

Основы машинного обучения

Поляк Марк Дмитриевич

2025

Классификация

Лекция 5

Логистическая регрессия

Классификация с помощью разделяющих поверхностей, принцип максимизации правдоподобия

Задача обучения классификации

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell, x_i \in \mathbb{R}^n, y_i = \varphi(x_i) \in \{-1, +1\}$

- Модель классификации – линейная с параметром $\theta_n \in \mathbb{R}^n$:

$$g(x, \theta) = \text{sign}\langle x, \theta \rangle = \text{sign} \sum_{j=1}^n \theta_j f_j(x)$$

- Функция потерь – бинарная или её **аппроксимация**:

$$\mathcal{L}(\theta, x) = [g(x, \theta)\varphi(x) < 0] = [\langle x, \theta \rangle \varphi(x) < 0] \leq L(\langle x, \theta \rangle \varphi(x))$$

- Метод обучения – минимизация эмпирического риска:

$$Q(\theta) = \sum_{i=1}^\ell \mathcal{L}(\theta, x_i) = \sum_{i=1}^\ell [\langle x_i, \theta \rangle y_i < 0] \leq \sum_{i=1}^\ell L(\langle x_i, \theta \rangle y_i) \rightarrow \min_{\theta}$$

- Проверка по тестовой выборке $X^k = (\tilde{x}_i, \tilde{y}_i)_{i=1}^k$:

$$\tilde{Q}(\theta) = \frac{1}{k} \sum_{i=1}^k [\langle \tilde{x}_i, \theta \rangle \tilde{y}_i < 0]$$

Задача многоклассовой классификации

Обучающая выборка: $X^\ell = (x_i, y_i)_{i=1}^\ell, x_i \in \mathbb{R}^n, y_i = \varphi(x_i) \in Y$

- Модель классификации – линейная, $\theta = (\theta_y : y \in Y)$:

$$g(x, \theta) = \arg \max_{y \in Y} \langle x, \theta_y \rangle$$

- Функция потерь – бинарная или её аппроксимация:

$$\mathcal{L}(\theta, x) = \sum_{z \neq \varphi(x)} [\langle x, \theta_{\varphi(x)} \rangle < \langle x, \theta_z \rangle] \leq \sum_{z \neq \varphi(x)} L(\langle x, \theta_{\varphi(x)} - \theta_z \rangle)$$

- Метод обучения – минимизация эмпирического риска:

$$Q(\theta) = \sum_{i=1}^\ell \sum_{z \neq y_i} L(\langle x, \theta_{\varphi(x)} - \theta_z \rangle) \rightarrow \min_{\theta}$$

- Проверка по тестовой выборке $X^k = (\tilde{x}_i, \tilde{y}_i)_{i=1}^k$

Разделяющие классификаторы (margin-based classifier)

Бинарный классификатор: $g(x, \theta) = \text{sign } h(x, \theta)$, $Y = \{-1, +1\}$

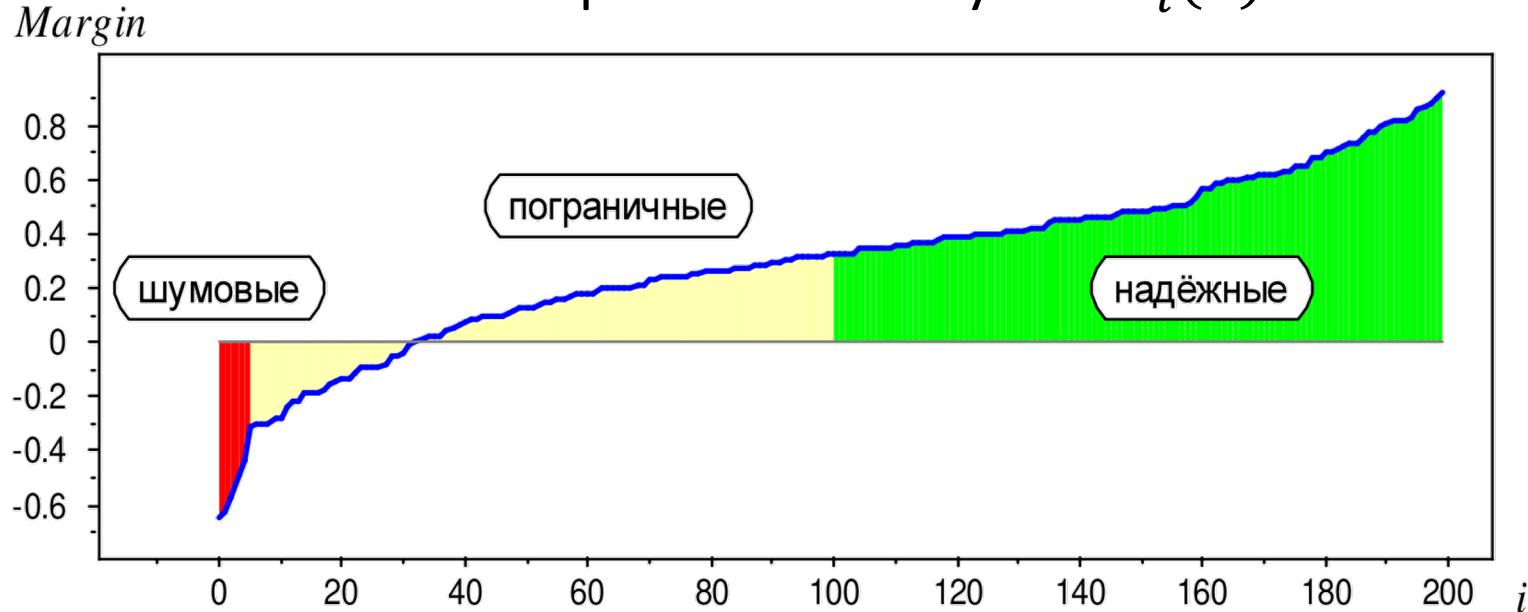
$h(x, \theta)$ – разделяющая (дискриминантная) функция

$x: h(x, \theta) = 0$ – разделяющая поверхность между классами

$M_i(\theta) = h(x_i, \theta)y_i$ – отступ (margin) объекта x_i

$M_i(\theta) < 0 \Leftrightarrow$ алгоритм $g(x, \theta)$ ошибается на x_i

Ранжирование объектов по возрастанию отступов $M_i(\theta)$:



Разделяющие классификаторы (margin-based classifier)

Многоклассовый классификатор:

$$g(x, \theta) = \arg \max_{y \in Y} h_y(x, \theta_y)$$

$h_y(x, \theta_y)$ – дискриминантная функция класса $y \in Y$

$x: h_y(x, \theta_y) = h_z(x, \theta_z)$ – разделяющая поверхность между y, z

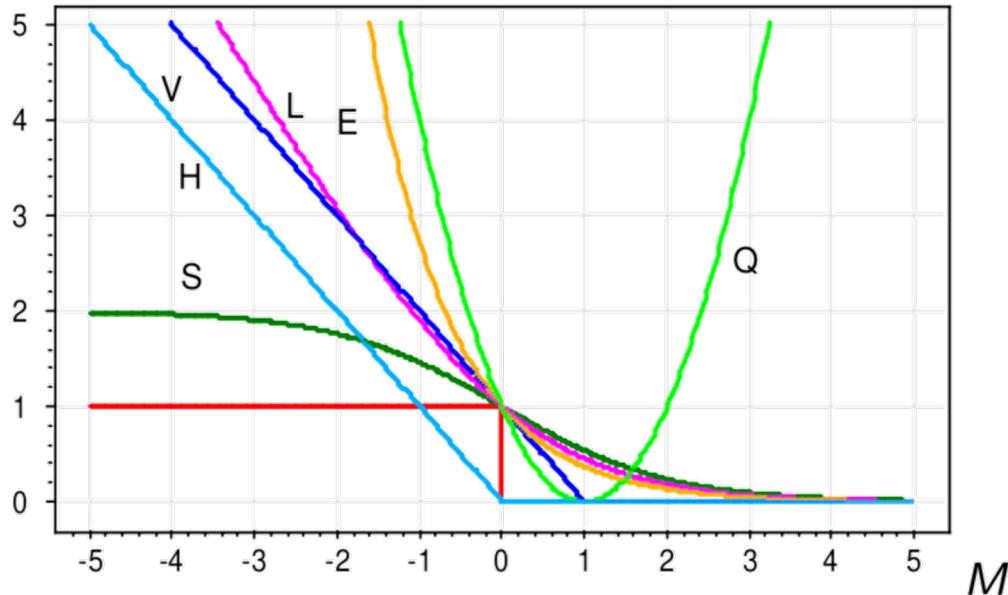
$M_{iy}(\theta) = h_{y_i}(x_i, \theta_{y_i}) - h_y(x_i, \theta_y)$ – отступ объекта x_i от класса y

$M_i(\theta) = \min_{y \neq y_i} M_{iy}(\theta)$ – отступ (margin) объекта x_i

$M_i(\theta) < 0 \Leftrightarrow$ алгоритм $g(x, \theta)$ ошибается на x_i

Непрерывные аппроксимации пороговой функции потерь

Часто используемые непрерывные функции потерь $L(M)$:



- | | |
|-----------------------------|-----------------------------------|
| $V(M) = (1 - M)_+$ | — кусочно-линейная (SVM); |
| $H(M) = (-M)_+$ | — кусочно-линейная (Hebb's rule); |
| $L(M) = \log_2(1 + e^{-M})$ | — логарифмическая (LR); |
| $Q(M) = (1 - M)^2$ | — квадратичная (FLD); |
| $S(M) = 2(1 + e^M)^{-1}$ | — сигмоидная (ANN); |
| $E(M) = e^{-M}$ | — экспоненциальная (AdaBoost); |
| $[M < 0]$ | — пороговая функция потерь. |

Двухклассовая (бинарная) логистическая регрессия

Линейная модель классификации для двух классов $Y = \{-1, +1\}$:

$$g(x, \theta) = \text{sign}(\theta, x), \quad x, \theta \in \mathbb{R}^n$$

Отступ $M = \langle \theta, x \rangle y$.

Логарифмическая функция потерь:

$$L(M) = \log(1 + e^{-M})$$

Модель условной вероятности:

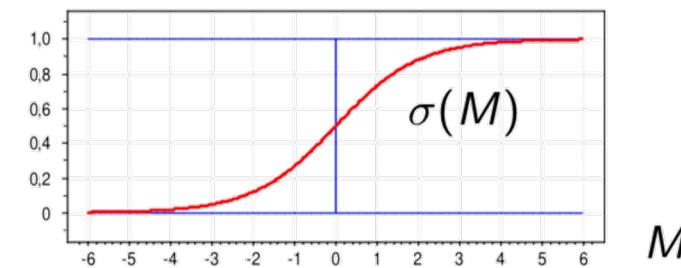
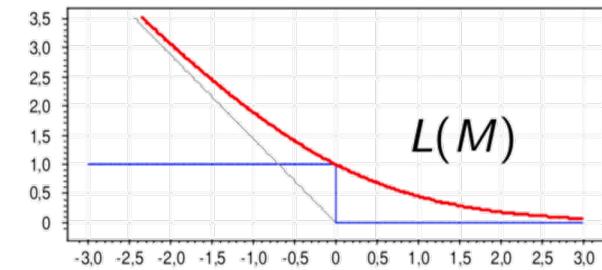
$$P(y = 1|x; \theta) = \sigma(M) = \frac{1}{1 + e^{-M}}$$

где $\sigma(M)$ – сигмоидная (**логистическая**) функция,

важное свойство: $\sigma(M) + \sigma(-M) = 1$.

Максимизация правдоподобия (logistic loss) с регуляризацией:

$$Q_{MAP}(\theta) = \sum_{i=1}^{\ell} \log(1 + \exp(-\langle x, \theta \rangle y_i)) + \frac{\tau}{2} \|\theta\|^2 \rightarrow \min_{\theta}$$



Многоклассовая логистическая регрессия

Линейный классификатор при произвольном числе классов $|Y|$:

$$g(x) = \arg \max_{y \in Y} \langle \theta_y, x \rangle, \quad x, \theta_y \in \mathbb{R}^n$$

Вероятность того, что объект x относится к классу y_k :

$$P(y = y_k | x; \theta) = \frac{\exp \langle \theta_{y_k}, x \rangle}{\sum_{z \in Y} \exp \langle \theta_z, x \rangle} = \text{SoftMax}_{y \in Y} \langle \theta_y, x \rangle$$

функция SoftMax: $\mathbb{R}^Y \rightarrow \mathbb{R}^Y$ переводит произвольный вектор в нормированный вектор дискретного распределения.

Максимизация правдоподобия (log-loss) с регуляризацией:

$$Q_{MAP}(\theta) = \sum_{i=1}^{\ell} \log P(y_i | x_i, \theta) - \frac{\tau}{2} \sum_{y \in Y} \|\theta_y\|^2 \rightarrow \max_{\theta}$$

Пример: бинаризация признаков и скоринговая карта

- Задача кредитного scoringа:
 - x_i – заемщики
 - $y_i = -1(\text{bad}), +1 (\text{good})$
- Бинаризация признаков $f_j(x)$:

$$b_{jk}(x) = [f_j(x) \text{ из } k\text{-го интервала}]$$

- Линейная модель классификации:

$$g(x, \theta) = \text{sign} \sum_{j,k} \theta_{jk} b_{jk}(x)$$

Вес признака θ_{jk} равен его вкладу в общую сумму баллов (score).

признак j	интервал k	θ_{jk}
Возраст	до 25	5
	25 - 40	10
	40 - 50	15
	50 и больше	10
Собственность	владелец	20
	совладелец	15
	съемщик	10
	другое	5
Работа	руководитель	15
	менеджер среднего звена	10
	служащий	5
	другое	0
Стаж	1/безработный	0
	1..3	5
	3..10	10
	10 и больше	15
Работа_мужа /жены	нет/домохозяйка	0
	руководитель	10
	менеджер среднего звена	5
	служащий	1

Бинарная классификация и вероятность

Подход 1:

Отступ $M = \langle \theta, x \rangle y$, где $y \in Y = \{0,1\}$. Вероятность того, что x относится к классу 1:

$$g_1(x, \theta) = P(y = 1|x; \theta) = \sigma(M) = \frac{1}{1 + e^{-M}} = \frac{1}{1 + \exp(-\sum_{j=0}^n \theta_j x_{i,j})}$$

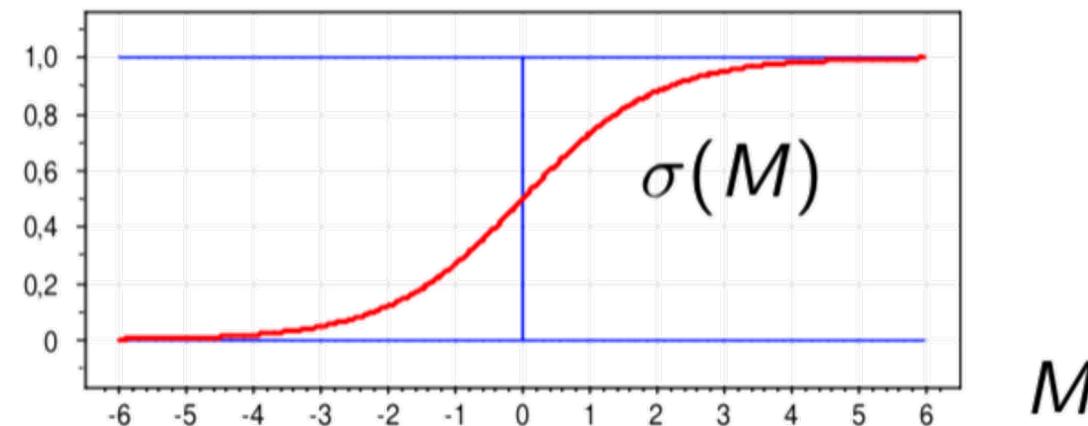
Для решения задачи классификации для двух классов $Y = \{0,1\}$ вводим порог, чаще всего 0.5, т.к. $g(x, \theta) \in [0,1]$: если $g(x, \theta) \geq 0.5$, то $y = 1$, иначе $y = 0$.

Наблюдение:

$$\sigma(M) \Big|_{M=0} = \sigma(0) = 0.5$$

$$\sigma(M) \Big|_{M<0} \in [0, 0.5)$$

$$\sigma(M) \Big|_{M>0} \in (0.5, 1]$$



Подход 2 (упрощение подхода 1):

Линейная модель классификации для двух классов $y \in Y = \{-1, +1\}$:

$$y = g_2(x, \theta) = \text{sign}\langle \theta, x \rangle, \quad x, \theta \in \mathbb{R}^n$$

Пример: классификация опухолей

- Задача медицинской диагностики:

- x_j – характеристики опухоли, напр.: x_1 – размер опухоли в сантиметрах
- $y_1 = 0$ – доброкачественная опухоль (т.н. "negative class" – отсутствие чего-либо)
- $y_2 = 1$ – злокачественная опухоль (т.н. "positive class" – наличие чего-либо)

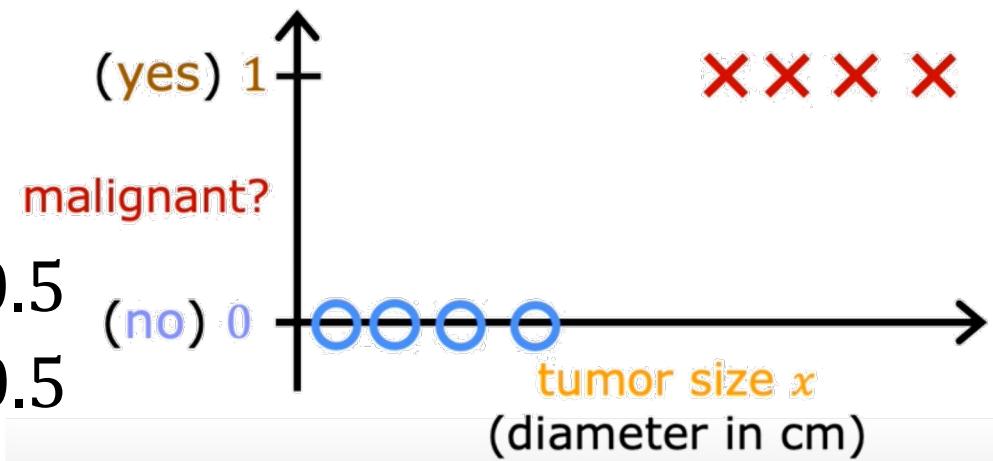
- Линейная модель классификации:

$$g(x, \theta) = \frac{1}{1 + e^{-\langle x, \theta \rangle}} = \frac{1}{1 + \exp(-\sum_{j=0}^n \theta_j x_{i,j})}$$

- $g(x, \theta)$ – "вероятность" того, что класс равен y_2 ,

- Для бинарной классификации можно

использовать порог: $y = \begin{cases} y_1, & \text{if } g(x, \theta) < 0.5 \\ y_2, & \text{if } g(x, \theta) \geq 0.5 \end{cases}$



Пример. Логистическая регрессия с двумя признаками

- Описание задачи:
 - $x = (x_0, x_1, x_2)$, где x_1, x_2 – признаки, $x_0 = 1$
 - $y \in Y = \{0,1\}$ – классы
- Вероятность того, что объект x относится к классу y :

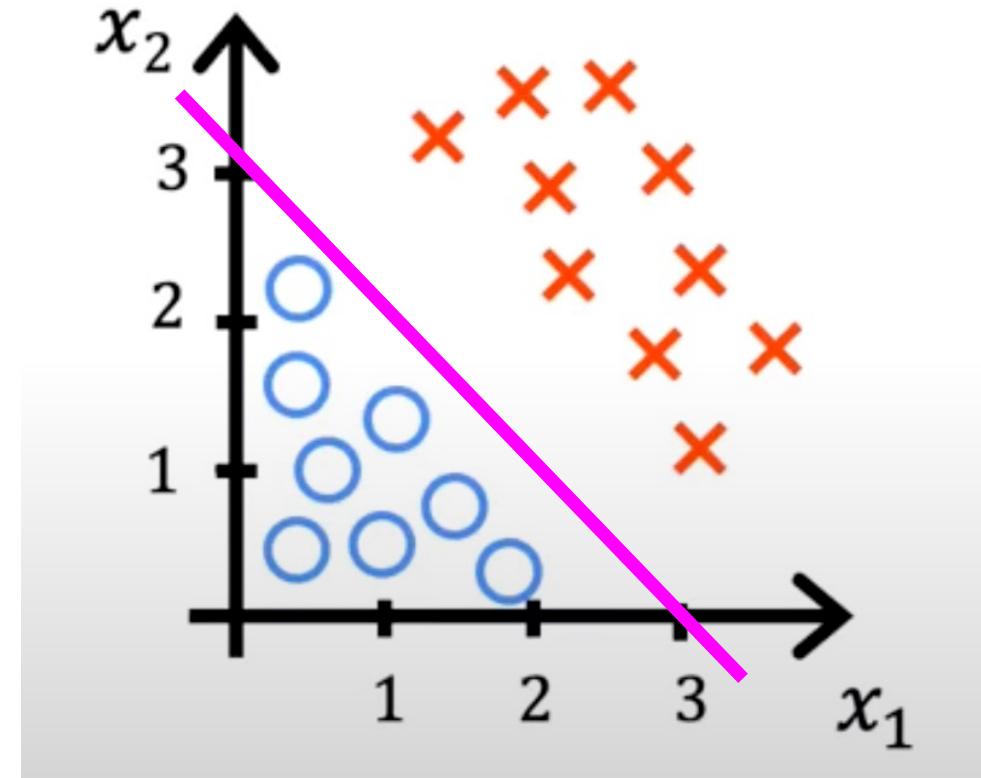
$$P(y = 1|x, \theta) = \sigma(M) = \frac{1}{1 + e^{-M}}$$

- Пусть $\theta = (\theta_0, \theta_1, \theta_2) = (-3, 1, 1)$ и
$$h(x, \theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- Тогда

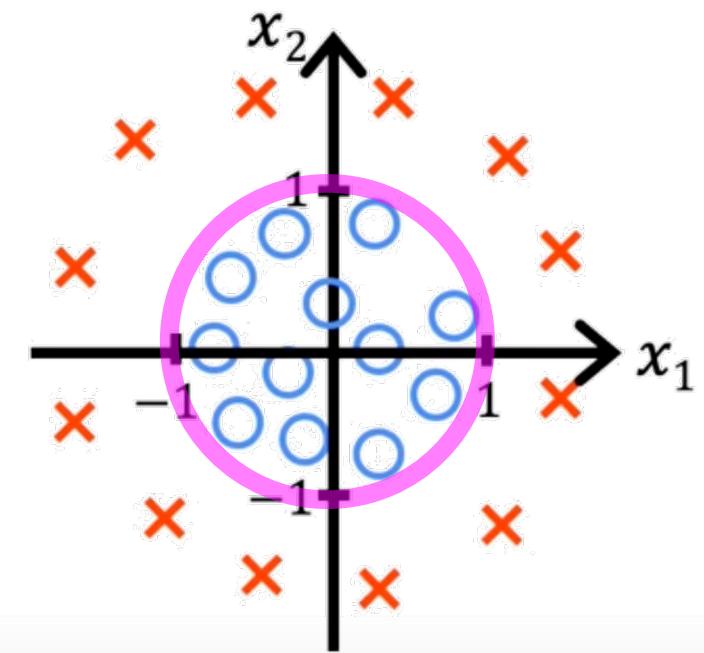
$$M = h(x, \theta) = \langle \theta, x \rangle y = -3 + x_1 + x_2$$

– разделяющая функция



Пример. Логистическая регрессия с двумя признаками

- Пусть классы линейно неразделимы и $h(x, \theta) = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2$
- Тогда, при $\theta = (\theta_0, \theta_1, \theta_2) = (-1, 1, 1)$, разделяющая поверхность $h(x, \theta) = 0$:
$$h(x, \theta) = -1 + x_1^2 + x_2^2 = 0$$
$$x_1^2 + x_2^2 = 1$$
- При $x_1^2 + x_2^2 \geq 1$ получим $\tilde{y} = 1$
- При $x_1^2 + x_2^2 < 1$ получим $\tilde{y} = 0$



Вероятностный подход. Принцип максимизации правдоподобия

- Пусть $X \times Y$ – вероятностное пространство с плотностью $p(x, y)$
- Пусть X^ℓ – простая (i.i.d., independent identically distributed) выборка:
 $(x_i, y_i)_{i=1}^\ell \sim p(x, y)$
- **Задача:** по выборке X^ℓ оценить плотность $p(x, y)$

$p(x, y) = P(y|x; \theta)p(x)$ – параметризация плотности:

$P(y|x; \theta)$ – условная вероятность класса y ;

$p(x)$ – неизвестное и непараметризуемое распределение на X .

- Максимум правдоподобия (Maximum Likelihood Estimate, MLE):

$$p(X^\ell, \theta) = \prod_{i=1}^{\ell} p(x_i, y_i) = \prod_{i=1}^{\ell} P(y_i|x_i; \theta) \cancel{p(x_i)} \rightarrow \max_{\theta}$$

- Максимум логарифма правдоподобия (log-likelihood, log-loss):

$$Q_{MLE}(\theta) = \sum_{i=1}^{\ell} \log P(y_i|x_i, \theta) \rightarrow \max_{\theta}$$

Связь правдоподобия и эмпирического риска

- Максимизация правдоподобия в задаче классификации, где $P(y|x, \theta)$ – модель условной вероятности класса y :

$$Q_{MLE}(\theta) = \sum_{i=1}^{\ell} \log P(y_i|x_i; \theta) \rightarrow \max_{\theta}$$

- Минимизация аппроксимированного эмпирического риска, где $h(x, \theta)$ – модель разделяющей поверхности, $Y = \{\pm 1\}$:

$$Q_{ERM}(\theta) = \sum_{i=1}^{\ell} L(y_i h(x_i, \theta)) \rightarrow \min_{\theta}$$

- Эти два принципа эквивалентны, если положить

$$-\log P(y_i|x_i; \theta) = L(y_i h(x_i, \theta))$$

модель $P(y_i|x_i; \theta)$



модель $h(x_i, \theta)$ и $L(M)$

Вероятностный смысл регуляризации

Двухуровневая модель порождения данных:

$P(y|x; \theta)$ – вероятностная модель данных;

$p(\theta; \gamma)$ – априорное распределение параметров модели;

γ – вектор гиперпараметров.

В этой модели **случайной (стохастической)** является не только выборка X^ℓ , но и вектор параметров θ , а значит и $g(x, \theta)$.

Совместное правдоподобие данных и модели:

$$p(X^\ell, \theta) = p(X^\ell | \theta)p(\theta; \gamma)$$

Принцип максимума апостериорной вероятности (Maximum a Posteriori Probability, MAP):

$$Q_{MAP}(\theta) = \ln p(X^\ell, \theta) = \underbrace{\sum_{i=1}^{\ell} \log P(y_i|x_i; \theta)}_{Q_{MLE}(\theta)} + \underbrace{\log p(\theta; \gamma)}_{\text{регуляризатор}} \rightarrow \max_{\theta}$$

Примеры L_1 и L_2 регуляризации

- Пусть параметры θ_j независимы, $E_{\theta_j} = 0$ и $D_{\theta_j} = C$
- Распределение Гаусса и квадратичный (L_2) регуляризатор:

$$p(\boldsymbol{\theta}; C) = \frac{1}{(2\pi C)^{n/2}} \exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2C}\right), \quad \|\boldsymbol{\theta}\|^2 = \sum_{j=1}^n \theta_j^2,$$

$$-\ln p(\boldsymbol{\theta}; C) = \frac{1}{2C} \|\boldsymbol{\theta}\|^2 + \text{const}$$

- Распределение Лапласа и абсолютный (L_1) регуляризатор:

$$p(\boldsymbol{\theta}; C) = \frac{1}{(2C)^n} \exp\left(-\frac{\|\boldsymbol{\theta}\|}{C}\right), \quad \|\boldsymbol{\theta}\| = \sum_{j=1}^n |\theta_j|,$$

$$-\ln p(\boldsymbol{\theta}; C) = \frac{1}{C} \|\boldsymbol{\theta}\| + \text{const}$$

- C – гиперпараметр, $\tau = \frac{1}{C}$ – коэффициент регуляризации

Метрические методы

Понятие расстояния между объектами, алгоритм k ближайших соседей (k -NN)

Обучение с учителем

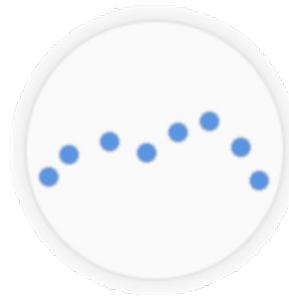
- **Задачи классификации и регрессии:**

X – объекты, Y – ответы;

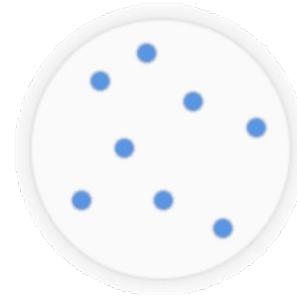
$X^\ell = (x_i, y_i)_{i=1}^\ell$ – обучающая выборка.

- **Гипотеза непрерывности** (для регрессии): *близким объектам соответствуют близкие ответы*

выполнена:

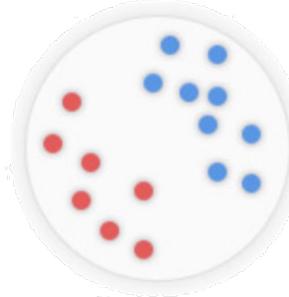


не выполнена:

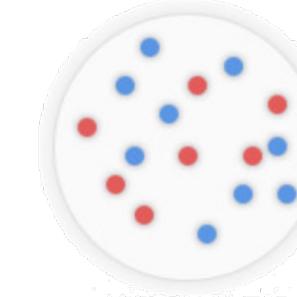


- **Гипотеза компактности** (для классификации): *близкие объекты, как правило, лежат в одном классе*

выполнена:



не выполнена:



Формализация понятия «расстояние» (distance)

- Евклидова метрика и обобщенная метрика Минковского:

$$\rho(x_i, x_k) = \left(\sum_{j=1}^n |x_{i,j} - x_{k,j}|^2 \right)^{1/2}, \quad \rho(x_i, x_k) = \left(\sum_{j=1}^n \theta_j |x_{i,j} - x_{k,j}|^p \right)^{1/p}$$

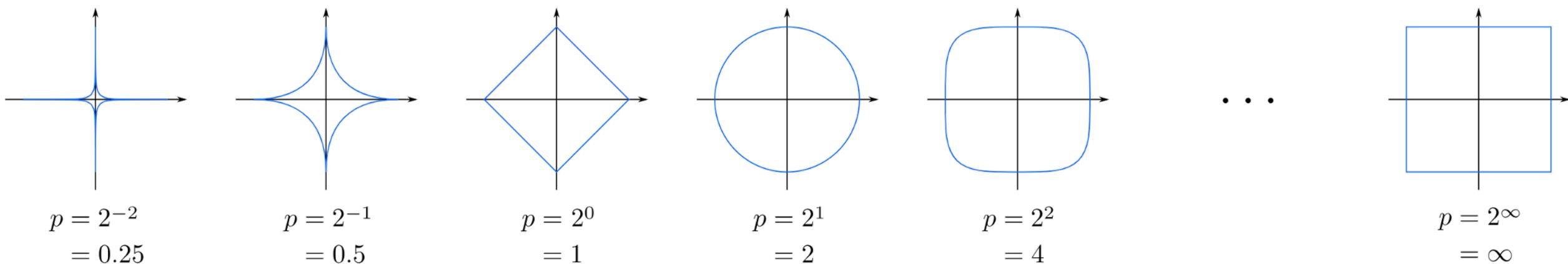
$x_i = (x_{i1}, \dots, x_{in})$ – вектор признаков объекта x_i

$x_k = (x_{k1}, \dots, x_{kn})$ – вектор признаков объекта x_k

$\theta_1, \dots, \theta_n$ – обучаемые веса (параметры) признаков, играющие две роли:

- нормировка, т.е. приведение к общему масштабу;
- задание степени важности (информативности) признаков.

- Линии уровня при различных p



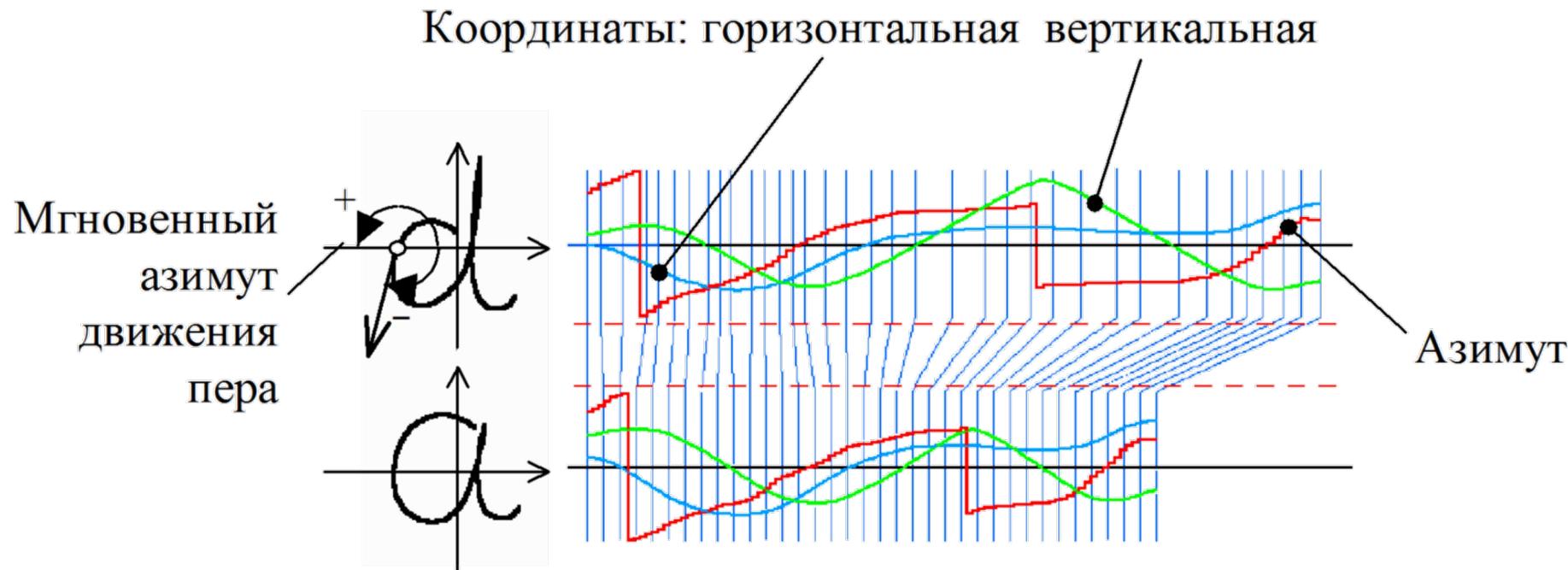
Расстояние между строками и сигналами

- Для строк – редакторское расстояние Левенштейна:

CTGGGСТАAAA~~GGT~~CCTTAGCC..TTTAGGAAAAA.GGCCATTAGG

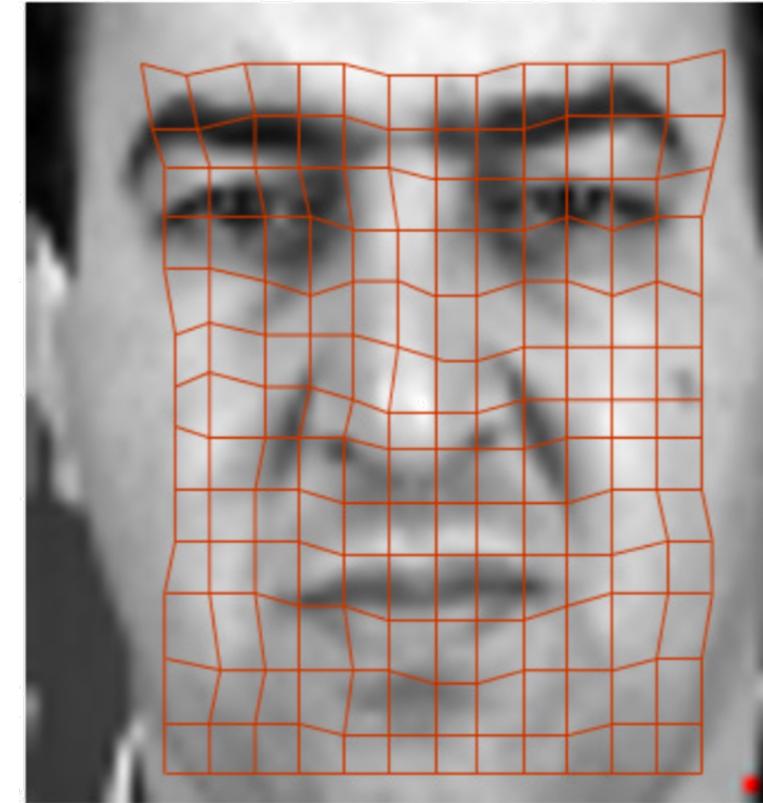
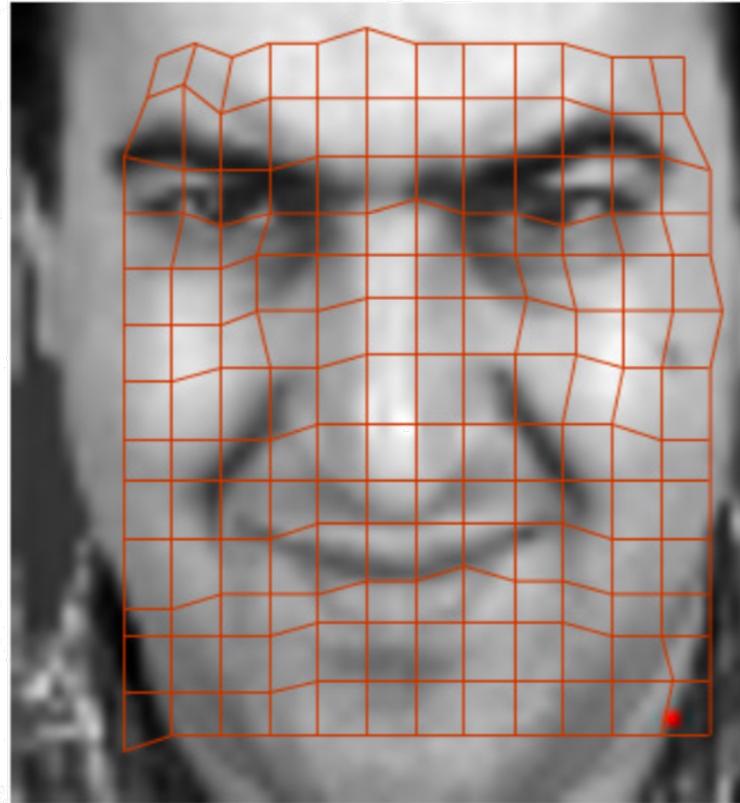
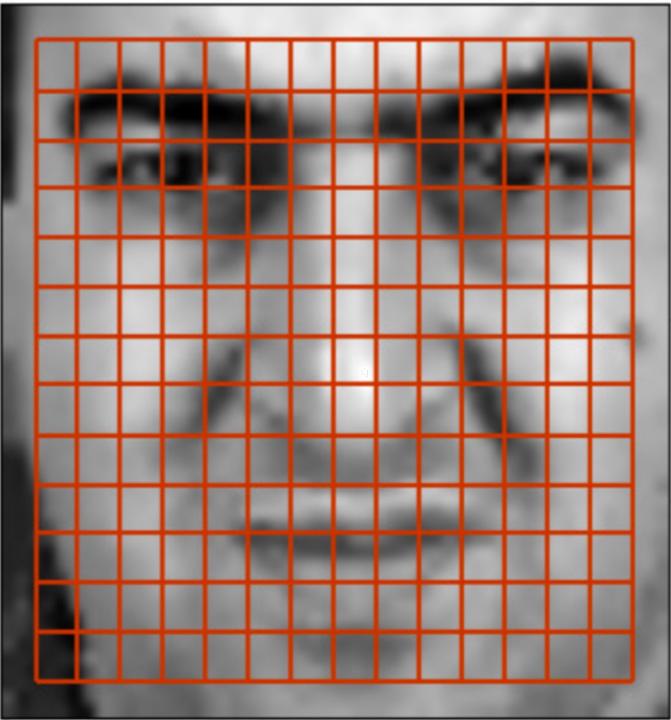
CTGGGACTAAA....CCTTAGCC~~T~~TTACAAAA~~T~~GGGCCATTAGG

- Для сигналов – энергия сжатий и растяжений



Расстояние между изображениями

- Расстояние между изображениями на основе выравнивания:



- Оценивается энергия растяжения прямоугольной сетки

Обобщенный метрический классификатор

- Для произвольного $x \in X$ отранжируем объекты x_1, \dots, x_ℓ :

$$\rho(x, x^{(1)}) \leq \rho(x, x^{(2)}) \leq \dots \leq \rho(x, x^{(\ell)})$$

$x^{(i)}$ – i -ый сосед объекта x среди x_1, \dots, x_ℓ ;

$y^{(i)}$ – ответ на i -м соседе объекта x .

- Метрический алгоритм классификации относит объект x к тому классу, которому принадлежат его ближайшие соседи:

$$g(x; X^\ell) = \arg \max_{y \in Y} \underbrace{\sum_{i=1}^{\ell} [y^{(i)} = y] w(i, x)}_{\Gamma_y(x)}$$

$w(i, x)$ – вес, степень близости к объекту x его i -го соседа, неотрицателен, не возрастает по i .

$\Gamma_y(x)$ – оценка близости объекта x к классу y .

Метод k ближайших соседей (k nearest neighbours, kNN)

$w(i, x) = [i \leq 1]$ – метод ближайшего соседа

$w(i, x) = [i \leq k]$ – метод k ближайших соседей

- Преимущества:

- простота реализации (lazy learning);
- параметр k можно оптимизировать по leave-one-out:

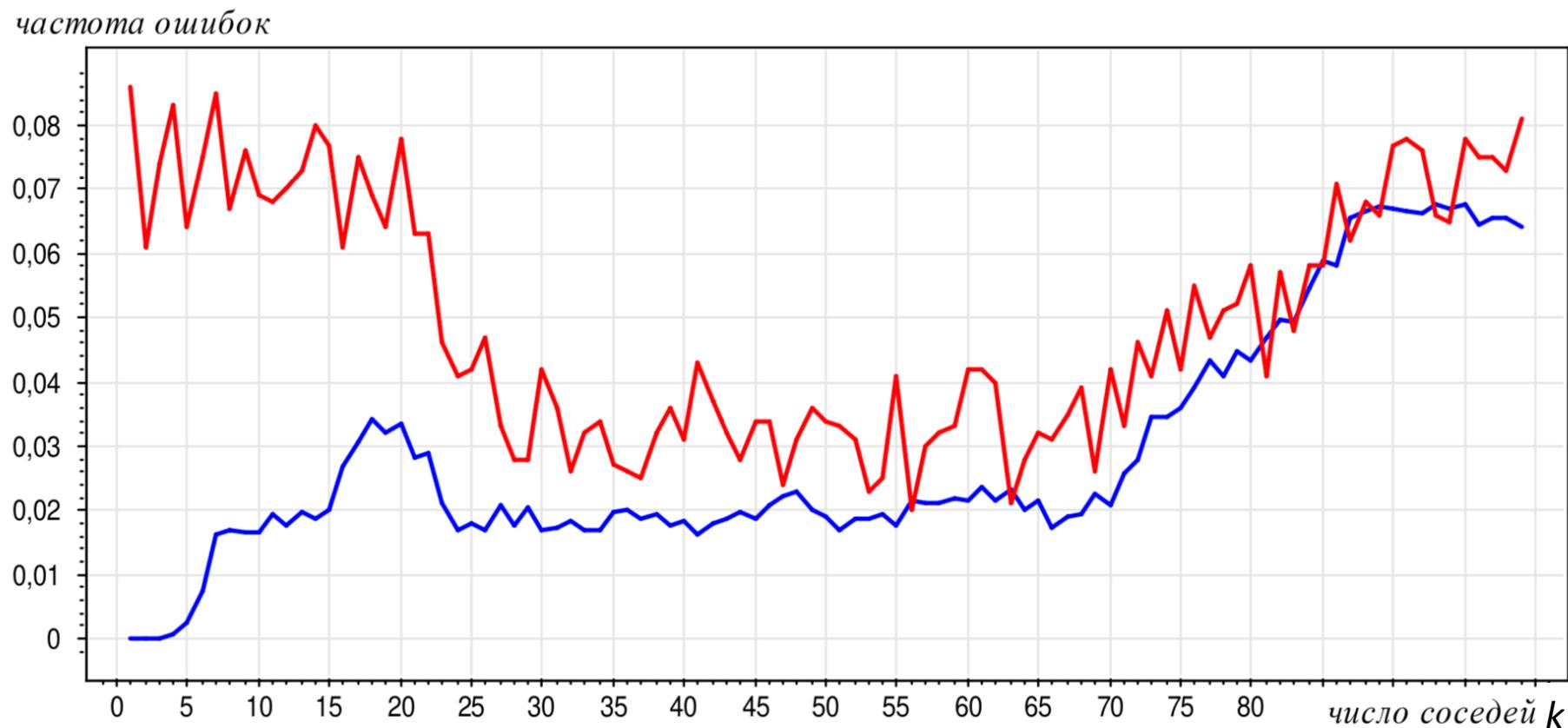
$$\text{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} [g(x_i; X^\ell \setminus \{x_i\}, k) \neq y_i] \rightarrow \min_k$$

- Недостатки:

- неоднозначность классификации при $\Gamma_y(x) = \Gamma_s(x), y \neq s$
- не учитываются значения расстояний



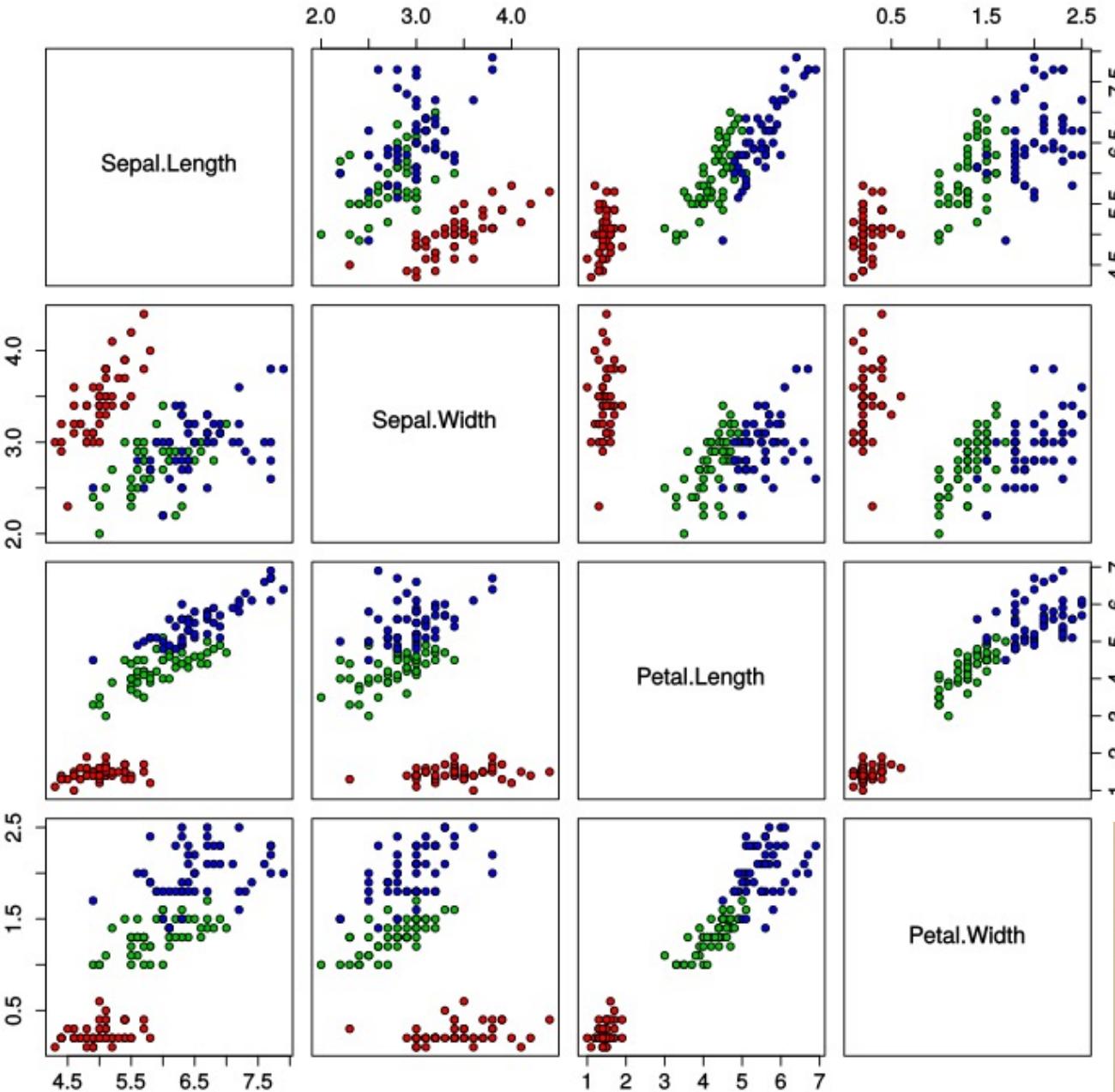
Пример. Ирисы Фишера



- смещённое число ошибок, когда объект учитывается как сосед самого себя
- несмешённое число ошибок LOO

Пример: классификация ирисов Фишера

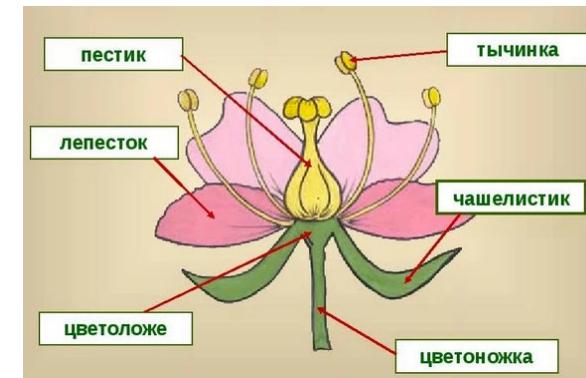
Iris Data (red=setosa,green=versicolor,blue=virginica)



$n = 4$ признака,
 $|Y| = 3$ класса,
длина выборки
 $\ell = 150$ объектов

petal – лепесток,
sepal – чашелистик

Данные собраны
Рональдом Фишером в
1936 году



Метод k взвешенных ближайших соседей

$$w(i, x) = [i \leq k] \theta_i,$$

θ_i – вес, зависящий только от номера соседа.

- Возможные эвристики:

$$\theta_i = \frac{k+1-i}{k} \text{ – линейное убывание веса;}$$

$$\theta_i = q^i \text{ – экспоненциально убывающие веса, } 0 < q < 1$$

- Проблемы:

- как более обоснованно задать веса?
- Возможно, было бы лучше, если бы вес $w(i, x)$ зависел не от порядкового номера соседа i , а от расстояния до него $\rho(x, x^{(i)})$



Метод окна Парзена

$w(i, x) = K \left(\frac{\rho(x, x^{(i)})}{h} \right)$, где h – ширина окна (bandwidth; радиус окрестности)

$K(r)$ – ядро (kernel), не возрастает и положительно на $[0, 1]$

- Метод парзеновского окна *фиксированной ширины*:

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

- Метод парзеновского окна *переменной ширины*:

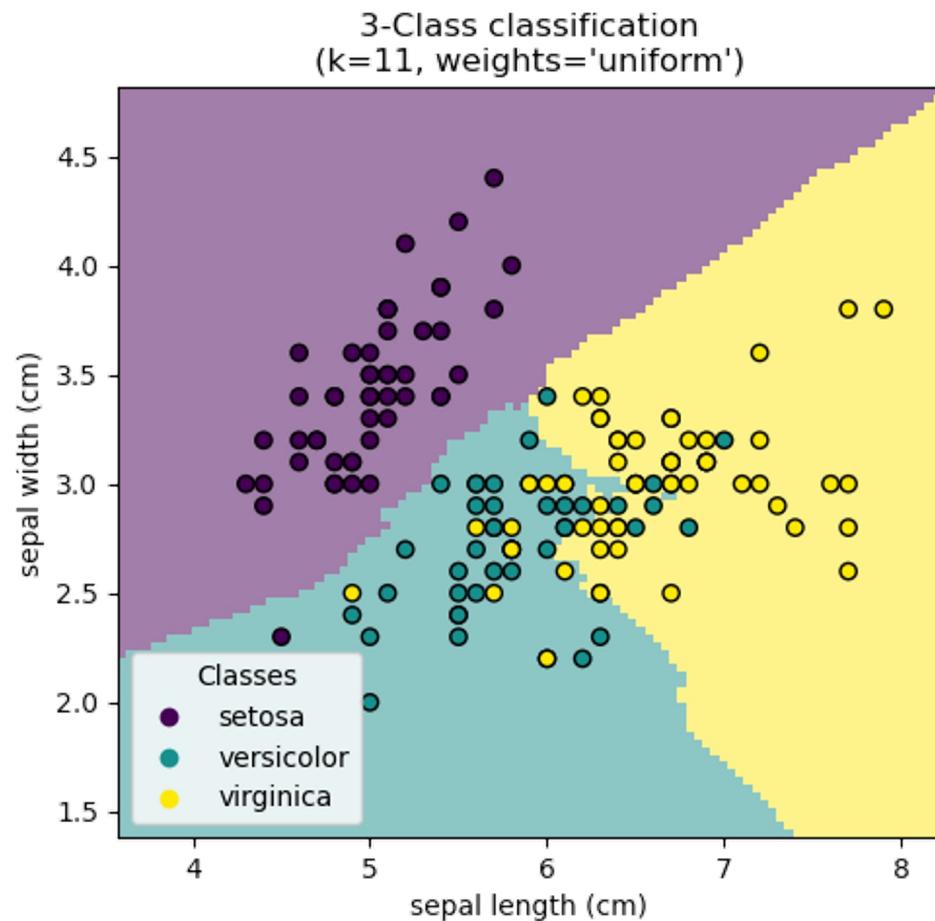
$$g(x; X^\ell, \textcolor{red}{k}, K) = \arg \max_{y \in Y} \sum_{i=1}^{\ell} [y_i = y] K \left(\frac{\rho(x, x^{(i)})}{\rho(x, x^{(k+1)})} \right)$$

- Оптимизация параметров – по критерию LOO:

- выбор ширины окна h или числа соседей k
- выбор ядра K

Пример. Ирисы Фишера, kNN и scikit-learn

$$\theta_i = 1, i = 1..n$$



$$\theta_i = \frac{1}{\rho(x, x_i)}, i = 1..n$$

