Введение в анализ данных

Лекция 14

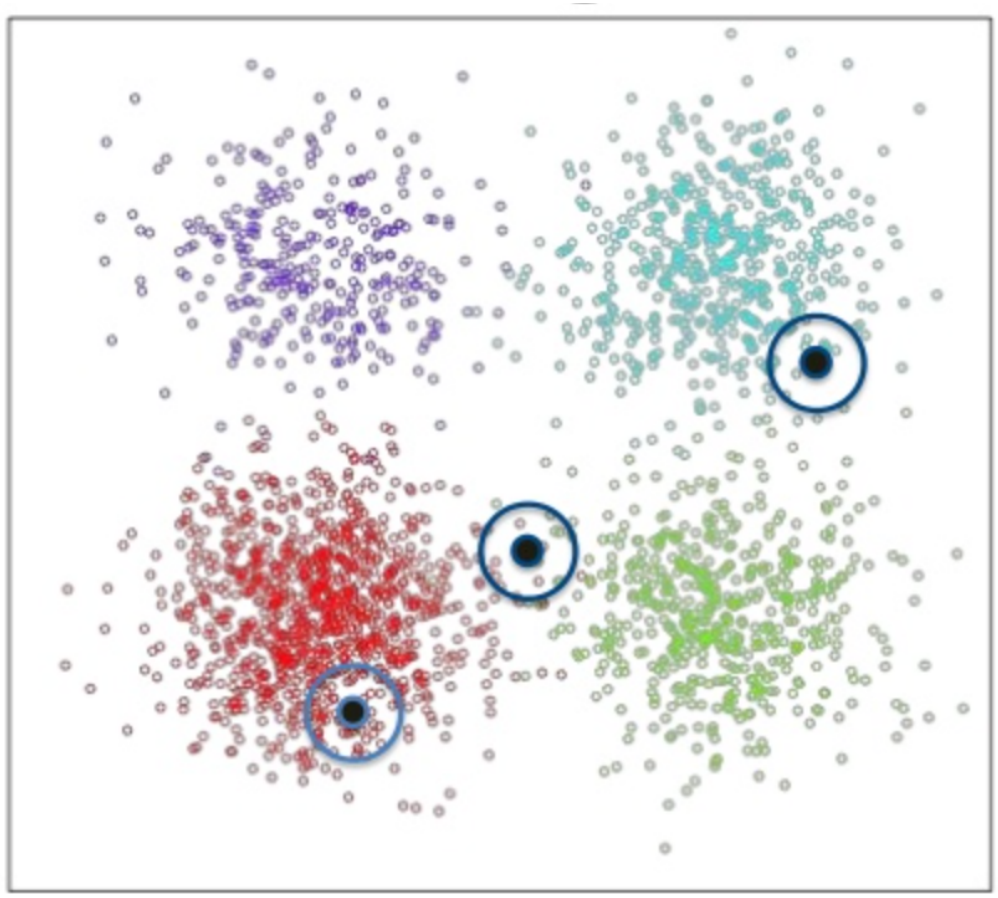
Метрические методы

Евгений Соколов

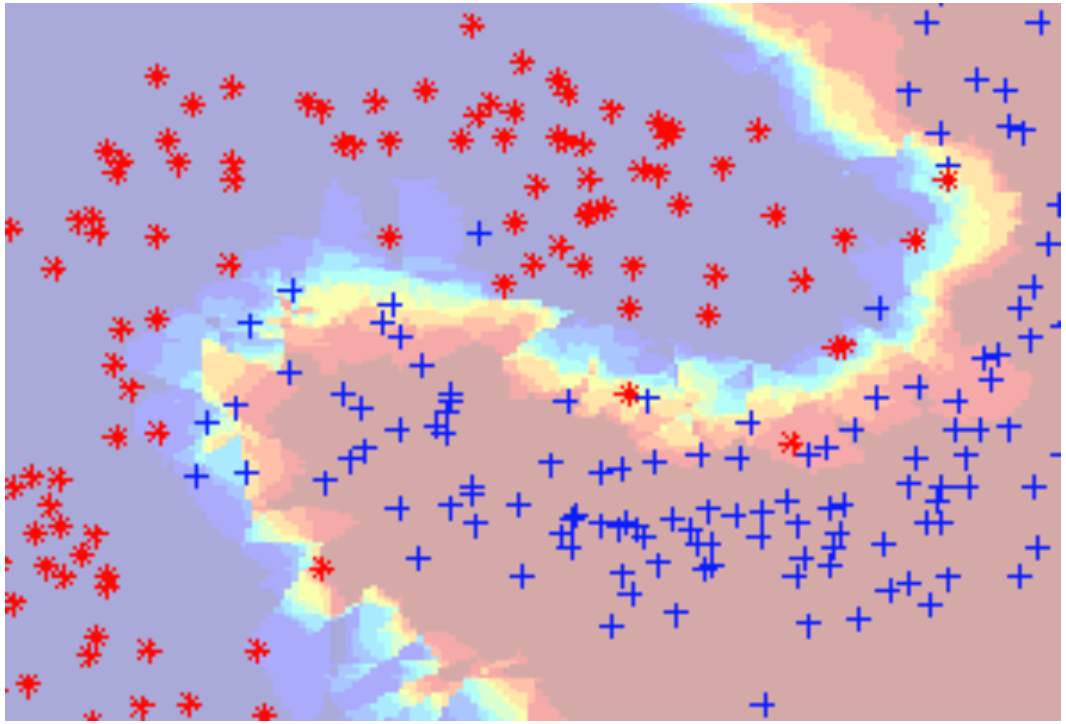
sokolov.evg@gmail.com

НИУ ВШЭ, 2016

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

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

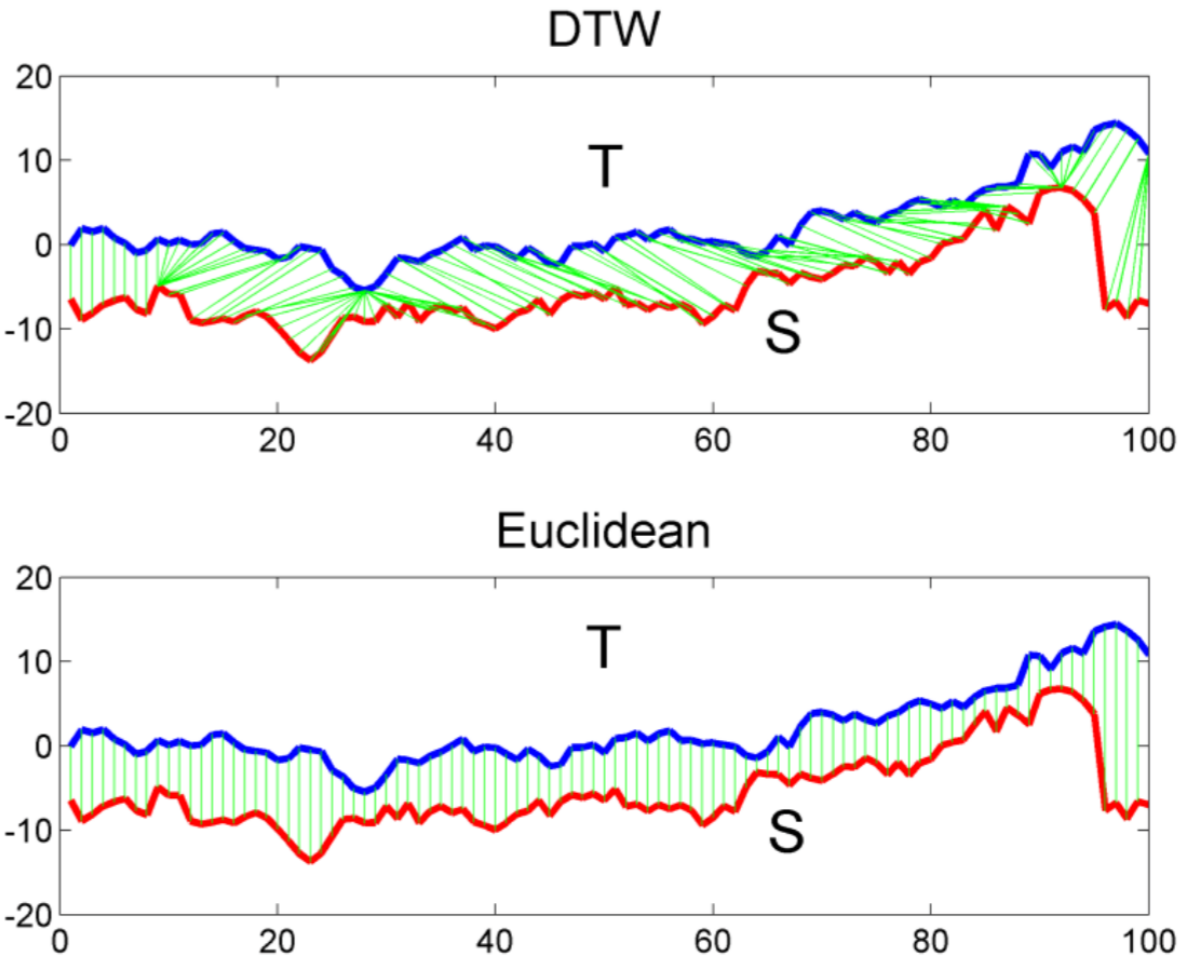
Расстояния на текстах

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

```
CTGGGCTAAAAGGTCCCTTAGCC..TTTAGAAAAA.GGGCCATTAGGAAATTGC
CTGGGACTAAA....CCTTAGCCTATTTACAAAAATGGGCCATTAGG...TTGC
```

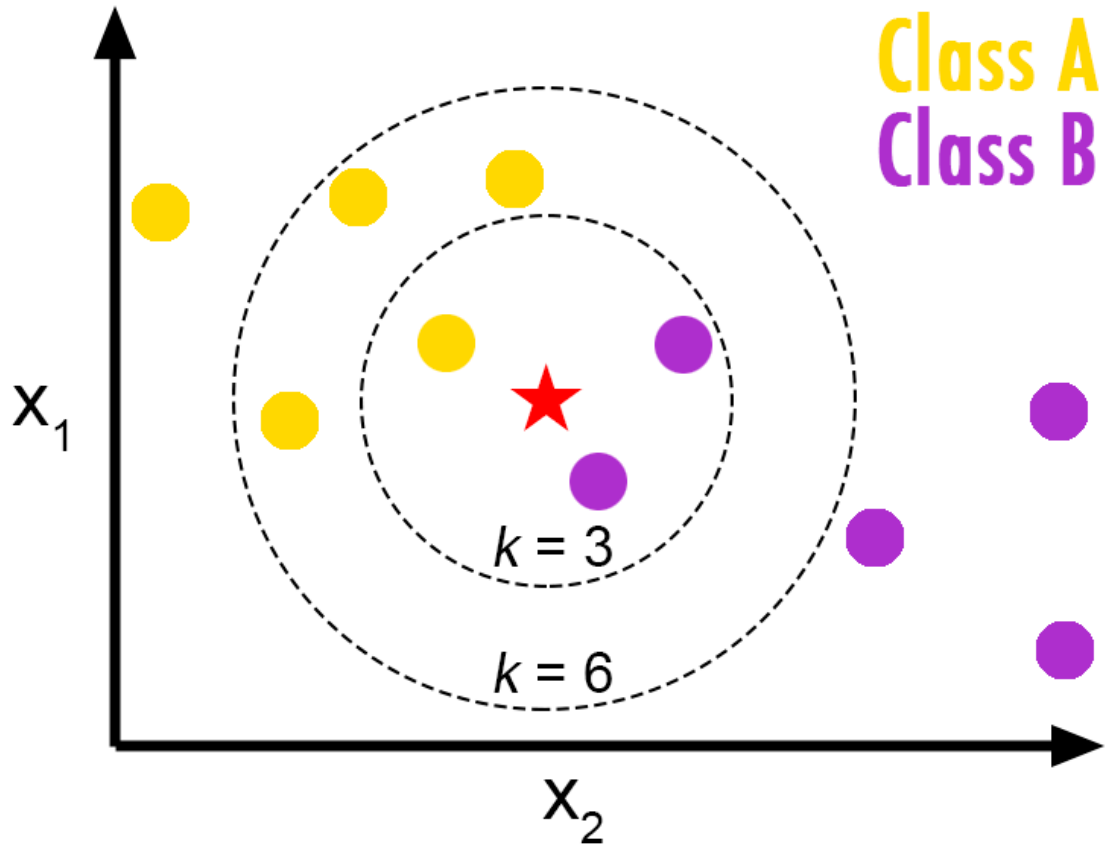
Расстояния на временных рядах

- Суммарное евклидово расстояние
- Dynamic time warping
- И другие



Метрические методы классификации

Метод k ближайших соседей



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:

$$\rho\left(x, x_{(1)}\right) \leq \cdots \leq \rho\left(x, x_{(\ell)}\right)$$

- Выбираем класс, наиболее популярный среди k ближайших соседей:

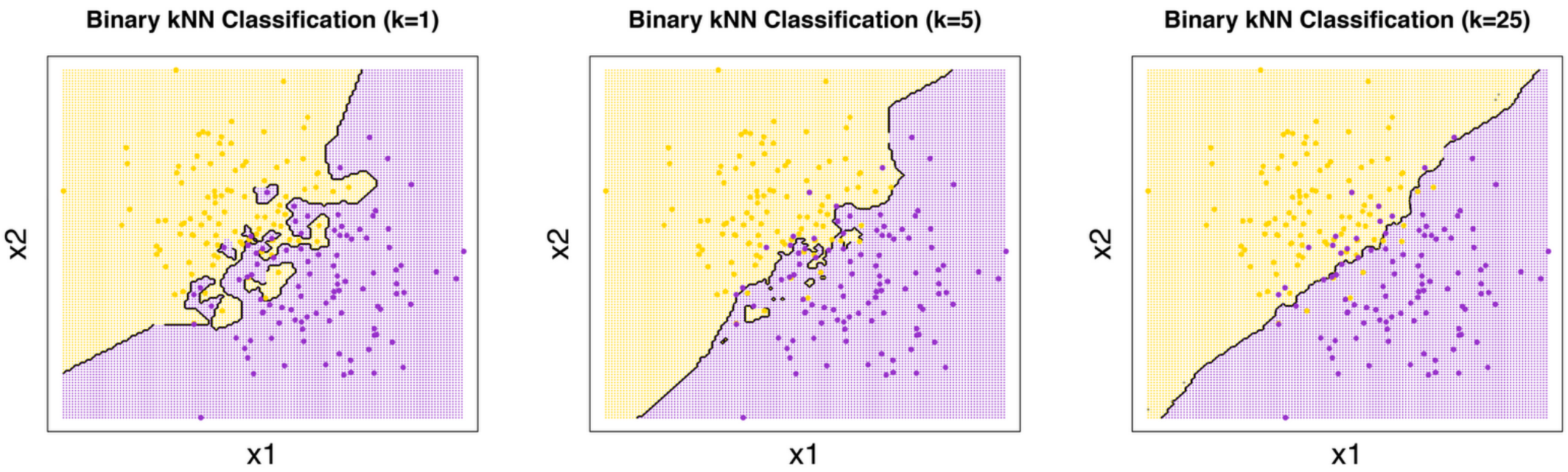
$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{\boldsymbol{k}} [y_{(i)} = y]$$

Метод k ближайших соседей

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

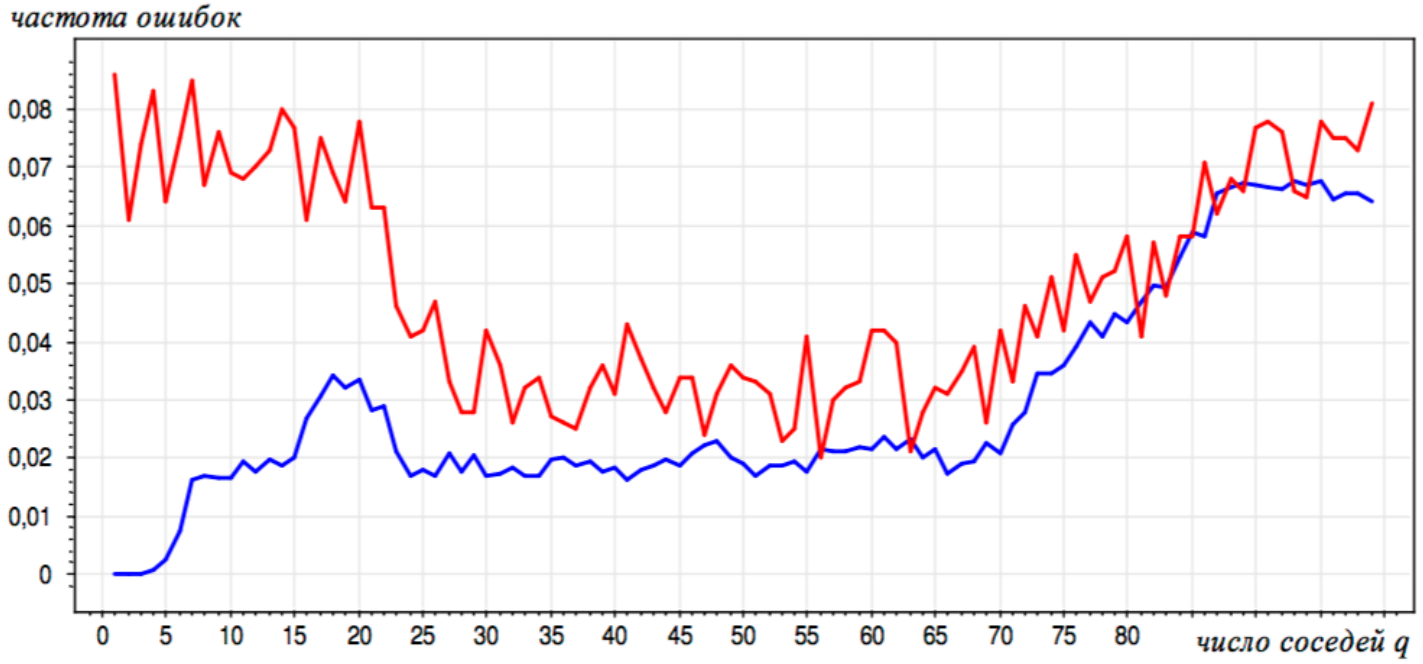
- k — гиперпараметр алгоритма
- Подбирается с помощью holdout-выборки или кросс-валидации
- Чем больше k, тем проще разделяющая поверхность

Выбор числа соседей

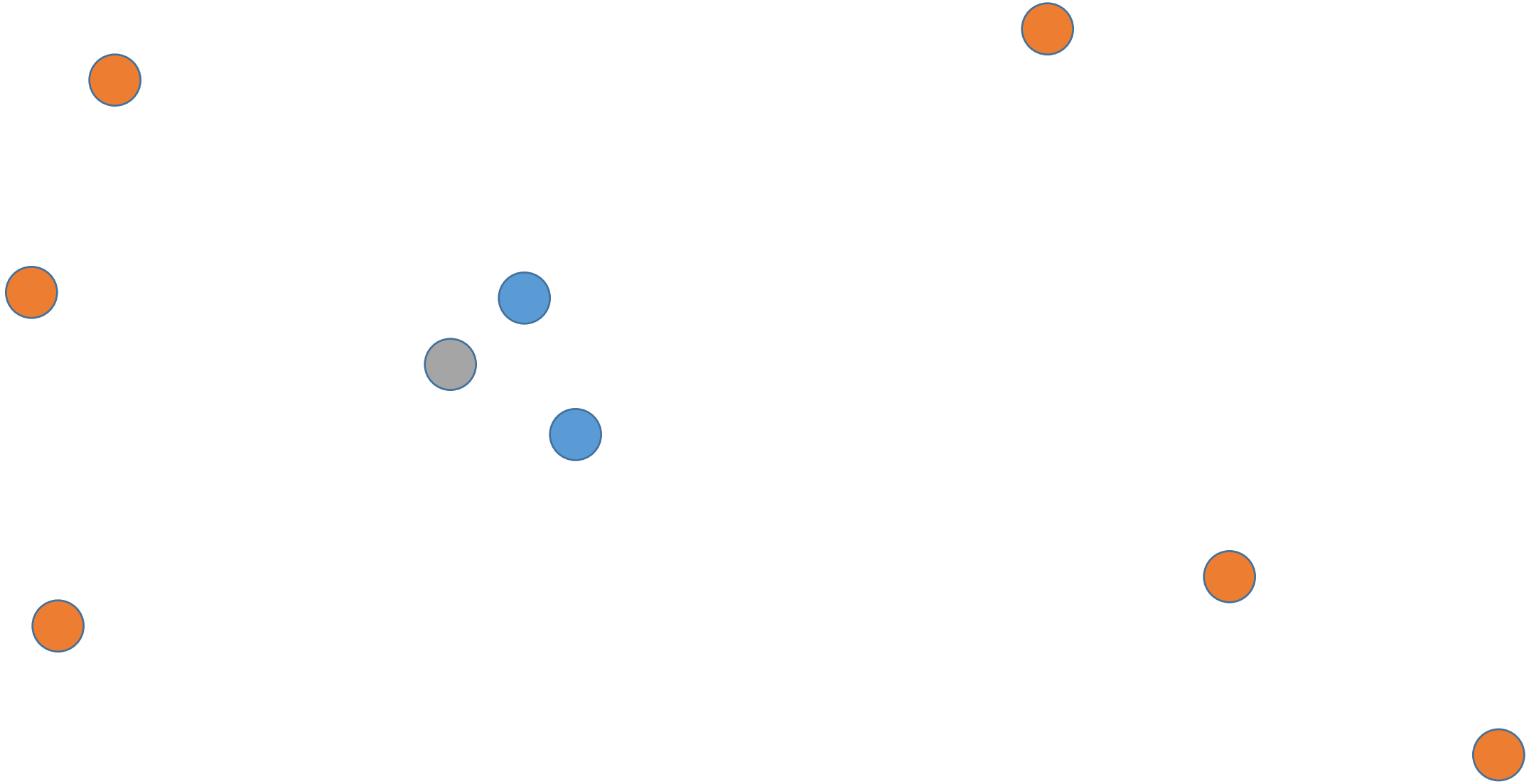


Выбор числа соседей

- 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:

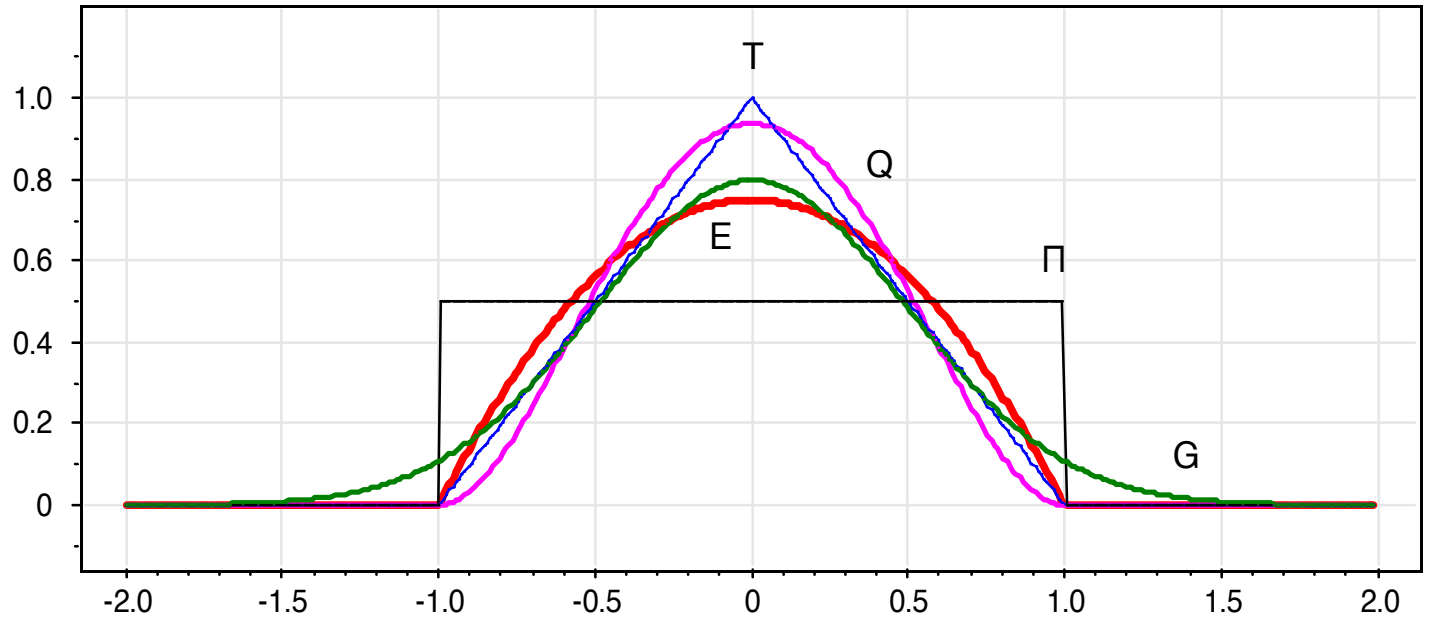
- $w_i = \frac{k+1-i}{k}$
- $w_i = q^i$
- Не учитывают сами расстояния

kNN с весами

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

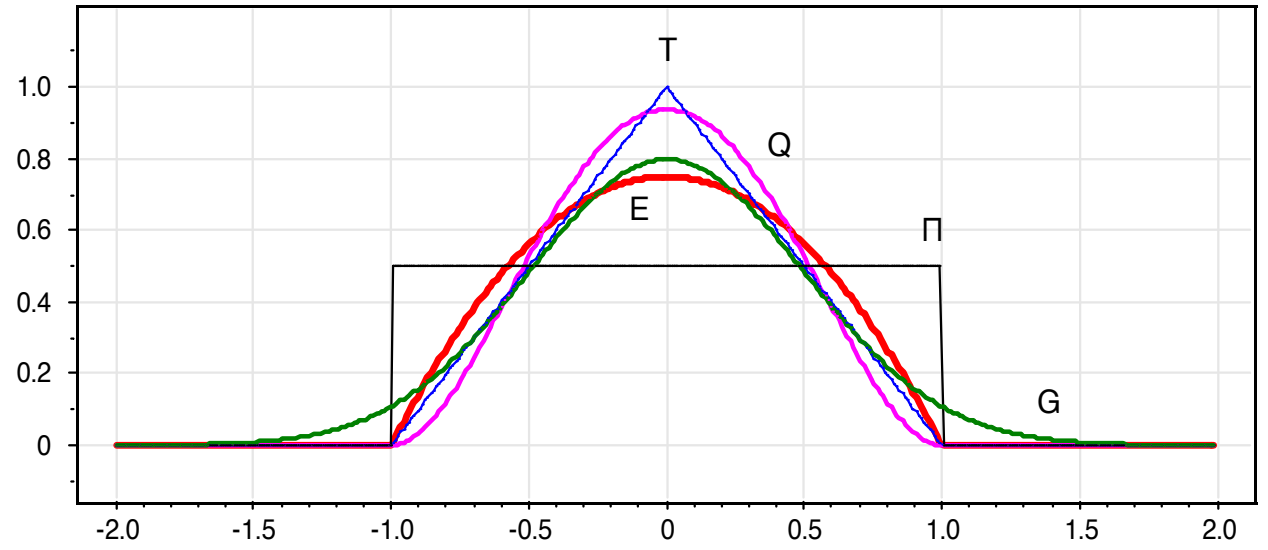
Парзеновское окно:

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

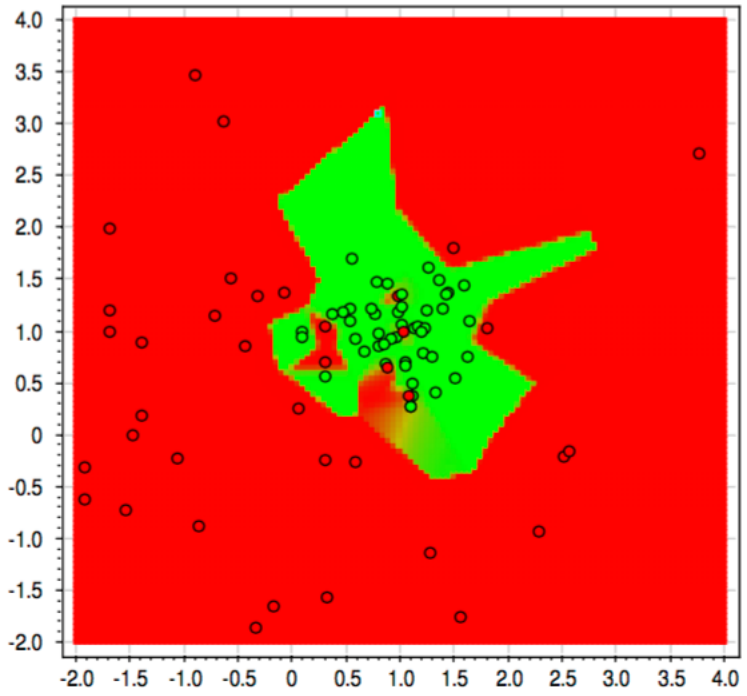


Ядра

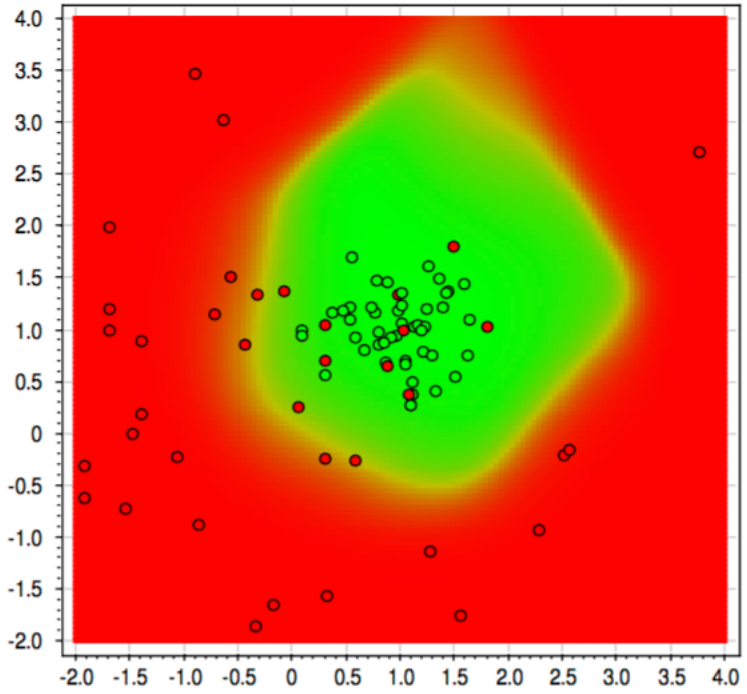
- Гауссовское ядро: $K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2}z\right)$
- И много других



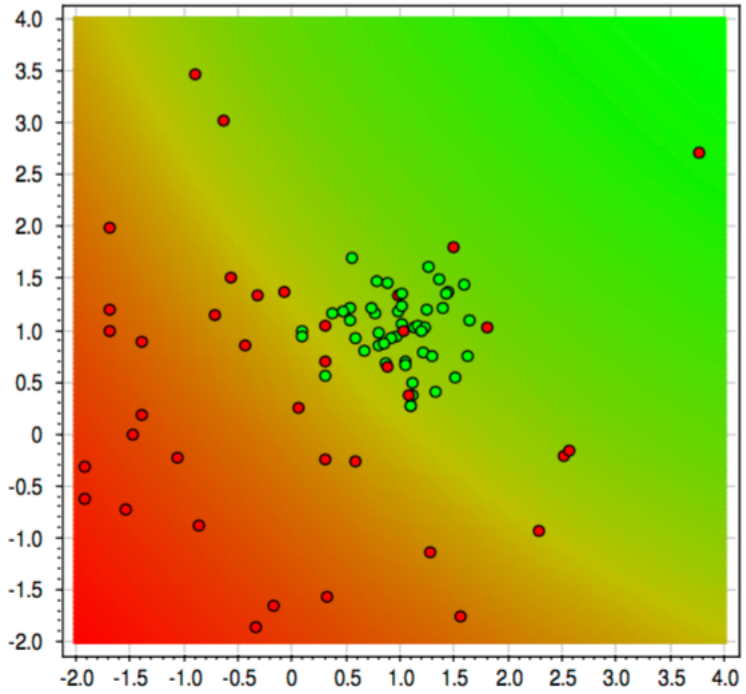
Ядра



$h = 0.05$



$h = 0.5$



$h = 5$

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 ℓd операций
- Существуют специальные методы для поиска ближайших соседей

Метрические методы регрессии

kNN для регрессии

- Классификация:

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

- Регрессия:

kNN для регрессии

- Классификация:

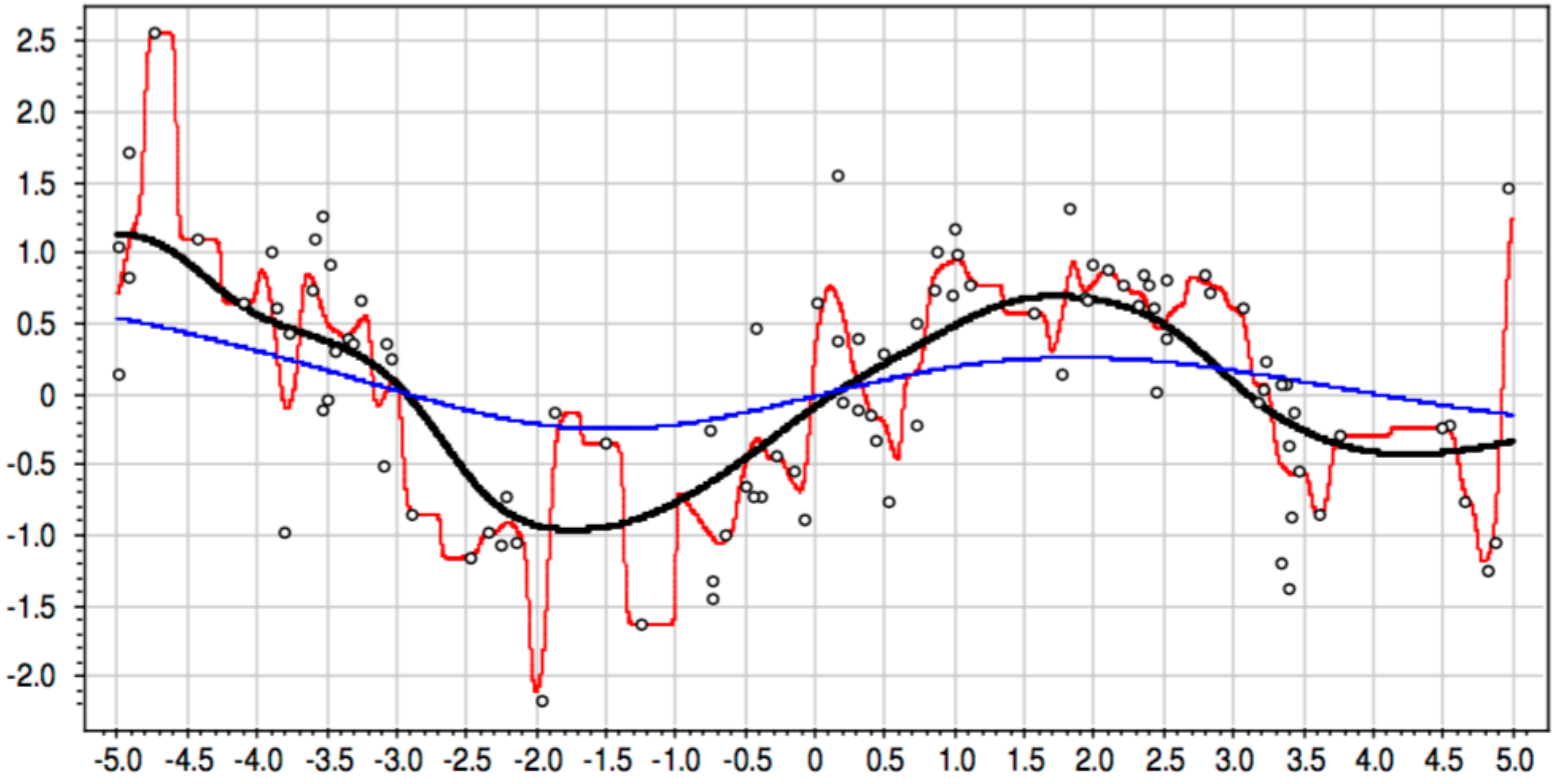
$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i [y_{(i)} = y]$$

- Регрессия:

$$a(x) = \frac{\sum_{i=1}^{k} w_i y_{(i)}}{\sum_{i=1}^{k} w_i}$$

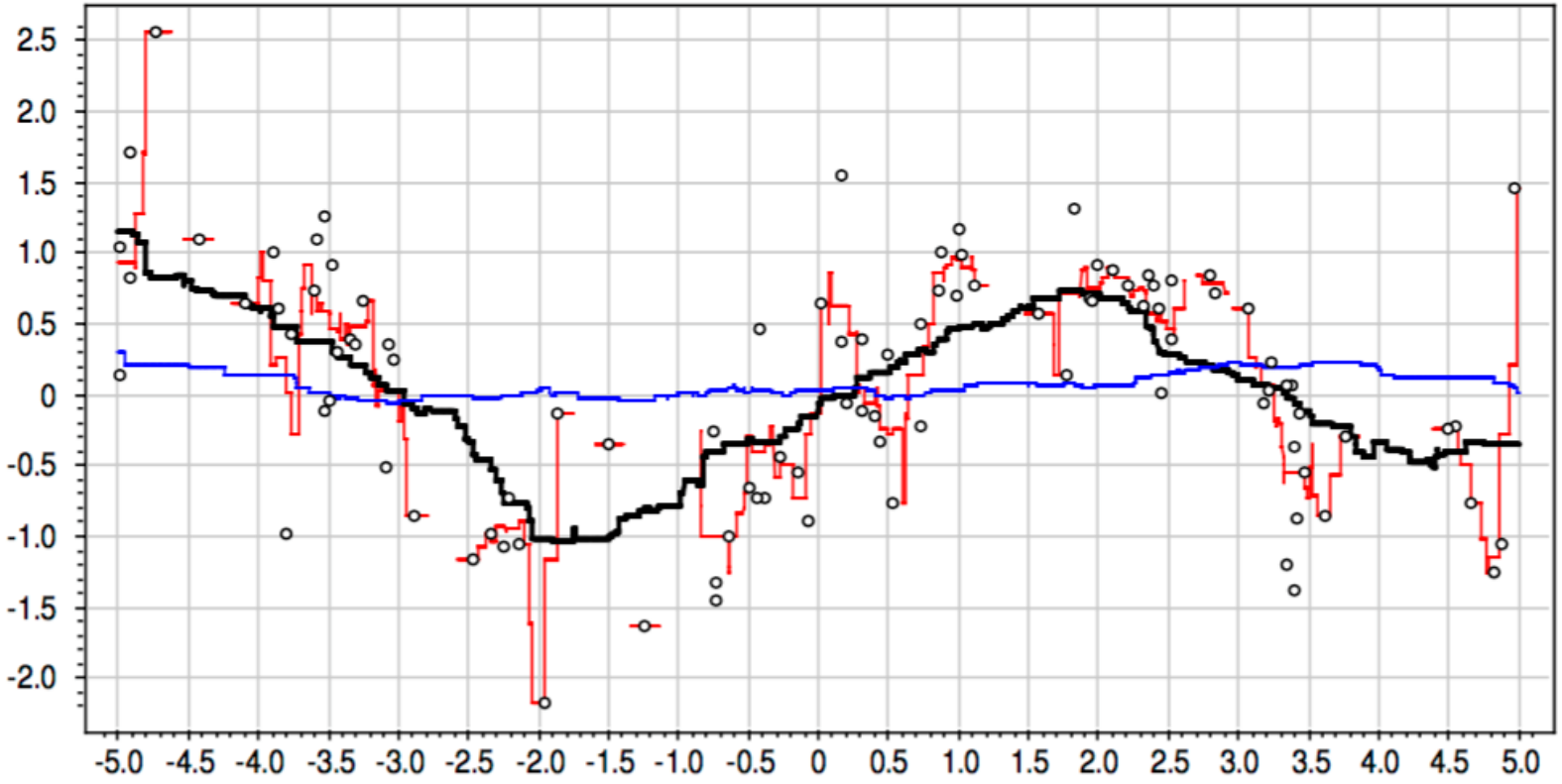
kNN для регрессии

- Гауссовское ядро
- $h \in \{0.1, 1.0, 3.0\}$



kNN для регрессии

- Прямоугольное ядро $K(z) = [|z| \leq 1]$
- $h \in \{0.1, 1.0, 3.0\}$



Функции расстояния

Евклидова метрика

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

- Более общий вариант — метрика Минковского:

$$\rho(x, z) = \left(\sum_{j=1}^{d} \left(x_j - z_j\right)^p\right)^{1/p}$$

Чувствительность к масштабу

- Задача: определение пола
- Признаки:
  - Рост
  - Экспрессия гена SRY (от 0 до 1) — у женщин ближе к нулю
- Обучающая выборка:
  - $x_1 = (180, 0.2)$
  - $x_2 = (172, 0.9)$
- Новый объект: $x = (178, 0.85)$

Чувствительность к масштабу

- Задача: определение пола
- Признаки:
  - Рост
  - Экспрессия гена SRY (от 0 до 1) — у женщин ближе к нулю
- Обучающая выборка:
  - $x_1 = (180, 0.2)$
  - $x_2 = (172, 0.9)$
- Новый объект: $x = (178, 0.85)$
- $\rho(x, x_1) = 2.1, \rho(x, x_2) = 5$

Чувствительность к масштабу

- Если признаки имеют разные масштабы, то будут учитываться лишь самые крупные
- Перед применением kNN выборку необходимо масштабировать!

Расстояние Джаккарда

- Измеряет расстояния между множествами
- Пример: каждый объект — набор слов или тэгов
- Метрика:

$$\rho(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Расстояние Джаккарда

- Пример 1:
  - $A = \{\text{комедия}, \text{триллер}, \text{США}\}$
  - $B = \{\text{триллер}, \text{ужасы}, \text{Великобритания}\}$
- $\rho(A, B) = 1 - \frac{1}{5} = 0.8$

- Пример 2:
  - $A = \{\text{комедия}, \text{США}\}$
  - $B = \{\text{комедия}, \text{США}\}$
- $\rho(A, B) = 1 - \frac{2}{2} = 0$

Резюме

- Метрические методы — одни из самых интуитивных в машинном обучении
- Простая процедура обучения
- Гиперпараметры:
  - функция расстояния
  - число соседей
  - ядро
  - ширина окна