

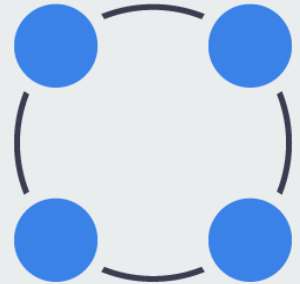


Машинное обучение

Лекция 1. Введение. k-NN

(05.02.2022)

Даниил Литвинов
Лаврентий Данилов



Машинное обучение



План курса



- Введение в машинное обучение. k-NN
- Логистическая регрессия
- Кластеризация
- Деревья решений
- Ансамбли
- Отбор и создание признаков
- Введение в нейронные сети
- Обучение нейронных сетей
- Сверточные нейронные сети
- Продвинутое обучение
- Transfer learning
- Сегментация. Детекция
- Explainability и interpretability
- Работа с последовательностями

Аттестация




Активности:

- Практические домашние работы на Python
- Прохождение онлайн курса на Stepik
- Участие в соревнованиях по анализу данных
- Квизы на лекциях **[опционально]**
- *Статья и разбор какой-то очень интересной темы*

Оценки:

- [65 - 70] % — 3
- (70 - 80] % — 4
- (80 - 100] % — 5

Правила сдачи и дедлайны

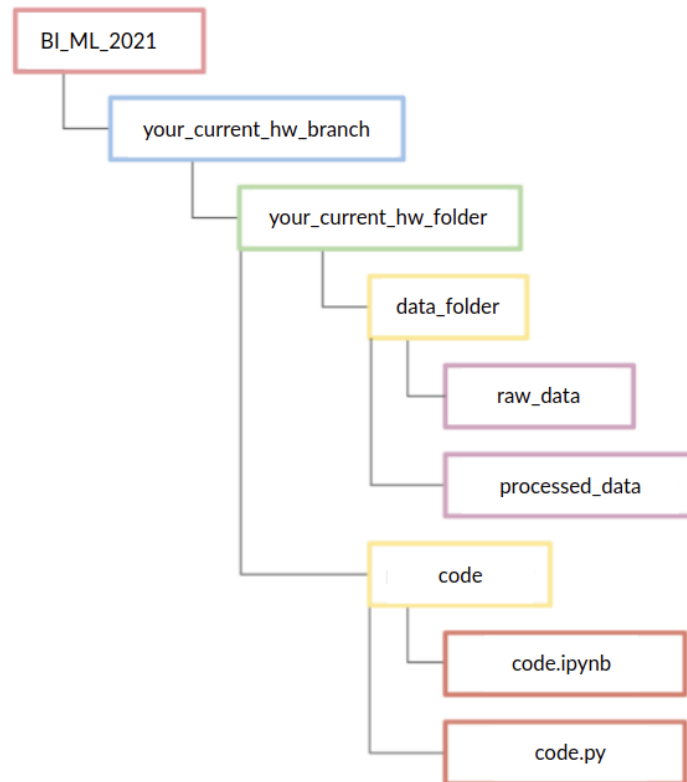


Дедлайны:

- 1 неделя (*ориентировочно*
старт в ВС 12:00 -> **конец** ВС 23:59) сдача в виде PR
- На соревнования примерно 1.5 месяца Kaggle

[notion](#) страница со всеми материалами

Репозиторий



Ближе к делу

Что хорошо бы знать

- Питон с его классами

```
my_string = "Hello, World!"  
print(my_string)
```

(а еще numpy и pandas)

- **Чуть-чуть** линейной алгебры

$$\begin{bmatrix} 2 & -3 & 1 \\ 5 & 4 & -2 \end{bmatrix} \times \begin{bmatrix} -7 & 5 \\ 2 & -1 \\ 4 & 3 \end{bmatrix} =$$

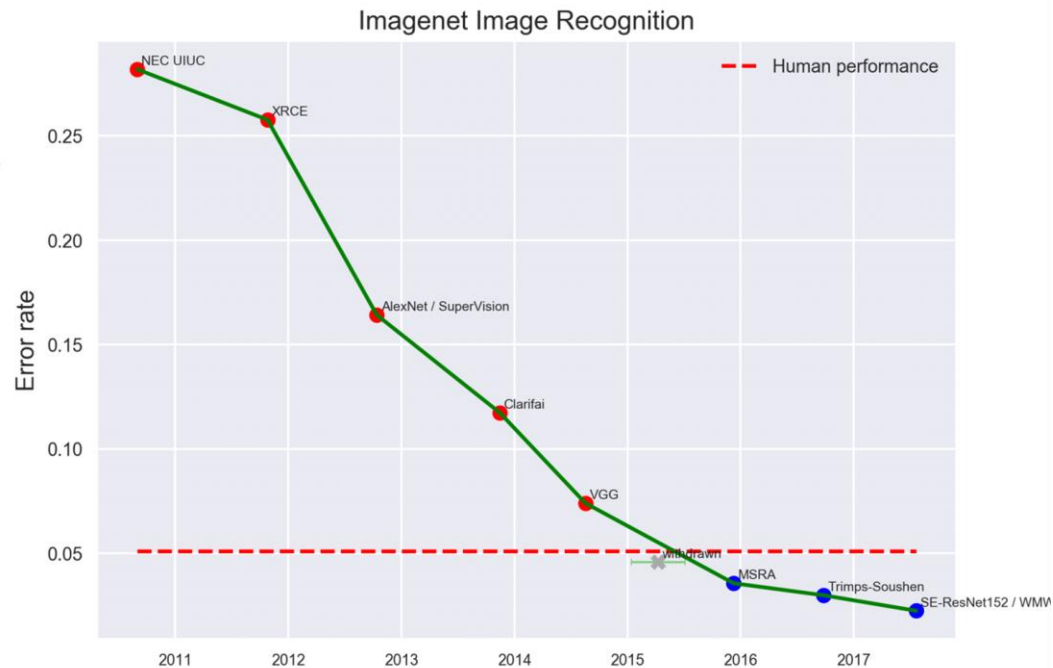
- **Чуть-чуть** матанализа

Например, градиент функции

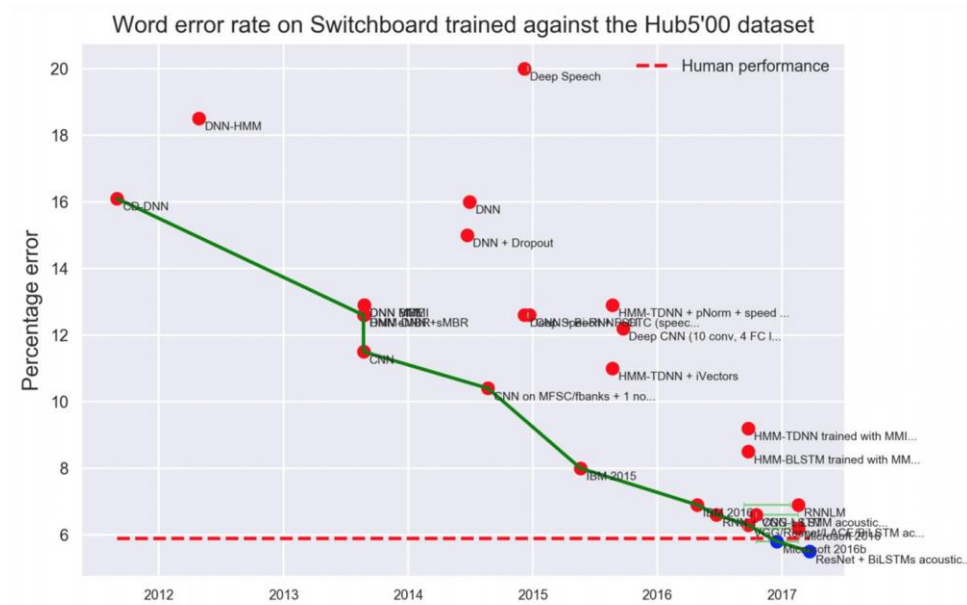
$\varphi(x, y, z) = 2x + 3y^2 - \sin z$ будет
представлять собой:

$$\nabla \varphi = \left(\frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y}, \frac{\partial \varphi}{\partial z} \right) = (2, 6y, -\cos z).$$

Распознавание изображений



Распознавание речи



Машинный перевод



Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, “Ngaje Ngai” in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

Kilimanjaro is a mountain of 19,710 feet covered with snow and is said to be the highest mountain in Africa. The summit of the west is called “Ngaje Ngai” in Masai, the house of God. Near the top of the west there is a dry and frozen dead body of leopard. No one has ever explained what leopard wanted at that altitude.

Веб-поиск



<https://habr.com/ru/company/yandex/blog/336094/>

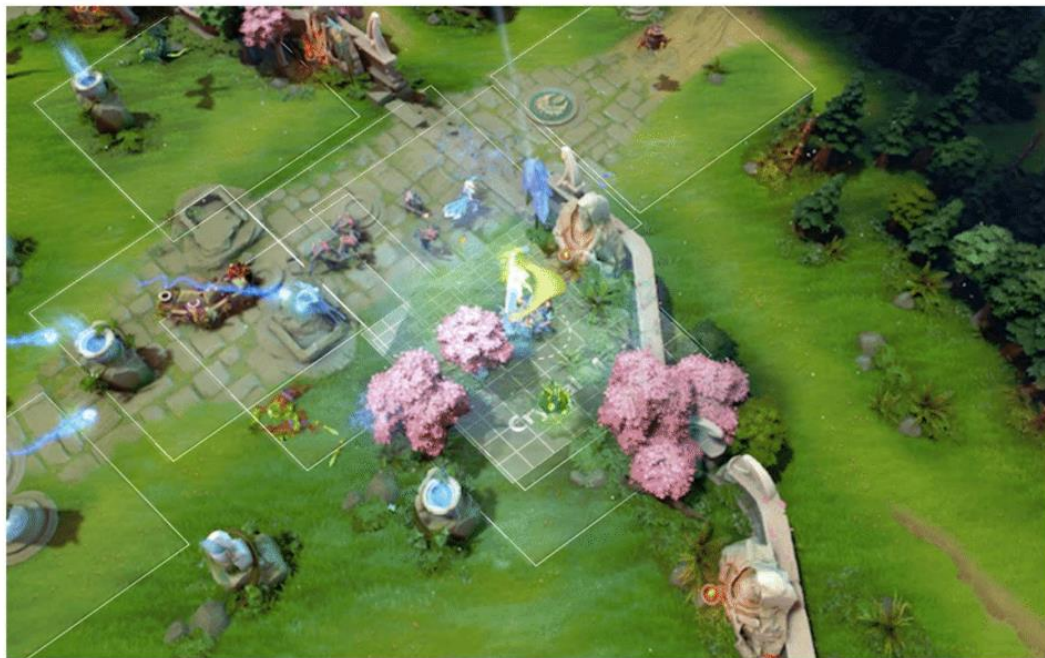
Создание нового



- [Выставка](#)
- [Кое-что интересное](#)
- [И еще](#)



Апогей



<https://openai.com/blog/openai-five/>

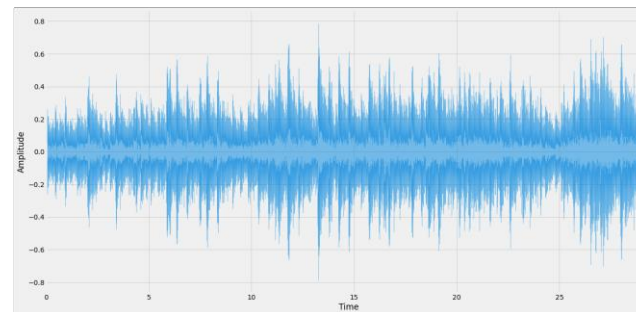
Что такое этот ваш ML?



Машинное обучение — это наука, изучающая алгоритмы, автоматически улучшающиеся благодаря опыту.

- Перевести текст с одного языка на другой
- Диагностировать болезнь по симптомам
- Сравнить, какой из двух документов в интернете лучше подходит под данный поисковый запрос
- Сказать, что изображено на картинке
- Оценить, по какой цене удастся продать квартиру

Данные



| | text | sentiment |
|---|---|-----------|
| 0 | For a movie that gets no respect there sure ar... | 0 |
| 1 | Bizarre horror movie filled with famous faces ... | 0 |
| 2 | A solid, if unremarkable film. Matthau, as Ein... | 0 |
| 3 | It's a strange feeling to sit alone in a theat... | 0 |
| 4 | You probably all already know this by now, but... | 0 |
| 5 | I saw the movie with two grown children. Altho... | 0 |
| 6 | You're using the IMDb. You've given some heft... | 0 |
| 7 | This was a good film with a powerful message o... | 0 |
| 8 | Made after QUARTET was, TRIO continued the qua... | 0 |
| 9 | For a mature man, to admit that he shed a tear... | 0 |

| | mean radius | mean texture | radius error | cancer |
|-----|-------------|--------------|--------------|--------|
| 334 | 12.30 | 19.02 | 0.1840 | 1 |
| 260 | 20.31 | 27.06 | 0.3977 | 0 |
| 396 | 13.51 | 18.89 | 0.2136 | 1 |
| 469 | 11.62 | 18.18 | 0.4101 | 1 |
| 264 | 17.19 | 22.07 | 0.4203 | 0 |
| 543 | 13.21 | 28.06 | 0.2351 | 1 |

Виды задач. Обучение с учителем. Supervised learning

X обученный - признаки $n \times d$ y - целевые переменные (target)

1) $y \in \mathbb{R}$ - регрессия
(0.2; 100)

2) $y = \{0, 1\}$ - бинарные кл.

3) $y = \{0, 1, \dots, k\}$ - многокл. кл.

4) $y = \{0, 1\}^k$ - multilabel кл.

5) y - укоресу. кл.
минимирование

| П. 344 | | | |
|----------|---|--|----------|
| пол. втр | | | пол. втр |
| x | 0 | | 0, 1, 2 |
| | 1 | | 3 |
| | 3 | | 10 |
| | 5 | | 1, 7 |

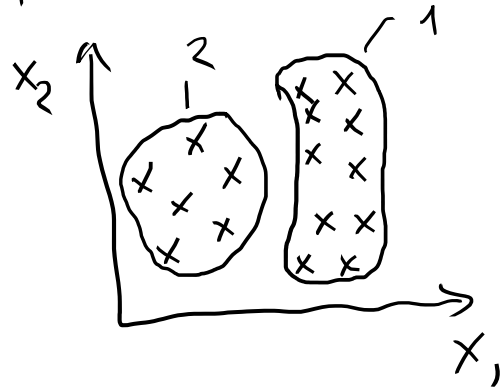
Виды задач. Обучение без учителя. Unsupervised learning

Есть множество X .

1) Сжатие разм. (PCA, tSNE, UMAP).

2) Ассоциации: OZON

3) Кластеризация:



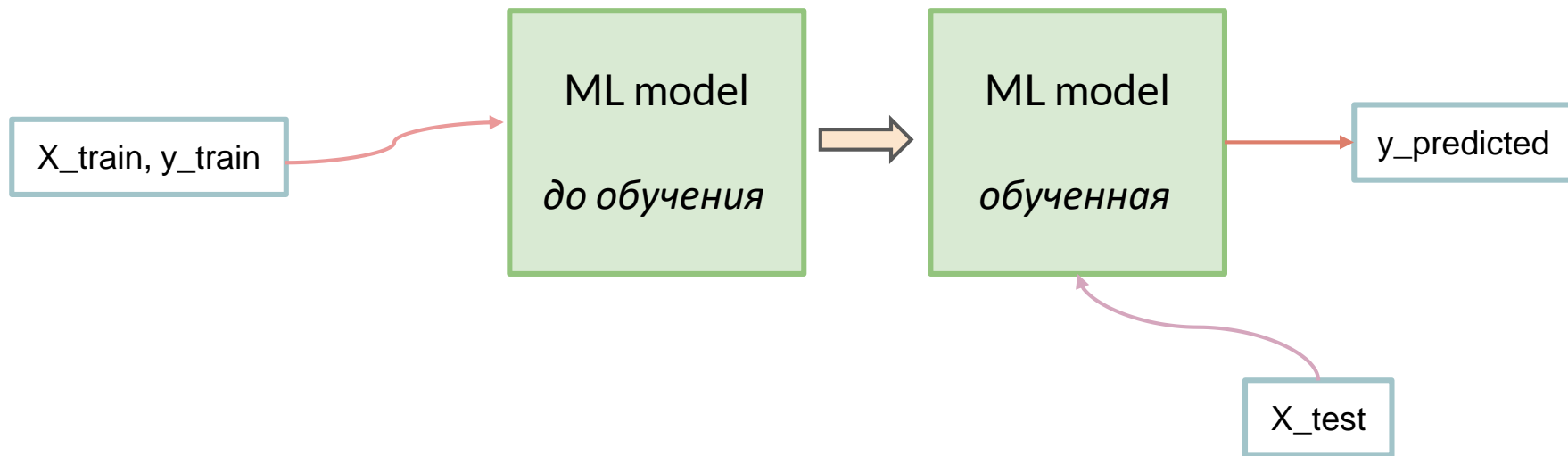
Вопрос

Определите, относятся ли следующие задачи к обучению с учителем, или без, или к чему-то другому?

1. Предсказание курса евро к доллару на следующий день
2. Стилизация текста. Например, детоксификация текста: «Ясно.»
→ «Ясно^^»
3. Поиск котов на изображениях
4. Поиск наборов продуктов, которые покупают посетители магазина
5. Обучение бота играть в компьютер



Что будет происходить?



А что если просто запомнить все ответы?



Перерыв \rightarrow k-NN

Алгоритм поиска ближайшего соседа (nearest neighbour)

Train:

просто все запомнить

Predict:

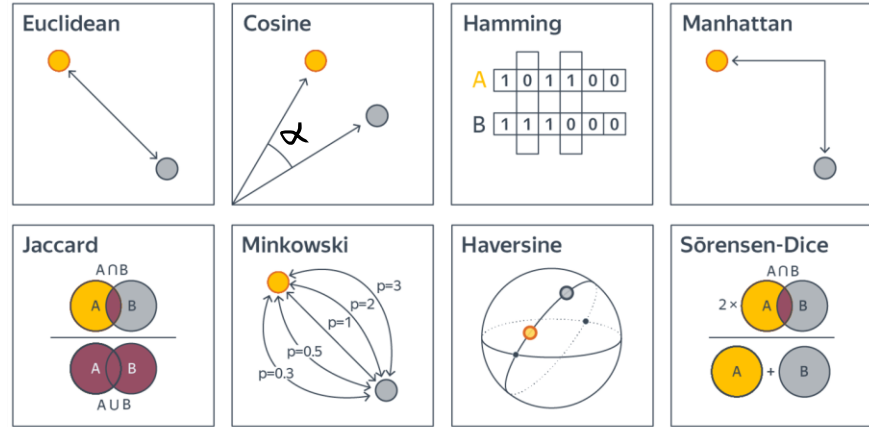
найти ближайший
и выдать его класс

$$L_2 = \sqrt{\sum_i (v_i - u_i)^2}$$

$$L_1 = \sum_i |v_i - u_i|$$

$v: [1.5, 2, 3]$
 $u: [2, 7, 10]$

$$(1.5 - 2)^2 + (2 - 7)^2 + \dots$$



$$\left[1 - \cos \alpha \right] - \cos \text{dist}$$
$$\left(\sum_{i=1}^N |v_i - u_i|^p \right)^{1/p}$$

Minkowski

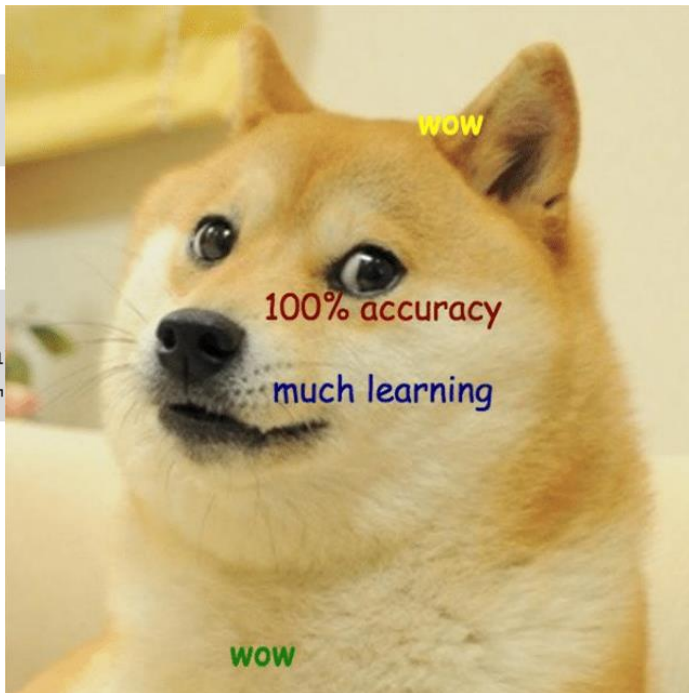
Алгоритм поиска ближайшего соседа (nearest neighbour)

Train:

просто все

Predict:

найти ближайшего
и выдать его



Как сделать лучше?



Как сделать лучше? → k-ближайших соседей (k-NN)

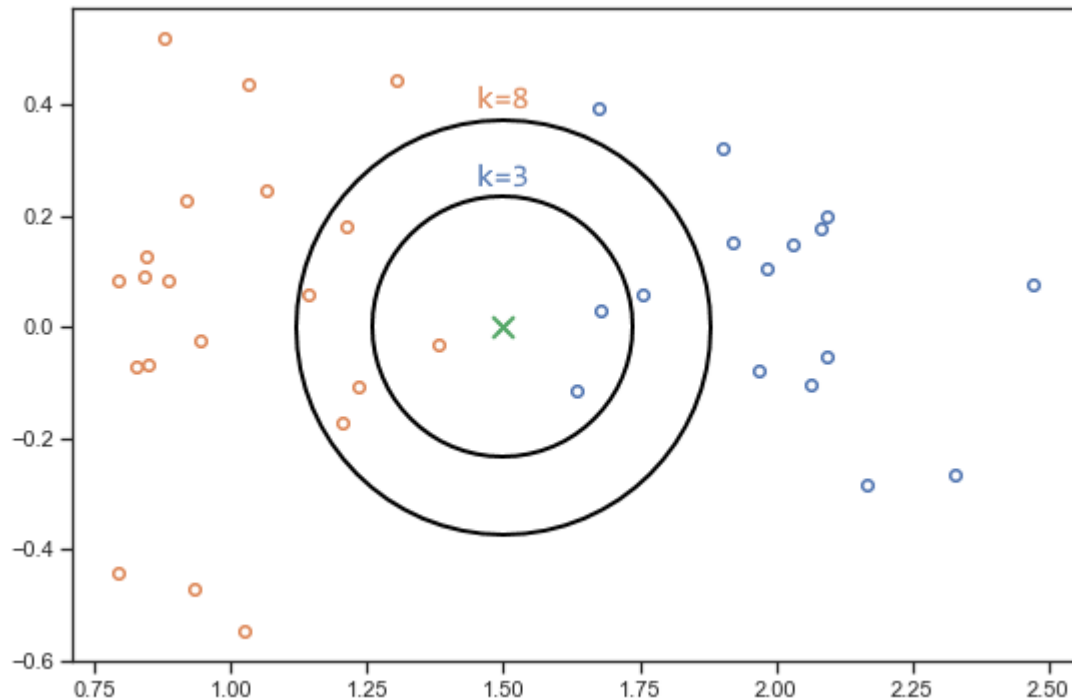
Плюсы:

- Простой
- Если подходящая задача, то работает хорошо

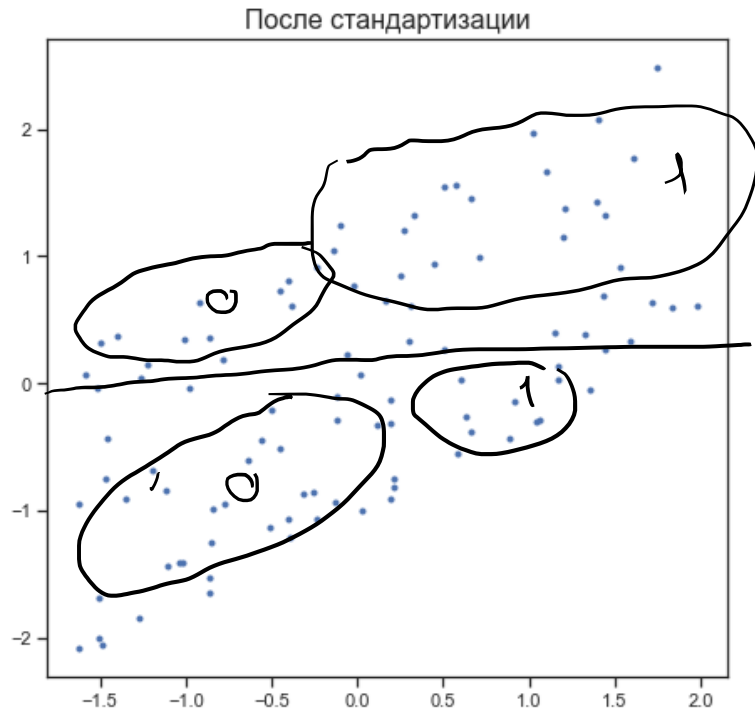
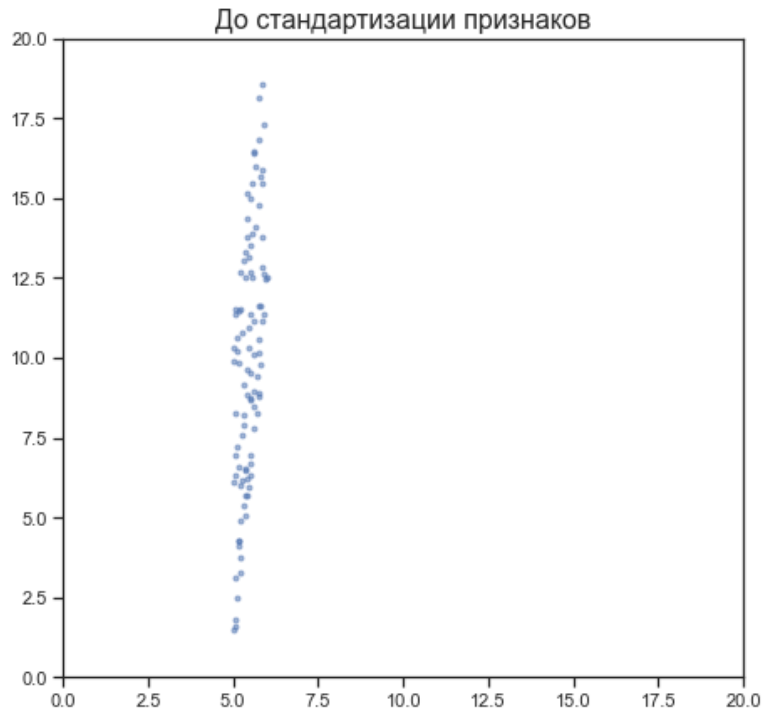
Минусы:

- Долгий
- Нужны **хорошие** признаки

Гиперпараметры



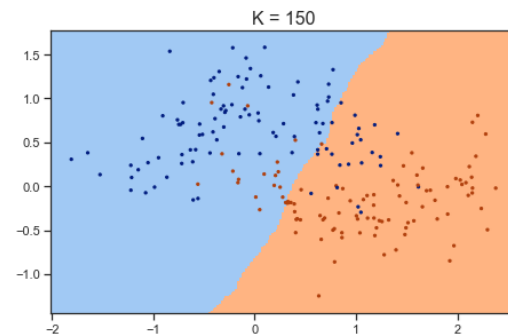
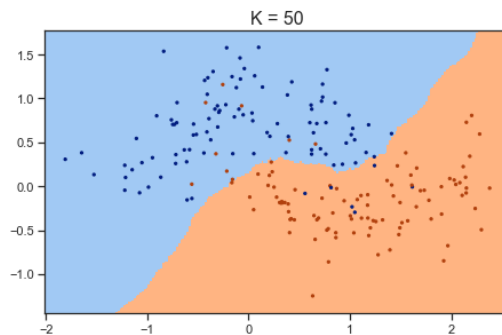
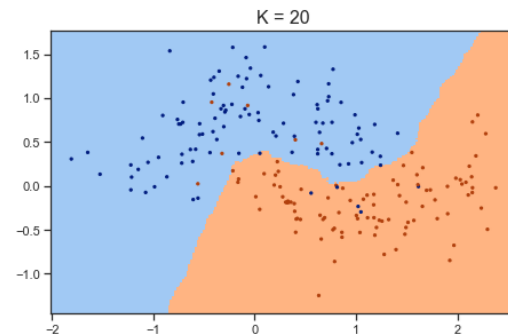
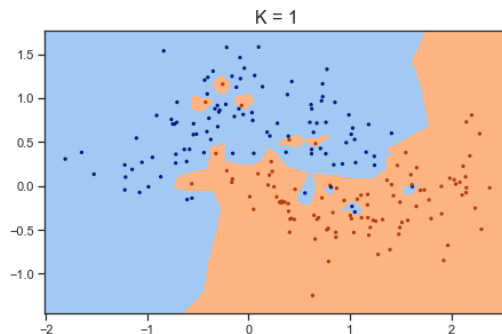
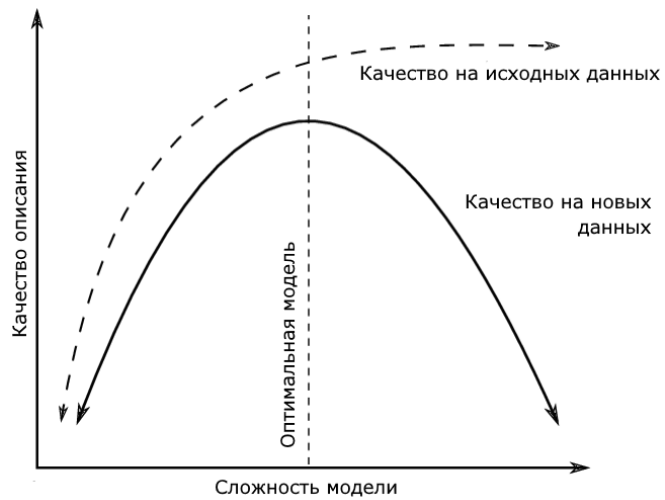
Предобработка данных



Сложность моделей



Сложность моделей



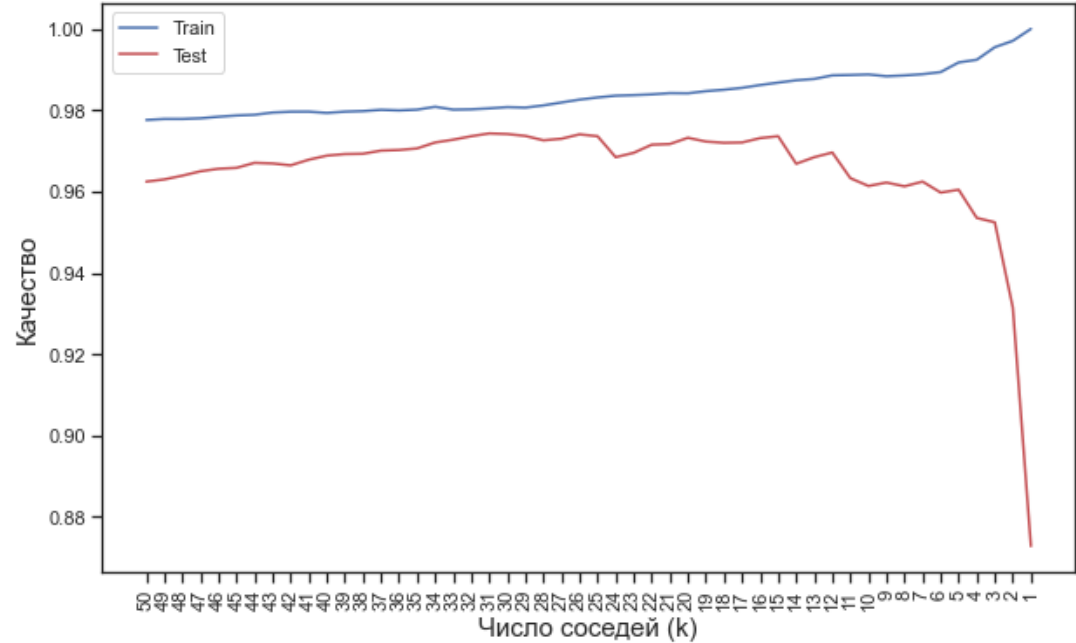
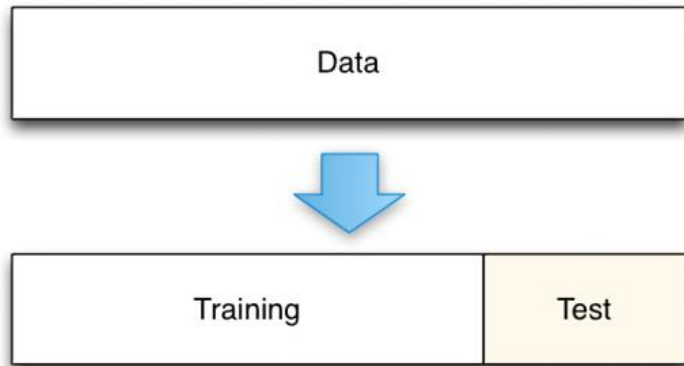
Классификация и регрессия



Как выбрать k ?



Как выбрать k? train-test split



Метрики качества классификации



$$y: \{0, 1\}$$

| | | |
|---|----|----|
| | 1 | 0 |
| 1 | TP | FN |
| 0 | FP | TN |

y-true

y-pred
↑

confusion matrix

- Обычный год в обычном терапевтическом отделении обычной больницы

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Определение болезни, которая жутким клеймом падёт на каждого, кому поставили такой диагноз

$$\text{precision} = \frac{TP}{TP + FP}$$

- Определение опасной и очень заразной болезни

$$\text{recall} = \frac{TP}{TP + FN}$$

$$f_1 = 2 \cdot \frac{pr \cdot rc}{pr + rc}$$

f_β

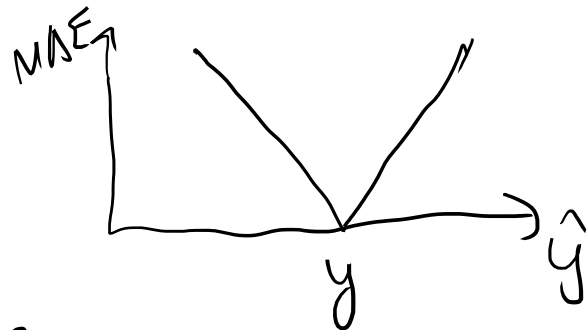
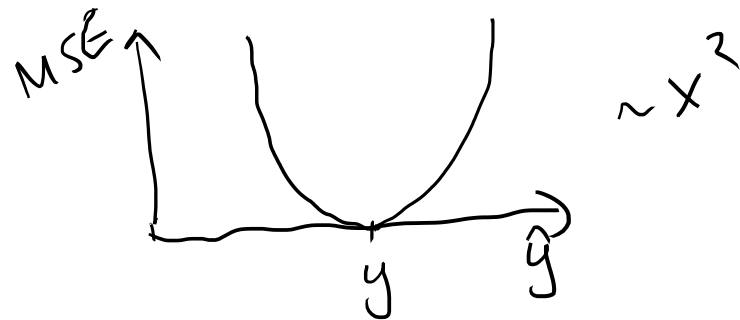
Метрики качества регрессии

MSE, MAE, R²...

$$1. \text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$2. \text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

$$3. R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$



$$R^2: (-\infty; 1]$$