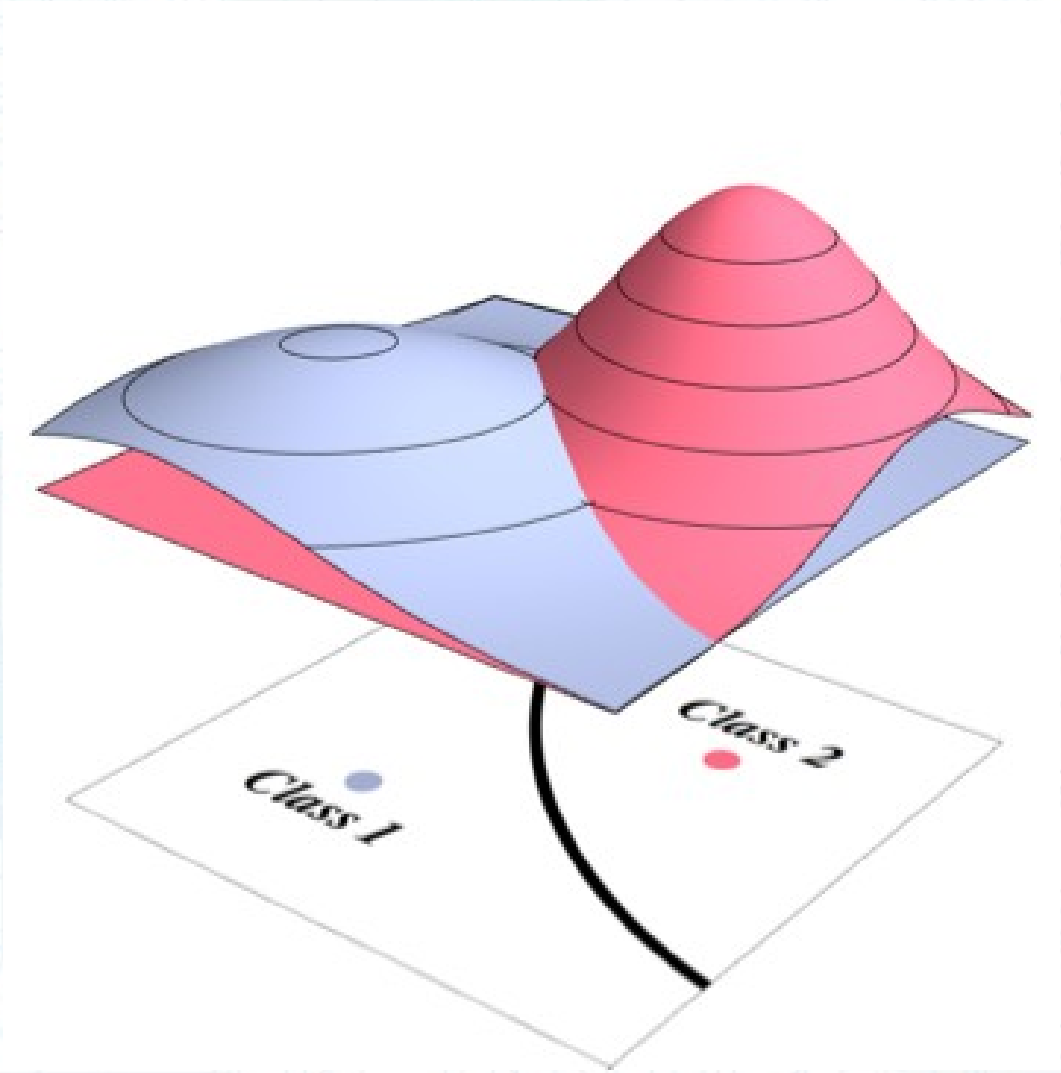
Машинное обучение
Лекция 4. Байесовский подход



https://yandexdataschool.ru/edu-process/courses/machine-learning

Содержание лекции

- Байесовский классификатор
- Восстановление плотности распределения
  - непараметрическое
  - параметрическое

Вероятностная постановка задачи

- P(x,y) – неизвестная точная плотность распределения на X×Y
- X_ℓ - выборка из случайных, независимых и одинаково распределенных прецедентов
- Найти: эмпирическую оценку плотности
- Классификатор с минимальной вероятностью ошибки:

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

- Классификатор с минимальным средним риском: $a(x) = \arg\min_y E_s \; \mathscr{L}(s, y)$

Пример

y	+	o
p(y|x)	0.3	0.7
- Классификатор с минимальной вероятностью ошибки: a(x) = o
- Классификатор с минимальным средним риском (пусть $\mathcal{L}$(+,o)=3, $\mathcal{L}$(o,+)=1):

  a(x) = +

Подходы к восстановлению плотности распределения

- Непараметрическое оценивание плотности:

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

- Параметрическое оценивание плотности:

$$\hat{p}(x) = \varphi(x, \theta)$$

Наивный байесовский классификатор

- Восстановление n одномерных плотностей — намного более простая задача, чем одной n-мерной.
- Допущение (наивное): признаки являются независимыми случайными величинами
- Тогда совместная плотность распределения представима в виде произведения частных плотностей:

$$p(x|y) = p_1(\xi_1|y) \cdots p_n(\xi_n|y), \quad x = (\xi_1, \ldots, \xi_n), \quad y \in Y.$$

Непараметрическое оценивание

- Определение плотности вероятности (одномерный случай):

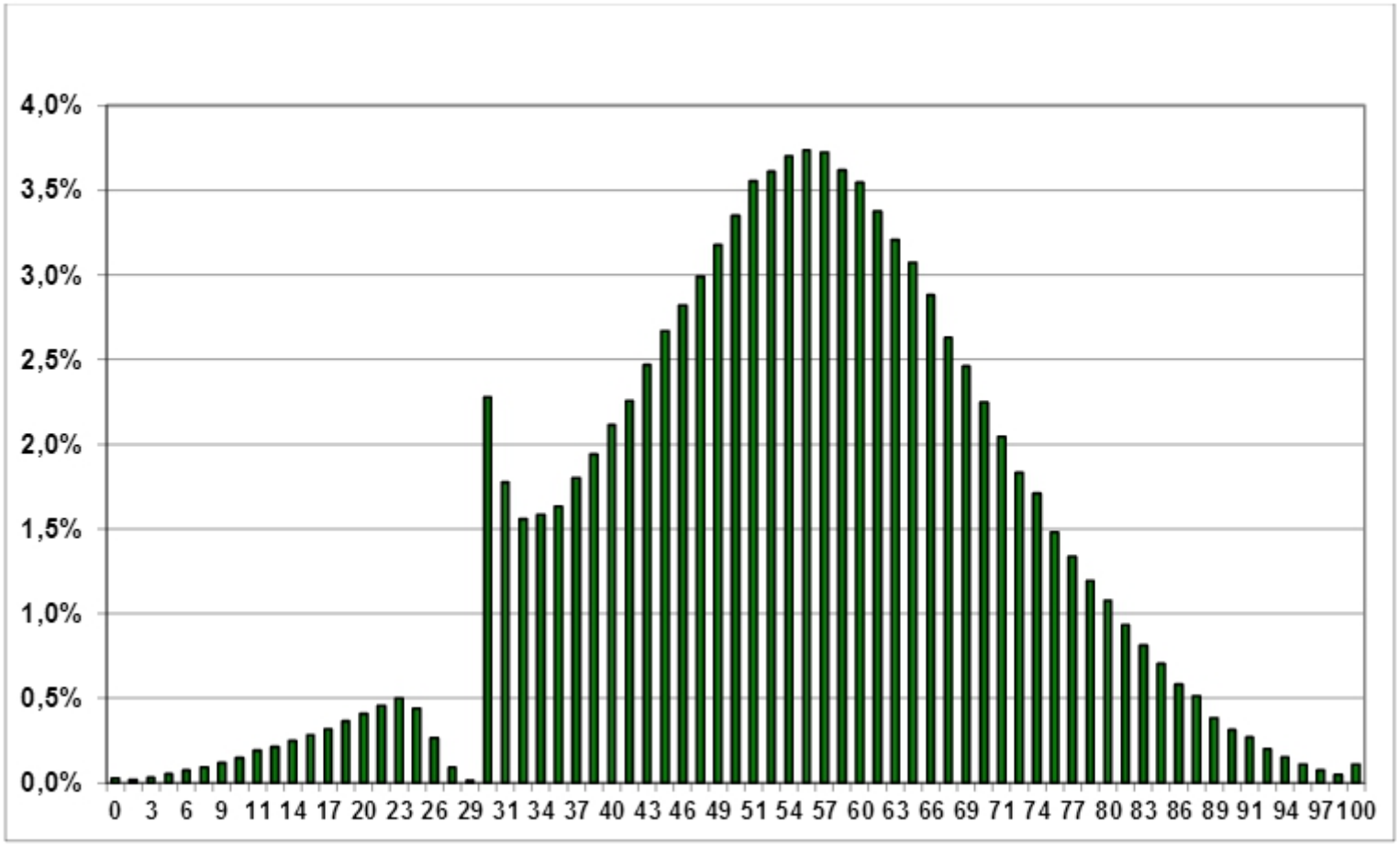
$$p(x) = \lim_{h \to 0} \frac{1}{2h} P[x - h, x + h]$$

- Эмпирическая оценка:

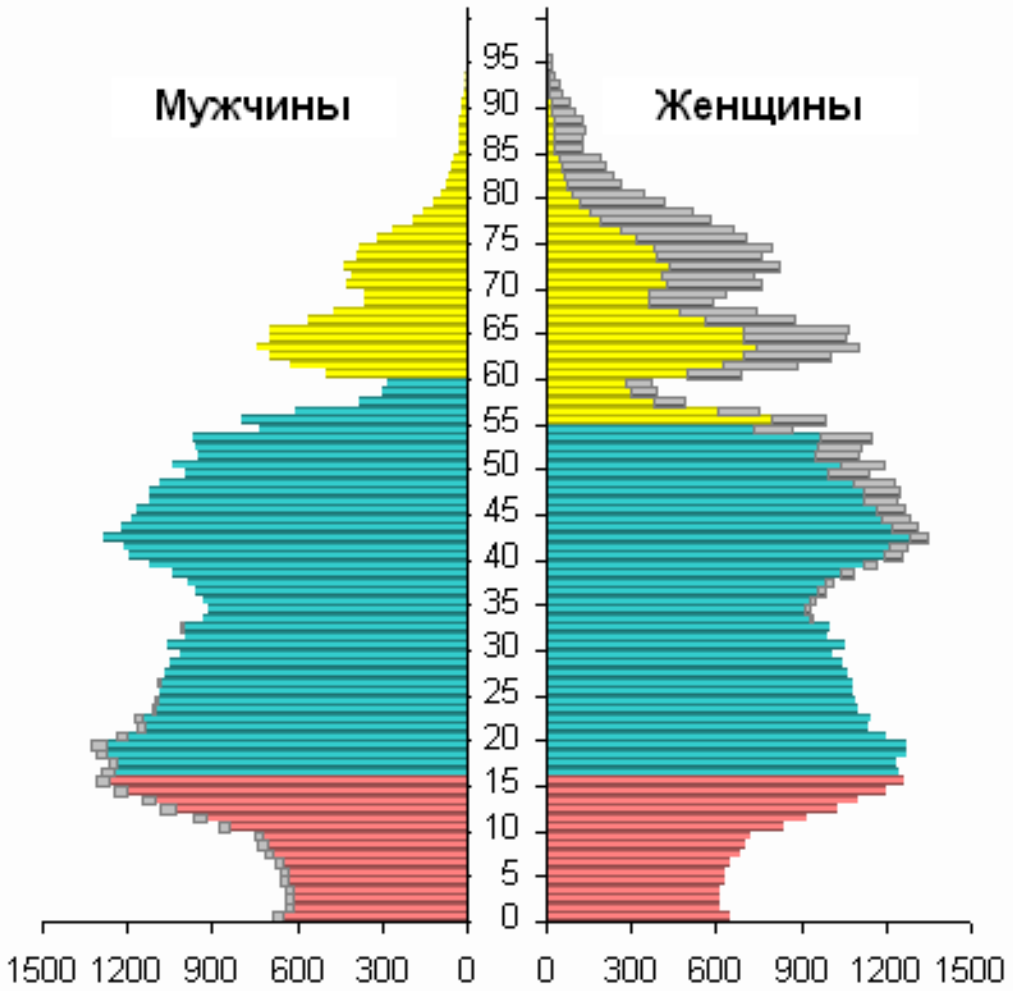
$$\hat{p}_h(x) = \frac{1}{2h} \frac{1}{\ell} \sum_{i=1}^{\ell} [|x - x_i| < h]$$

Пример – гистограмма оценок

2.1. Poziom podstawowy



Пример – гистограмма возрастов



Локальная непараметрическая оценка Парзена-Розенблатта

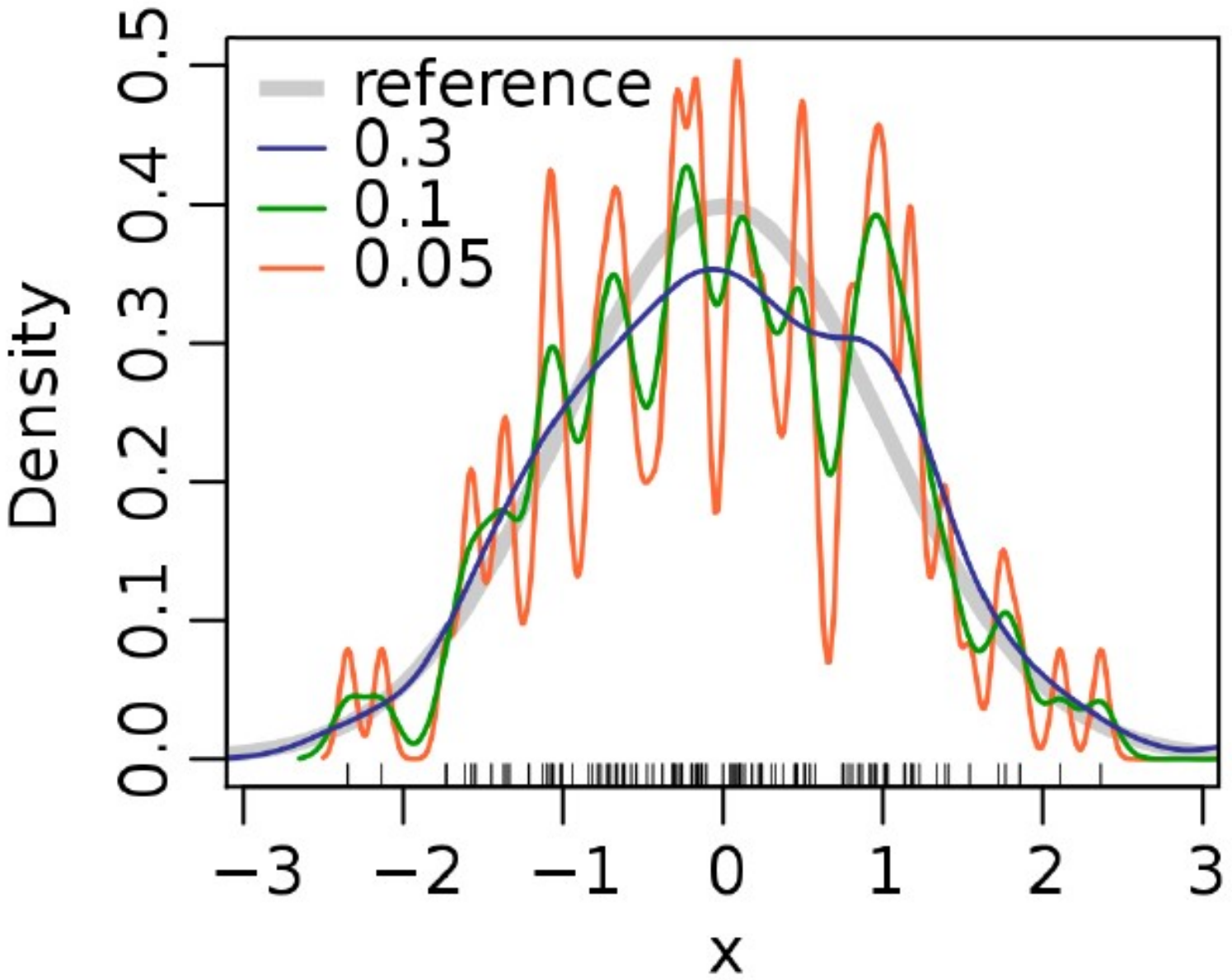
$$\hat{p}_h(x) = \frac{1}{\ell\, h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

K(z) — функция, называемая ядром, чётная и нормированная:

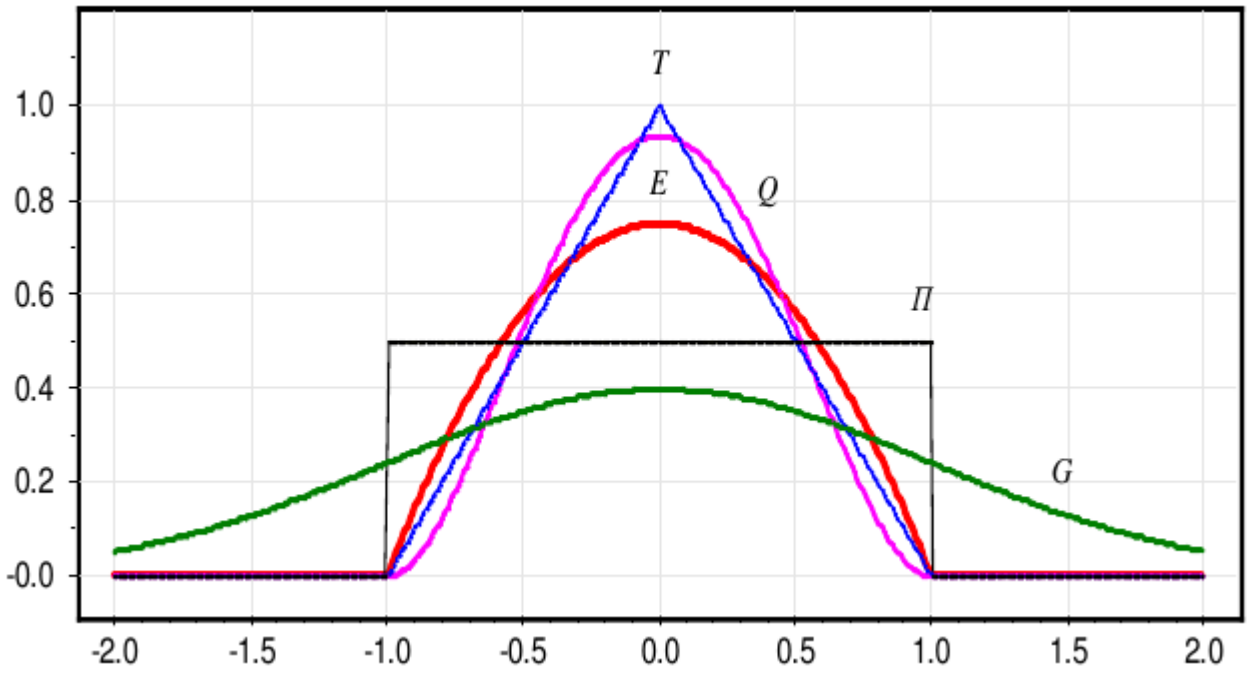
$$\int K(z)\, dz = 1$$

$\hat{p}_h$ сходится к p при $h \to 0$, $\ell \to \infty$, $h\ell \to \infty$

Зависимость от h



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

Функционал качества восстановления плотности:

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

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

В таблице представлены асимптотические значения отношения J(K*)/J(K) при m → ∞, причём это отношение не зависит от p(x).

Параметрическое оценивание плотности

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

- Принцип максимума правдоподобия:

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

- Необходимое условие оптимума:

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

Многомерное нормальное распределение

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

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

где $\mu_y \in \mathbb{R}^n$ — вектор матожидания (центр) класса $y \in Y$
$\Sigma_y \in \mathbb{R}^{n \times n}$ — ковариационная матрица класса $y \in Y$

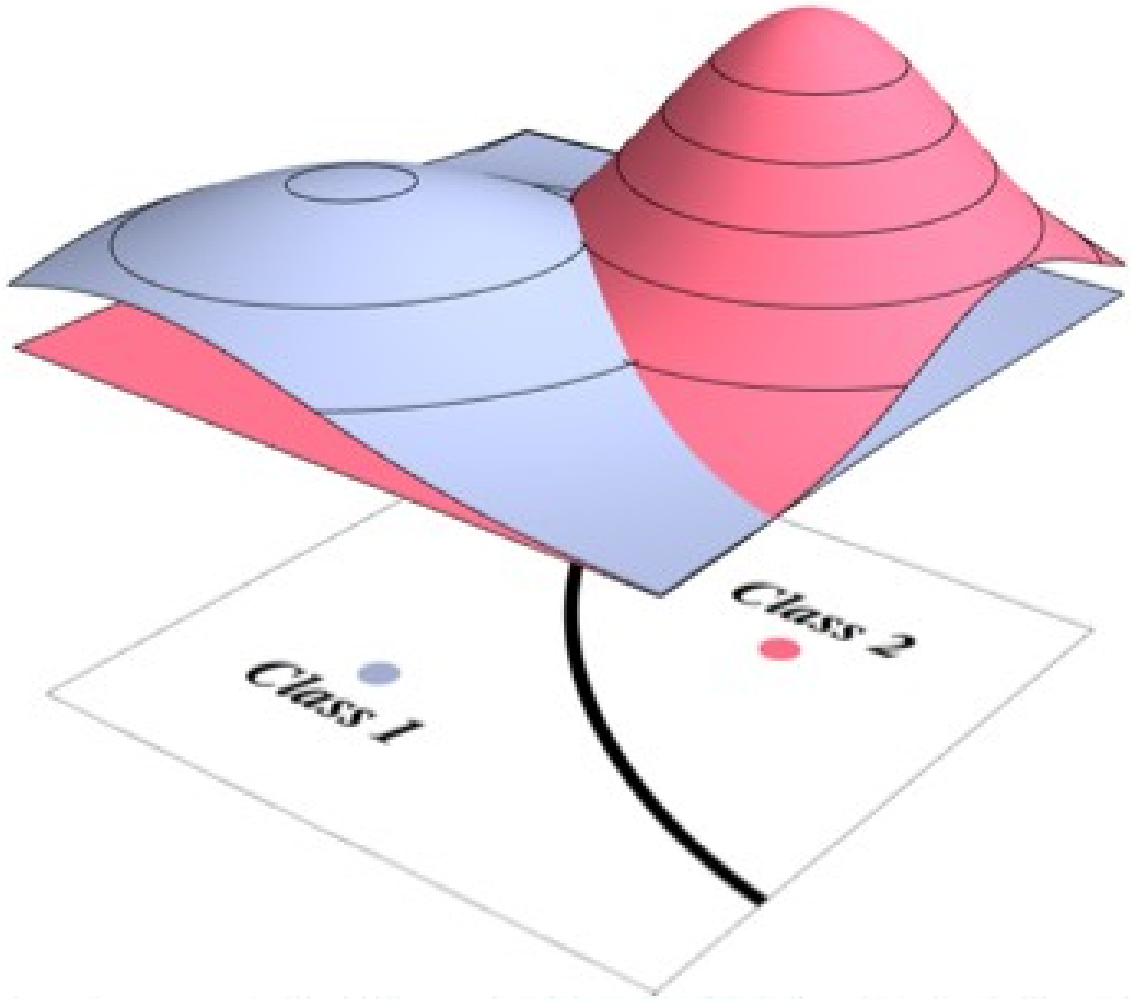
Принцип максимума правдоподобия:

$$L(\theta, X^\ell) = \prod_{i=1}^{\ell} \varphi(x_i, y_i, \theta) \to \max_{\theta}$$

Решение – подстановочный алгоритм:

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

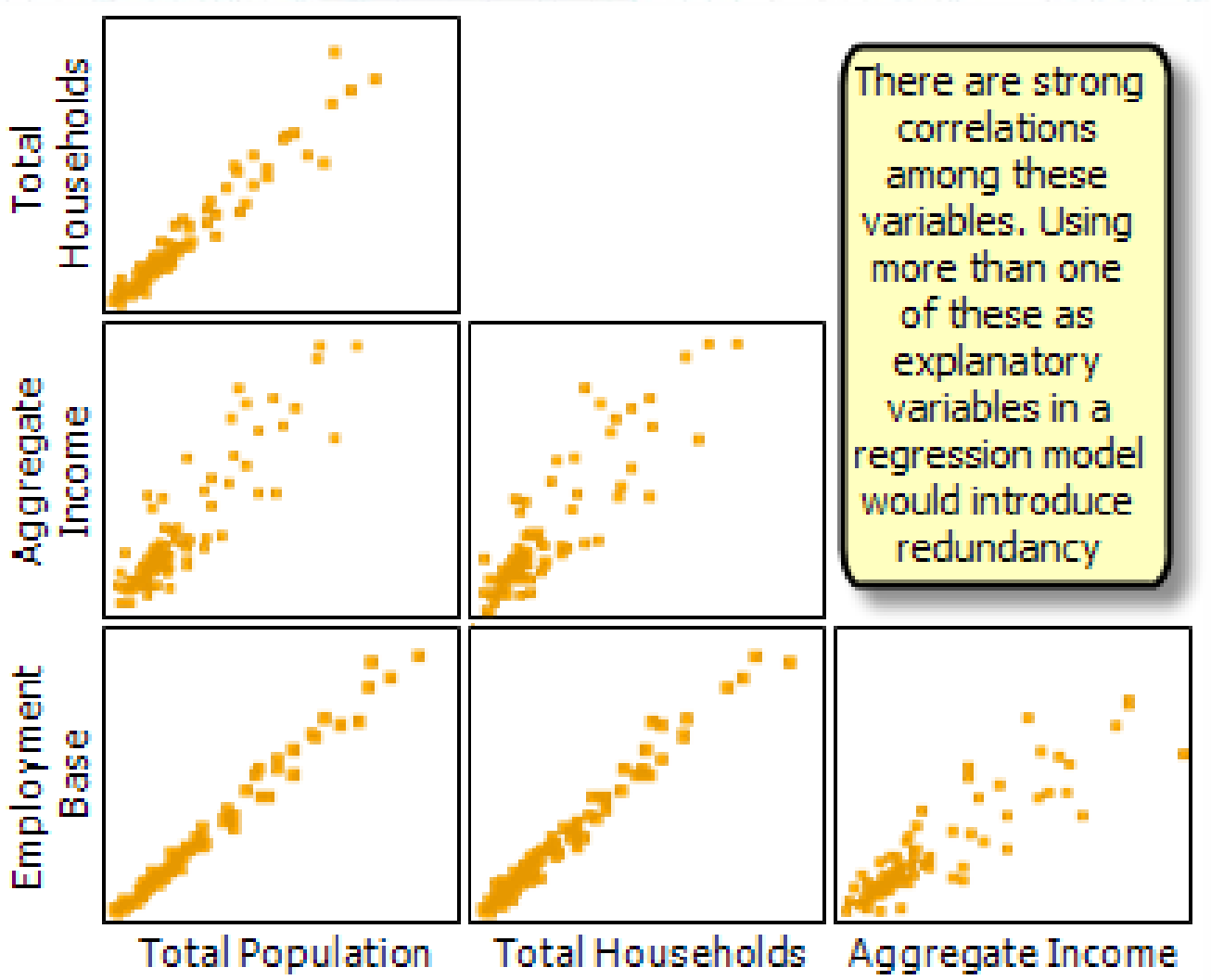
Многомерное нормальное распределение



Недостатки подстановочного алгоритма

- Функции правдоподобия классов могут существенно отличаться от гауссовских.
- Проблема мультиколлинеарности: на практике встречаются задачи, в которых признаки «почти линейно зависимы». Тогда матрица Σ_y^{-1} является плохо обусловленной. Она может непредсказуемо и сильно изменяться при незначительных вариациях исходных данных.
- Выборочные оценки чувствительны к нарушениям нормальности распределений, в частности, к редким большим выбросам.

Мультиколлинеарность признаков



Методы устранения мультиколлинеарности

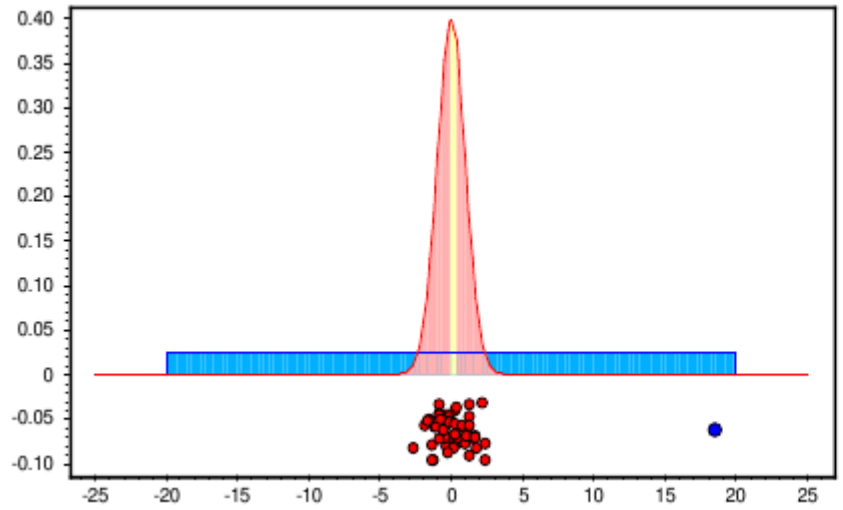
- Регуляризация ковариационной матрицы: обращение $\Sigma+\tau$ вместо Σ
- Диагонализация ковариационной матрицы - нормальный наивный байесовский классификатор:

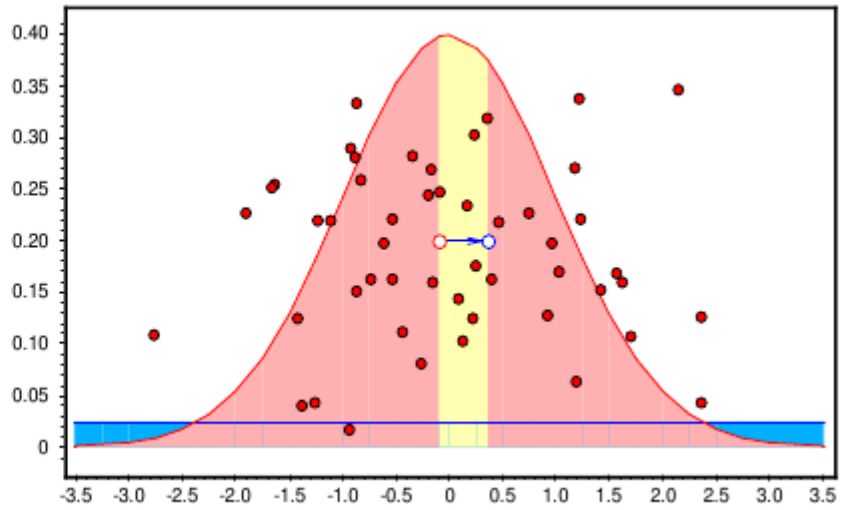
$$p_j(\xi|y) = \frac{1}{\sqrt{2\pi}\sigma_{yj}} \exp\left(-\frac{(\xi-\mu_{yj})^2}{2\sigma_{yj}^2}\right)$$

$$p_y(x) = p_{y1}(\xi_1)\cdots p_{yn}(\xi_n)$$

Проблема выбросов

- Эмпирическое среднее является оценкой матожидания, неустойчивой к редким большим выбросам.





Отсев выбросов

- Идея: решать задачу два раза
  - В первый – найти и исключить выбросы
  - Во второй – построить более точное решение по выборке без выбросов
- Критерий крутого склона: