

# Основы машинного обучения

Лекция 2  
Метод k ближайших соседей

Евгений Соколов

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

НИУ ВШЭ, 2026

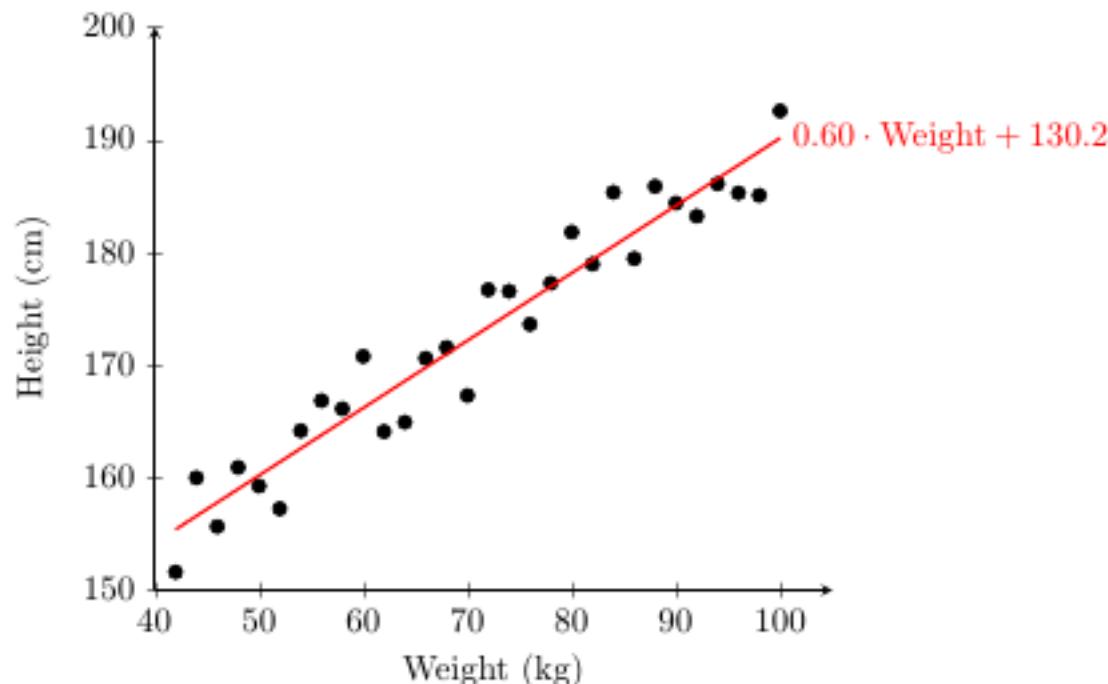
# Напоминание

- $\mathbb{X}$  — 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

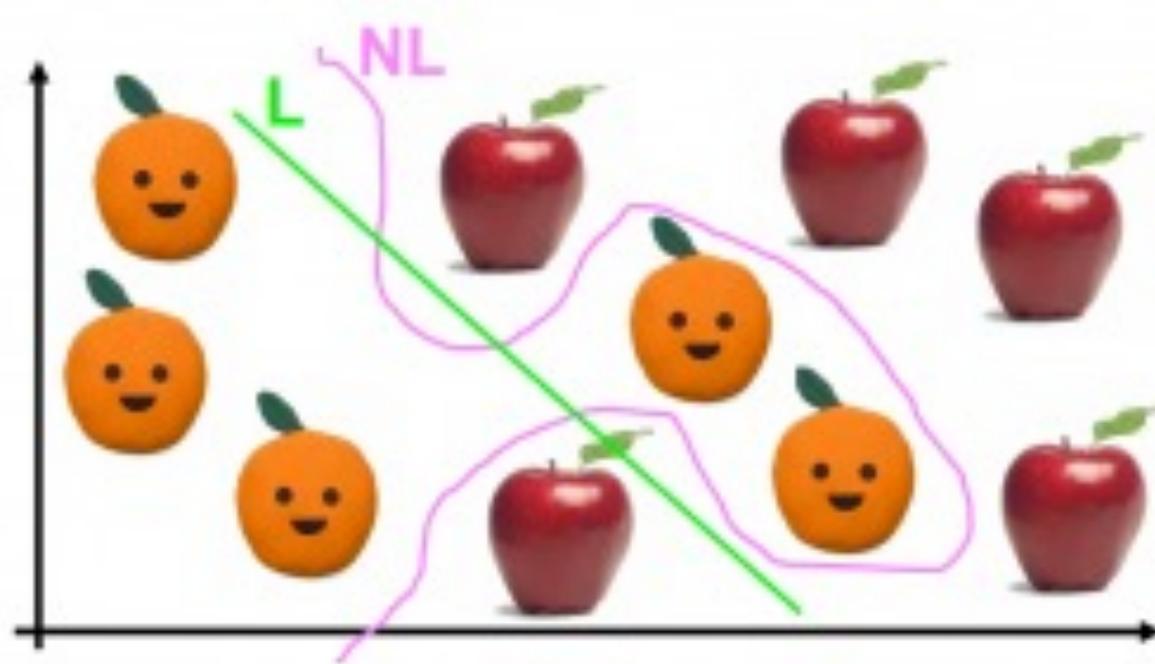
# Регрессия

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



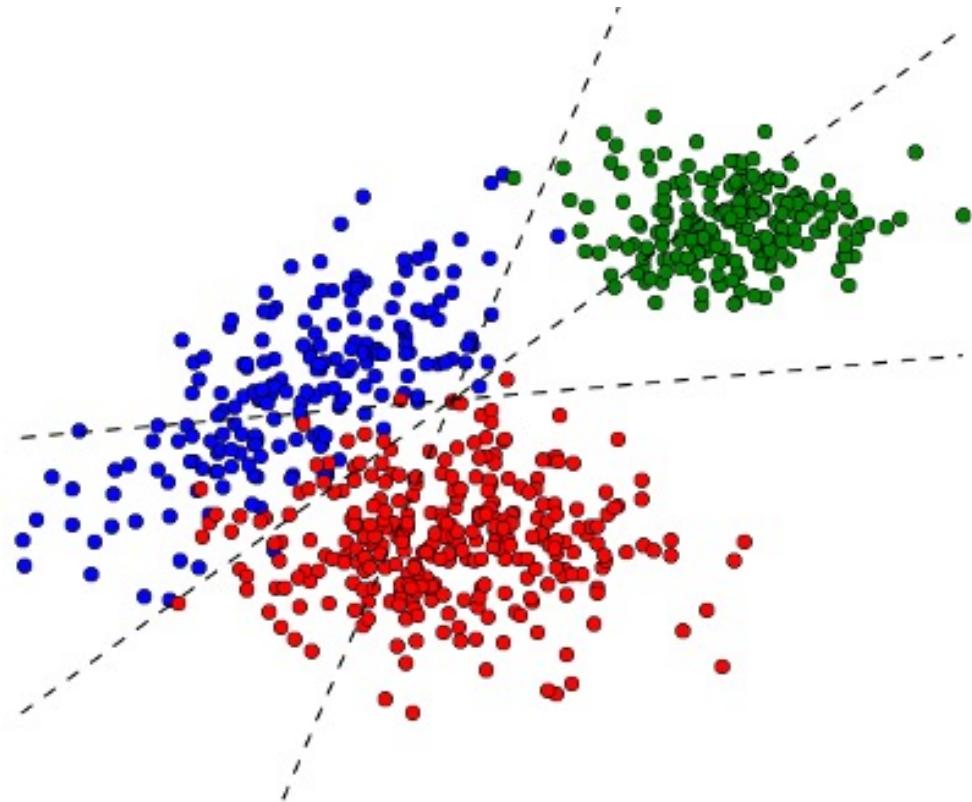
# Классификация

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



# Классификация

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



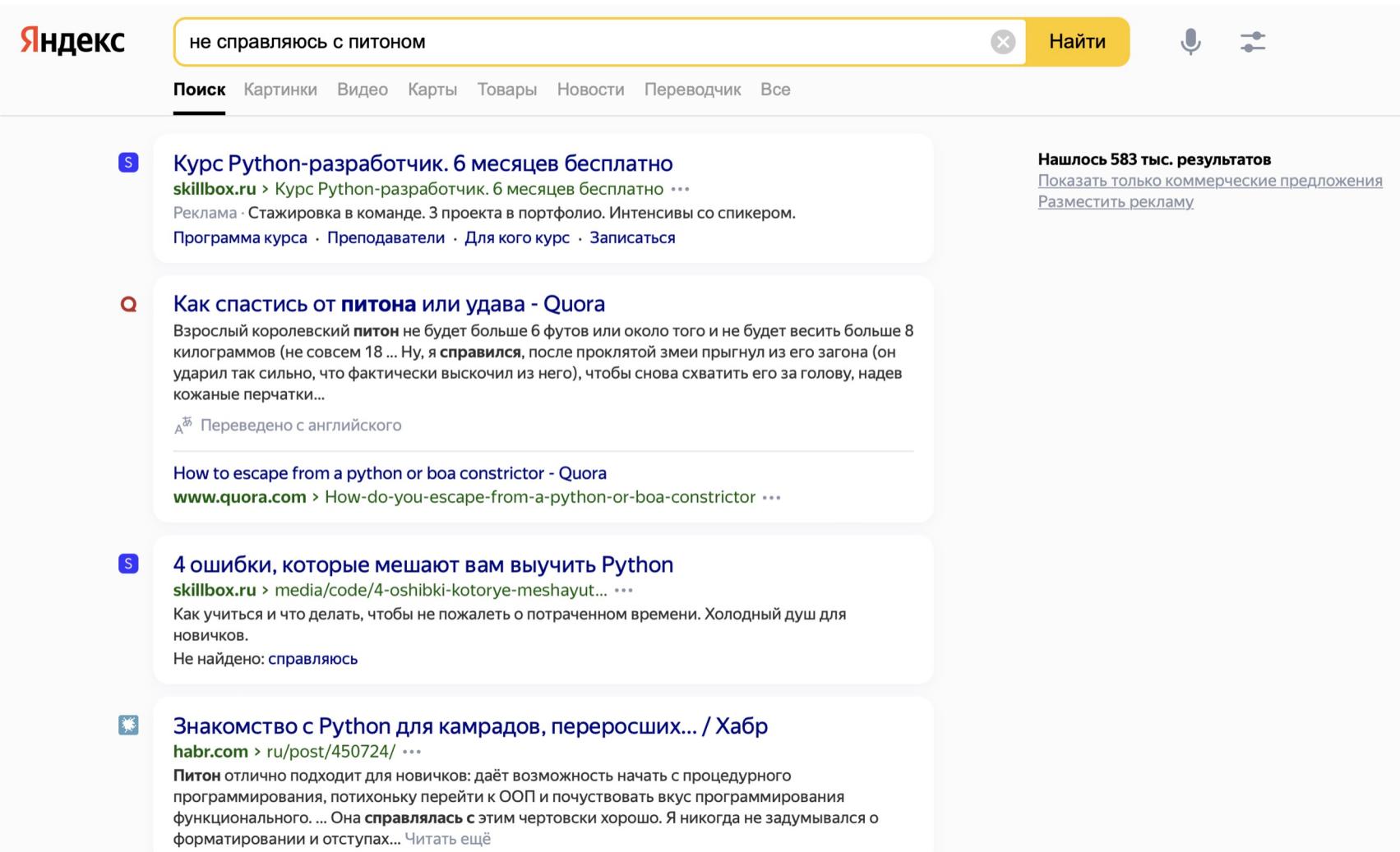
# Классификация

- 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

не справляюсь с питоном

Поиск Картинки Видео Карты Товары Новости Переводчик Все

**S** Курс Python-разработчик. 6 месяцев бесплатно  
[skillbox.ru](#) > Курс Python-разработчик. 6 месяцев бесплатно ...  
Реклама · Стажировка в команде. З проекта в портфолио. Интенсивы со спикером.  
Программа курса · Преподаватели · Для кого курс · Записаться

**Q** Как спастись от питона или удава - Quora  
Взрослый королевский питон не будет больше 6 футов или около того и не будет весить больше 8 килограммов (не совсем 18 ... Ну, я справился, после проклятой змеи прыгнул из его загона (он ударил так сильно, что фактически выскочил из него), чтобы снова схватить его за голову, надев кожаные перчатки...  
Переведено с английского

How to escape from a python or boa constrictor - Quora  
[www.quora.com](#) > How-do-you-escape-from-a-python-or-boa-constrictor ...

**S** 4 ошибки, которые мешают вам выучить Python  
[skillbox.ru](#) > media/code/4-oshibki-kotorye-meshayut... ...  
Как учиться и что делать, чтобы не пожалеть о потраченном времени. Холодный душ для новичков.  
Не найдено: справляюсь

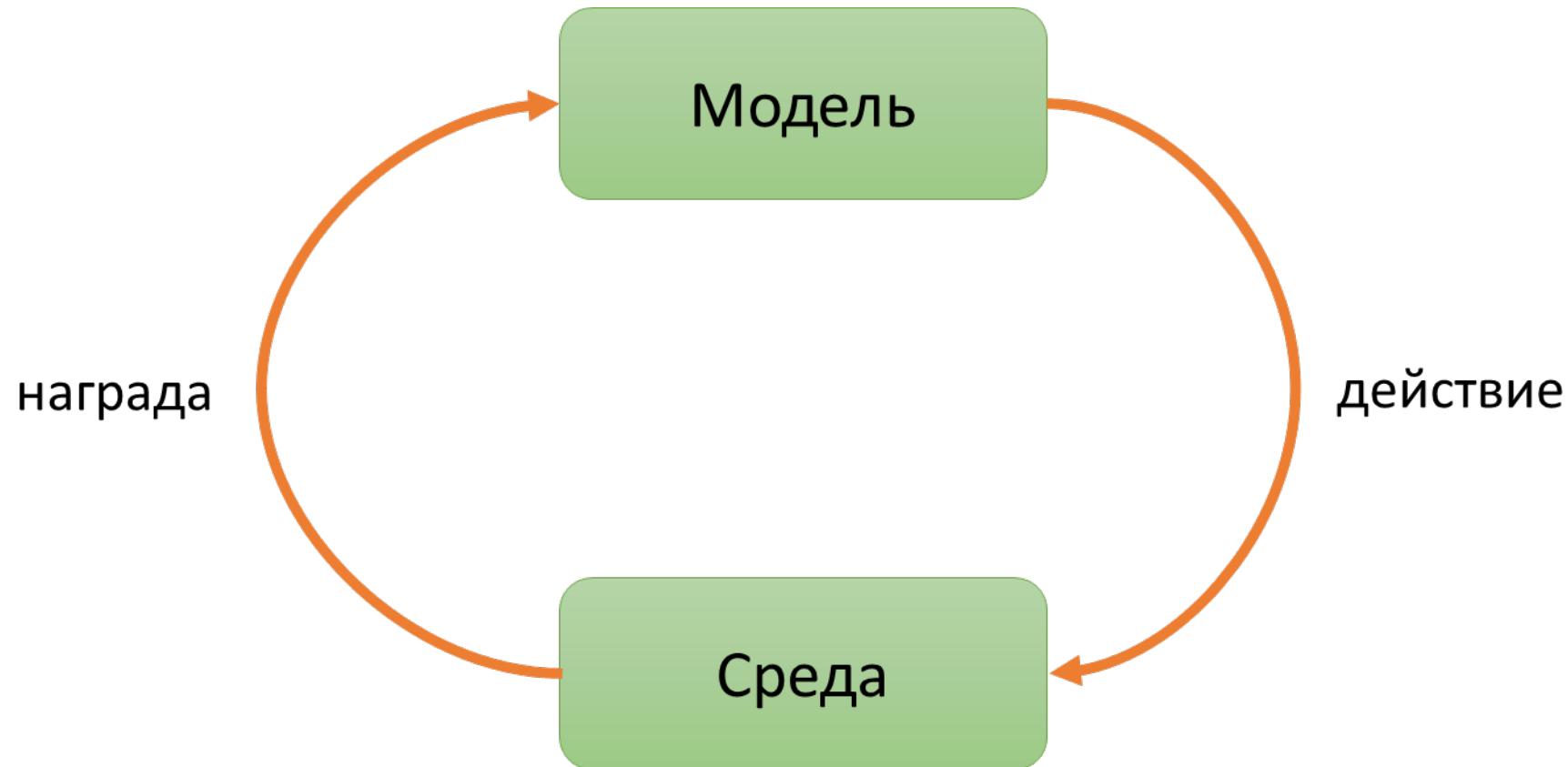
**✳** Знакомство с Python для камрадов, переросших... / Хабр  
[habr.com](#) > ru/post/450724/ ...  
Питон отлично подходит для новичков: даёт возможность начать с процедурного программирования, потихоньку перейти к ООП и почувствовать вкус программирования функционального. ... Она справилась с этим чертовски хорошо. Я никогда не задумывался о форматировании и отступах... Читать ещё

Нашлось 583 тыс. результатов  
Показать только коммерческие предложения  
Разместить рекламу

# Кластеризация

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

# Обучение с подкреплением



# Типы задач

- Регрессия
- Классификация
- Кластеризация
- Много других: ранжирование, поиск аномалий и т.д.

# 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

# Типы признаков

- Бинарные
- Числовые
- Категориальные и порядковые
- Есть и более сложные: тексты, изображения, звук и т.д.

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

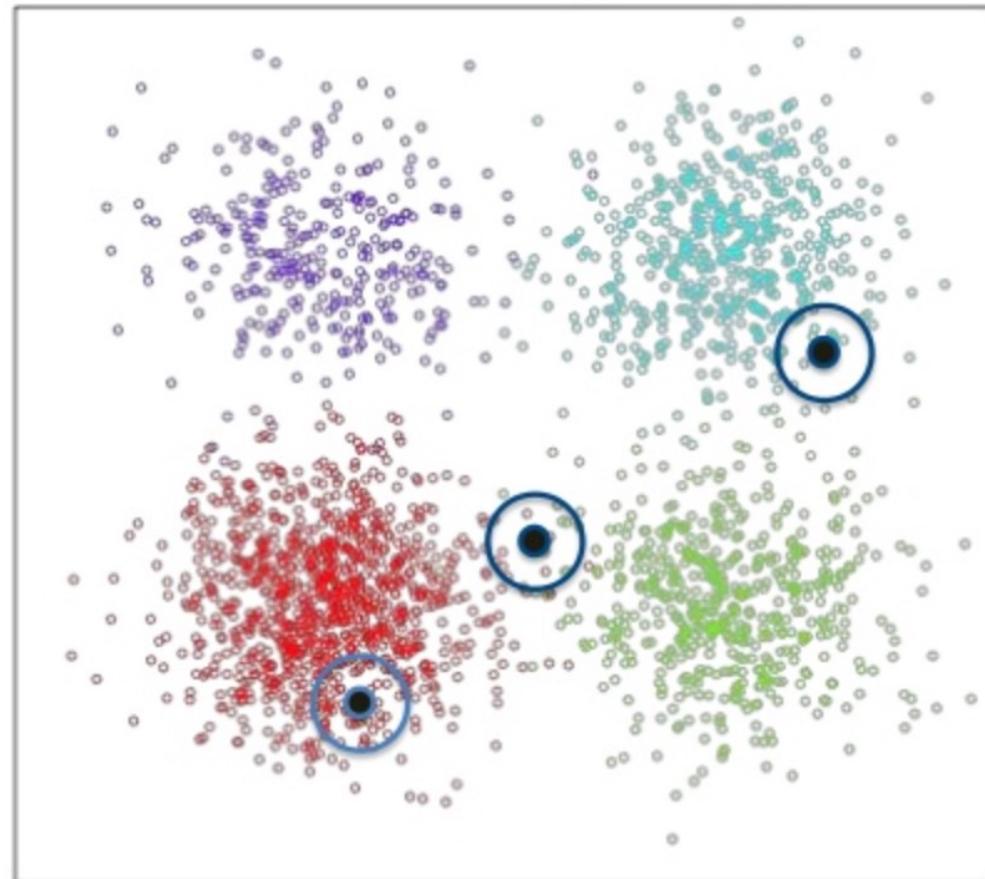


Скорее всего ель

# 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

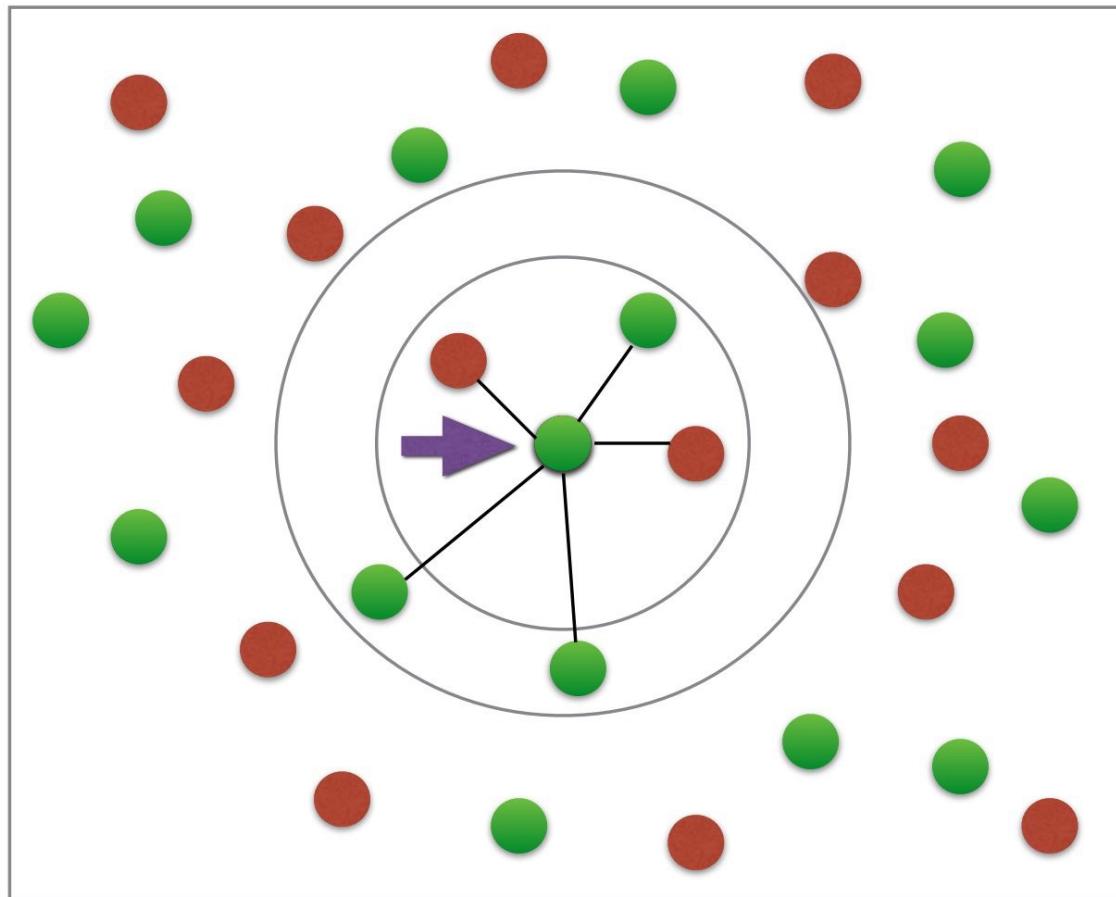
Дано: новый объект  $x$

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

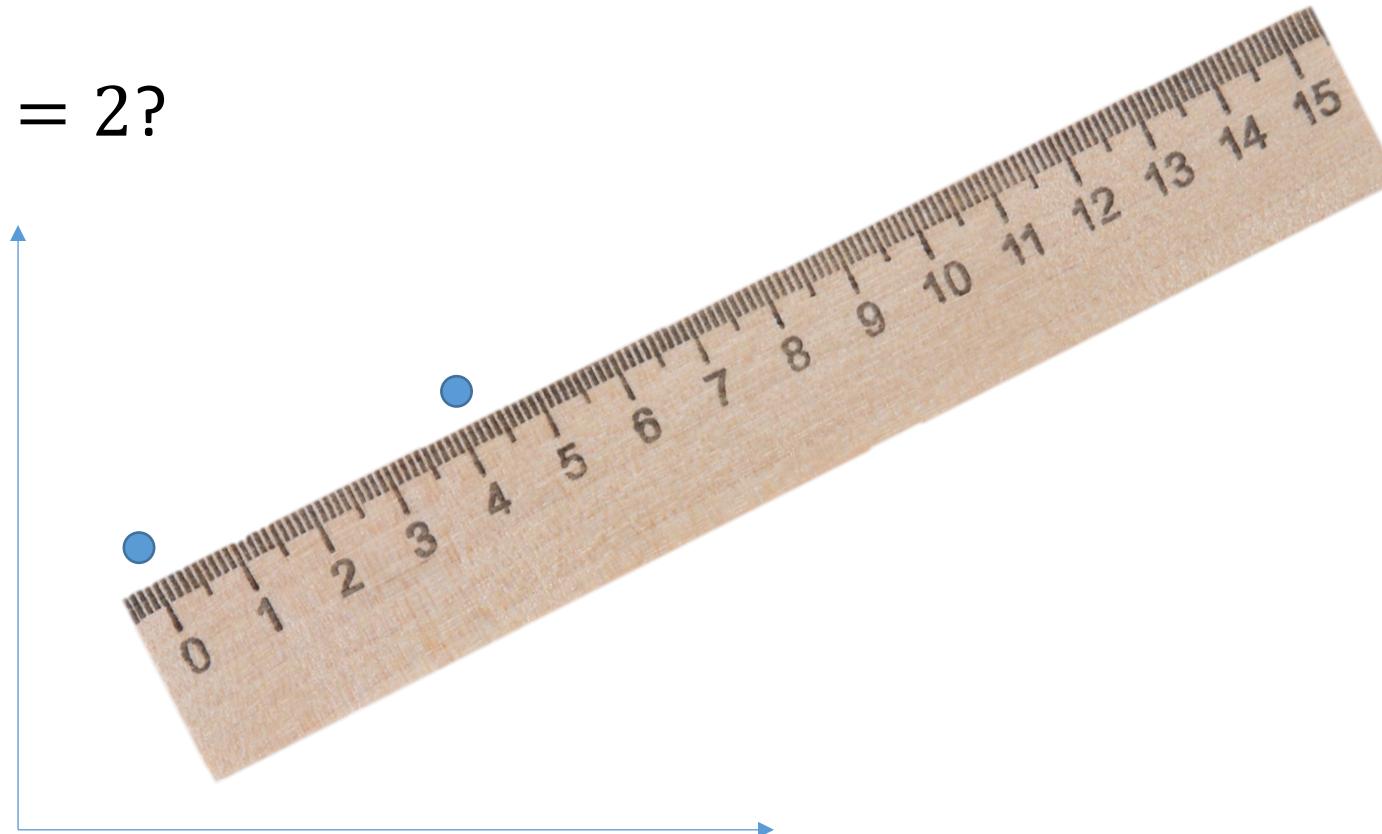
Сколько раз в день вызывает такси	Средние расходы на такси в день	Как часто вызывал комфорт	Возраст	Согласился повысить категорию?
2	400	0.3	29	да
0.3	80	0	28	нет
...	...	...	...	...

# 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

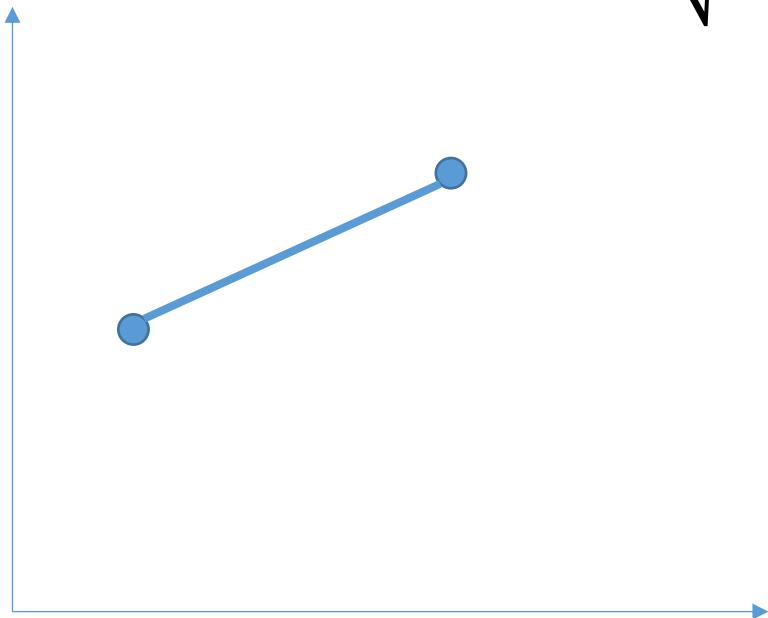
Метрика — обобщение расстояния на многомерные пространства

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

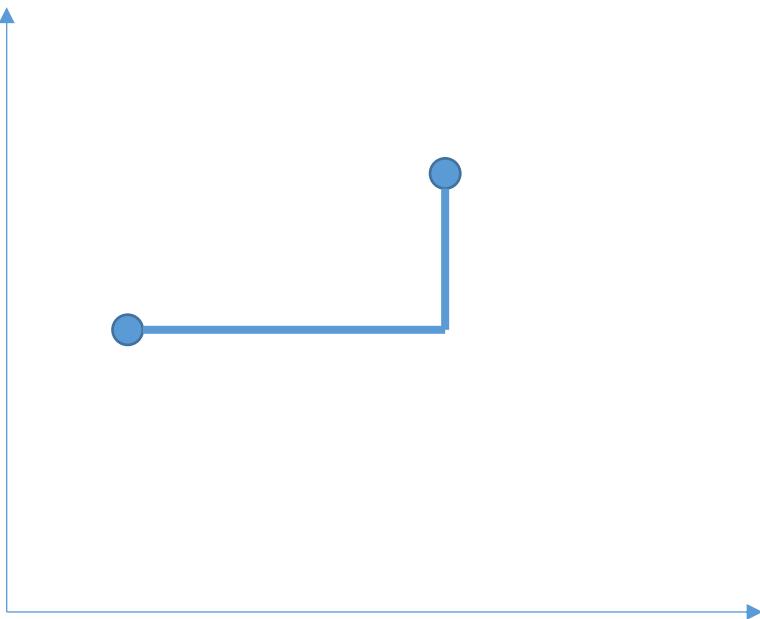
# Евклидова метрика

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

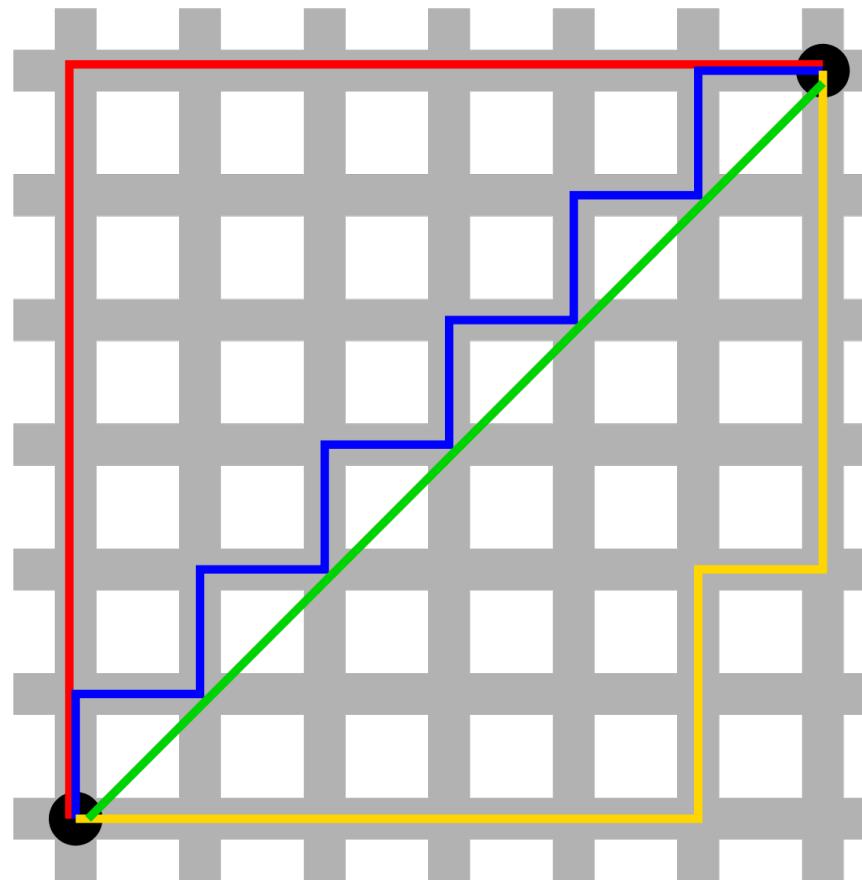


# Манхэттенская метрика

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



# 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

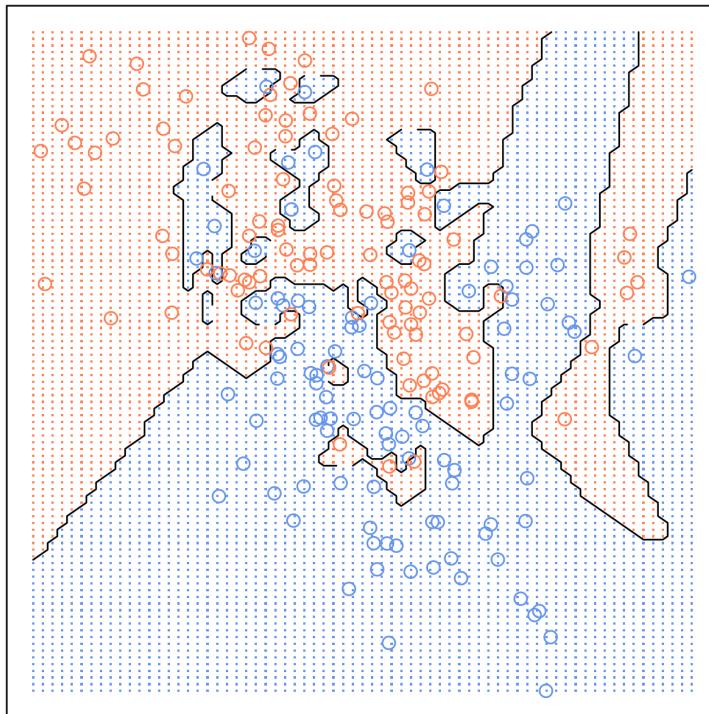
Применяем модель:

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

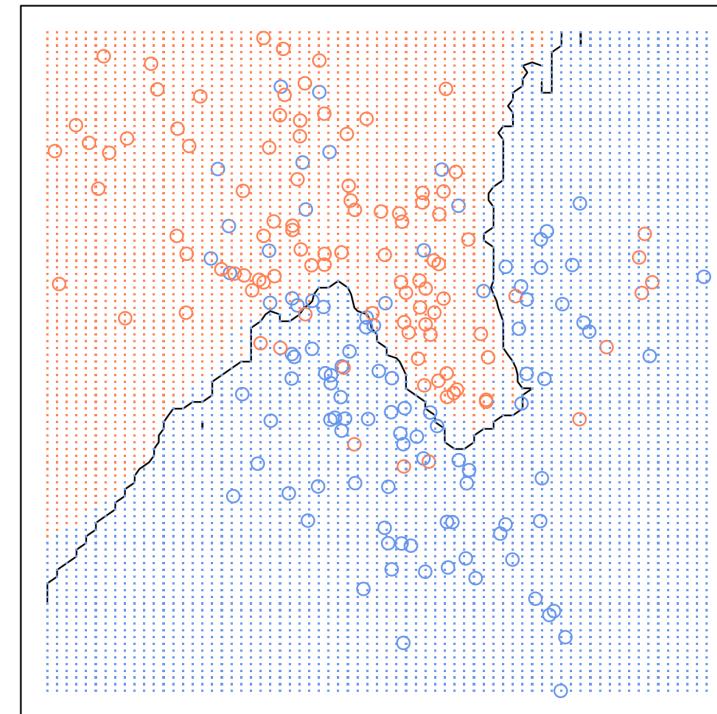
С точки зрения качества на обучающей выборке лучший выбор  $k = 1$

# 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