

# Введение в анализ данных

Лекция 2

Метод k ближайших соседей

Евгений Соколов

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

НИУ ВШЭ, 2021

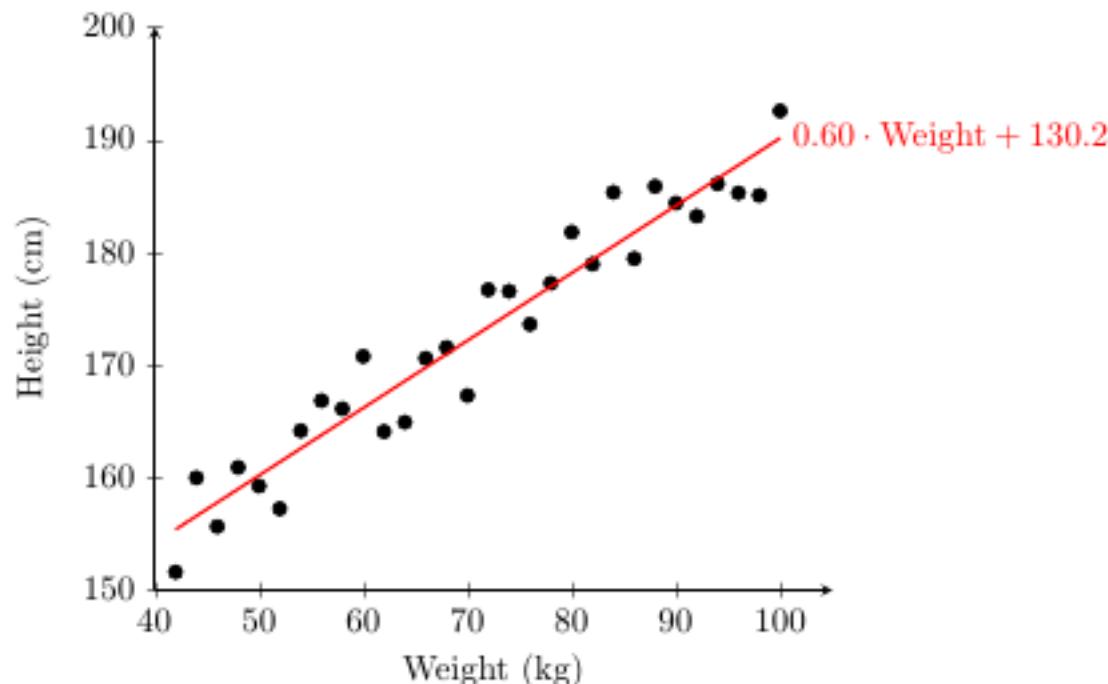
# Напоминание

- $\mathbb{X}$  — пространство объектов,  $\mathbb{Y}$  — пространство ответов
- $x = (x^1, \dots, x^d)$  — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка
- $a(x)$  — алгоритм, модель
- $Q(a, X)$  — функционал ошибки алгоритма  $a$  на выборке  $X$
- Обучение:  $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

# Типы ответов

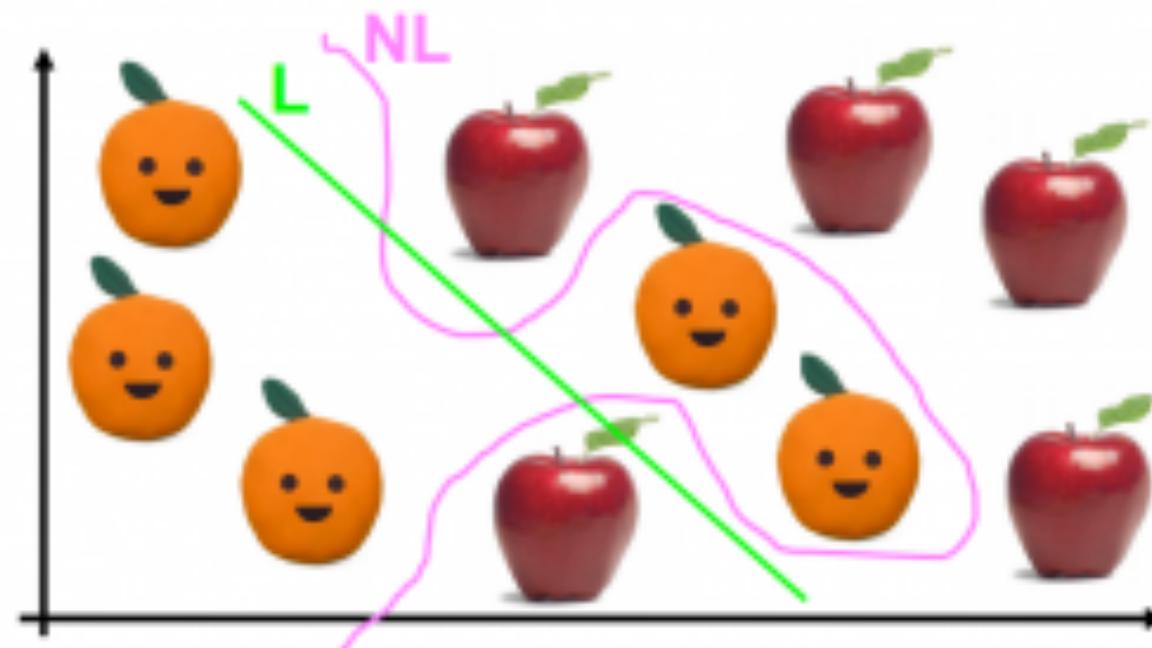
# Регрессия

- Вещественные ответы:  $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



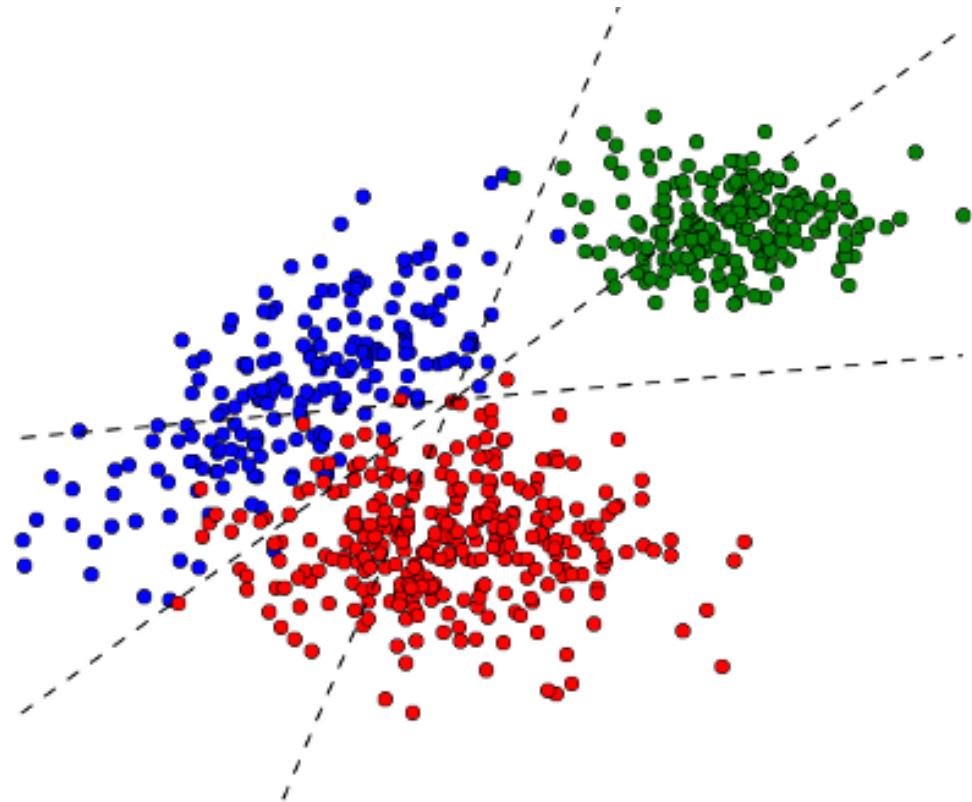
# Классификация

- Конечное число ответов:  $|\mathbb{Y}| < \infty$
- Бинарная классификация:  $\mathbb{Y} = \{-1, +1\}$



# Классификация

- Многоклассовая классификация:  $\mathbb{Y} = \{1, 2, \dots, K\}$



# Классификация

- Классификация с пересекающимися классами:  $\mathbb{Y} = \{0, 1\}^K$ 
  - (multi-label classification)
- Ответ — набор из  $K$  нулей и единиц
- $i$ -й элемент ответа — принадлежит ли объект  $i$ -му классу
- Какие темы присутствуют в статье?
- (математика, биология, экономика)

# Ранжирование

- Набор документов  $d_1, \dots, d_n$
- Запрос  $q$
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$  — оценка релевантности

# Ранжирование

Яндекс

картинки с котиками — 5 млн ответов



Найти

Поиск

[Картинки с кошками | Fun Cats — Забавные коты](#)

[funcats.by > pictures/](#) ▾

Картинки с кошками. Прикольные коты. 777 изображений. ... 32 изображения. Кошки

Стамбула. 41 изображение. Веселые котята.

Картинки

Видео

[Уморные котики \(57 фото\) » Бяки.нет | Картинки](#)

[byaki.net > Картинки > 14026-umornye-kotiki-57...](#) ▾

Бяки нет! . NET. Уморные котики (57 фото). 223. Коментариев:9Автор:4ertonok

Просмотров:161 395 Картинки28-10-2008, 00:03.

Карты

Маркет

Ещё

[Смешные картинки кошек с надписями | Лолкот.Ру](#)

[lolkot.ru](#) ▾

Смешные картинки для новых приколов! Сделать свой прикол очень просто. ... Котик

верит в чудеса. Он в носке подарок ищет...

[Красивые картинки и фото кошек, котят и котов](#)

[foto-zverey.ru > Кошки](#) ▾

Фото и картинки кошек и котят потрясающей красоты и нежности. Здесь мы собрали

такие изображения, которые всегда вызывают море положительных эмоций...

[Обои для рабочего стола Котята | картинки на стол Котята](#)

[7fon.ru > Чёрные обои и картинки > Обои котята](#) ▾

Картинки Котята с 1 по 15. Обои для рабочего стола Котята. ... Скачать Картинки Котята

на рабочий стол бесплатно.

# Кластеризация

- $\mathbb{Y}$  — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?
- Пример: сегментация пользователей мобильного оператора

# Типы признаков

# Типы признаков

- $D_j$  — множество значений признака

# Бинарные признаки

- $D_j = \{0, 1\}$
- Доход клиента выше среднего по городу?
- Цвет фрукта — зеленый?

# Вещественные признаки

- $D_j = \mathbb{R}$
- Возраст
- Площадь квартиры
- Количество звонков в колл-центр

# Категориальные признаки

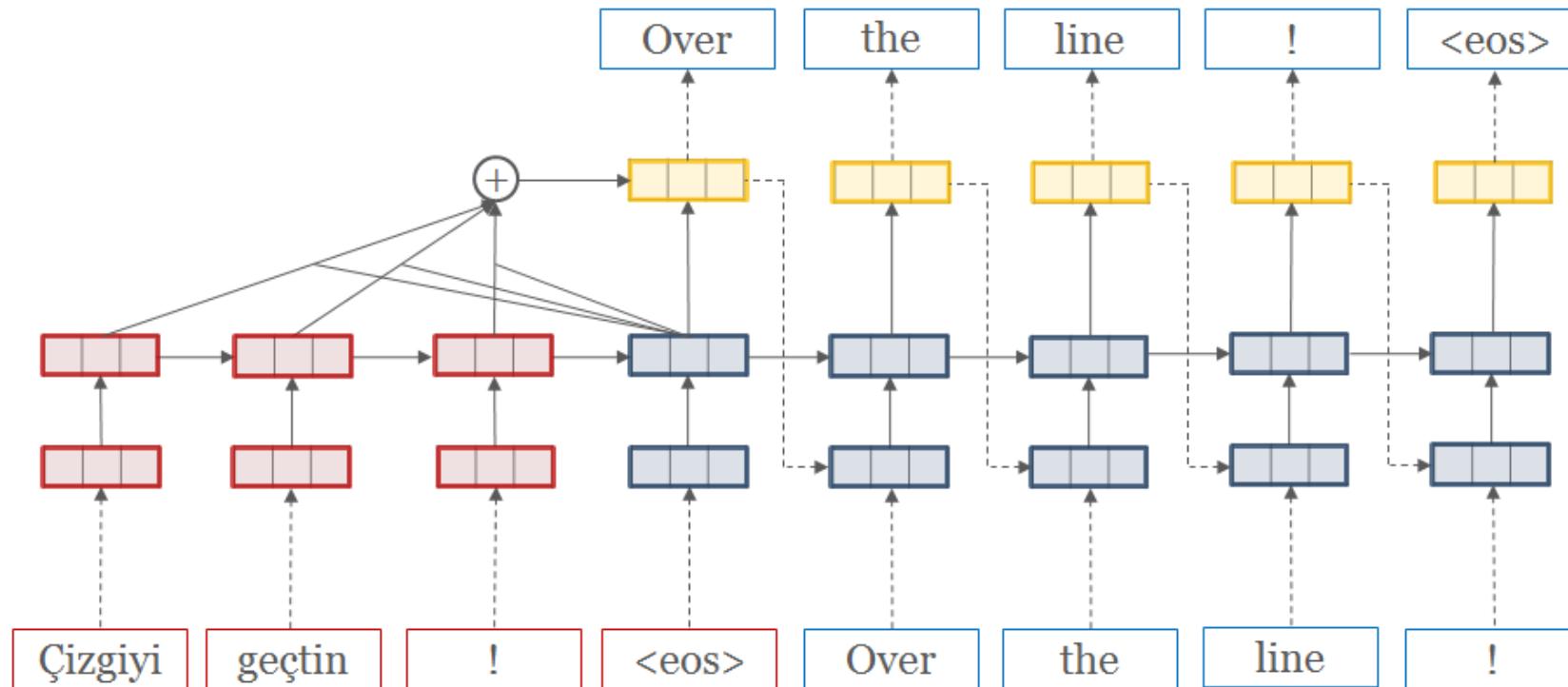
- $D_j$  — неупорядоченное множество
- Цвет глаз
- Город
- Образование (может быть упорядоченным)
  
- Очень трудны в обращении

# Порядковые признаки

- $D_j$  — упорядоченное множество
- Воинское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

Зачем это нужно?

# Машинный перевод

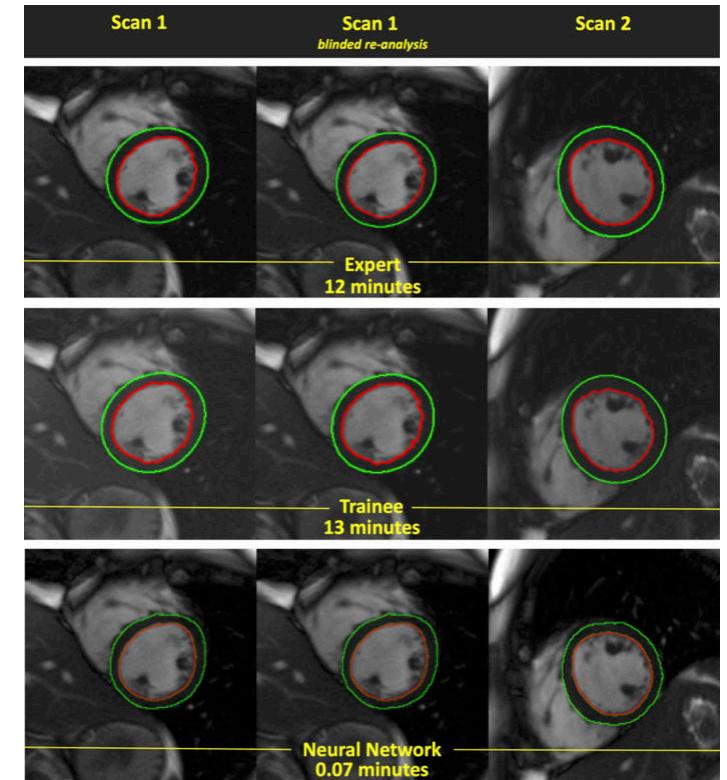


# Генерация текста

- GPT-3 от OpenAI
- <https://arxiv.org/abs/2005.14165>
- <https://talktotransformer.com>

# Биоинформатика и медицина

- Поиск связей между ДНК и заболеваниями (23andme и другие)
- Таргетные лекарства
- Анализ медицинских снимков



# Сельское хозяйство

- Робототехника
- Мониторинг посевов и почвы
- Прогнозирование болезней и урожайности



# Рекомендательные системы

- Полки рекомендаций на Amazon генерируют 35% от всех покупок
- Рекомендации на основе машинного обучения и анализа больших объёмов данных

**Frequently Bought Together**

Price For All Three: \$86.01

Add all three to Cart    Add all three to Wish List

Show availability and shipping details

This item: Machine Learning for Hackers by Drew Conway Paperback \$33.87

Machine Learning in Action by Peter Harrington Paperback \$25.75

Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback \$26.39

---

**Customers Who Bought This Item Also Bought**

Page 1 of 17

Item	Author	Type	Price
Programming Collective Intelligence: Building Smart Web 2.0 Applications	Toby Segaran	Paperback	\$26.39
Machine Learning in Action	Peter Harrington	Paperback	\$25.75
Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and More Using Python	Matthew A. Russell	Paperback	\$26.36
Data Analysis with Open Source Tools	Philipp K. Janert	Paperback	\$24.05
R Cookbook (O'Reilly Cookbooks)	Paul Teator	Paperback	\$32.43
The Art of R Programming: A Tour of Statistical Analysis and Computation Using R	Norman Matloff	Paperback	\$25.06

Are any of these items inappropriate for this page? [Let us know](#)

# Зачем это нужно?

- Это круто
  - Сложные задачи
  - Движение к искусственному интеллекту
- Это полезно
  - Извлечение прибыли из данных
  - Data-driven companies

# Как можно заниматься анализом данных?

- Data scientist
  - Работа с данными
  - Знание инструментов и методов
  - Опыт решения задач
- Менеджер
  - Понимание, как работает машинное обучение
  - Понимание узких мест, оценивание сроков
- Заказчик
  - Метрики качества
  - Требования к данным
  - Ограничения современных подходов

Гипотеза компактности и knn

# Как отличить ель от сосны?



# Как отличить ель от сосны?



# Как отличить ель от сосны?



Ель:

- Ветки смотрят вверх
- Ствол не видно
- Густые иголки
- Цвет ближе к зелёному



Сосна:

- Ветки параллельны земле
- Ствол видно
- Иголки более редкие
- Цвет ближе к жёлтому

# Как отличить ель от сосны?



Ветки вверх  
Ствол не видно  
Густые иголки  
Цвет ближе к синему

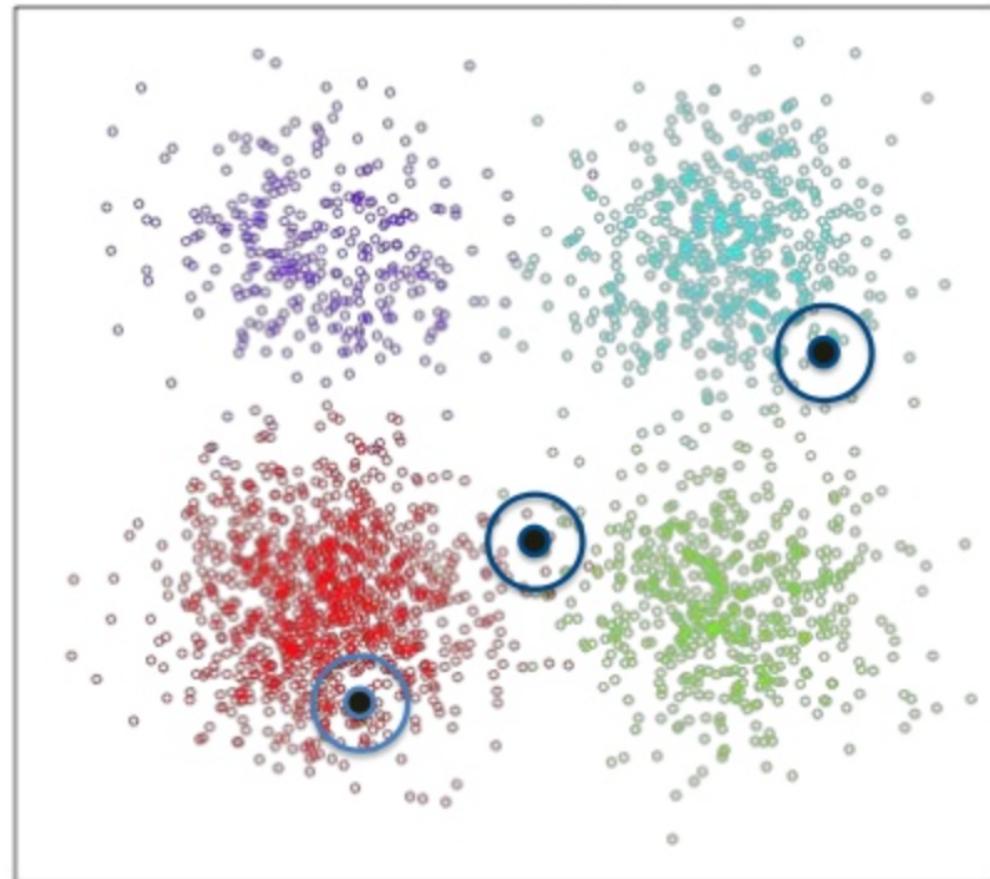


Скорее всего ель

# Что такое обучение?

- Запоминаем примеры (объекты и ответы)
- Когда приходит новый объект, сравниваем с запомненными примерами
- Выдаём ответ от наиболее похожего примера

# Гипотеза компактности



# Гипотеза компактности



# Гипотеза компактности

Если два объекта похожи друг на друга, то ответы на них  
тоже похожи

# kNN: обучение

- Дано: обучающая выборка  $X = (x_i, y_i)_{i=1}^{\ell}$
- Задача классификация (ответы из множества  $\mathbb{Y} = \{1, \dots, K\}$ )
- Обучение модели:
  - Запоминаем обучающую выборку  $X$

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

# kNN: применение

Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

# kNN: применение

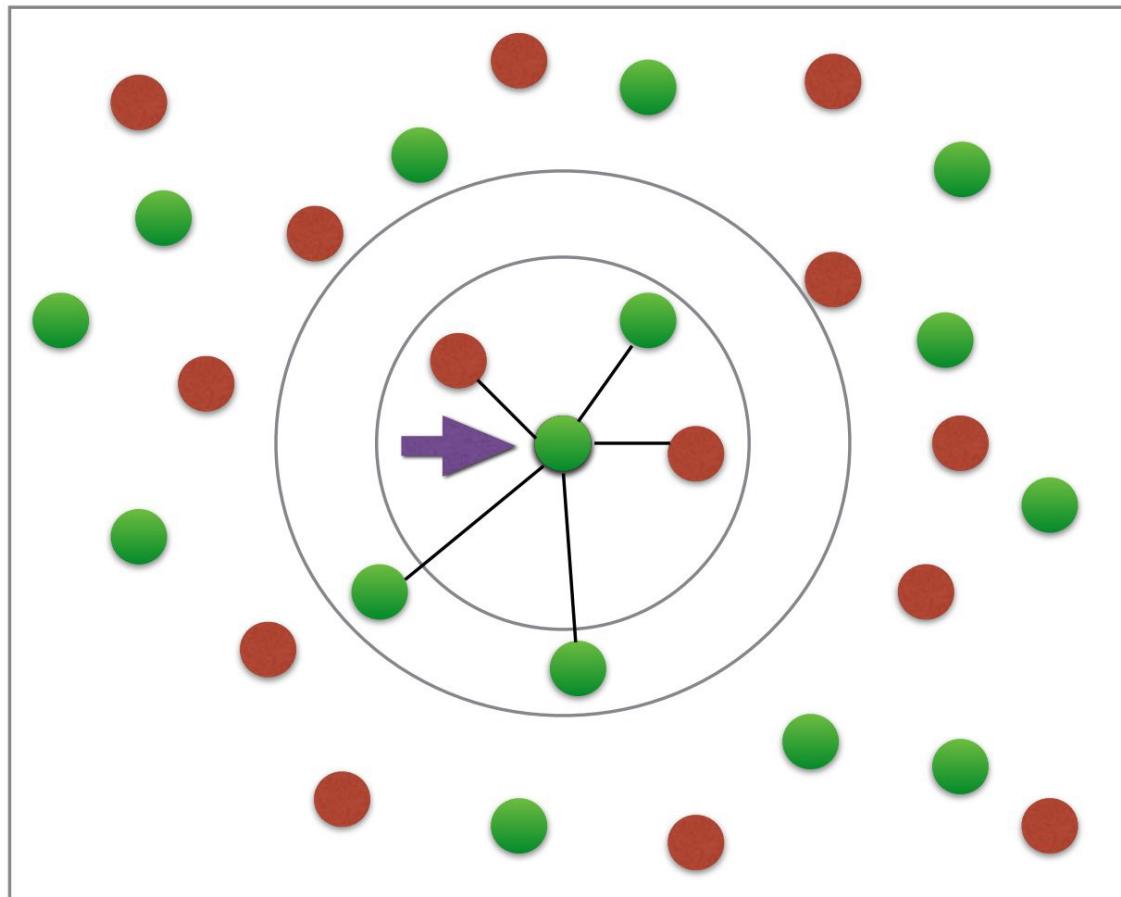
Дано: новый объект  $x$

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:  
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем  $k$  ближайших объектов:  $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

# kNN: применение



# Сравнение объектов и метрики

# Числовые данные

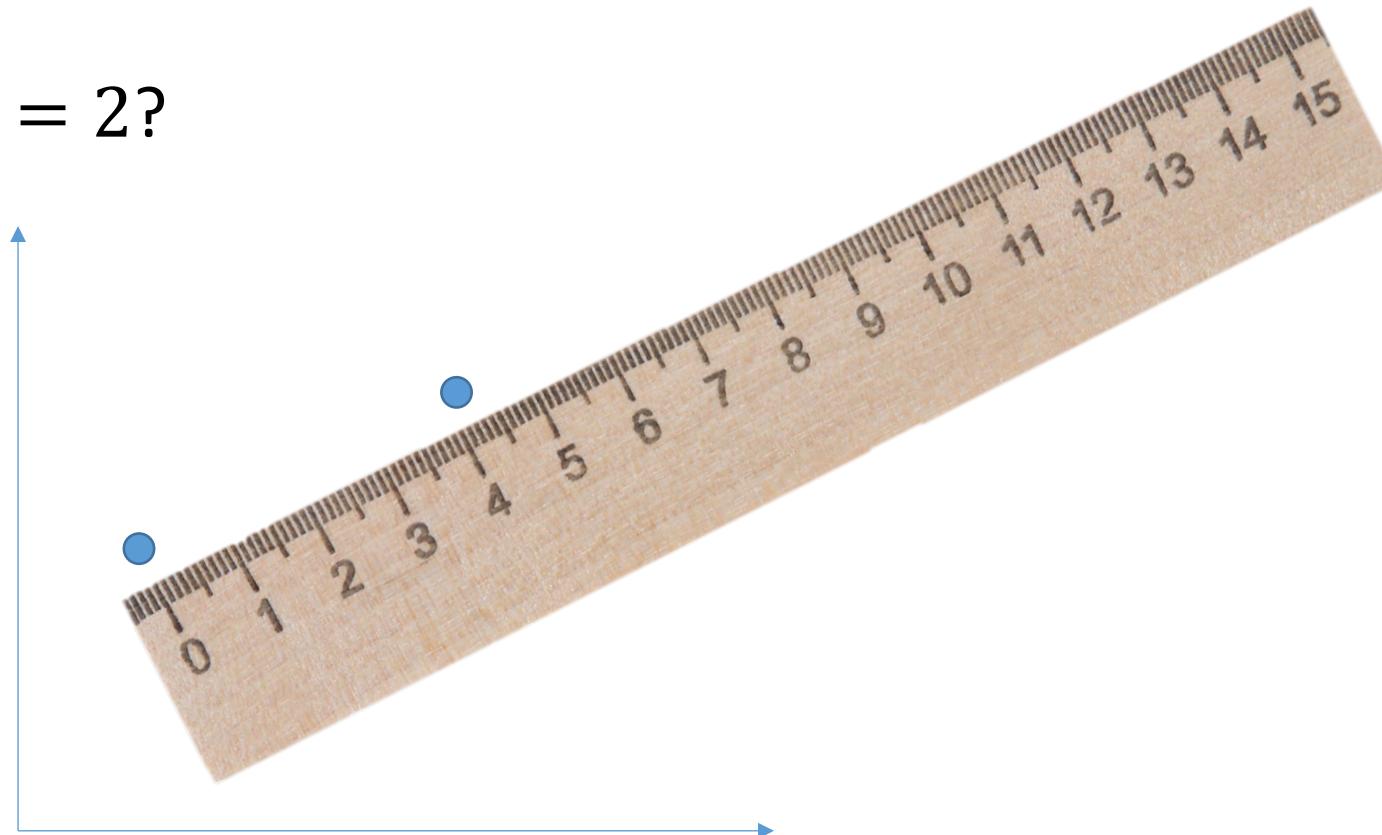
Сколько раз в день вызывает такси	Средние расходы на такси в день	Как часто вызывал комфорт	Возраст	Согласился повысить категорию?
2	400	0.3	29	да
0.3	80	0	28	нет
...	...	...	...	...

# Числовые данные

- Каждый объект описывается набором из  $d$  чисел — **вектором**
- Если  $x$  — вектор, то  $x_i$  — его  $i$ -я координата
- Если  $x_i$  — вектор, то  $x_{ij}$  — его  $j$ -я координата

# Числовые данные

- Каждый объект описывается набором из  $d$  чисел — **вектором**
- Что, если  $d = 2$ ?



# Метрика

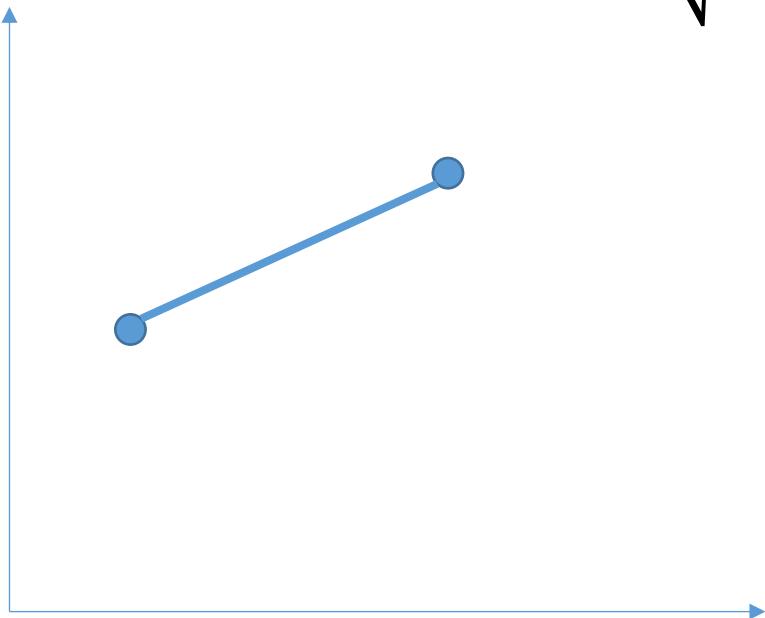
Метрика — обобщение расстояния на многомерные пространства

Метрика — это функция  $\rho$  с двумя аргументами, удовлетворяющая трём требованиям:

- $\rho(x, z) = 0$  тогда и только тогда, когда  $x = z$
- $\rho(x, z) = \rho(z, x)$
- $\rho(x, z) \leq \rho(x, v) + \rho(v, z)$  — неравенство треугольника

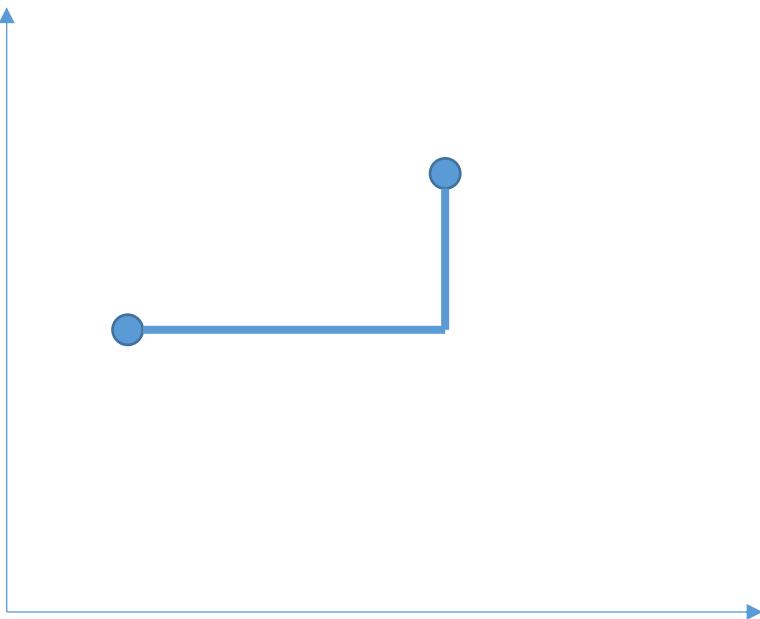
# Евклидова метрика

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

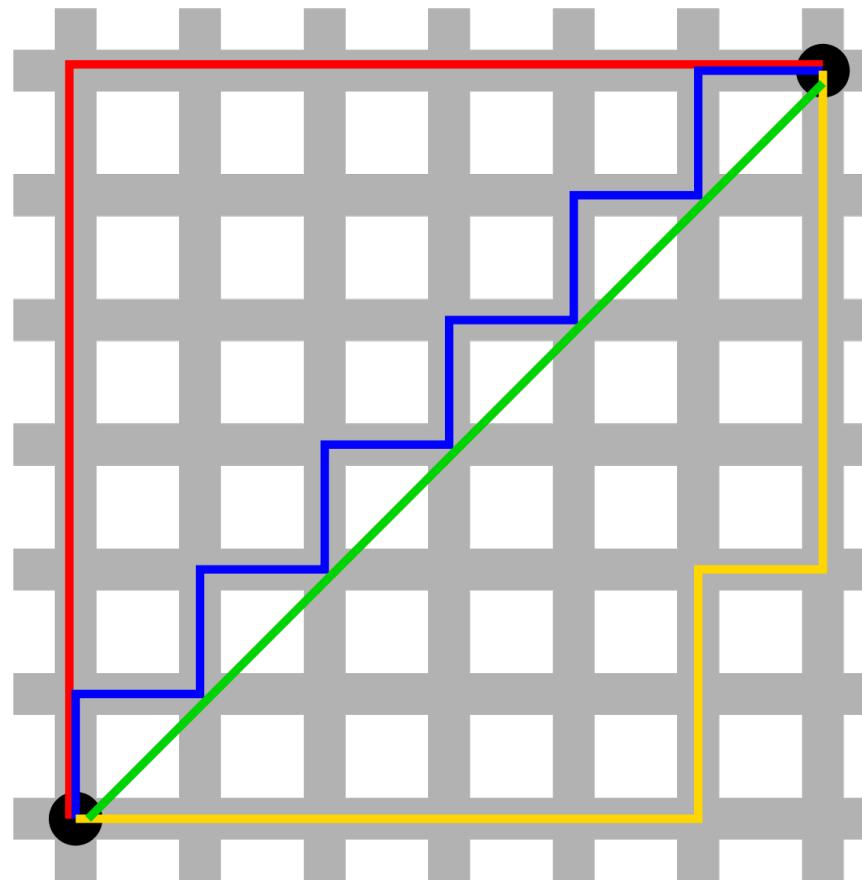


# Манхэттенская метрика

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



# Сравнение



# Обобщение

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

- Метрика Минковского
- Можно подбирать  $p$  под конкретную задачу

# Категориальные данные

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
...	...	...	...

# Считывающая метрика

- Простейшая метрика: подсчёт различий

$$\rho(x, z) = \sum_{j=1}^d [x_j \neq z_j]$$

# Что ещё?

- Текстовые данные — чуть-чуть изучим в курсе, подробно потом
- Изображения — потом

# Измерение ошибки модели

# Вопросы

- Как сравнить две модели?
- Как подобрать  $k$  и метрику?

# ФУНКЦИЯ ПОТЕРЬ ДЛЯ КЛАССИФИКАЦИИ

- Частый выбор — бинарная функция потерь

$$L(y, a) = [a \neq y]$$

- Функционал ошибки — доля ошибок (error rate)

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

- Нередко измеряют долю верных ответов (accuracy):

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

# Функция потерь для классификации

ВАЖНО

Accuracy — не точность!

# Accuracy

$a(x)$	$y$
-1	-1
+1	+1
-1	-1
+1	-1
+1	+1

# Accuracy

$a(x)$	$y$
-1	-1
+1	+1
-1	-1
+1	-1
+1	+1

Доля ошибок: 0.2

Доля верных ответов: 0.8

# Accuracy

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

Решаем задачу выявления редкого заболевания

- 950 здоровых ( $y = +1$ )
- 50 больных ( $y = -1$ )

Модель:  $a(x) = +1$

**Доля ошибок: 0.05**

# Accuracy

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

- Всегда смотрите на баланс классов!
- Доля верных ответов не обязательно меняется от 0.5 до 1 для разумных моделей

# Как выбрать k?

Обучающая выборка

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
Комфорт	Строгино	Карта	да

Применяем модель:

Эконом	Таганская	Карта	?
--------	-----------	-------	---

# Как выбрать k?

Обучающая выборка

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
Комфорт	Строгино	Карта	да

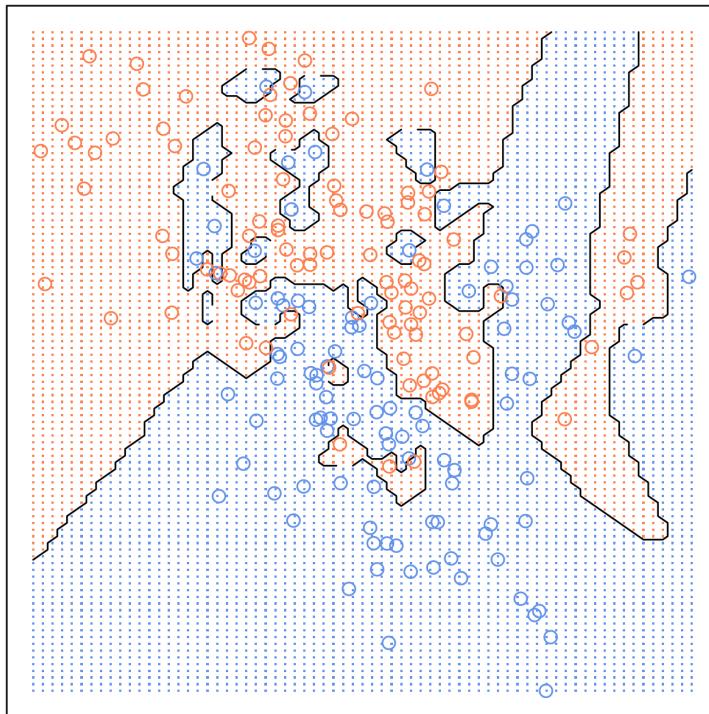
Применяем модель:

Эконом	Таганская	Карта	да
--------	-----------	-------	----

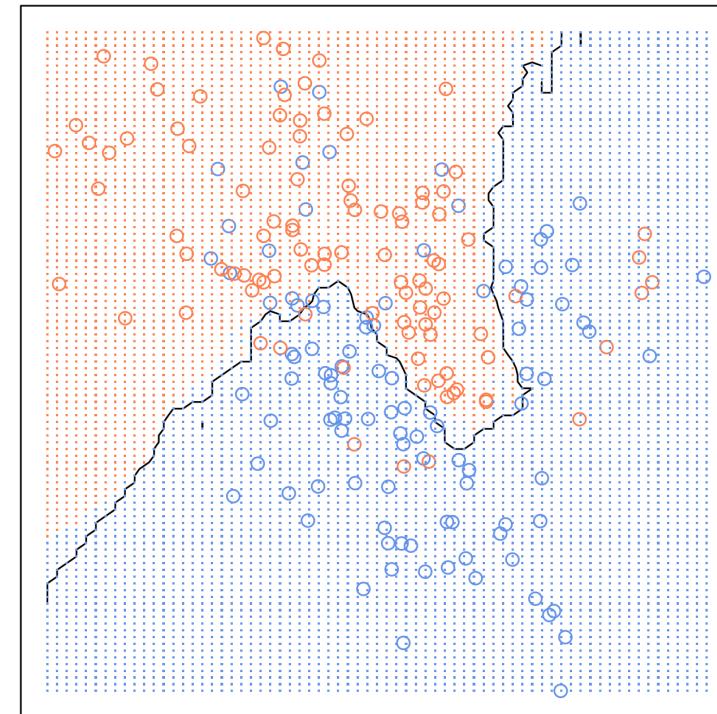
С точки зрения качества на обучающей выборке лучший выбор  $k = 1$

# Как выбрать k?

1-nearest neighbours



20-nearest neighbours



<https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

# Гиперпараметры

- Нельзя подбирать  $k$  по обучающей выборке — **гиперпараметр**
- Нужно использовать дополнительные данные

Обобщающая способность

# Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с  
занятий

Разобраться в предмете и  
усвоить алгоритмы решения  
задач

# Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с  
занятий

Переобучение (overfitting)

Разобраться в предмете и  
усвоить алгоритмы решения  
задач

Обобщение (generalization)

# Обобщающая способность

## Как готовиться к экзамену?

Заучить все примеры с занятий

Переобучение (overfitting)

Хорошее качество на обучении  
Низкое качество на новых данных

Разобраться в предмете и усвоить алгоритмы решения задач

Обобщение (generalization)

Хорошее качество на обучении  
Хорошее качество на новых данных

# Отложенная выборка



Обучение



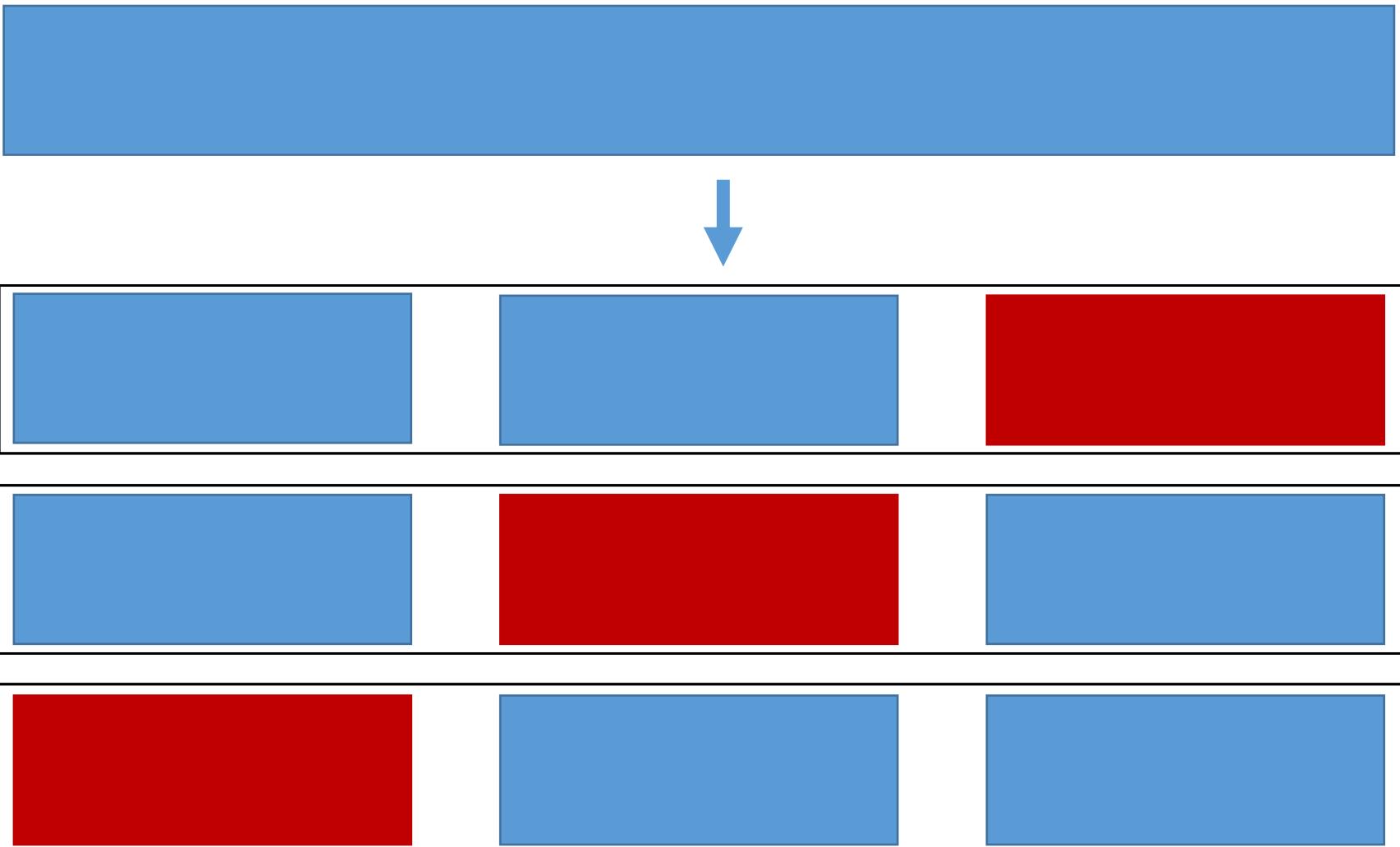
Тест

# Отложенная выборка



- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не сможет обучиться
- Обычно: 70/30, 80/20

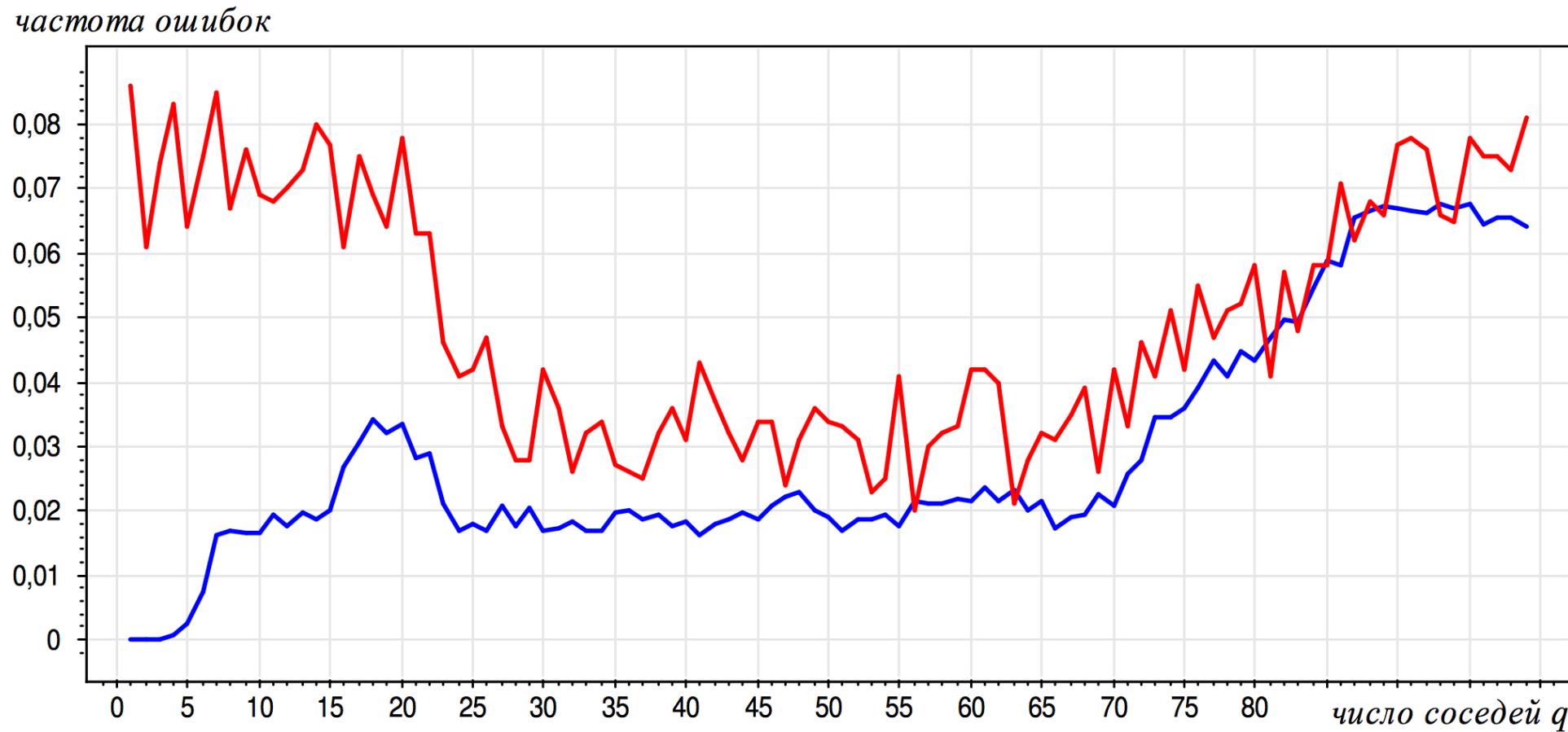
# Кросс-валидация



# Кросс-валидация

- Надёжнее отложенной выборки, но медленнее
- Параметр — количество разбиений  $n$  (фолдов, folds)
- Хороший, но медленный вариант —  $n = \ell$  (leave-one-out)
- Обычно:  $n = 3$  или  $n = 5$  или  $n = 10$

# Подбор числа соседей



# Чуть больше терминов

- После подбора всех гиперпараметров стоит проверить на совсем новых данных, что модель работает
- Обучающая выборка — построение модели
- Валидационная выборка — подбор гиперпараметров модели
- Тестовая выборка — финальная оценка качества модели