# Меры близости и KNN

**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: 

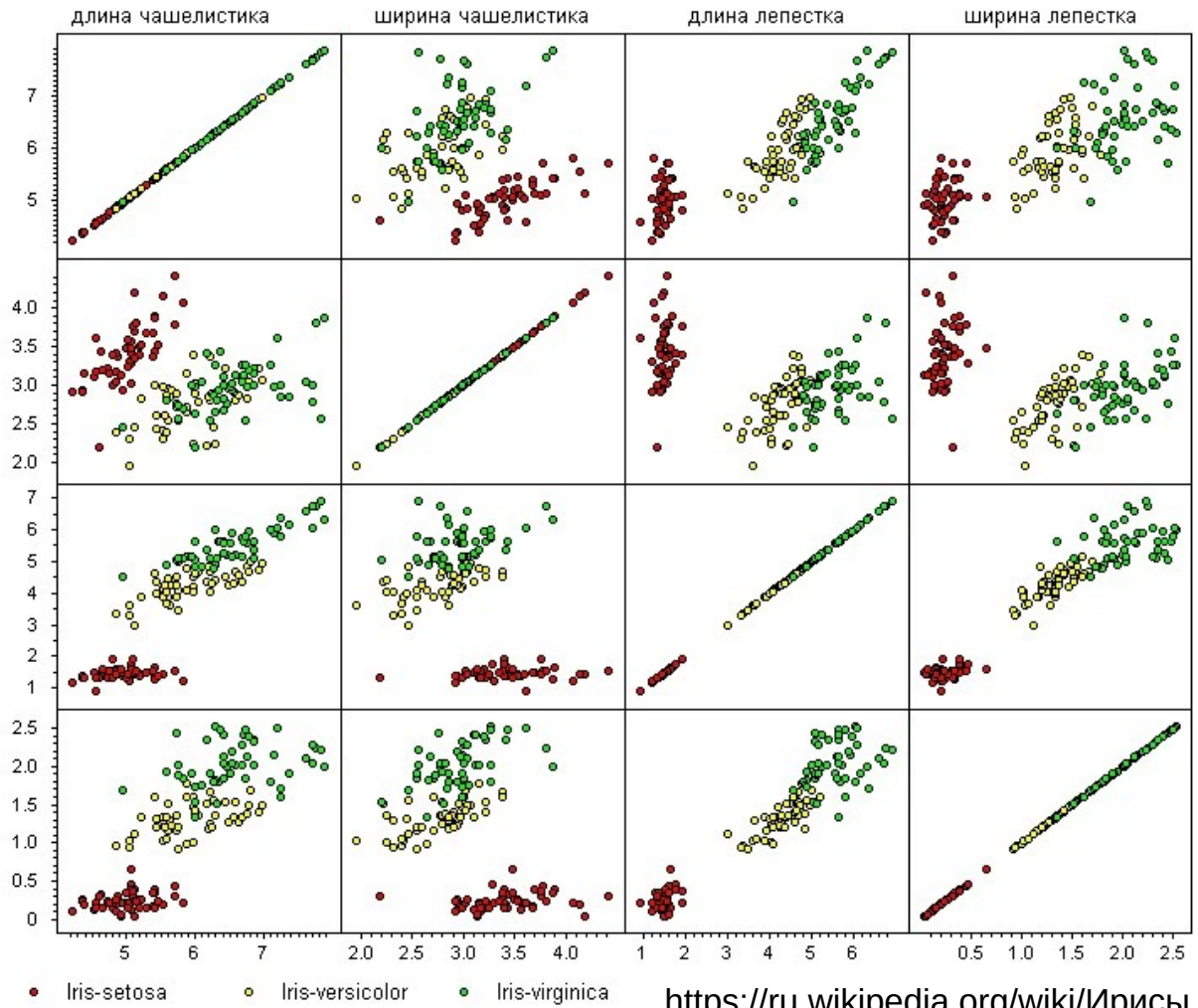
не выполнена: 

**Гипотеза компактности** (для классификации):

*близкие объекты, как правило, лежат в одном классе.*

выполнена: 

не выполнена: 




# Классификация ирисов Фишера

Как, воспользовавшись допущениями, которые мы сделали, решить задачу классификации/регрессии?

https://ru.wikipedia.org/wiki/Ирисы_Фишера?ysclid=mv17zm78sj268027885

Представим, что мы нашли неизвестный фрукт. Мы измерили:

- массу — 140 граммов;
- сладость — 7 баллов из 10;
- красноту — 8 баллов из 10.

Как понять, яблоко это или апельсин?

Один из самых естественных способов рассуждать так:

> **Посмотрим, на какие уже известные фрукты он больше всего похож.**

Если рядом с нашим объектом в основном находятся яблоки, назовём его яблоком. Если апельсины — апельсином.

> Именно эту идею использует алгоритм **K Nearest Neighbors**, или **KNN**, — алгоритм k ближайших соседей.

# Формализация понятия “близости”

1. Что значит «рядом»?
2. Как измерять расстояние между объектами?
3. Что делать, если признаки измеряются в разных единицах?
4. Сколько соседей нужно учитывать?

Компьютер не понимает слова «яблоко», «картинка» или «ученик». Объект нужно представить набором чисел — **признаками**, или **фичами**.

Например, ученика можно описать так:

$$x = (170,\ 15,\ 4)$$

где:

- 170 — рост в сантиметрах;
- 15 — возраст;
- 4 — число тренировок в неделю.

Такой объект можно воспринимать как точку в пространстве.

> Количество признаков называют **размерностью пространства**.

**Метри́ческое простра́нство** — множество вместе со способом измерения расстояния между его элементами — ***метрикой*** — двуместной числовой неотрицательной функцией на нём, отвечающей интуитивным представлениям о свойствах расстояния между точками➤.

Метрика - **способ измерения расстояния** между объектами.

## Определения [ править | править код ]

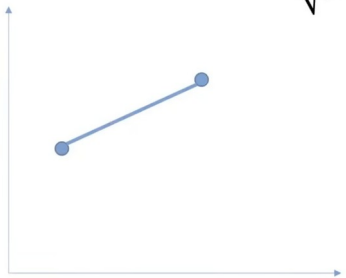
Пара $(M,\ d)$, состоящая из множества $M$ и функции $d\colon M \times M \to \mathbb{R}$, называется метрическим пространством, если[2]:

- $d(x, y) = 0 \Leftrightarrow x = y$ (*аксиома тождества*);
- $d(x, y) \geqslant 0$ (*аксиома положительности*);
- $d(x, y) = d(y, x)$ (*аксиома симметричности*);
- $d(x, z) \leqslant d(x, y) + d(y, z)$ (*аксиома треугольника* или неравенство треугольника).

Какое свойство лишнее??

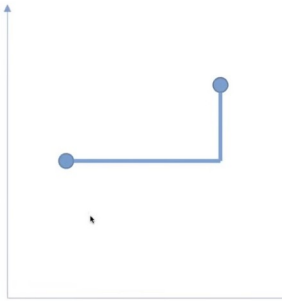
## Евклидова метрика

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



## Манхэттенская метрика

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



Эту метрику называют:

- манхэттенским расстоянием;
- расстоянием городских кварталов;
- $L_1$-расстоянием.

## Обобщение

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

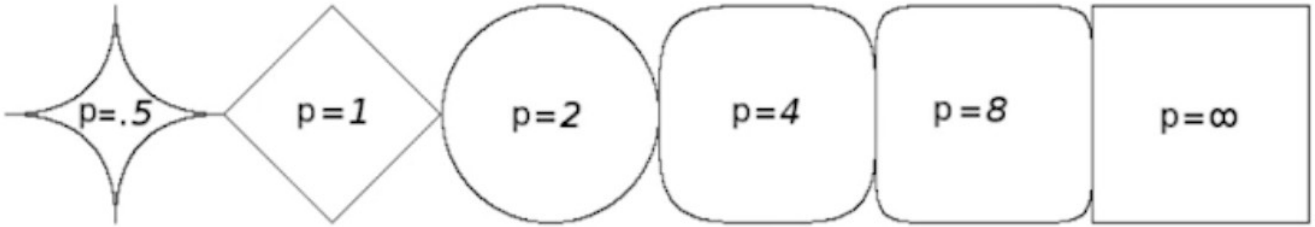
- Метрика Минковского
- Можно подбирать $p$ под конкретную задачу

Евклидова метрика и обобщённая метрика Минковского:

$$\rho(x, x_i) = \left( \sum_{j=1}^{n} \left| x^j - x_i^j \right|^2 \right)^{1/2} \qquad \rho(x, x_i) = \left( \sum_{j=1}^{n} w_j \left| x^j - x_i^j \right|^p \right)^{1/p}$$

$x = (x^1, \ldots, x^n)$ — вектор признаков объекта $x$,
$x_i = (x_i^1, \ldots, x_i^n)$ — вектор признаков объекта $x_i$,

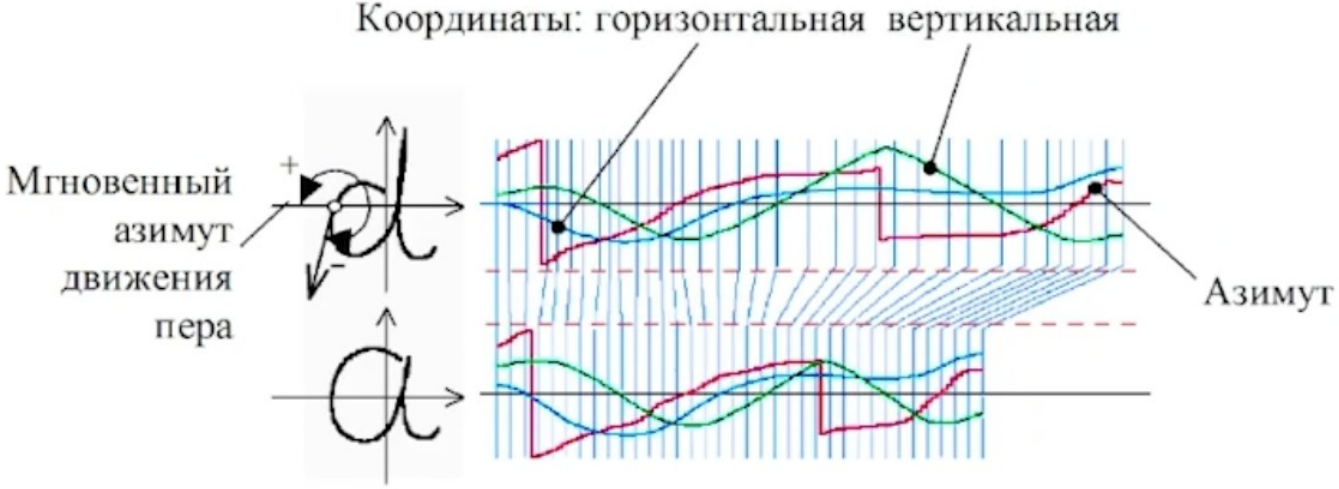
$w_1, \ldots, w_n$ — веса признаков, которые можно обучать.



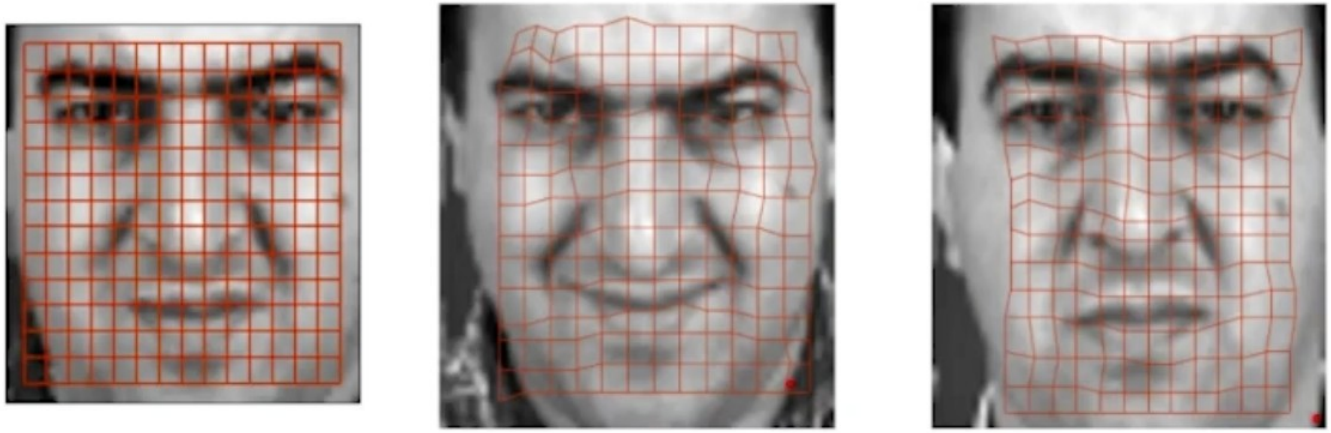
**Для строк** — редакторское расстояние Левенштейна:

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

**Для сигналов** — энергия сжатий и растяжений:



Расстояние **между изображениями** на основе выравнивания:



Оценивается энергия растяжения прямоугольной сетки

Как тут можно посчитать расстояние между двумя объектами?

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

Простейшая метрика: подсчёт различий

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

Как оценить качество модели?

# 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):

не переводить как точность

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

| $a(x)$ | $y$ |
|---|---|
| $-1$ | $-1$ |
| $+1$ | $+1$ |
| $-1$ | $-1$ |
| $\mathbf{+1}$ | $\mathbf{-1}$ |
| $+1$ | $+1$ |

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**

Тут цены ошибок не равнозначны из-за дисбаланса классов.

Это было для классификации, а если имеем модель регрессии?

Можно ли использовать ту же акураси?

Что еще?

# Возвращаемся к KNN

Пусть есть обучающие объекты с известными классами:

- красные точки — класс A;
- синие точки — класс B.

Появляется новая точка, класс которой неизвестен.

Алгоритм:

1. Посчитать расстояние от новой точки до всех обучающих объектов.
2. Выбрать $k$ объектов с наименьшим расстоянием.
3. Посмотреть на их классы.
4. Выбрать класс, который встречается чаще всего.

Например, при $k = 5$ среди соседей оказалось:

- 3 синих;
- 2 красных.

Предсказание KNN — синий класс.

Это вариант для классификации объекта, как звучит алгоритм для регрессии?

# Мы обсудили, как можно считать расстояние между объектами, но….

Рассмотрим два признака:

- масса фрукта — от 100 до 200 граммов;
- сладость — от 0 до 10 баллов.

Есть два известных объекта:

$$A = (100, 0), \qquad B = (200, 10)$$

И новый объект:

$$Q = (180, 1)$$

Посчитаем евклидово расстояние без нормализации.

До $A$:

$$d(Q, A) = \sqrt{(180 - 100)^2 + (1 - 0)^2} = \sqrt{6401} \approx 80$$

До $B$:

$$d(Q, B) = \sqrt{(180 - 200)^2 + (1 - 10)^2} = \sqrt{481} \approx 21.9$$

Получается, что $Q$ намного ближе к $B$.

Если заменить граммы килограммами, числа изменятся в тысячу раз, хотя сами фрукты останутся теми же. Предсказание алгоритма может измениться только из-за выбора единиц измерения.

# Нормализация признаков

Min-max нормализация

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Если в данных есть выброс, например фрукт массой 10 000 граммов, почти все обычные значения могут сжаться около нуля.

Стандартизация

$$x' = \frac{x - \mu}{\sigma}$$

где:

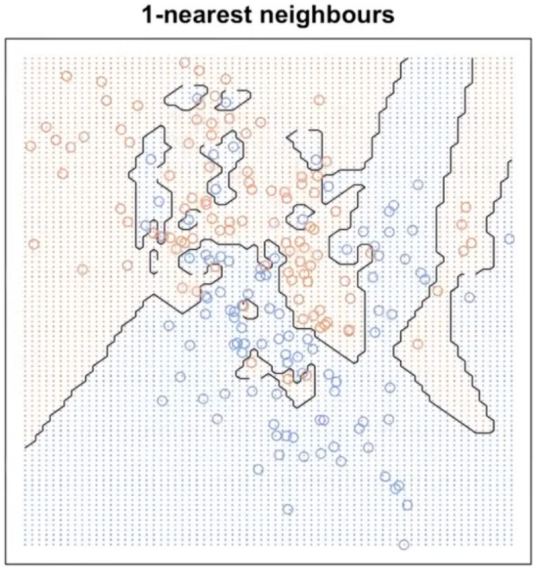
- $\mu$ — среднее значение признака;
- $\sigma$ — стандартное отклонение.

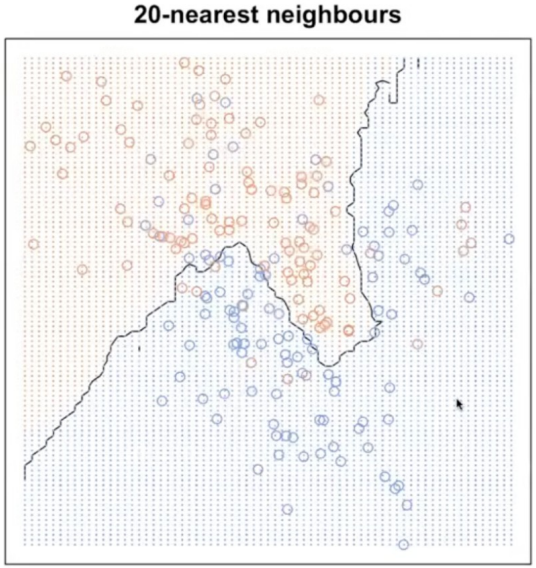
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

**Ошибка!!! Нормализовать весь датасет до разделения**

# Как выбрать k для KNN?

k — количество соседей, которые участвуют в голосовании.




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

С точки зрения качества на обучающей выборке лучший выбор $\boldsymbol{k = 1}$

Нельзя подбирать k по обучающей выборке — **гиперпараметр**

Нужно использовать дополнительные данные

# Обобщающая способность

Как готовиться к экзамену?

| Заучить все примеры с занятий | Разобраться в предмете и усвоить алгоритмы решения задач |
| --- | --- |

переобучение!

# Отложенная выборка

Как можно еще?

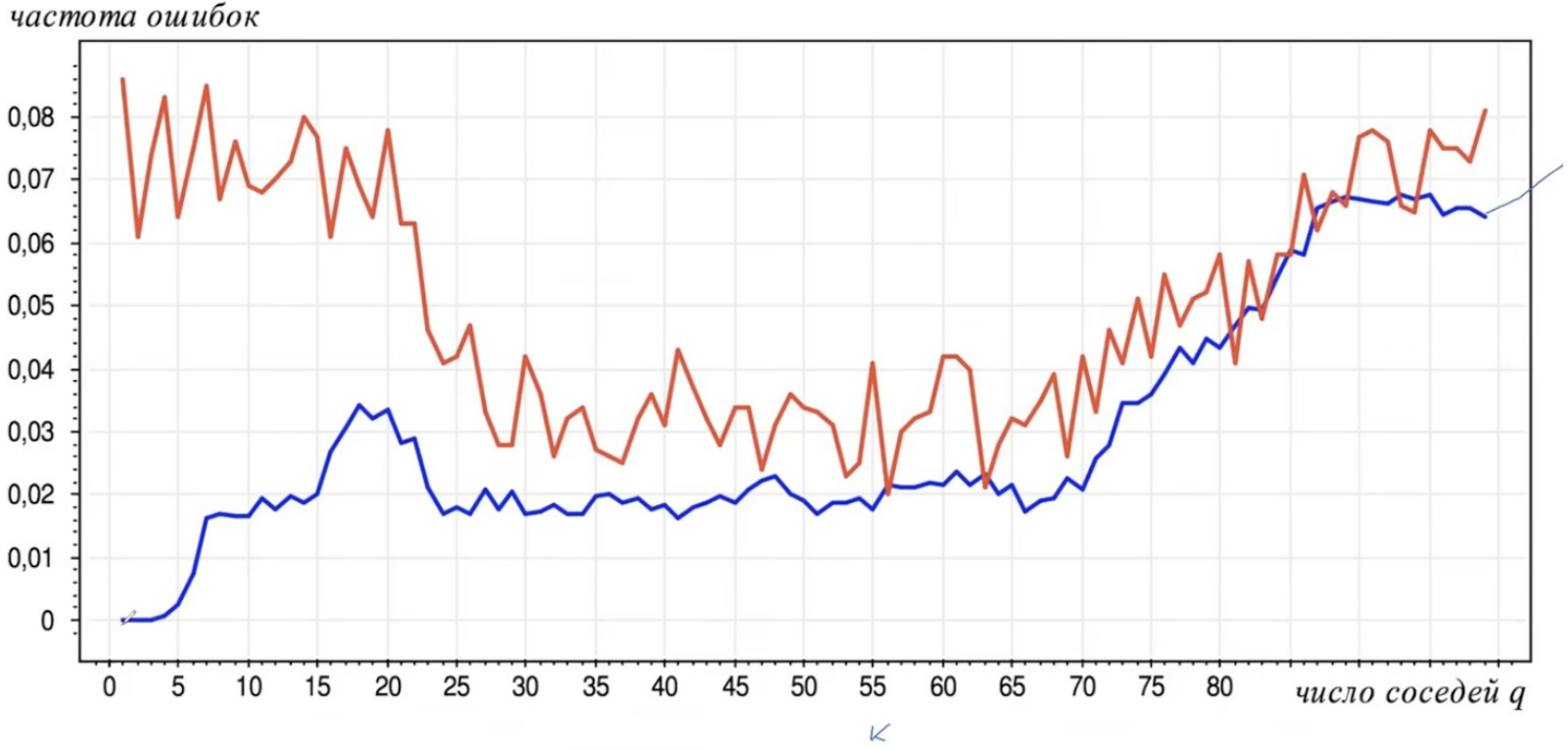


Обучение

Тест

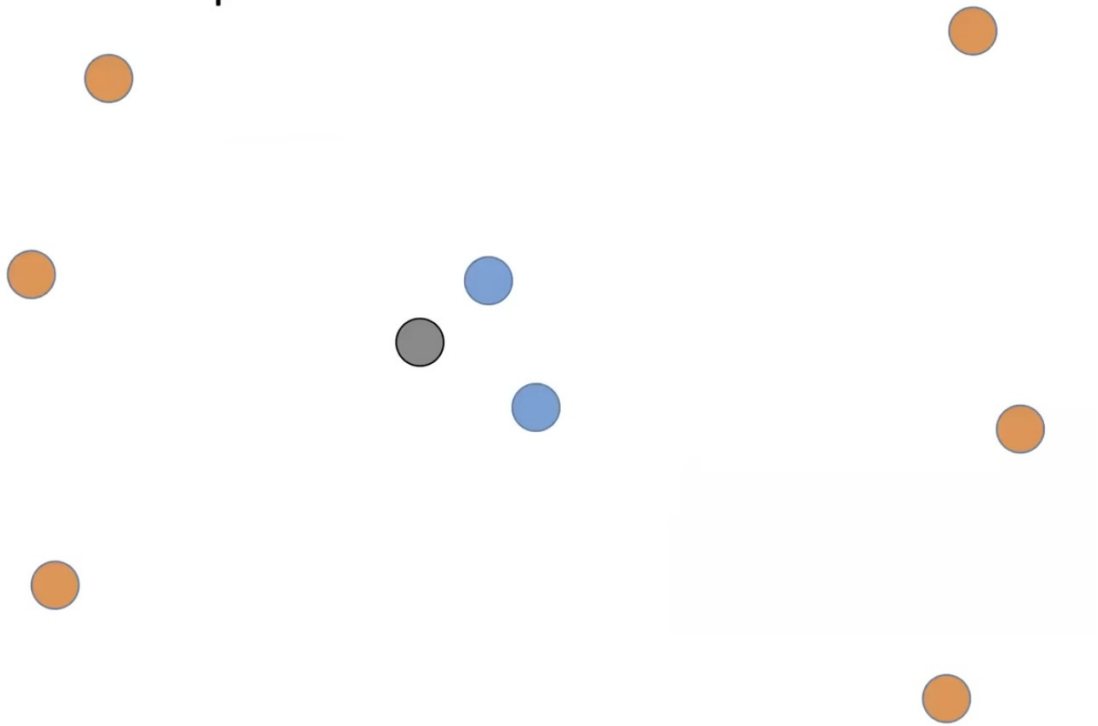


# Подбор числа соседей




http://www.machinelearning.ru/wiki/index.php?title=MO

# Проблема с расстояниями



# Взвешенный knn

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

Варианты:

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

# Откуда чуть-чуть тырились слайды

https://www.youtube.com/watch?v=UZB2t43B1dI лекция Евгения Соколова (ФКН ВШЭ)

https://www.youtube.com/watch?v=mCJcRDIY4TI лекция Воронцова (ШАД, МГУ)

Савватеев про метрики https://youtu.be/C6zYL6hGyp8?si=PD26SHSLzUqj7BNs

## Что на самом деле «обучает» KNN?

KNN почти ничего не вычисляет заранее. Он в основном запоминает обучающие объекты.

Поэтому его называют **ленивым алгоритмом**.

## Проблема высокой размерности

Представим, что у объектов не 2 признака, а 10 000.

В пространствах большой размерности происходит необычная вещь: почти все точки становятся далёкими друг от друга, а расстояния до ближайших и далёких объектов начинают отличаться не так сильно.

Это называют **проклятием размерности**.

Кроме того, ненужные признаки добавляют в расстояние шум.