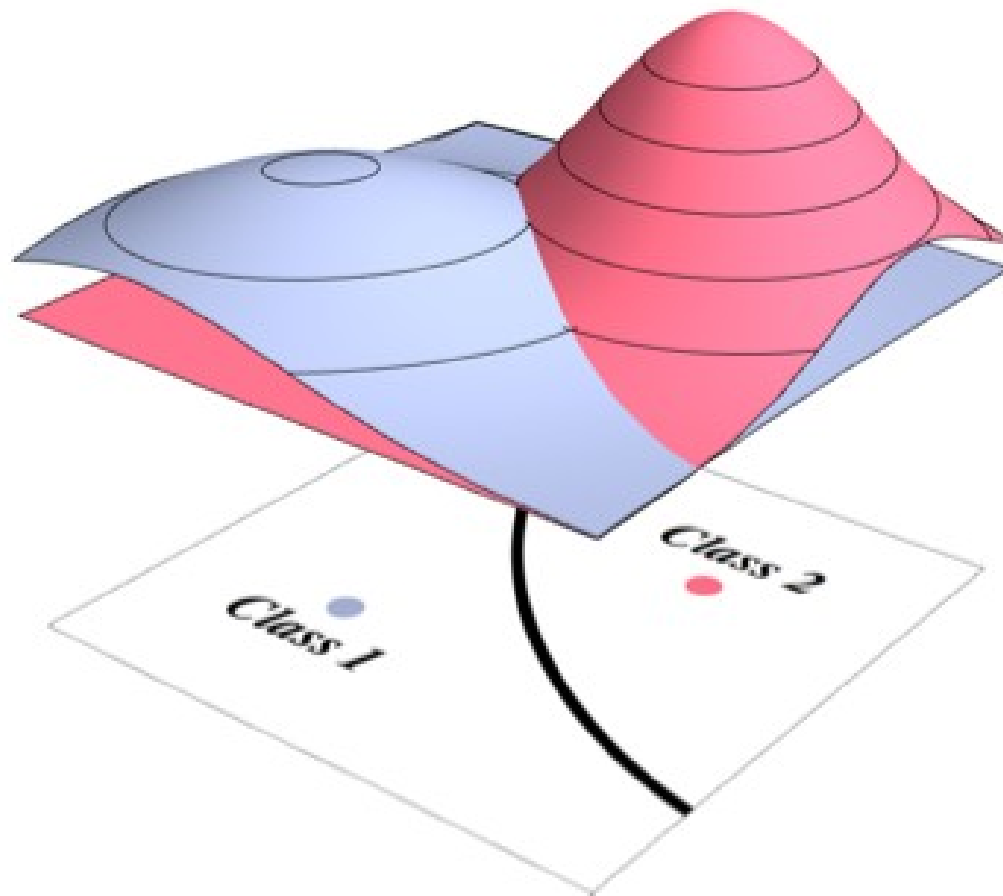


Машинное обучение

Вероятностный подход



Содержание лекции

- Байесовский классификатор
- Восстановление плотности распределения
 - непараметрическое
 - параметрическое

Вероятностная постановка задачи

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

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

- Классификатор с минимальным средним риском:

$$a(x) = \arg \min_s E_y \mathcal{L}(s, y)$$

Пример

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

Пример

- | | | |
|----------|-----|-----|
| y | $+$ | o |
| $p(y x)$ | 0.3 | 0.7 |
- Классификатор с минимальной вероятностью ошибки: $a(x) = o$
- Классификатор с минимальным средним риском Пусть
 $\mathcal{L}(a=+, y=o)=1, \quad \mathcal{L}(a=o, y=+)=3$
 $a(x) = + \quad (\text{т.к. } 1 \cdot 0.7 < 3 \cdot 0.3)$

Подходы к восстановлению плотности распределения

- Непараметрическая оценка плотности:

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

- Параметрическая оценка 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:
$$p(x|y) = p_1(\xi_1|y) \cdots p_n(\xi_n|y), \quad x = (\xi_1, \dots, \xi_n), \quad y \in Y.$$

Непараметрическая оценка

- Определение плотности вероятности (одномерный случай):

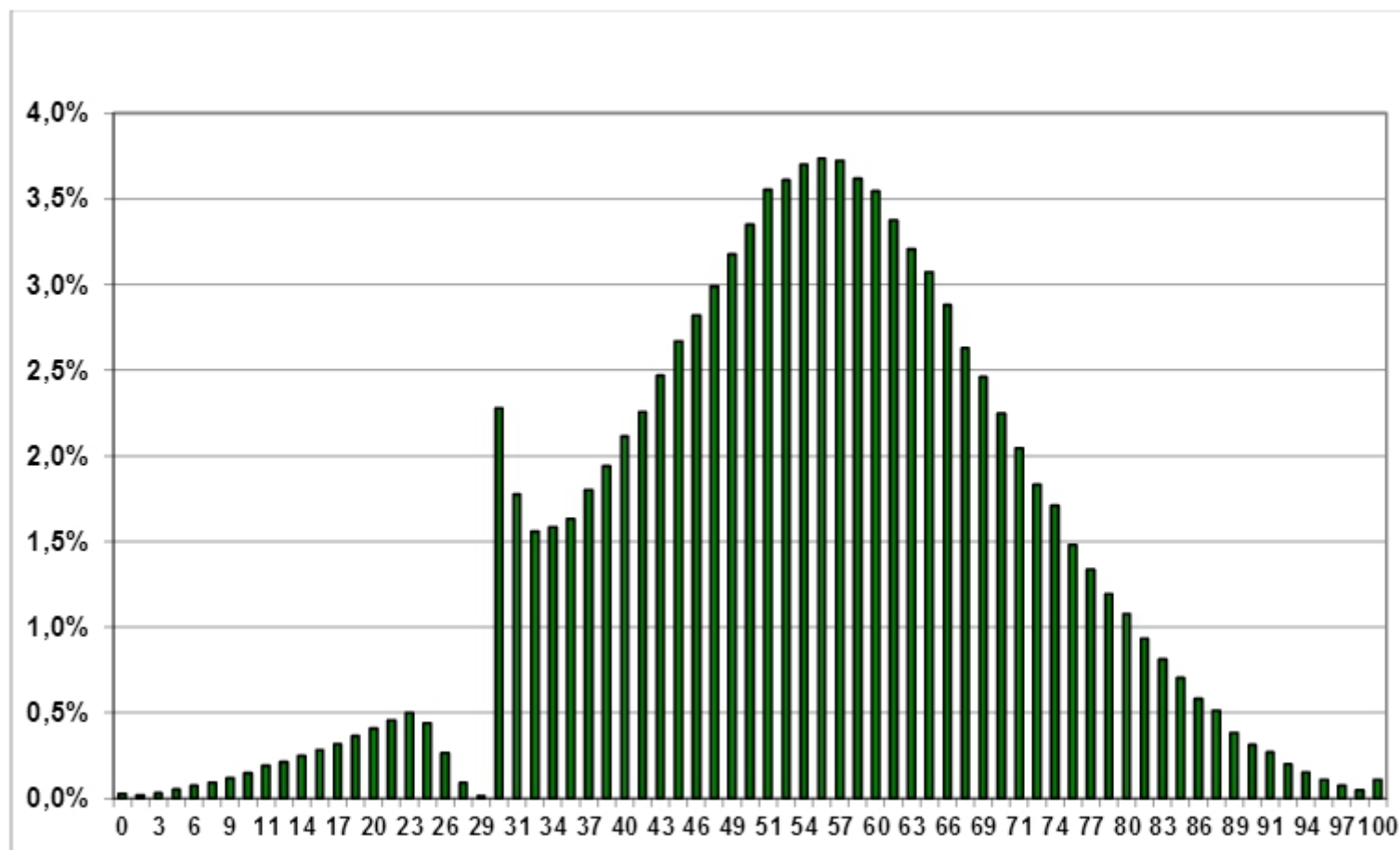
$$p(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P[x - h, x + h]$$

- 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 (Россия 2012г)

Мужчины

Женщины



Локальная непараметрическая оценка Парзена-Розенблатта

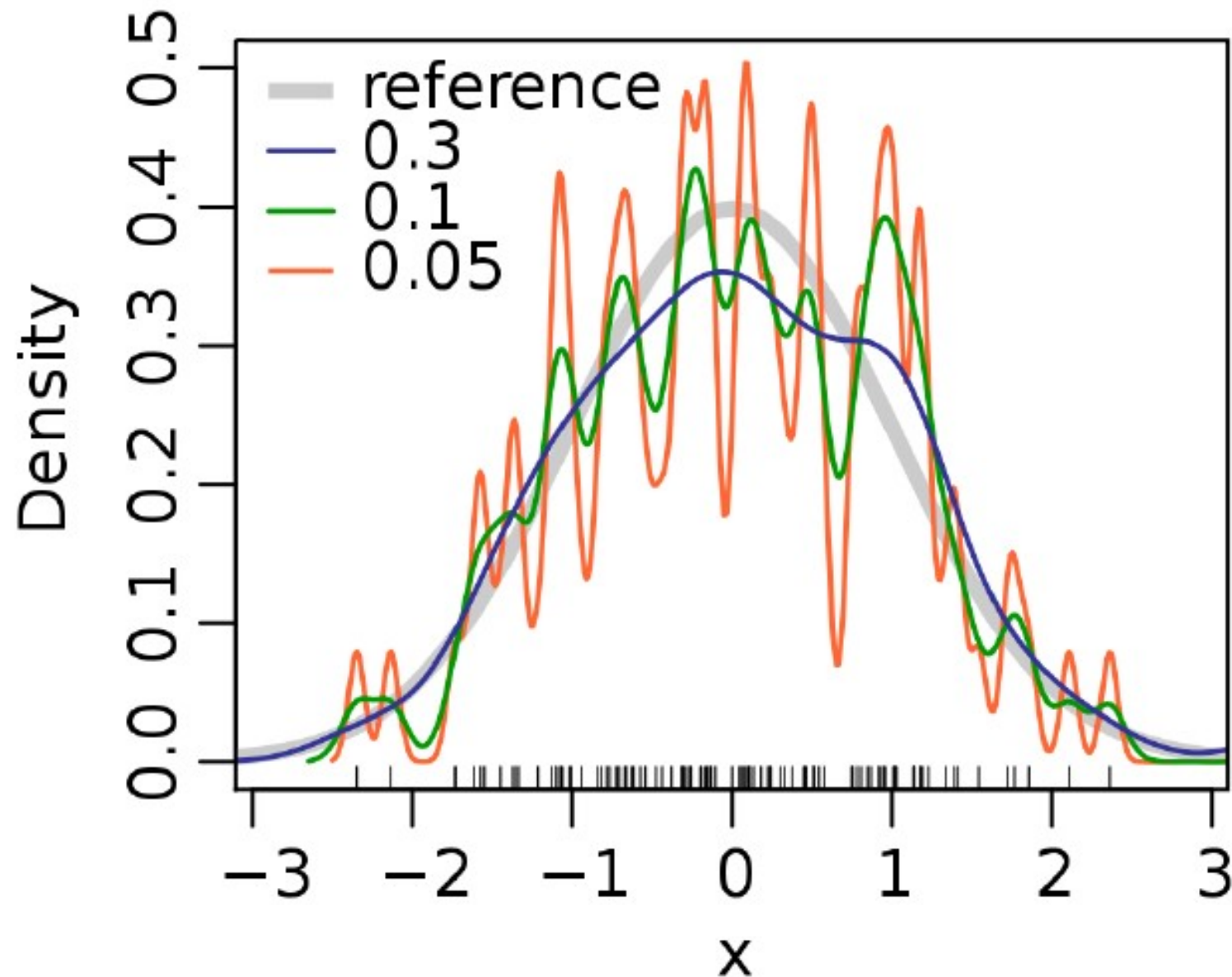
$$\hat{p}_h(x) = \frac{1}{\ell h} \sum_{i=1}^{\ell} K\left(\frac{x - x_i}{h}\right)$$

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

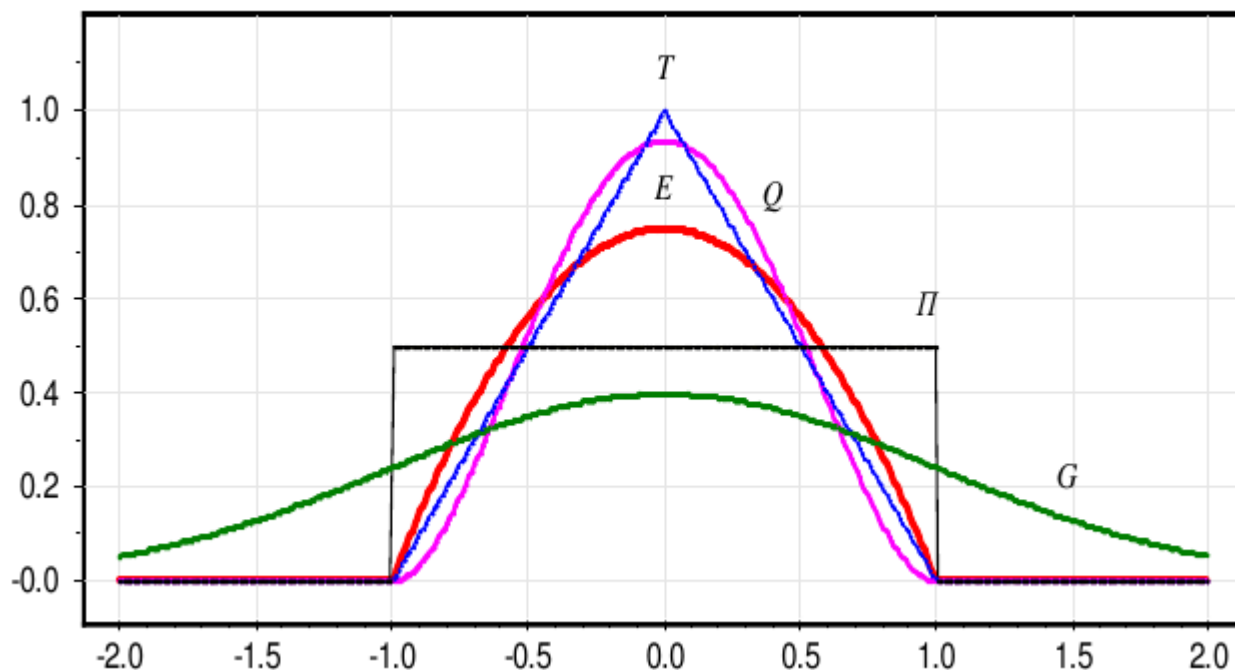
$$\int K(z) dz = 1$$

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

Зависимость от h



Выбор ядра



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

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

$T(r) = (1 - |r|)[|r| \leq 1]$ — 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

Функционал качества восстановления плотности:

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

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 $m \rightarrow \infty$, причём это отношение не зависит от $p(x)$.

Параметрическая оценка плотности

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

- Принцип максимума правдоподобия:

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

- 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

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

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

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

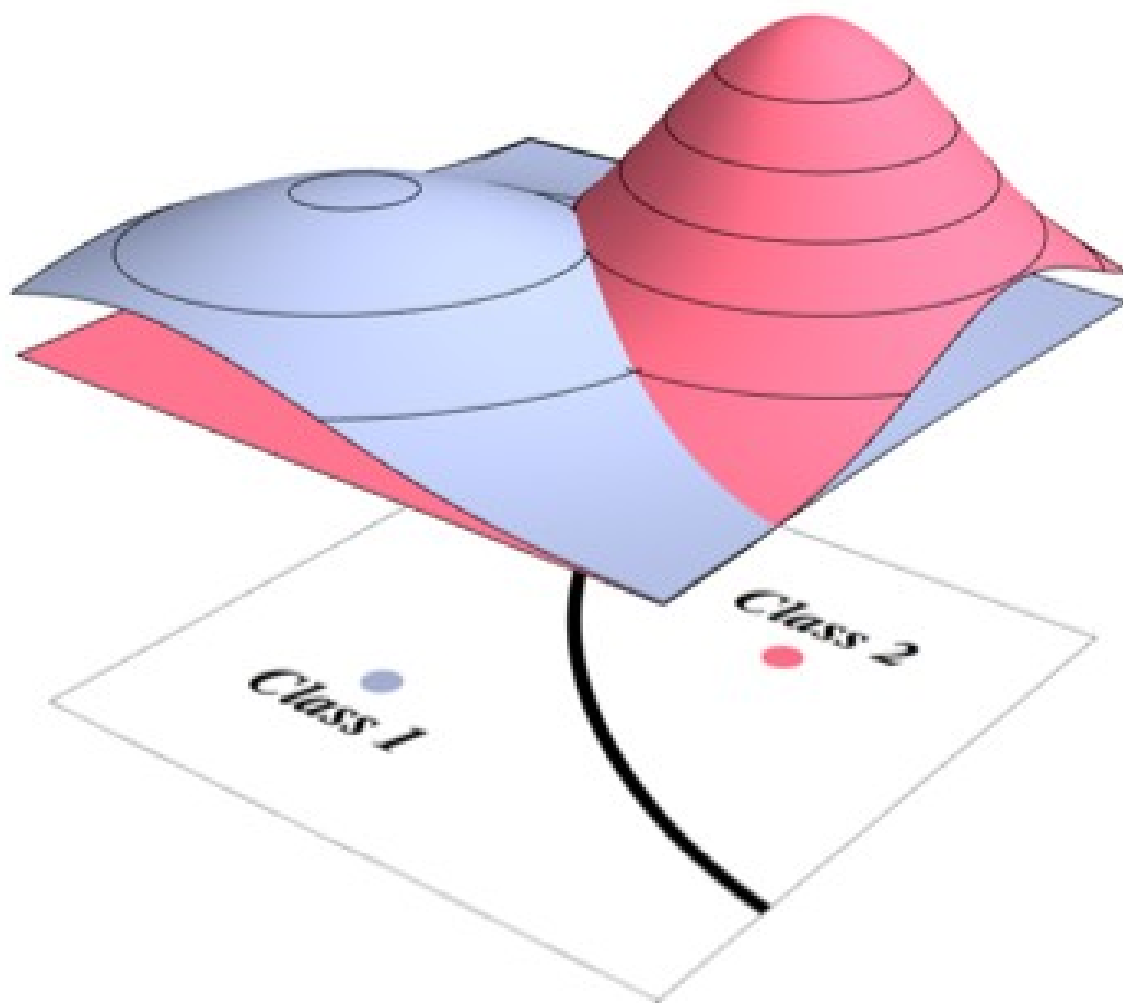
Принцип максимума правдоподобия:

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

Решение — подстановочный алгоритм:

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

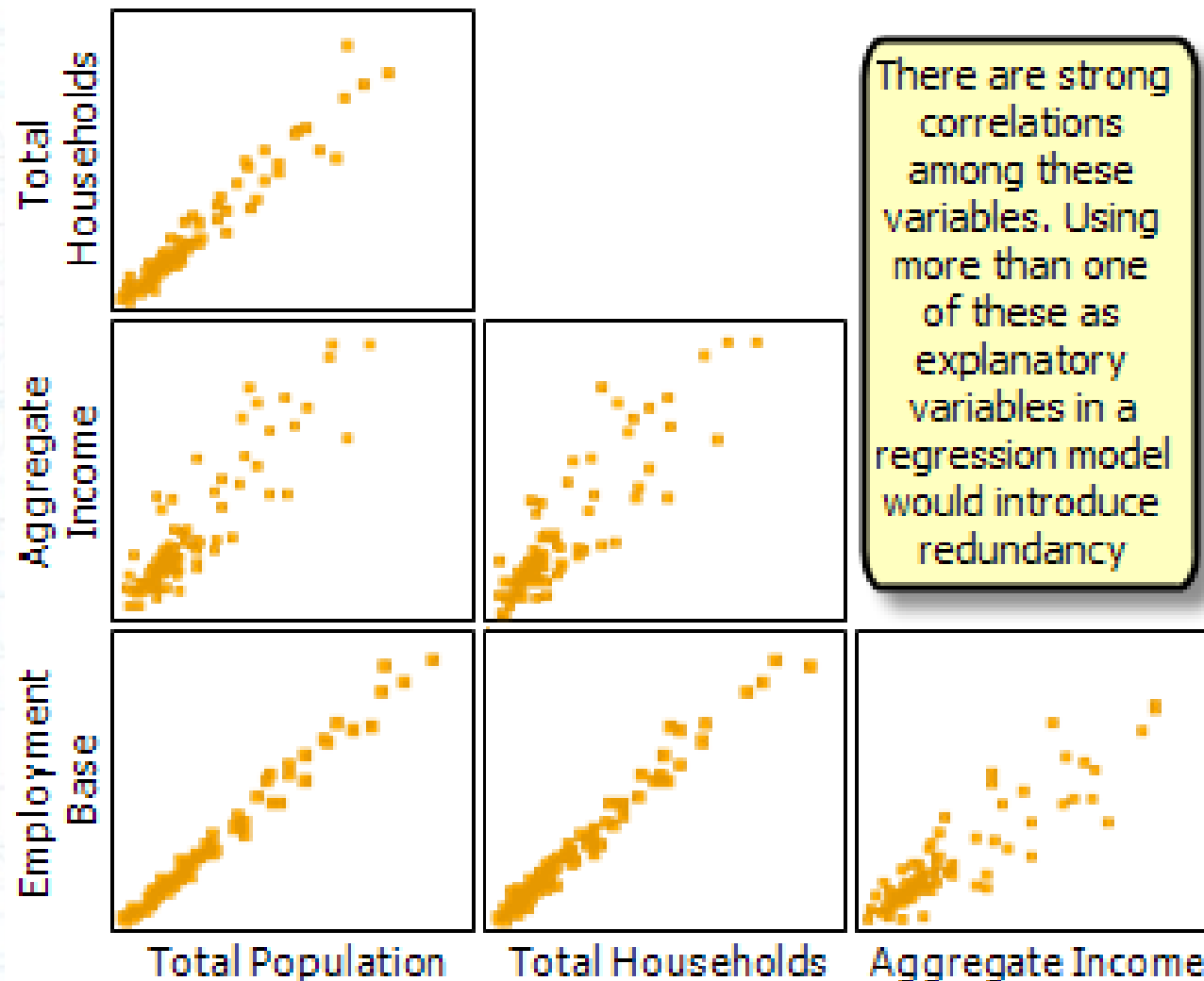
Многомерное нормальное распределение



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

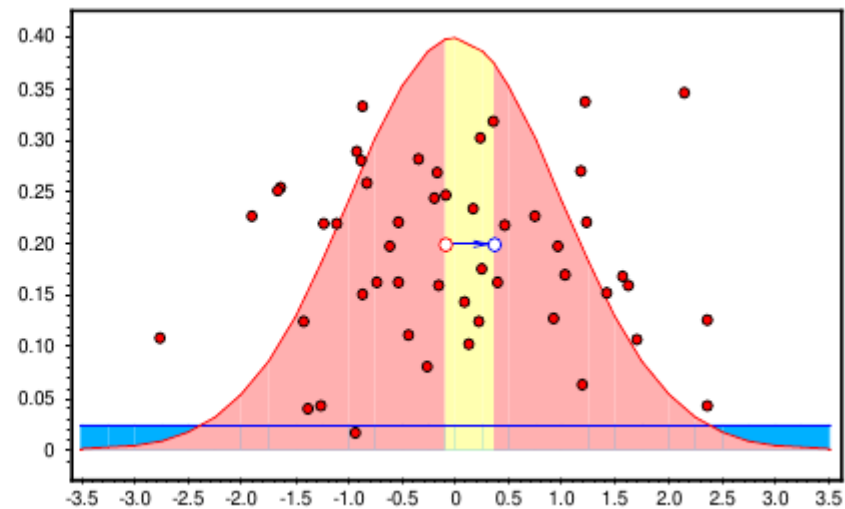
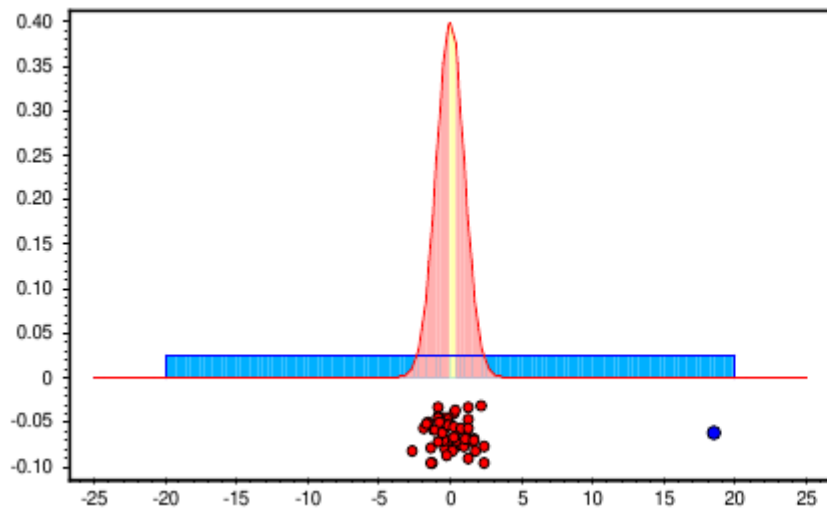
- Регуляризация ковариационной матрицы: обращение $\Sigma + \tau$ вместо Σ
- Диагонализация ковариационной матрицы - нормальный наивный байесовский классификатор:

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

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

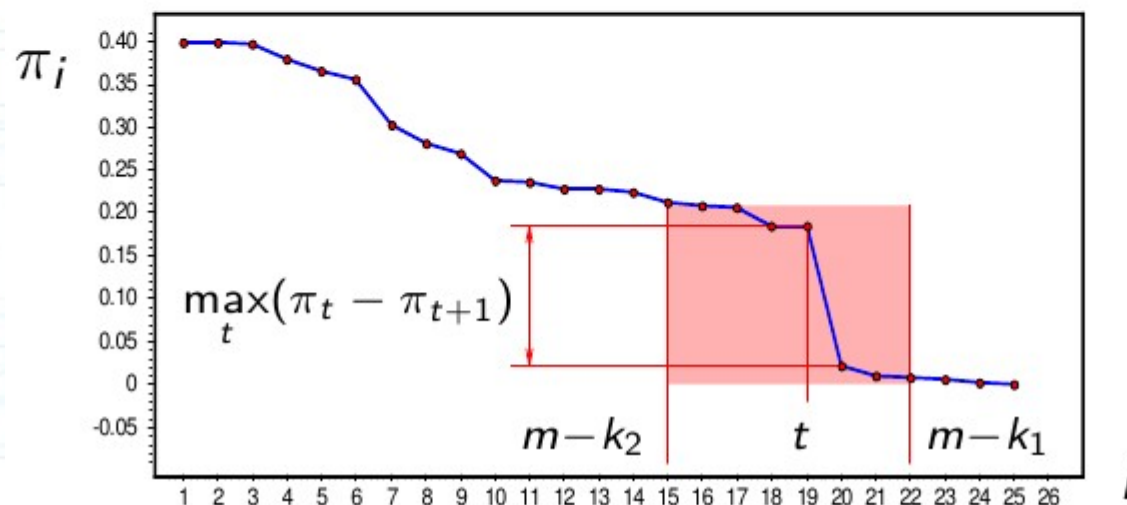
Проблема выбросов

- 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:



Вопросы для самоконтроля

- Сгенерируйте случайную выборку из 10 объектов с классами +1 и -1 и двумя бинарными (или тернарными) признаками: x_1 , x_2 . По заданной выборке вычислите все $p(x_1, x_2|y)$, а также $p(y)$. Примените формулу Байеса и вычислите все $p(y|x_1, x_2)$. Проверьте ваш ответ со значениями $p(y|x_1, x_2)$, вычисленными подсчетом частот.
- Примените наивный байесовский подход к вышеуказанной выборке и вычислите с его помощью $p(x_1, x_2|y)$