

Лекция 1

Обучение и валидация модели

Кантонистова Елена Олеговна

elena.kantonistova@yandex.ru

ekantonistova@hse.ru

ВШЭ, 2021

ФУНКЦИЯ ПОТЕРЬ

Функция потерь – функция, измеряющая ошибку на одном объекте.

- Пусть y – истинный ответ на объекте x
- $a(x)$ – предсказание алгоритма на объекте x

Как измерить ошибку предсказания?

ФУНКЦИЯ ПОТЕРЬ

Функция потерь – функция, измеряющая ошибку на одном объекте.

- Пусть y – истинный ответ на объекте x
- $a(x)$ – предсказание алгоритма на объекте x

Как измерить ошибку предсказания?

Пример (квадратичная функция потерь):

$$L(y, a(x)) = (a(x) - y)^2$$

ФУНКЦИОНАЛ ОШИБКИ

- Как измерить ошибку алгоритма на всех объектах выборки?

ФУНКЦИОНАЛ ОШИБКИ

- Как измерить ошибку алгоритма на всех объектах выборки?

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

ФУНКЦИОНАЛ ОШИБКИ

- Как измерить ошибку алгоритма на всех объектах выборки?

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2$$

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i – истинные ответы

ФУНКЦИОНАЛ ОШИБКИ

Функционал ошибки – функционал, измеряющий качество работы алгоритма.

Пример (среднеквадратичная ошибка, MSE):

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2 \rightarrow \text{min}$$

X – объекты, l – количество объектов

a – алгоритм, $a(x_i)$ – ответ алгоритма на объекте x_i

y_i – истинные ответы

При обучении алгоритма мы минимизируем функционал ошибки.

ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* у по его *площади* (x_1) и *количеству комнат* (x_2).



ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* у по его *площади* (x_1) и *количеству комнат* (x_2).

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A .



ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* у по его *площади (x_1)* и *количеству комнат (x_2)*.

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A .

Используем линейную модель для предсказания стоимости.

Она будет выглядеть так:

$$a(x) = w_0 + w_1x_1 + w_2x_2,$$

где w_0, w_1, w_2 -

параметры модели (*веса*).



ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* у по его *площади (x_1)* и *количеству комнат (x_2)*.

Как правило, алгоритм $a(x)$ выбирают из некоторого семейства алгоритмов A .

Используем линейную модель для предсказания стоимости.

Она будет выглядеть так:

$$a(x) = w_0 + w_1x_1 + w_2x_2,$$

где w_0, w_1, w_2 -

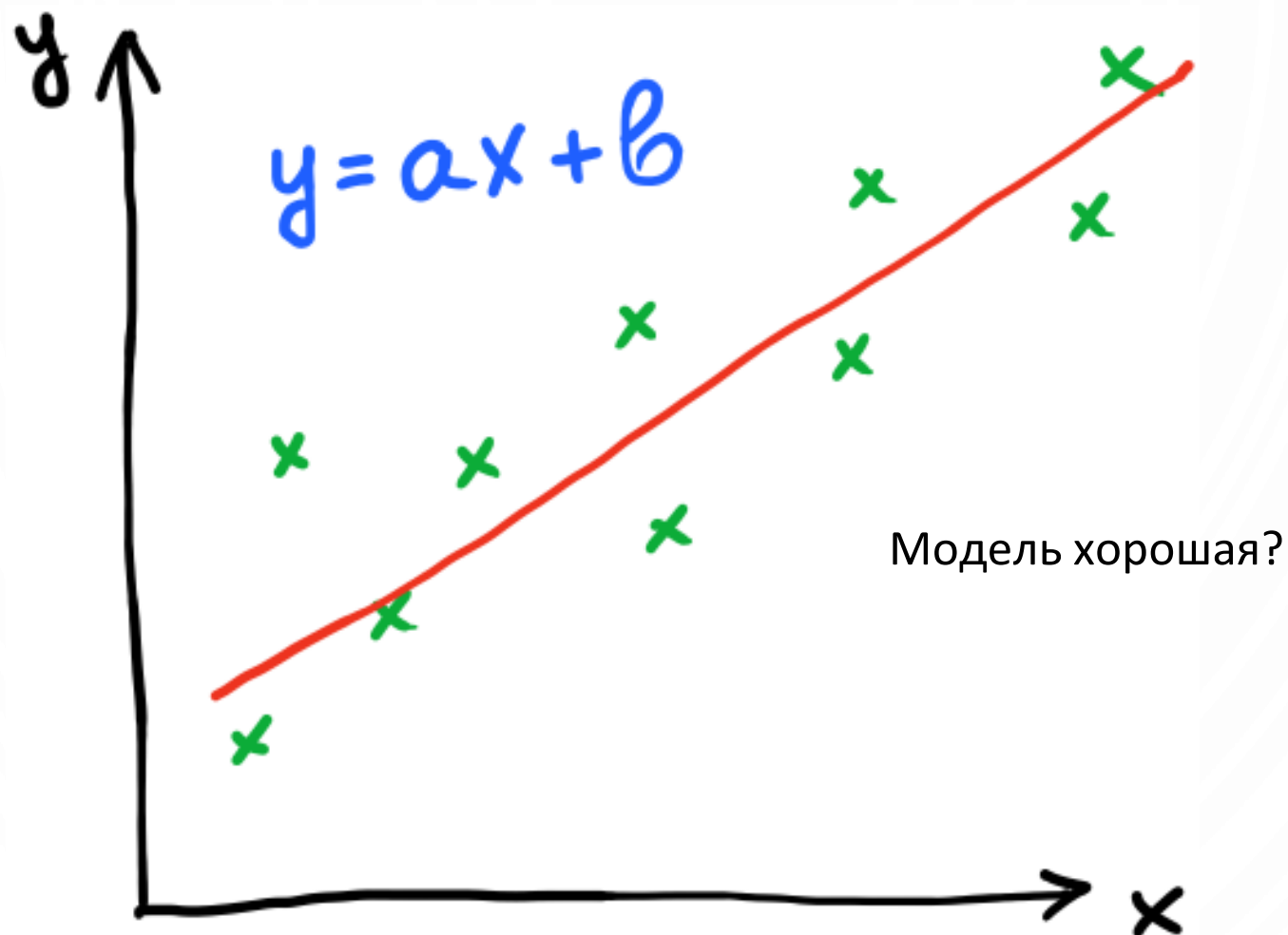
параметры модели (*веса*).

Общий вид линейных моделей:

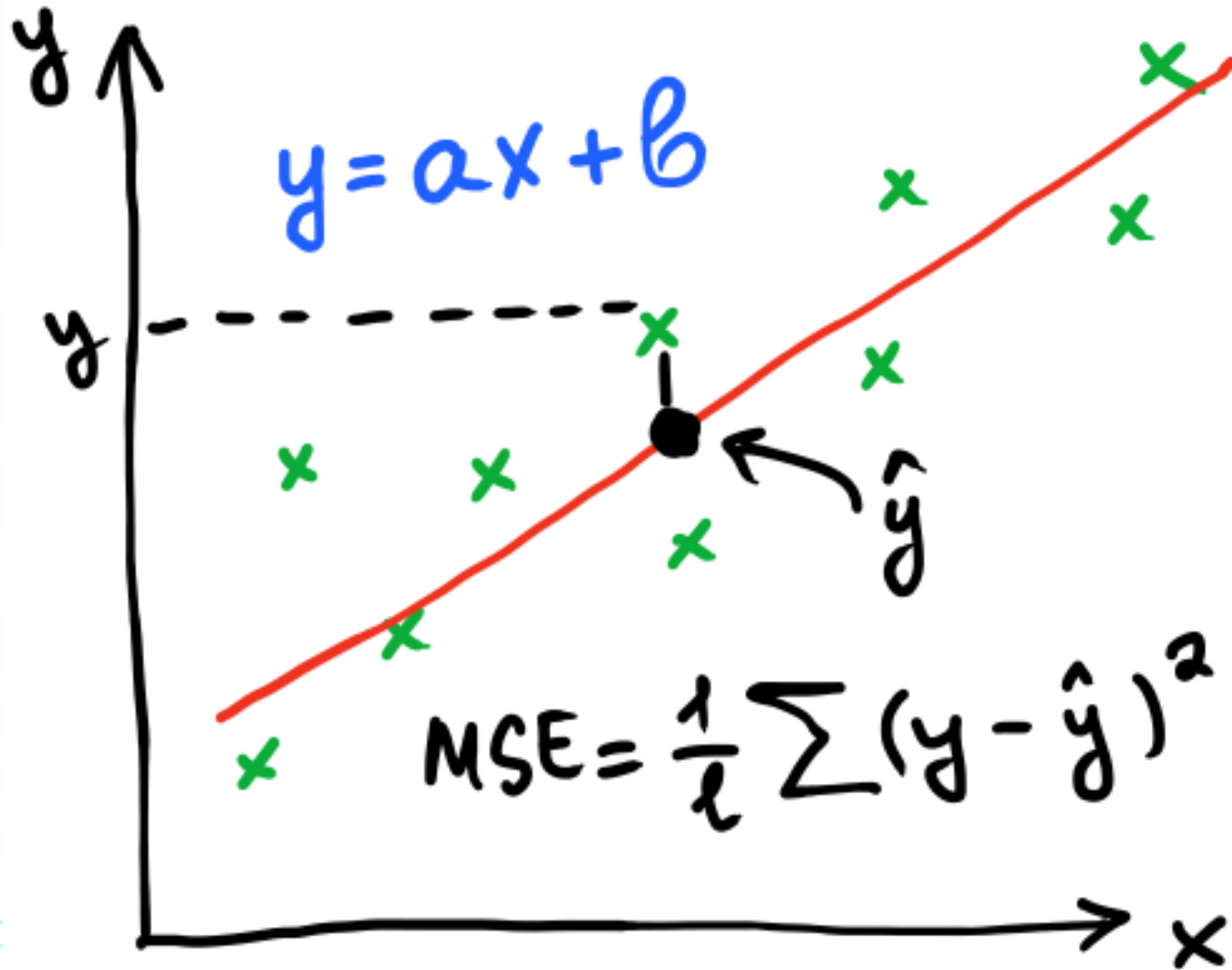
$$A = \{a(x) = w_0 + w_1x_1 + \dots + w_dx_d | w_0, w_1, \dots, w_d \in \mathbb{R}\}$$



ОБУЧЕНИЕ АЛГОРИТМА



ОБУЧЕНИЕ АЛГОРИТМА



ОБУЧЕНИЕ АЛГОРИТМА

Пример (семейство линейных моделей):

$$A = \{a(x) = w_0 + w_1x_1 + \dots + w_dx_d \mid w_0, w_1, \dots, w_d \in \mathbb{R}\}$$

Функционал ошибки:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2$$

ОБУЧЕНИЕ АЛГОРИТМА

Пример (семейство линейных моделей):

$$A = \{a(x) = w_0 + w_1x_1 + \dots + w_dx_d \mid w_0, w_1, \dots, w_d \in \mathbb{R}\}$$

Функционал ошибки:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2$$

Функционал ошибки для линейной модели стоимости дома:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (w_0 + w_1x_1 + w_2x_2 - y_i)^2$$

ОБУЧЕНИЕ АЛГОРИТМА

Параметры w_0, w_1, w_2 подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке):

Функционал ошибки для линейной модели стоимости дома:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (w_0 + w_1 x_1 + w_2 x_2 - y_i)^2 \rightarrow \min_{w_0, w_1, w_2}$$

ОБУЧЕНИЕ АЛГОРИТМА (ОБЩИЙ ВИД ЛИНЕЙНОЙ РЕГРЕССИИ)

Параметры w_0, \dots, w_n подбираются так, чтобы на них достигался минимум функции потерь (на обучающей выборке):

Функционал ошибки для линейной модели:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l \left(w_0 + \sum_{j=1}^d w_j x_{ij} - y_i \right)^2 \rightarrow \min_{w_0, \dots, w_d}$$

ОБУЧЕНИЕ АЛГОРИТМА

Процесс поиска оптимального алгоритма
(оптимального набора параметров или *весов*)
называется **обучением**.

МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются *метрики качества*.

МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются *метрики качества*.

Примеры:

- Среднеквадратичная ошибка – для регрессии

$$MSE(a, X) = \frac{1}{l} \sum_{i=1}^l (a(x_i) - y_i)^2$$

МЕТРИКИ КАЧЕСТВА

В задачах машинного обучения для оценки качества моделей и сравнения различных алгоритмов используются ***метрики качества***.

Примеры:

- Среднеквадратичная ошибка – для регрессии
- **Доля правильных ответов** – для классификации

$$\text{accuracy}(a, X) = \frac{1}{l} \sum_{i=1}^l [a(x_i) = y_i]$$

ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

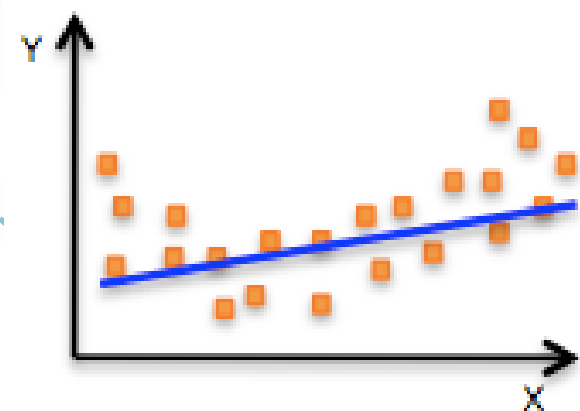
- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).



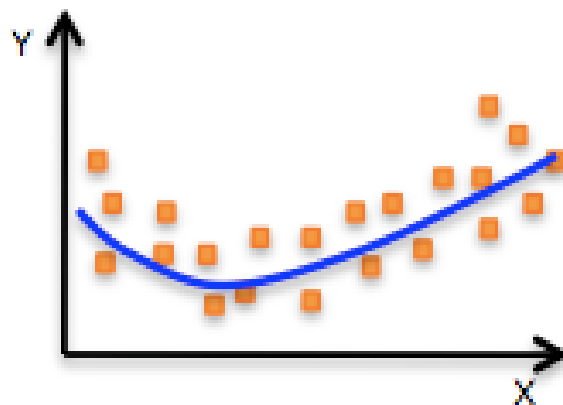
ОТЛОЖЕННАЯ ВЫБОРКА

- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).
- Тогда можно измерить качество построенной модели на отложенной выборке и оценить ее предсказательную силу.

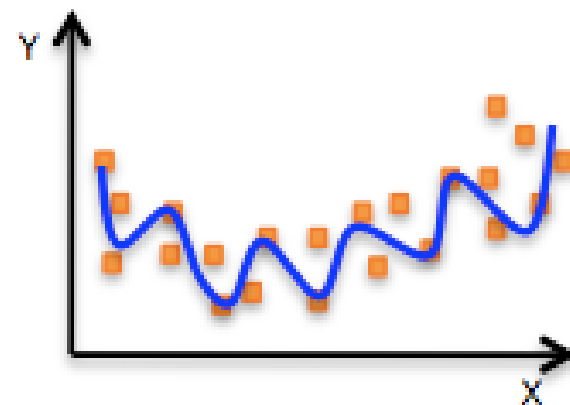
ПЕРЕОБУЧЕНИЕ И НЕДООБУЧЕНИЕ



Underfitting



Just right!

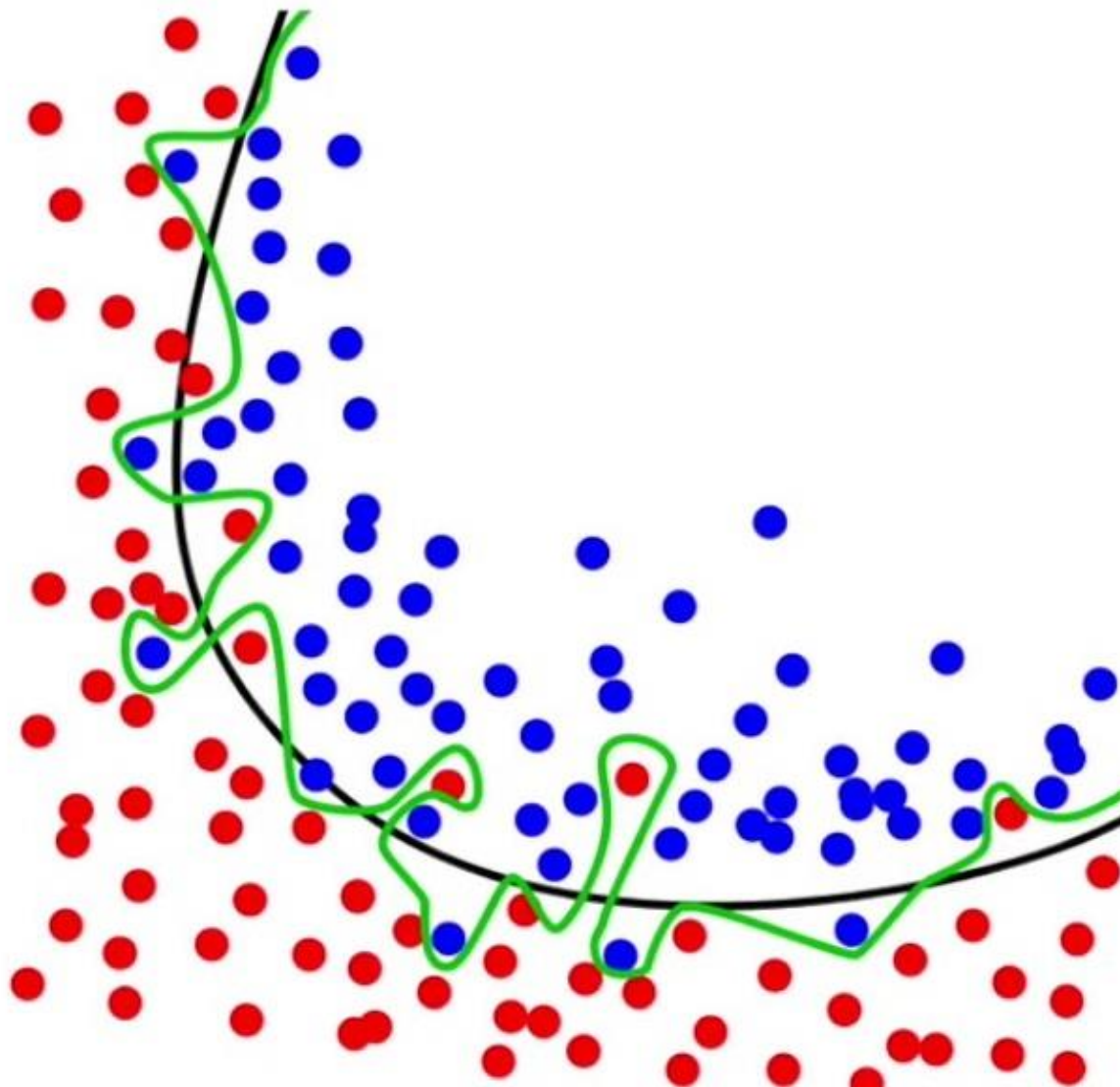


overfitting

ИЗ-ЗА ЧЕГО ВОЗНИКАЕТ ПЕРЕОБУЧЕНИЕ

- Избыточная сложность пространства параметров Ω , лишние степени свободы в модели $a(x, w)$ “тратятся” на чрезмерно точную подгонку под обучающую выборку.
- Переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке.

ПРИМЕР ПЕРЕОБУЧЕНИЯ В ЗАДАЧЕ КЛАССИФИКАЦИИ



ПРИЗНАК ПЕРЕОБУЧЕНИЯ

- *Если качество на отложенной выборке сильно ниже качества на обучающих данных, то происходит переобучение*



АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи






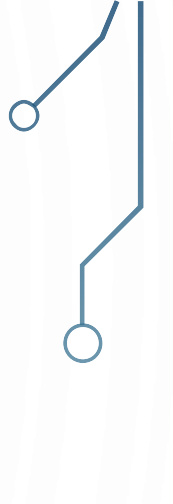
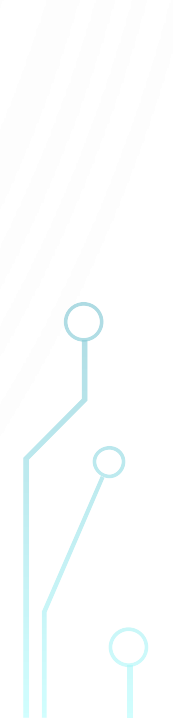
АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи

2. Выделение признаков


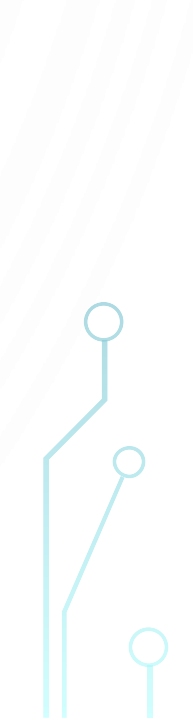


АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
 2. Выделение признаков
 3. Формирование выборки
- 
- 
- 


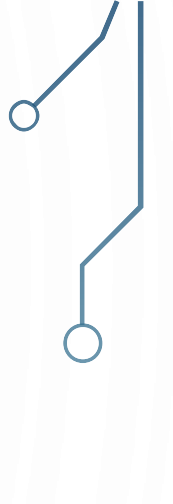
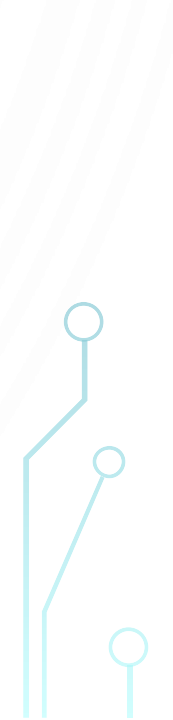


АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
 2. Выделение признаков
 3. Формирование выборки
 4. Выбор функции потерь и метрики качества
- 
- 


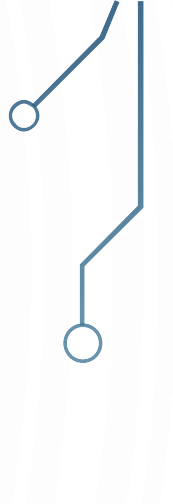
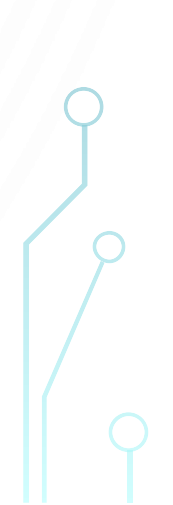


АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
 2. Выделение признаков
 3. Формирование выборки
 4. Выбор функции потерь и метрики качества
 5. Предобработка данных
- 
- 
- 

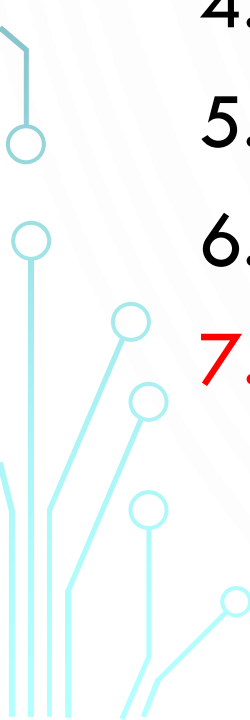
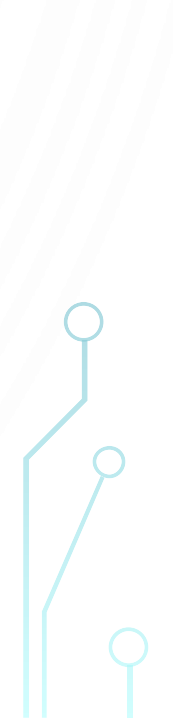


АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
 2. Выделение признаков
 3. Формирование выборки
 4. Выбор функции потерь и метрики качества
 5. Предобработка данных
 6. Построение модели
- 
- 
- 



АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

1. Постановка задачи
 2. Выделение признаков
 3. Формирование выборки
 4. Выбор функции потерь и метрики качества
 5. Предобработка данных
 6. Построение модели
 7. Оценивание качества модели
- 
- 

СТАДИИ РАЗРАБОТКИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

