предположении о независимости признаков