

# Введение в анализ данных

## Лекция 3

### Линейная алгебра и метрические методы

Евгений Соколов

[esokolov@hse.ru](mailto:esokolov@hse.ru)

НИУ ВШЭ, 2018

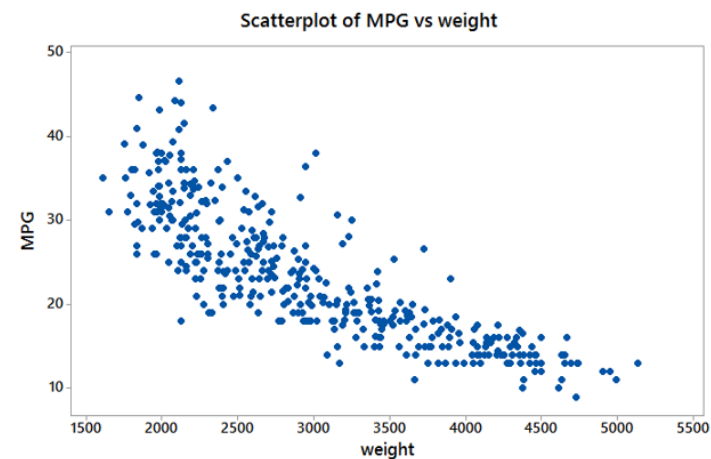
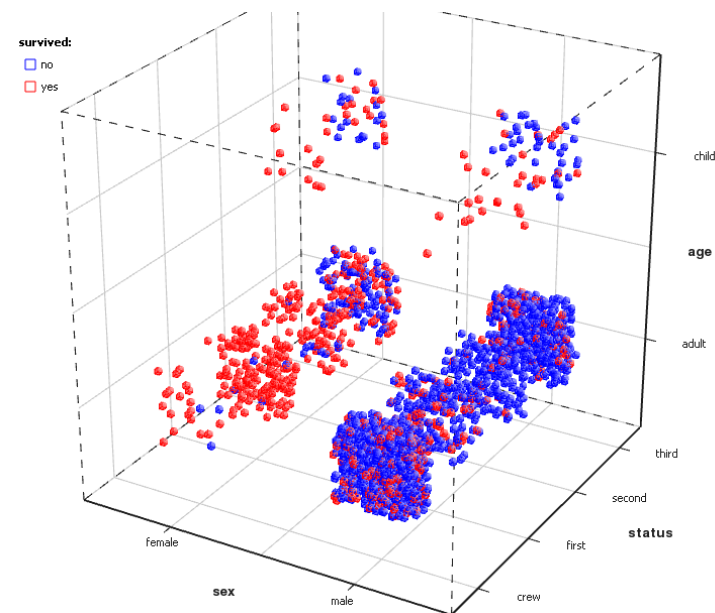
# Векторы и матрицы

# Вектор

- $x = (x^1, \dots, x^d)$  — признаковое описание
- $x^1, \dots, x^d$  — вещественные числа
- $x$  — набор из  $d$  чисел — **вектор**

# Вектор

- 5 — число
- (5, 3) — точка на плоскости
- (5, 3, 9) — точка в пространстве
- (5, 3, 9, 1) — точка в четырехмерном пространстве
- ...



# Векторное пространство

- Что мы будем называть вектором?
- Векторное пространство  $V$  — множество, состоящее из векторов
- Например,  $V$  — все наборы из  $d$  вещественных чисел

# Векторное пространство

- Какие операции над векторами нам нужны?
- Простейшие: сложение и умножение на число
- Как их ввести? Какими свойствами они должны обладать?

# Аксиомы

1.  $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ , для любых  $\mathbf{x}, \mathbf{y} \in V$  (коммутативность сложения);
2.  $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ , для любых  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$  (ассоциативность сложения);
3. существует такой элемент  $\mathbf{0} \in V$ , что  $\mathbf{x} + \mathbf{0} = \mathbf{x}$  для любого  $\mathbf{x} \in V$  (существование нейтрального элемента относительно сложения), называемый **нулевым вектором** или просто **нулём** пространства  $V$ ;
4. для любого  $\mathbf{x} \in V$  существует такой элемент  $-\mathbf{x} \in V$ , что  $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ , называемый вектором, **противоположным** вектору  $\mathbf{x}$ ;
5.  $\alpha(\beta\mathbf{x}) = (\alpha\beta)\mathbf{x}$  (ассоциативность умножения на скаляр);
6.  $1 \cdot \mathbf{x} = \mathbf{x}$  (унитарность: умножение на нейтральный (по умножению) элемент поля  $F$  сохраняет вектор).
7.  $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$  (дистрибутивность умножения на вектор относительно сложения скаляров);
8.  $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$  (дистрибутивность умножения на скаляр относительно сложения векторов).

# Векторное пространство

Множество  $V$  называется векторным пространством, если:

- На нем заданы операции  $+$  (сложение) и  $*$  (умножение на число)
- Оно замкнуто относительно этих операций
- Для этих операций выполнены 8 аксиом



# Евклидово пространство

- Пространство наборов из  $d$  вещественных чисел — евклидово пространство  $\mathbb{R}^d$
- Бывают пространства с более сложными элементами: многочленами, уравнениями, функциями

# Евклидово пространство

- Сложение и умножение на число — покомпонентно
- Сложение
  - $a = (a_1, \dots, a_d)$
  - $b = (b_1, \dots, b_d)$
  - $a + b = (a_1 + b_1, \dots, a_d + b_d)$
- Умножение на число:
  - $a = (a_1, \dots, a_d)$
  - $\beta \in \mathbb{R}$
  - $\beta a = (\beta a_1, \dots, \beta a_d)$

# Матрицы

- Вектор описывает один объект
- А если объектов несколько?

# Матрицы

The diagram shows a matrix with 16 rows and 8 columns. A blue box highlights the first column, labeled 'Признак' (Feature) with an upward arrow. Another blue box highlights the third row, labeled 'Объект' (Object) with a leftward arrow. The matrix contains numerical data, with some cells containing decimal values and others containing integers.

36,18	2	1	2	3	59090	1
46,47671233	0	1	4	3	14773	1
45,13424658	0	1	3	3	19376	2
25,88767123	1	1	4	3	16098	0
25,70410959	1	1	3	3	20338	0
33,03	0	1	3	1	501667	2
46,44931507	3	1	2	1	26100	0
51,24383562	0	0	4	2	20727	0
46,8739726	0	1	1	3	27861	0
39,8630137	3	1	4	2	33495	1
37,09	0	1	4	3	55825	1
38,14	3	1	3	2	60000	1
45,46849315	3	1	1	1	40000	1
42,99726027	3	1	4	2	40343	0
29,98082192	3	0	4	2	27583,78	2
46,20547945	3	1	2	2	45385	1

# Матрицы

- Матрица — таблица с числами
- Пример:

$$A = \begin{pmatrix} 1 & 2 & 5 & 1 \\ 5 & 3 & 9 & 0 \\ 0 & 7 & 1 & 4 \end{pmatrix}$$

- Два индекса: строка и столбец
- $a_{11} = 1$
- $a_{23} = 9$

# Матрицы

- Матрица — таблица с числами
- Пример:

$$A = \begin{pmatrix} 1 & 2 & 5 & 1 \\ 5 & 3 & 9 & 0 \\ 0 & 7 & 1 & 4 \end{pmatrix}$$

- Два индекса: строка и столбец
- $a_{11} = 1$
- $a_{23} = 9$
- Пространство матриц 3 на 4:  $\mathbb{R}^{3 \times 4}$  (тоже векторное!)

# Матрицы

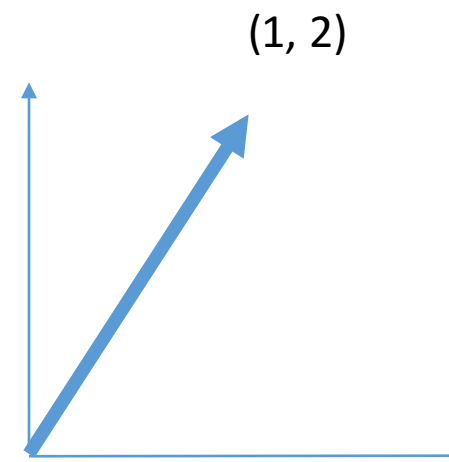
- Выборка объектов описывается матрицей «объекты-признаки»
- По строкам — объекты
- По столбцам — признаки

$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

# Операции в векторных пространствах

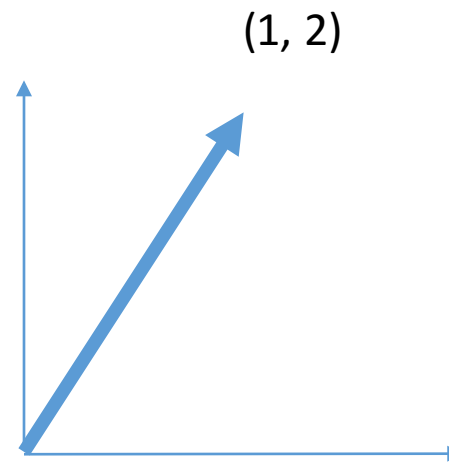


- Вектор — точка и стрелка, идущая к ней из нуля

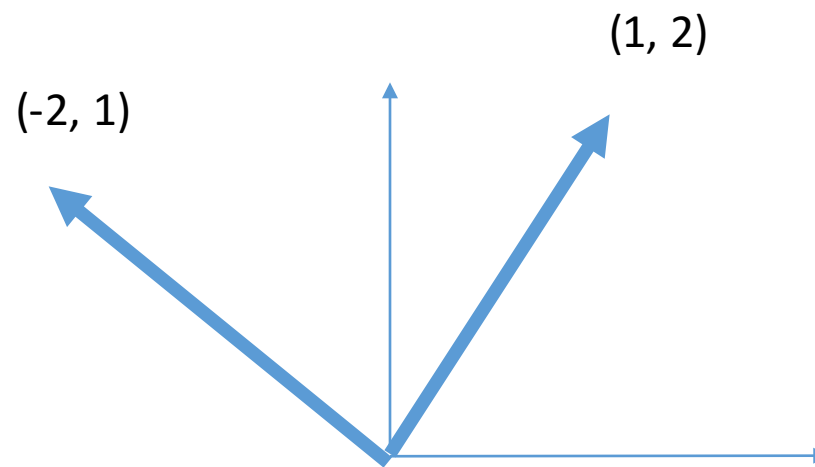


- Длина вектора:

$$\sqrt{1^2 + 2^2} = \sqrt{5}$$

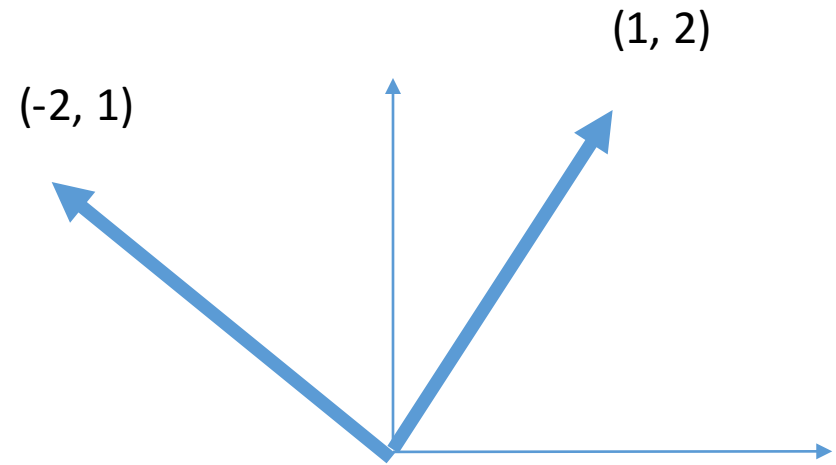


- Можем измерить угол с помощью транспортира: 90 градусов



- Можем измерить расстояние между точками:

$$\sqrt{(1 - (-2))^2 + (2 - 1)^2} = \sqrt{10}$$



# Норма

- Обобщение понятия длины вектора
  - Функция  $\|x\|$  от вектора
  - Если  $\|x\| = 0$ , то  $x = 0$
  - $\|x + y\| \leq \|x\| + \|y\|$
  - $\|\alpha x\| = |\alpha| \|x\|$
- 
- Векторное пространство с нормой — нормированное

# Примеры норм

- Евклидова норма:

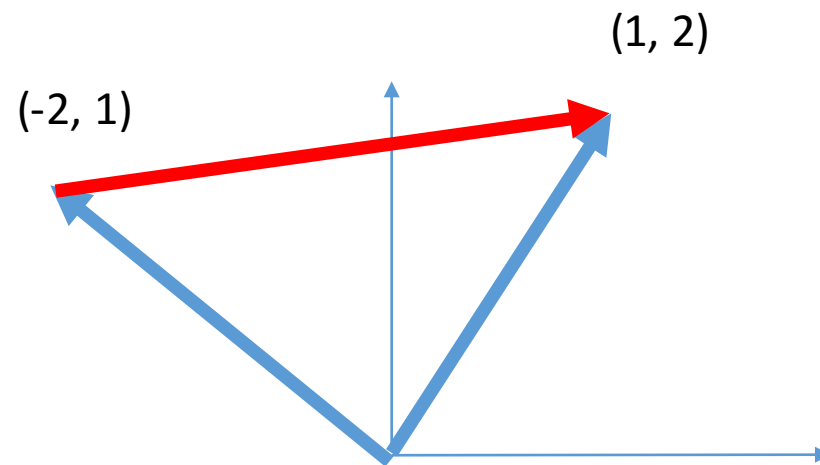
$$\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$$

- Манхэттенская норма:

$$\|x\|_1 = \sum_{i=1}^d |x_i|$$

# Метрика

- Обобщение понятия расстояния
- $\rho(x, y) = \|x - y\|$
- Соответствует геометрическим представлениям
- Векторное пространство с метрикой — метрическое



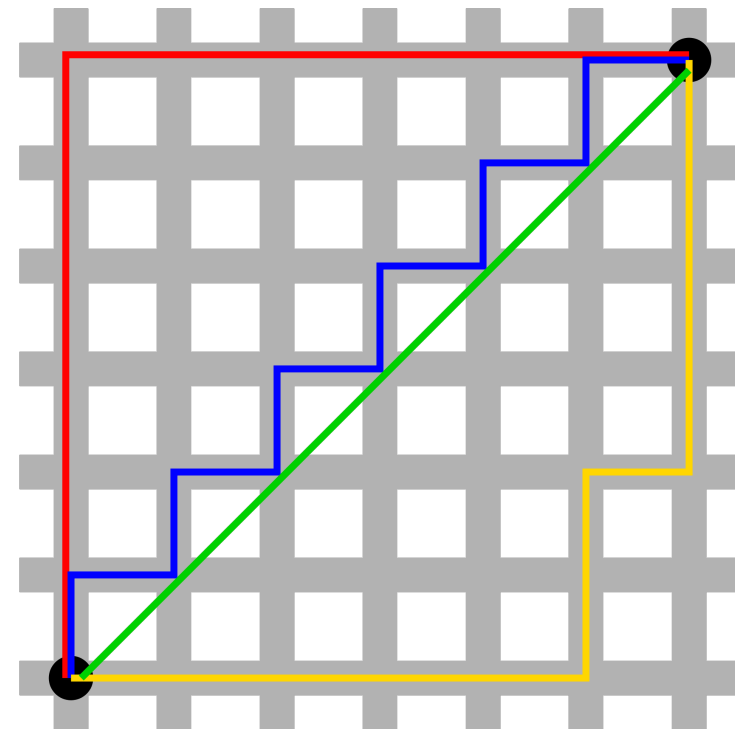
# Примеры метрик

- Евклидова метрика:

$$\rho_2(x, z) = \sqrt{\sum_{i=1}^d (x_i - z_i)^2}$$

- Манхэттенская метрика:

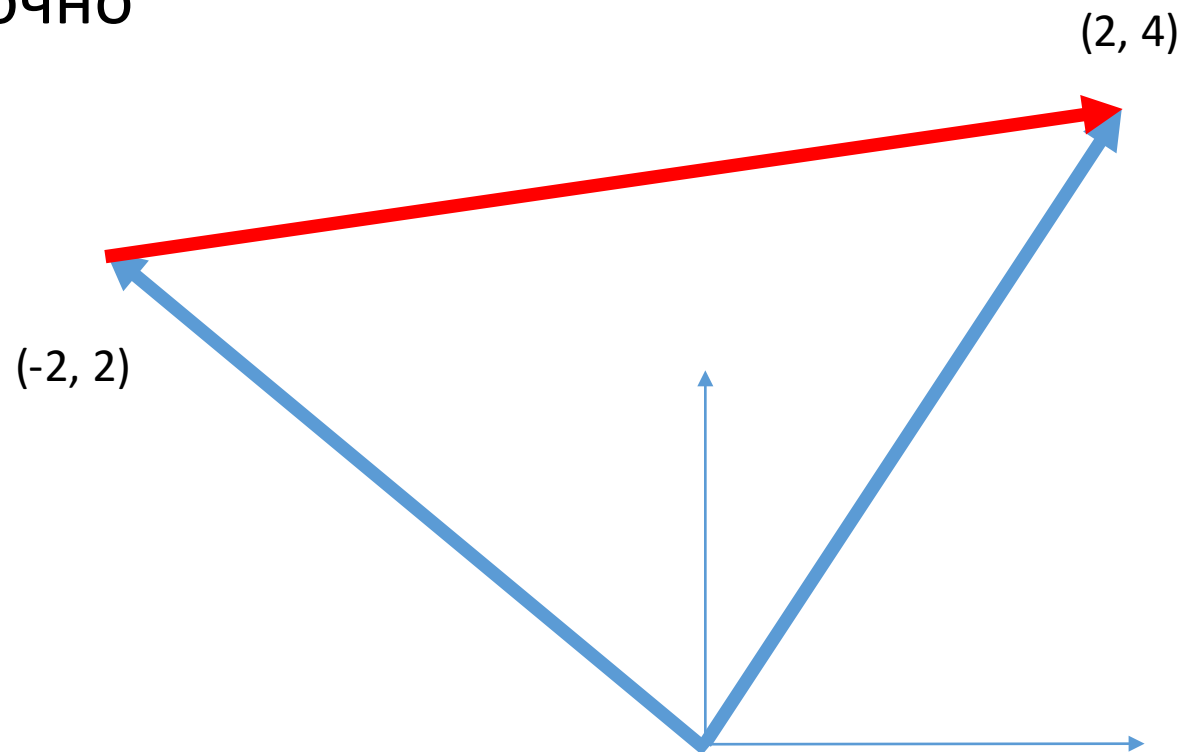
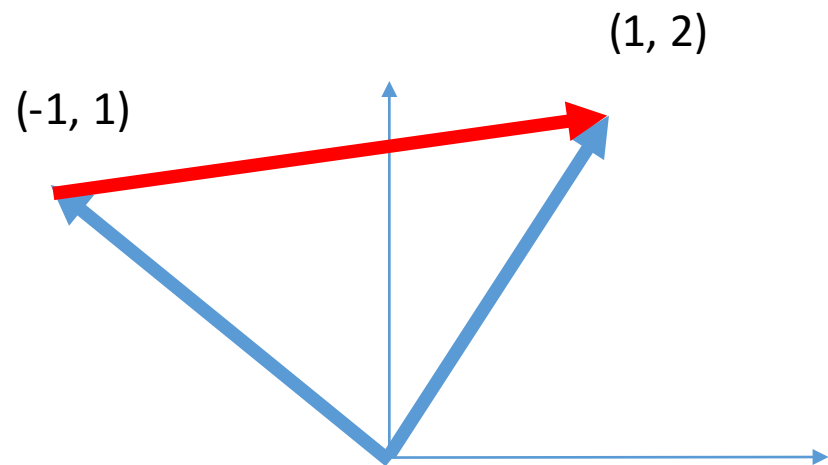
$$\rho_1(x, z) = \sum_{i=1}^d |x_i - z_i|$$





# Как искать углы?

- Нормы и метрики недостаточно



# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

- Норма:  $\|x\|_2 = \sqrt{\langle x, x \rangle}$

# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

- Норма:  $\|x\|_2 = \sqrt{\langle x, x \rangle}$
- Расстояние:  $\rho_2(x, z) = \|x - z\| = \sqrt{\langle x - z, x - z \rangle}$

# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

- Норма:  $\|x\|_2 = \sqrt{\langle x, x \rangle}$
- Расстояние:  $\rho_2(x, z) = \|x - z\| = \sqrt{\langle x - z, x - z \rangle}$
- Угол?

# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

- Важное соотношение:  $\langle x, y \rangle = \|x\| \|y\| \cos \angle(x, y)$

# Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i$$

- Важное соотношение:  $\langle x, y \rangle = \|x\| \|y\| \cos \angle(x, y)$

Косинус угла



# Скалярное произведение

- Косинус угла:  $\cos \angle(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$



# Скалярное произведение

- Косинус угла:  $\cos \angle(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$
- Мера сонаправленности векторов
- Для параллельных векторов  $\cos \angle(x, y) = 1$
- Для перпендикулярных векторов  $\cos \angle(x, y) = 0$

Функционал ошибки для  
классификации

# Ошибка классификации

- Доля **неправильных** ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Нотация Айверсона:
  - [истина] = 1
  - [ложь] = 0

# Ошибка классификации

$a(x)$	$y$
-1	-1
+1	+1
-1	-1
<b>+1</b>	<b>-1</b>
+1	+1

- Доля неправильных ответов:

?

# Ошибка классификации

$a(x)$	$y$
-1	-1
+1	+1
-1	-1
<b>+1</b>	<b>-1</b>
+1	+1

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

# Accuracy

- Доля **правильных** ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- На английском: **accuracy**

# Accuracy

- Доля **правильных** ответов:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) = y_i]$$

- На английском: **accuracy**
- ВАЖНО: не переводите это как «точность»!

Оценивание обобщающей  
способности



# Как оценить качество?

- Как алгоритм будет вести себя на новых данных?
- Какая у него будет доля ошибок?
- ...или другая метрика качества
- По обучающей выборке нельзя это оценить

# Отложенная выборка

- Разбиваем выборку на две части
  - Обучающая выборка
  - Отложенная выборка
- На первой обучаем алгоритм
- На второй измеряем качество



# Пропорции разбиения

- Маленькая отложенная часть
  - (+) Обучающая выборка репрезентативная
  - (-) Оценка качества ненадежная
- Большая отложенная часть
  - (+) Оценка качества надежная
  - (-) Оценка качества смещенная
- Обычно: 70/30, 80/20, 0.632/0.368

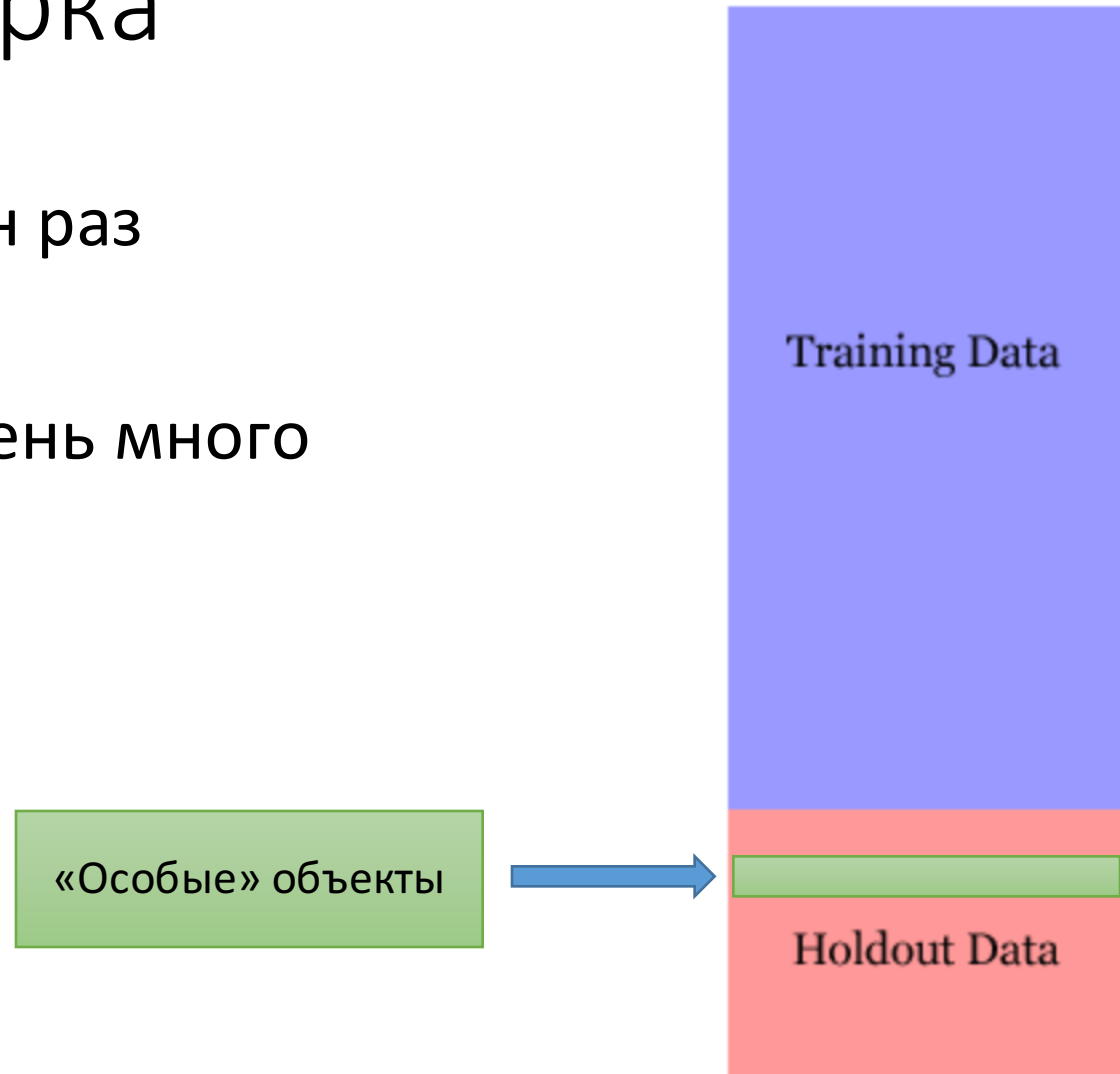
# Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



# Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



# Много отложенных выборок

- Улучшение: разбиваем выборку на две части  $n$  раз
- Усредняем оценку качества



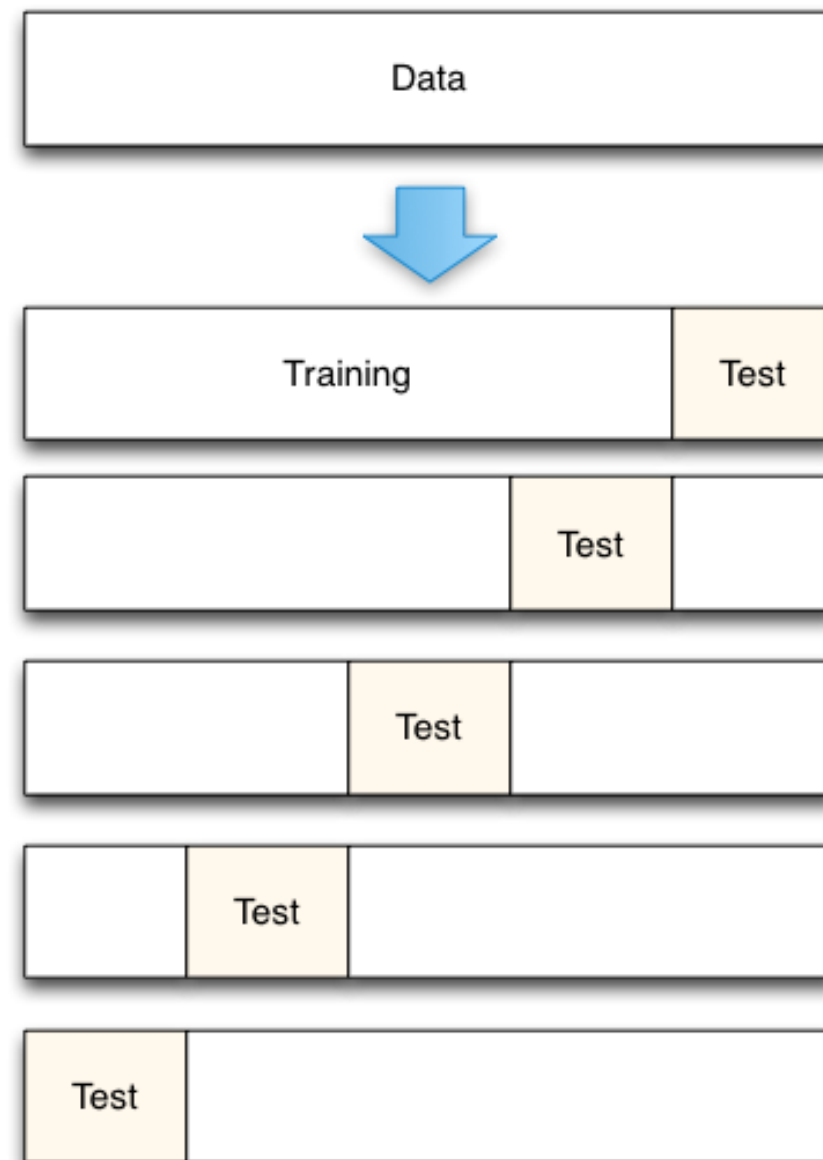
# Много отложенных выборок

- Нет гарантий, что каждый объект побывает в обучении



# Кросс-валидация

- Разбиваем выборку на  $k$  блоков
- Каждая по очереди выступает как тестовая





# Число блоков

- Мало блоков
  - Тестовая выборка всегда большая — (+) надежные оценки
  - Обучение маленькое — (-) смещенные оценки
- Много блоков
  - (-) Ненадежные оценки
  - (+) Несмещенные оценки

# Число блоков

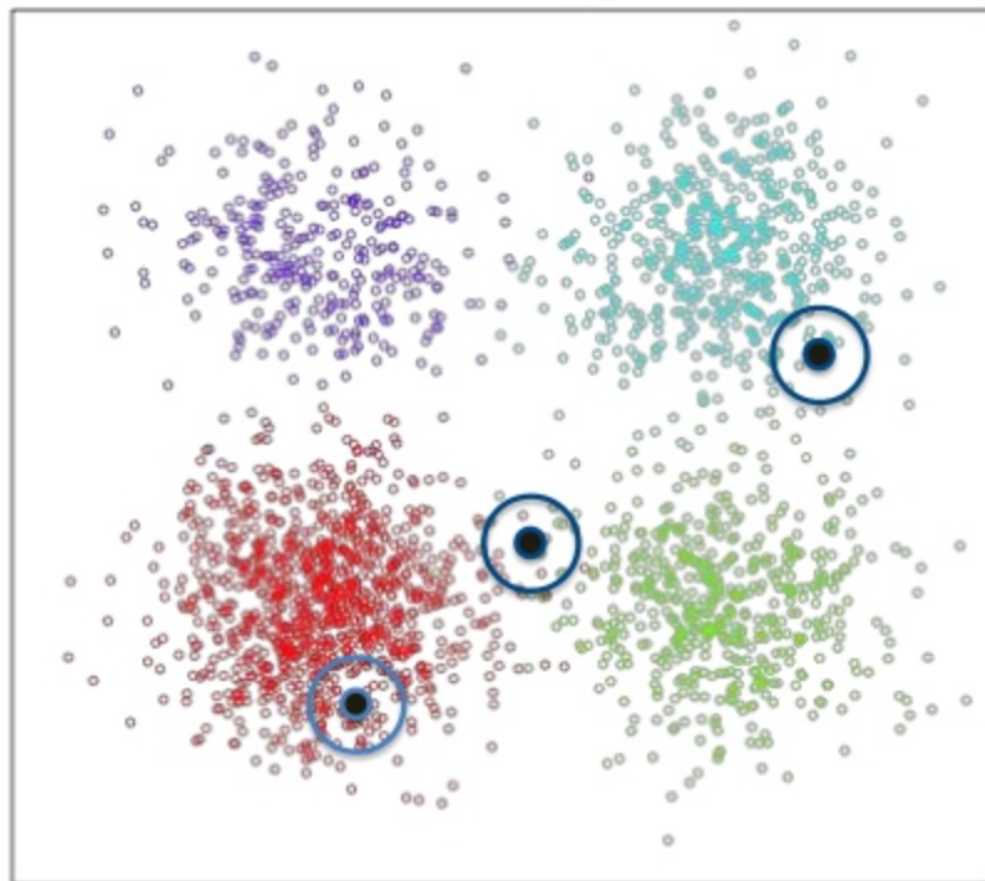
- Обычно:  $k = 3, 5, 10$
- Чем больше выборка, тем меньше нужно  $k$
- Чем больше  $k$ , тем больше раз надо обучать алгоритм

# Совет

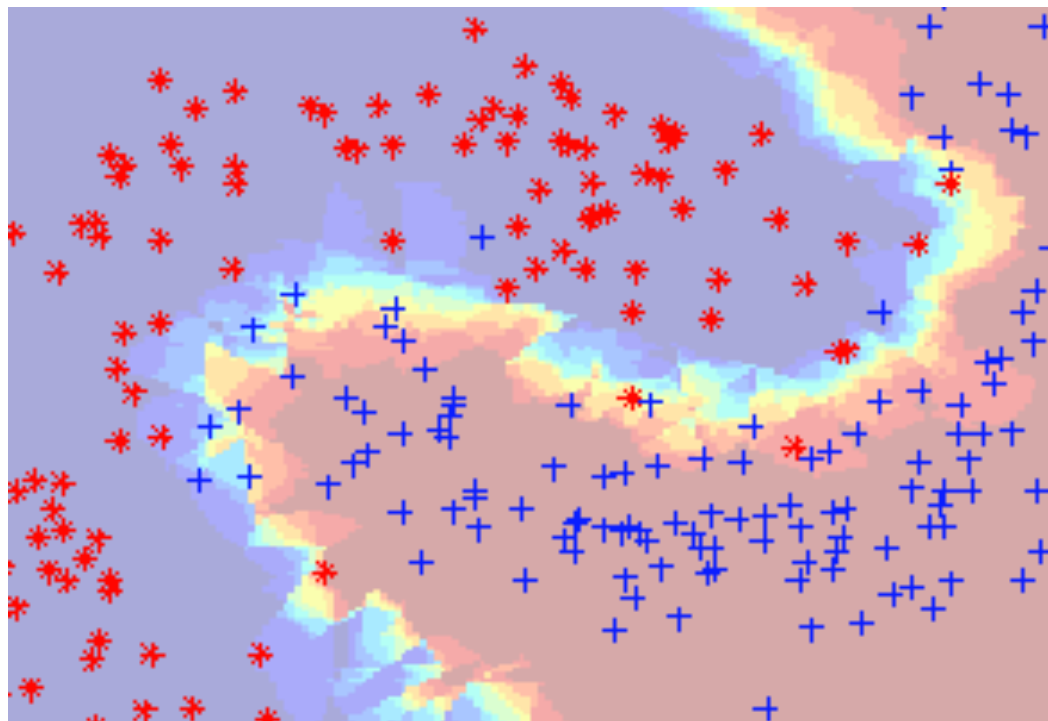
- Перемешивайте выборку!
- Объекты могут быть отсортированы
- При разбиении в обучении могут оказаться только мальчики, в контроле — только девочки

Гипотеза компактности

# Гипотеза компактности



# Гипотеза компактности



# Гипотеза компактности



# Гипотеза компактности

- Для классификации: близкие объекты, как правило, лежат в одном классе
- Для регрессии: близким объектам соответствуют близкие ответы
- Что такое «близкие объекты»?



# Измерение сходства

- Необходимо ввести расстояние между объектами
- $\rho(x, z)$  — функция расстояния (не обязательно метрика)
- Типичный пример: евклидова метрика

$$\rho(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$

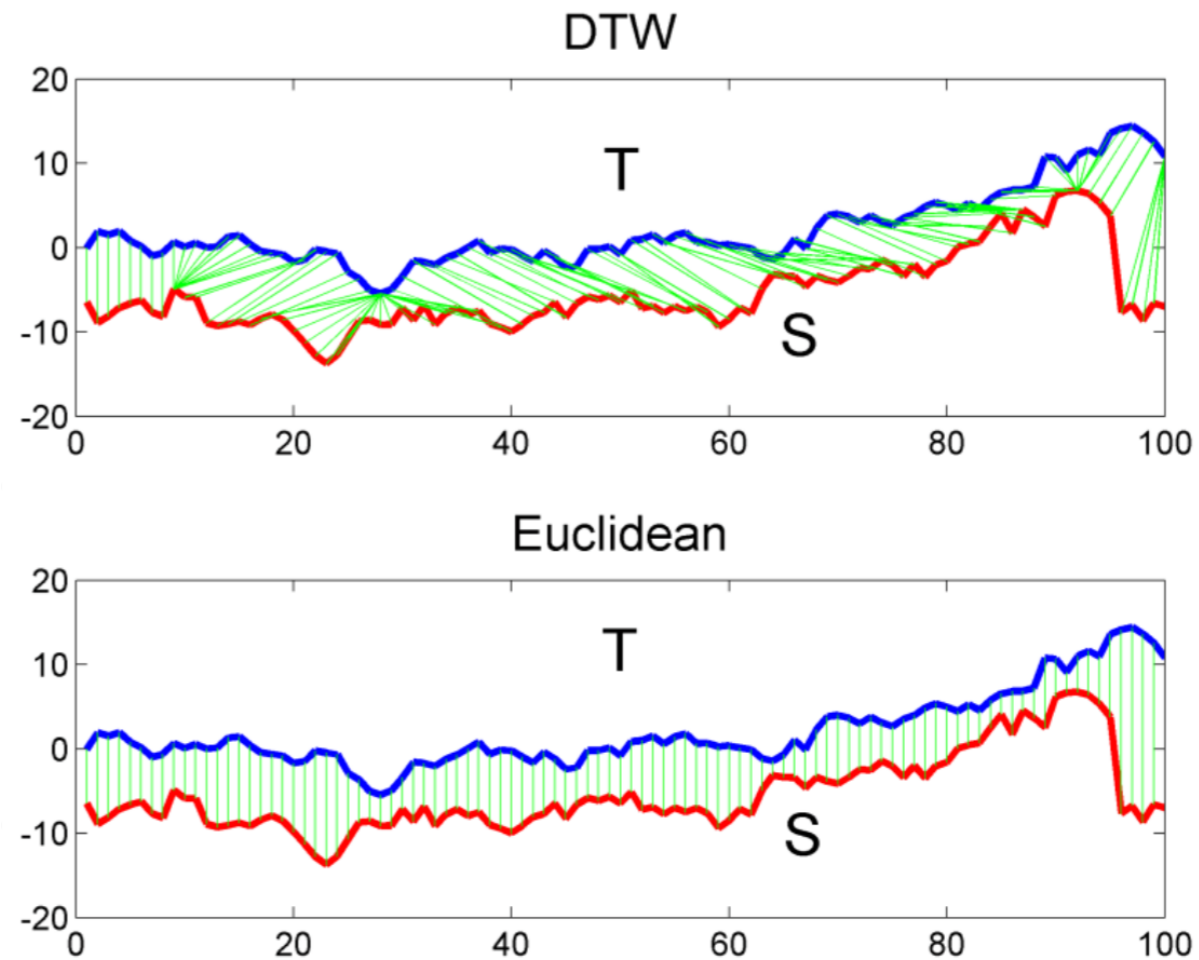
# Расстояния на текстах

- Расстояние Левенштейна
- Количество вставок и удалений символов, необходимое для преобразования одной строки в другую

СТGGGCTAAAAGGTCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC  
СТGGGACTAAA....CCTTAGCCTATTACAAAAATGGGCCATTAGG...TTGC

# Расстояния на временных рядах

- Суммарное евклидово расстояние
- Dynamic time warping
- И другие



# Метрические методы классификации

# Метод k ближайших соседей

- k nearest neighbors (kNN)
- Задача классификации
- Дано: выборка  $X = (x_i, y_i)_{i=1}^{\ell}$
- Этап обучения: запоминаем выборку  $X$

# Метод k ближайших соседей

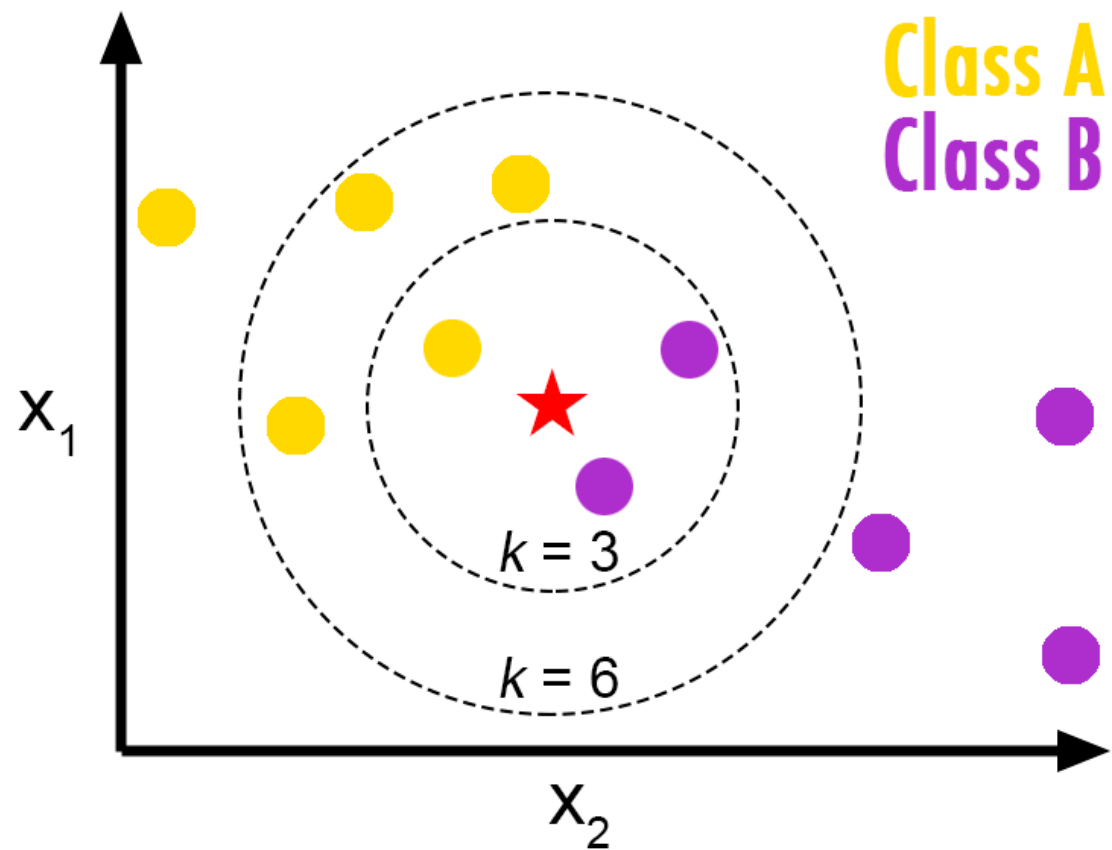
- Новый объект  $x$
- Сортируем объекты обучающей выборки по расстоянию до  $x$ :

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

- Выбираем класс, наиболее популярный среди  $k$  ближайших соседей:

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

# Метод k ближайших соседей



# Метод k ближайших соседей

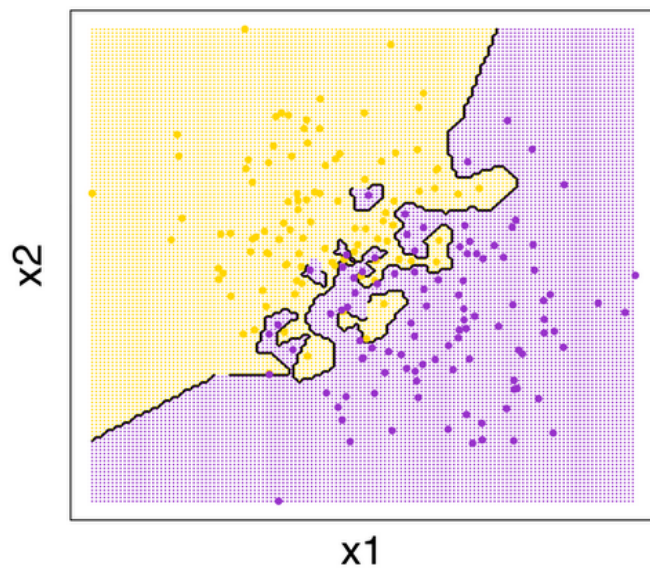
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

- $k$  — гиперпараметр алгоритма
- Подбирается с помощью holdout-выборки или кросс-валидации
- Чем больше  $k$ , тем проще разделяющая поверхность

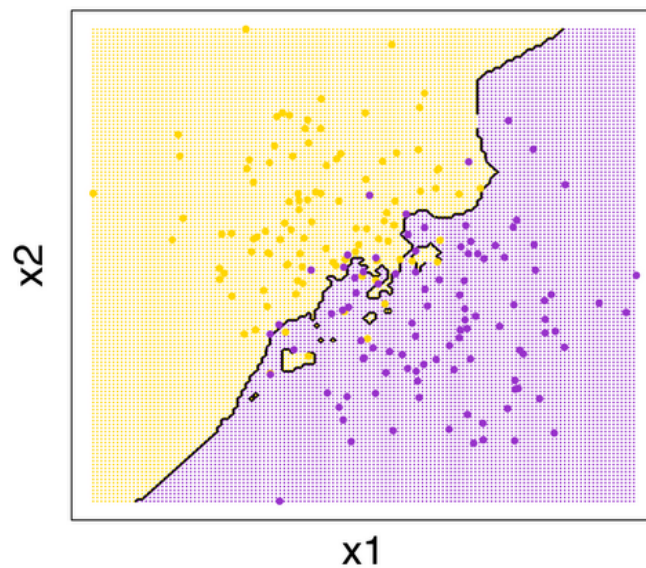


# Выбор числа соседей

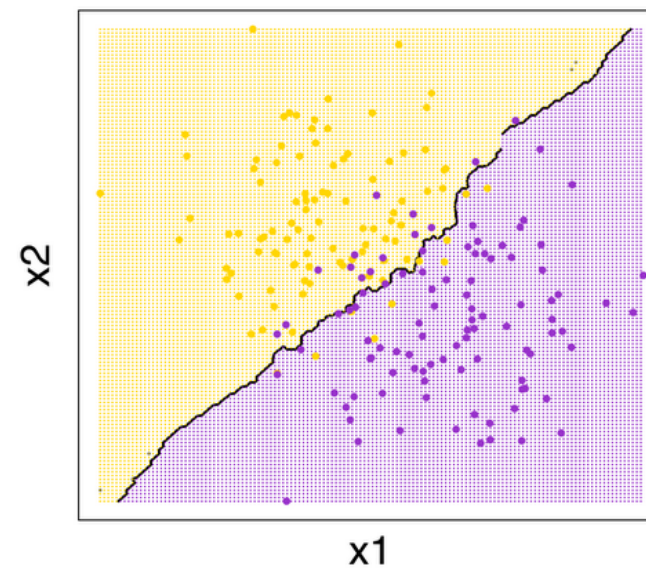
Binary kNN Classification (k=1)



Binary kNN Classification (k=5)

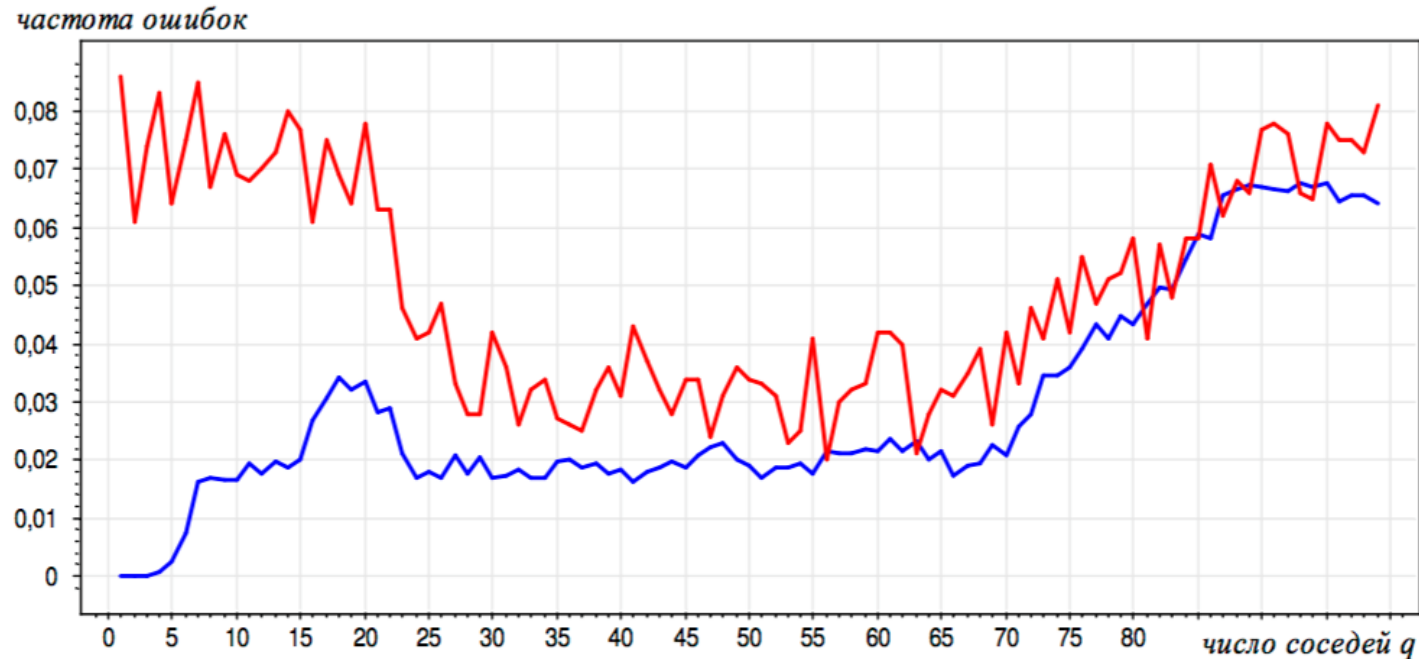


Binary kNN Classification (k=25)

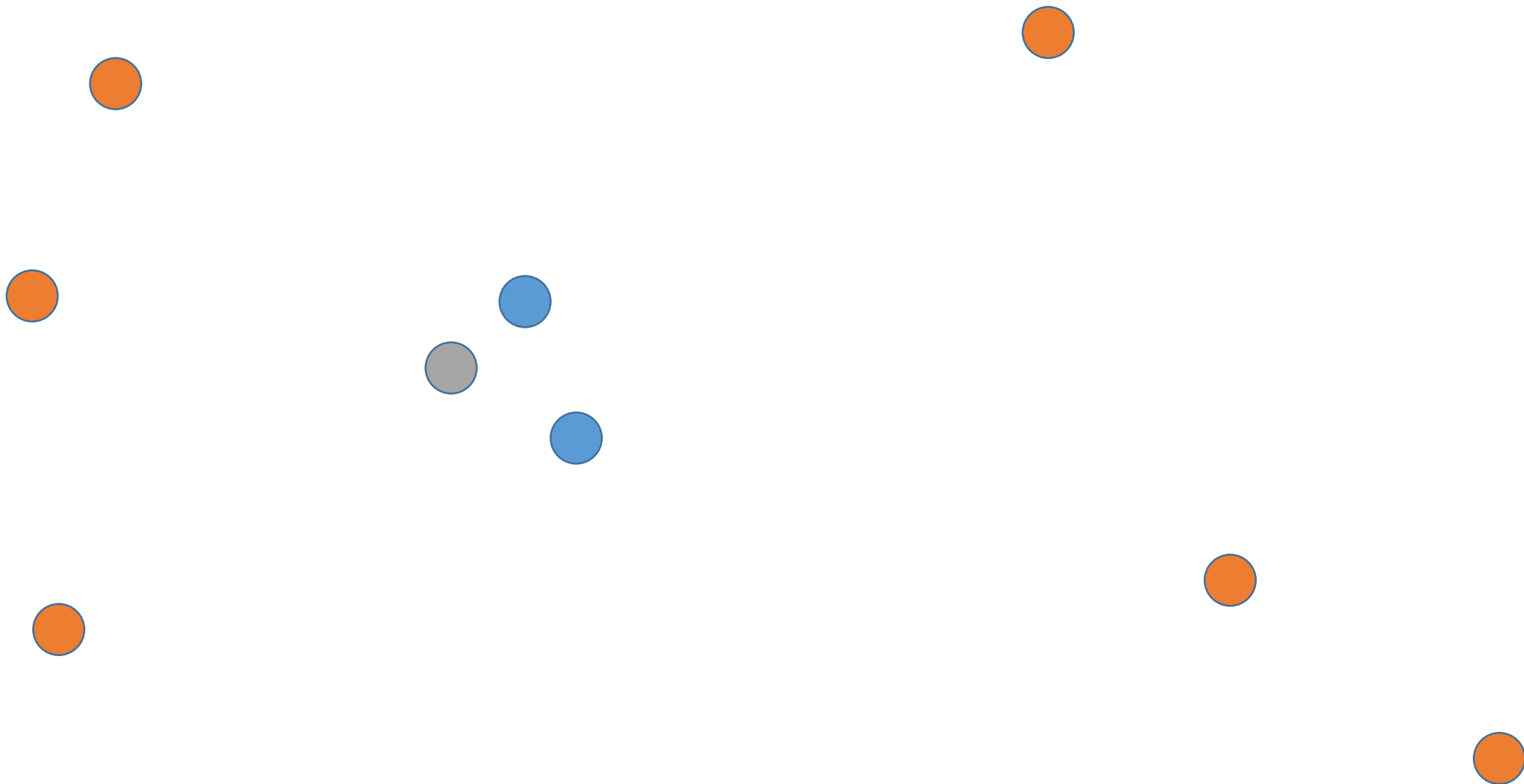


# Выбор числа соседей

- Синий — ошибка на обучении
- Красный — ошибка на кросс-валидации



# Проблема kNN



# Проблема kNN

- Никак не учитываются расстояния до  $k$  ближайших соседей
- Более близкие соседи должны быть важнее

# kNN с весами

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

Варианты:

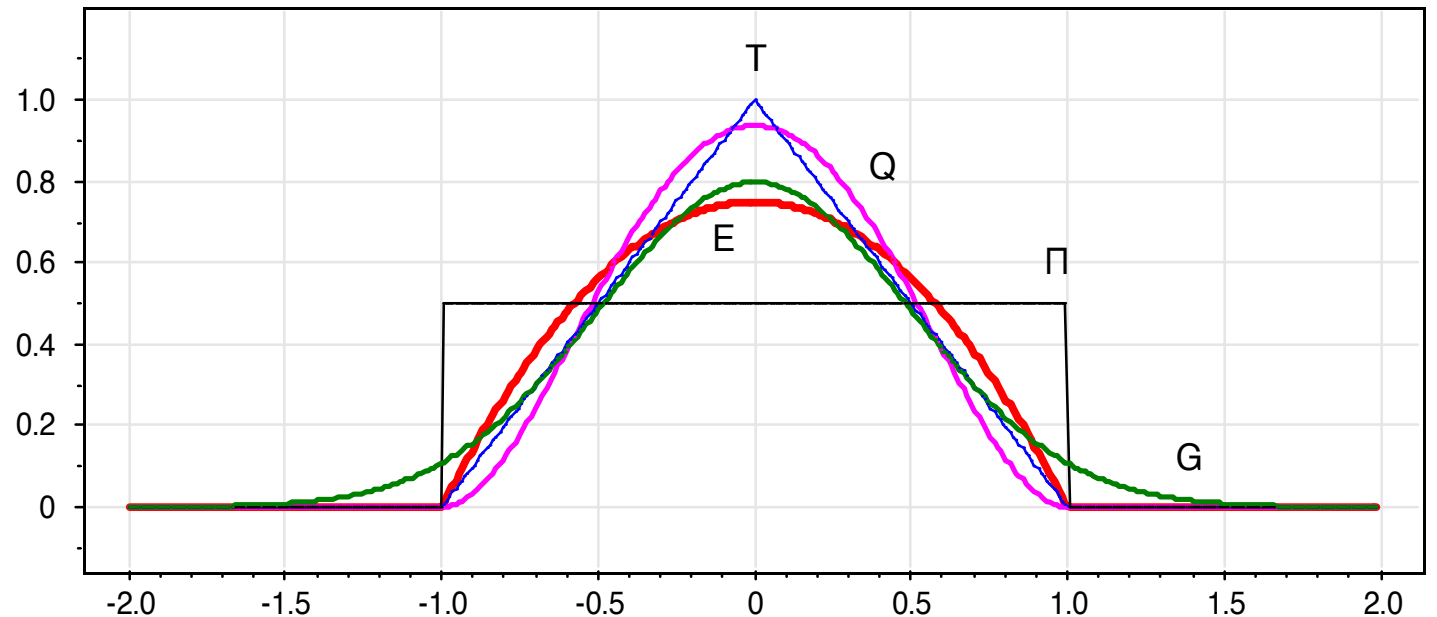
- $w_i = \frac{k+1-i}{k}$
- $w_i = q^i$
- Не учитывают сами расстояния

# kNN с весами

$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

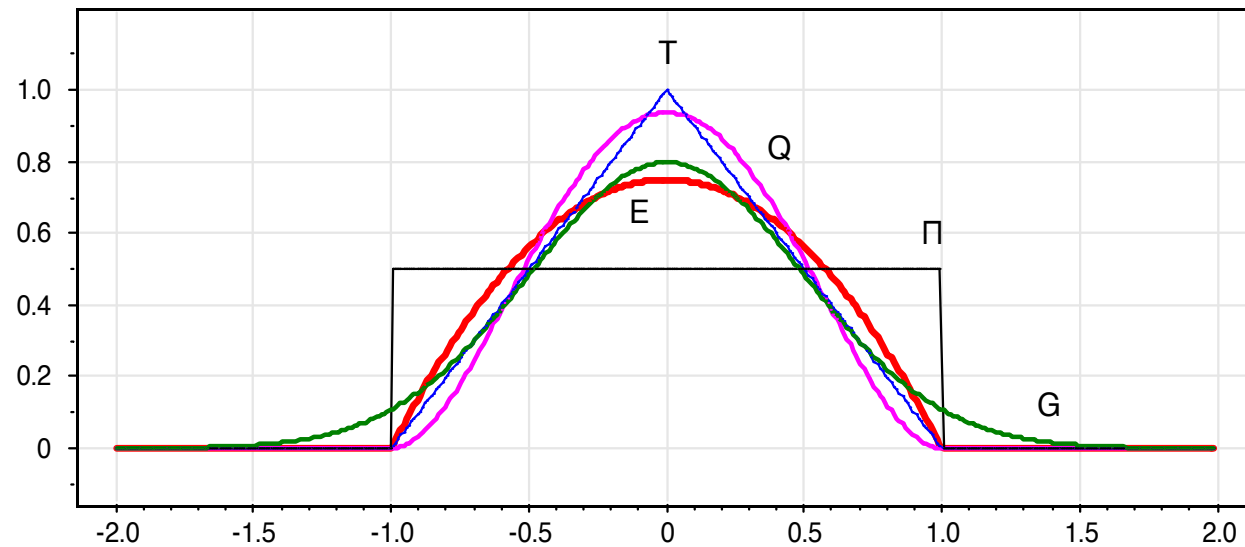
Парзеновское окно:

- $w_i = K \left( \frac{\rho(x, x_{(i)})}{h} \right)$
- $K$  — ядро
- $h$  — ширина окна

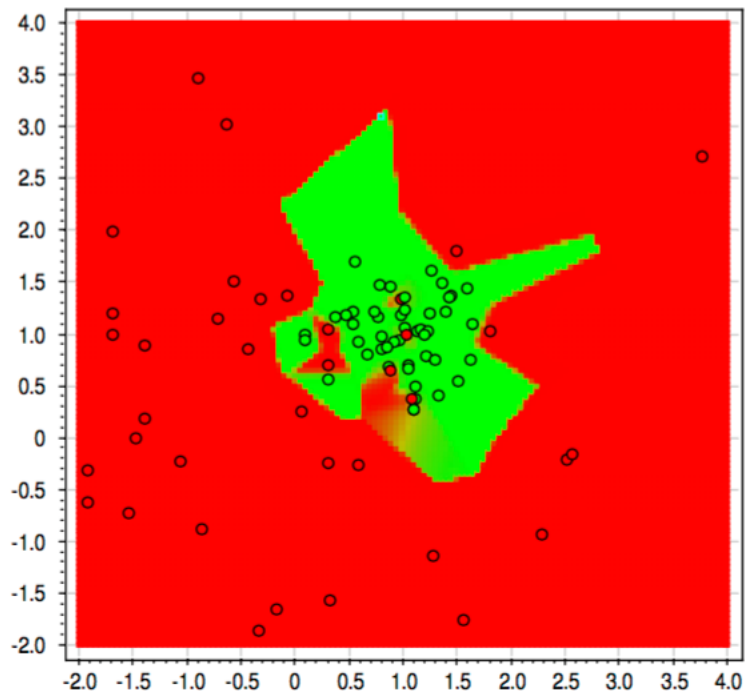


# Ядра

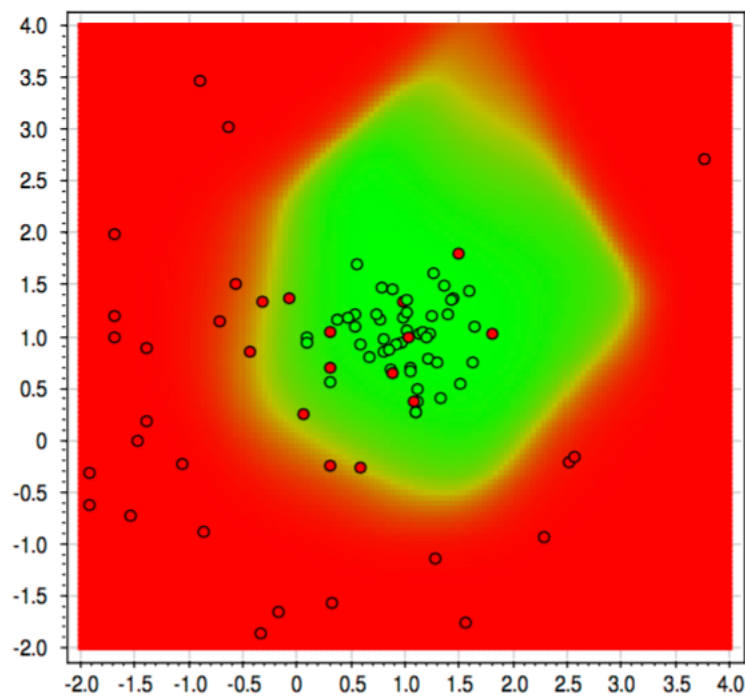
- Гауссовское ядро:  $K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2}z^2\right)$
- И много других



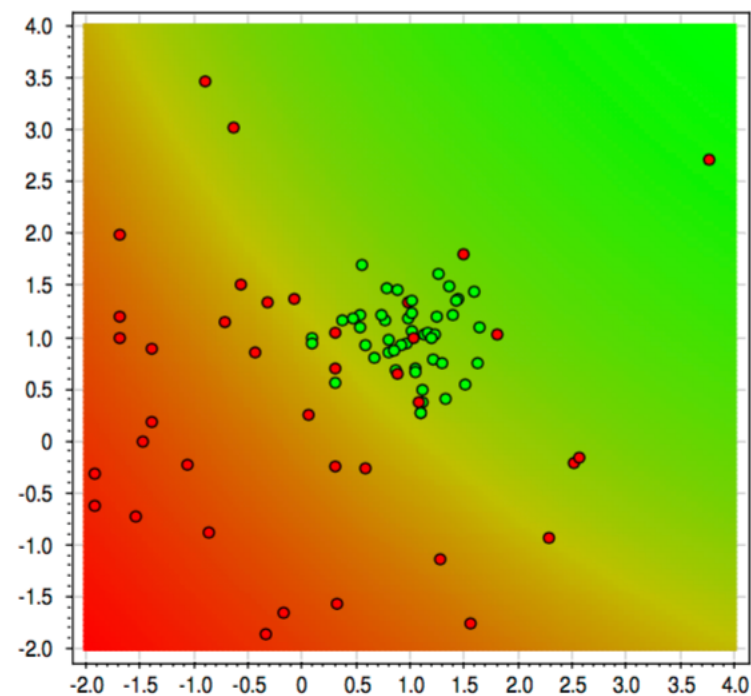
# Ядра



$$h = 0.05$$



$$h = 0.5$$



$$h = 5$$



# Особенности kNN

- Обучение как таковое отсутствует — нужно лишь запомнить обучающую выборку
- Для применения модели необходимо вычислить расстояния от нового объекта до всех обучающих объектов
- Применение требует  $\ell d$  операций
- Существуют специальные методы для поиска ближайших соседей