



Лекция 1

Введение в машинное обучение

Кантонистова Елена Олеговна

elena.kantonistova@yandex.ru

ekantonistova@hse.ru

ВШЭ, 2020

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД – ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО
ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД – ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ – ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД – ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ – ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ – РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД – ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ – ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ – РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта

с помощью правил

Слишком много ручного труда

для создания системы



ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

ИСТОРИЯ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

1956 ГОД — ПЕРВЫЙ СЕМИНАР ПО ПРОБЛЕМАМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Задача: моделирование интеллекта человека математическими методами

50-Е – 70-Е ГОДЫ — ПРОСТЕЙШИЕ СИСТЕМЫ ИИ

Системы дедукции для доказательства теорем

Робот-психотерапевт ELIZA

80-Е ГОДЫ — РАЗВИТИЕ ЭКСПЕРТНЫХ СИСТЕМ

Моделирование работы эксперта с помощью правил

Слишком много ручного труда для создания системы

90-Е ГОДЫ — РАЗВИТИЕ МАШИННОГО ОБУЧЕНИЯ КАК ОБЛАСТИ ИИ

Нейронные сети

Генетические алгоритмы

Автоматический поиск сложных закономерностей

НАЧАЛО 21 ВЕКА — ГЛУБИННОЕ ОБУЧЕНИЕ (DEEP LEARNING)

Решение сложных задач распознавания с точностью, близкой к человеку

ЧТО ТАКОЕ МАШИННОЕ ОБУЧЕНИЕ

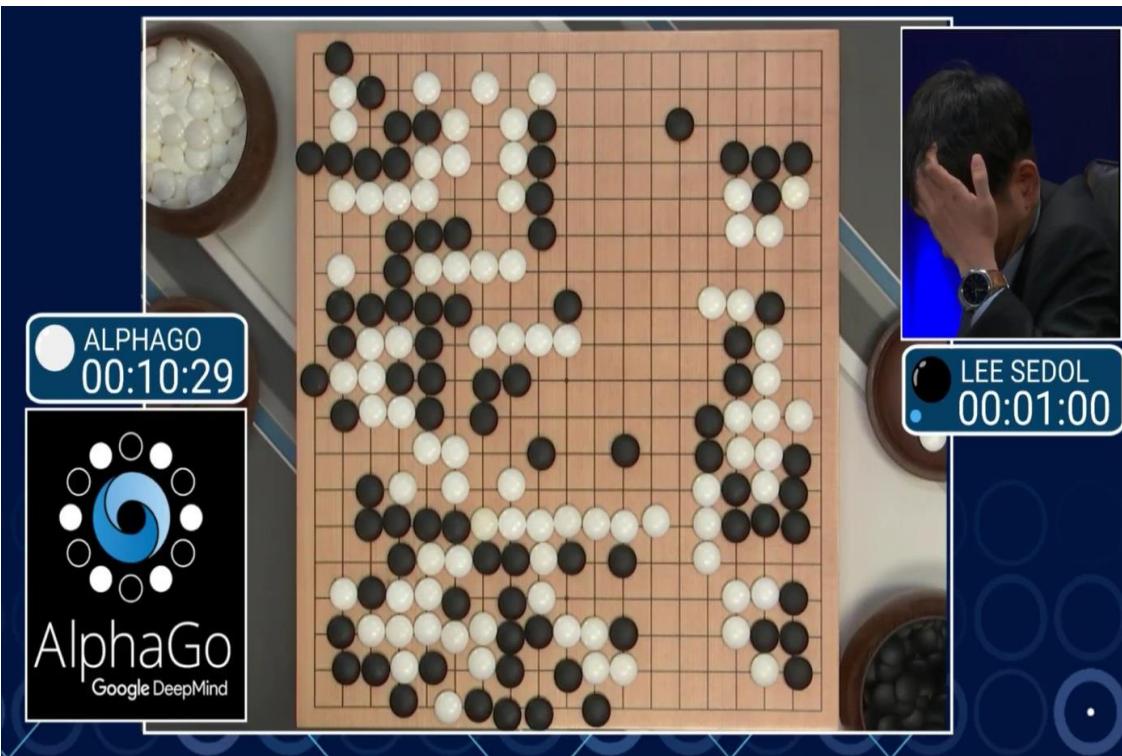
Машинное обучение – набор способов воспроизведения связей между событиями и результатом.

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных учиться.

Machine learning – the field of study that gives computers the ability to learn without being explicitly programmed.

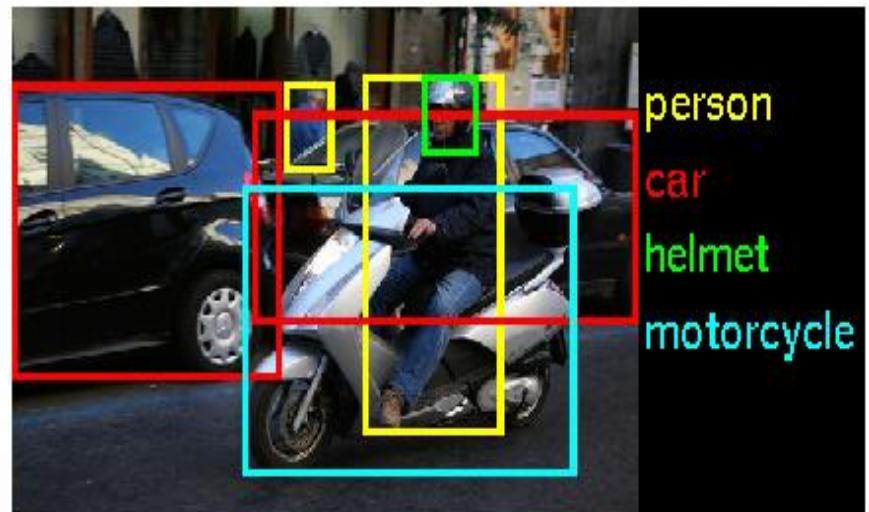
ПРИМЕРЫ

- Нейронная сеть, играющая в Го
- Март 2016 – победа над мировым чемпионом
- Нейронная сеть обучалась, играя сама с собой для увеличения объёмов входных данных (принцип обучения с подкреплением, reinforcement learning)



ПРИМЕРЫ

- **ImageNet** — задача распознавания объектов на изображении
- Решается с помощью нейронных сетей с точностью, превышающей точность работы человека



ПРИМЕРЫ

- Аннотирование изображений



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with legos toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

ПРИМЕРЫ

- Чтение по губам

Google Deepmind в 2017 году создали модель, обученную на телевизионном датасете, которая смогла превзойти профессионального lips reader'a с канала BBC.



BERT ДЛЯ РЕШЕНИЯ ЗАДАЧ NLP

В октябре **2018** года Google

выпустила модель кодирования текстовых
данных под названием BERT –

*Bidirectional Encoder Representations
from Transformers.*

Такой способ кодировать тексты
даёт state-of-the-art результаты во многих задачах
машиинного обучения, связанных с обработкой естественного
языка:

- *Определение тональности текста*
- *Перевод с одного языка на другой*
- *Определение связности предложений в тексте и др.*



ПРИМЕР: BERT ДЛЯ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Применение модели с помощью BERT-кодирования текстов
для анализа тональности:

Query	Score
How good is the iPhone 11	0.6 (positive)
How can a person get mental peace	0.4 (negative)
My boyfriend is not talking to me	0.7 (negative)
How to download video from youtube	0.0 (neutral)

ПРИМЕР: BERT ДЛЯ АНАЛИЗА СУЩНОСТЕЙ

Запрос: “what is the age of Selena Gomez?”

Ответ Google с использованием BERT-кодирования:

The screenshot shows a Google search results page. The search query "what is the age of selena gomez" is entered in the search bar. Below the search bar, there are navigation links for All, News, Images, Videos, Shopping, More, Settings, and Tools. It indicates about 98,200,000 results found in 1.27 seconds. The top result is a card for "Selena Gomez / Age" showing her age as 27 years, born on July 22, 1992, with a photo of her. Below this, there is a brief biography mentioning she was named after Tejano singer Selena Quintanilla-Perez and was born in Grand Prairie, Texas on July 22, 1992. There is also a link to "How Old Is Selena Gomez and When Did She Start Acting?" and a section for "People also search for" featuring Justin Bieber, Ariana Grande, and Taylor Swift. To the right, there is a detailed card for "Selena Gomez" listing her availability on YouTube, Spotify, and Apple Music, along with a "More music services" link. The card also provides her birth date (July 22, 1992), height (5' 5"), net worth (\$50 million as of September 2018), and parents (Mandy Teefey and Ricardo Joel Gomez). A "Wikipedia" link is also present.

Google

what is the age of selena gomez

All News Images Videos Shopping More Settings Tools

About 98,200,000 results (1.27 seconds)

Selena Gomez / Age

27 years

July 22, 1992

Obvi. Named for the late Tejano singer, Selena Quintanilla-Perez, Selena Maria Gomez was born in Grand Prairie, Texas on **July 22, 1992**. Now 26 years old, the exotic beauty of mixed Italian-Mexican descent was raised by a single mom who was a part-time stage actress. Feb 21, 2019

How Old Is Selena Gomez and When Did She Start Acting?
<https://www.cheatsheet.com/entertainment/how-old-is-selena-gomez-and-...>

People also search for

Justin Bieber 25 years Ariana Grande 26 years Taylor Swift 29 years

Selena Gomez

American singer

Available on

YouTube Spotify Apple Music

More music services

Selena Marie Gomez is an American singer, songwriter, actress, and television producer. After appearing on the children's series *Barney & Friends*, she received wider recognition for her portrayal of ...
[Wikipedia](#)

Born: July 22, 1992 (age 27 years), Grand Prairie, TX
Height: 5' 5"
Net worth: US \$50 million (September 2018)
Parents: Mandy Teefey, Ricardo Joel Gomez

ПРИМЕР: BERT ДЛЯ ОТВЕТОВ НА ВОПРОСЫ ПО ТЕКСТУ

- Context: *Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.*
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*

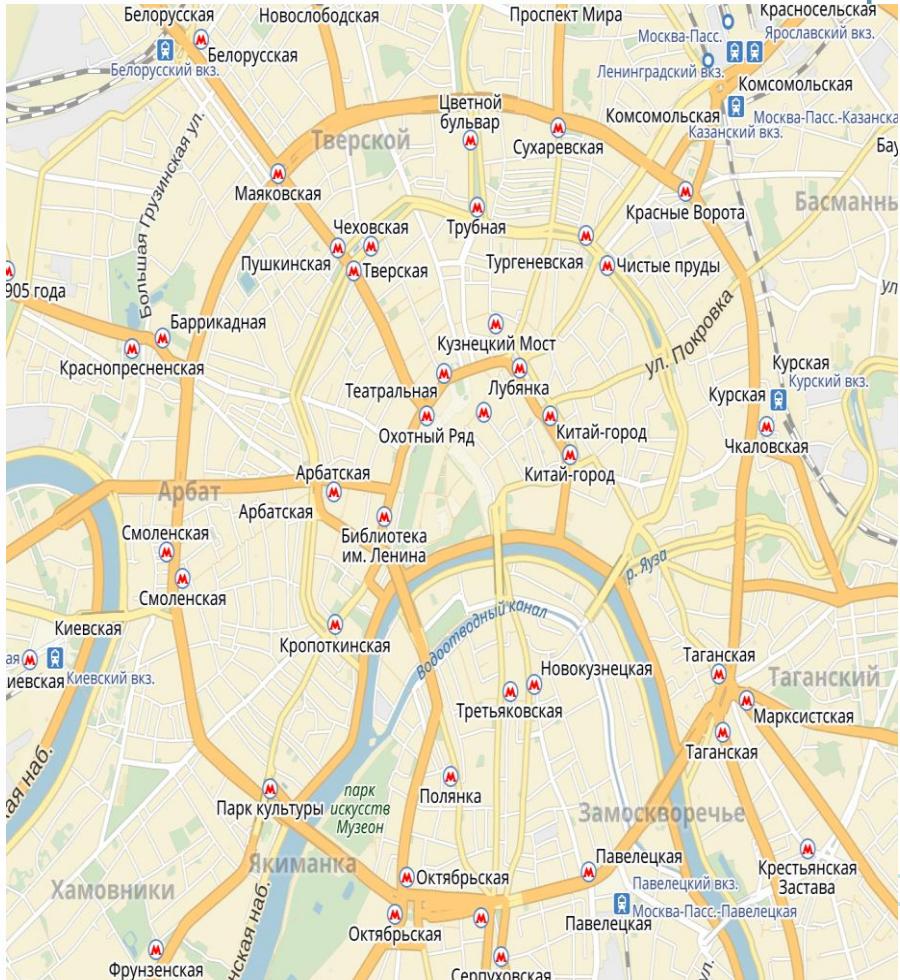
ПРИМЕР: BERT ДЛЯ ОТВЕТОВ НА ВОПРОСЫ ПО ТЕКСТУ

- Context: *Architecturally, the school has a Catholic character. Atop the Main Building's gold dome is a golden statue of the Virgin Mary. Immediately in front of the Main Building and facing it, is a copper statue of Christ with arms upraised with the legend "Venite Ad Me Omnes". Next to the Main Building is the Basilica of the Sacred Heart. Immediately behind the basilica is the Grotto, a Marian place of prayer and reflection. It is a replica of the grotto at Lourdes, France where the Virgin Mary reputedly appeared to Saint Bernadette Soubirous in 1858. At the end of the main drive (and in a direct line that connects through 3 statues and the Gold Dome), is a simple, modern stone statue of Mary.*
- Question: *The Basilica of the Sacred heart at Notre Dame is beside to which structure?*
- Answer: *start_position: 49, end_position: 51*

ОСНОВНЫЕ ПОНЯТИЯ МАШИННОГО ОБУЧЕНИЯ

ПЕРВЫЙ ПРИМЕР: ЗАДАЧА О РЕСТОРАНАХ

- Сеть ресторанов
- Хотим открыть еще один
- Несколько вариантов размещения
- Какой из вариантов принесет максимальную прибыль?



ФОРМАЛИЗАЦИЯ

X – множество объектов

Y – множество ответов

$a: X \rightarrow Y$ – неизвестная зависимость

Дано:

$\{x_1, \dots, x_n\} \subset X$ – обучающая выборка

$\{y_1, \dots, y_n\}, y_i = y(x_i)$ – известные ответы

Найти:

$a: X \rightarrow Y$ – алгоритм (решающую функцию),

приближающую y на всем множестве X

ПРИЗНАКОВОЕ ОПИСАНИЕ ОБЪЕКТОВ

Признаки объекта x можно записать в виде вектора
 $(f_1(x), \dots, f_n(x))$

Матрица “объекты-признаки”:

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

СТАНДАРТНАЯ ПОСТАНОВКА ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

ПРИЗНАКИ

ОБЪЕКТЫ

X_{train}



ЗНАЕМ - ОБУЧЕНИЕ

ЕЩЕ ОБЪЕКТЫ

X_{test}



НЕ ЗНАЕМ, ХОТИМ
ПРЕДСКАЗАТЬ

СХЕМА ПОЛУЧЕНИЯ ПРЕДСКАЗАНИЯ

В задачах обучения с известными классами (обучение по прецедентам) всегда есть два этапа:

- Этап обучения (*training*):

по выборке $X = \{(x_i, y_i)\}$ строим алгоритм a

- Этап применения (*testing*):

алгоритм a для новых объектов x выдает ответы
 $a(x)$

ОПРЕДЕЛЕНИЯ

- **Признаки, факторы (features)** – количественные характеристики объекта
- **Обучающая выборка (training set)** – конечный набор объектов, для которых известны значения целевой переменной

Пример: набор ресторанов, открытых более года назад, для которых известна их прибыль за первый год

- **Объекты** – абстрактные сущности (но компьютеры работают только с числами)
- **Признаки** описывают объекты с помощью чисел

Специалист по анализу данных не является экспертом в предметной области – вся необходимая информация содержится в обучающей выборке. Эксперты нужны при формировании признаков.

ВИДЫ ПРИЗНАКОВ

- Числовые
- Бинарные (0/1)
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой
(изображение)

ВИДЫ ДАННЫХ

- Таблицы (-xls, -csv и другие форматы, содержащие данные)
- Текстовые данные
- Изображения
- Звук
- Логи

Большинство алгоритмов машинного обучения работает с числовыми данными, поэтому все виды данных необходимо переводить в числовые.

ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

- Задачи медицинской диагностики (пациент здоров или болен)
- Задачи кредитного scoringа (выдаст банк кредит данному клиенту или нет)
- Задача предсказания оттока клиентов (уйдет клиент в следующем месяце или нет)
- Предсказание поведения пользователя (кликнет пользователь по данному баннеру или нет)
- Классификация изображений (на изображении кошка или собака)

ПРИМЕРЫ ЗАДАЧ КЛАССИФИКАЦИИ

Мультиклассовая классификация

- Определение типа объекта на изображении



Pedestrian



Car



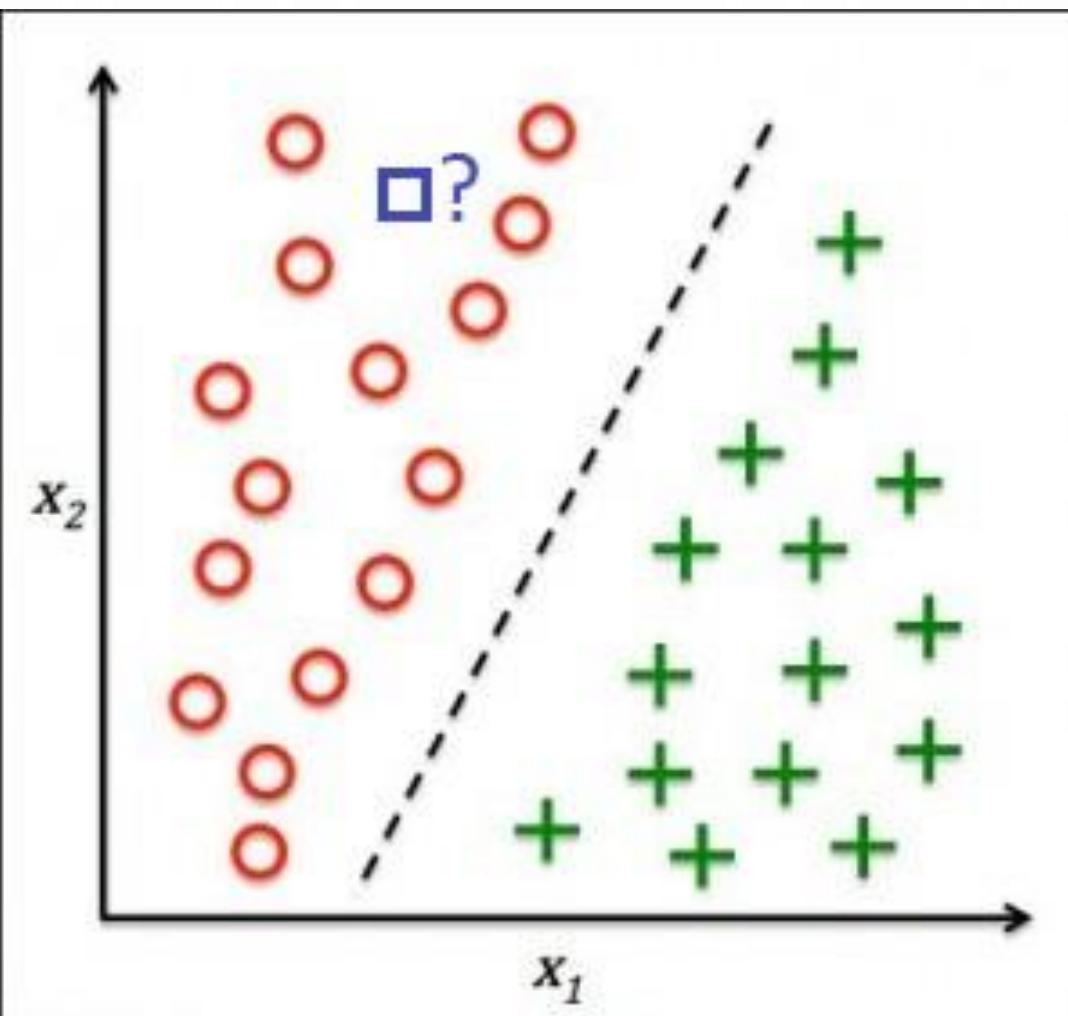
Motorcycle



Truck

- Определение наиболее подходящей профессии для данного кандидата

ЗАДАЧА КЛАССИФИКАЦИИ



ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

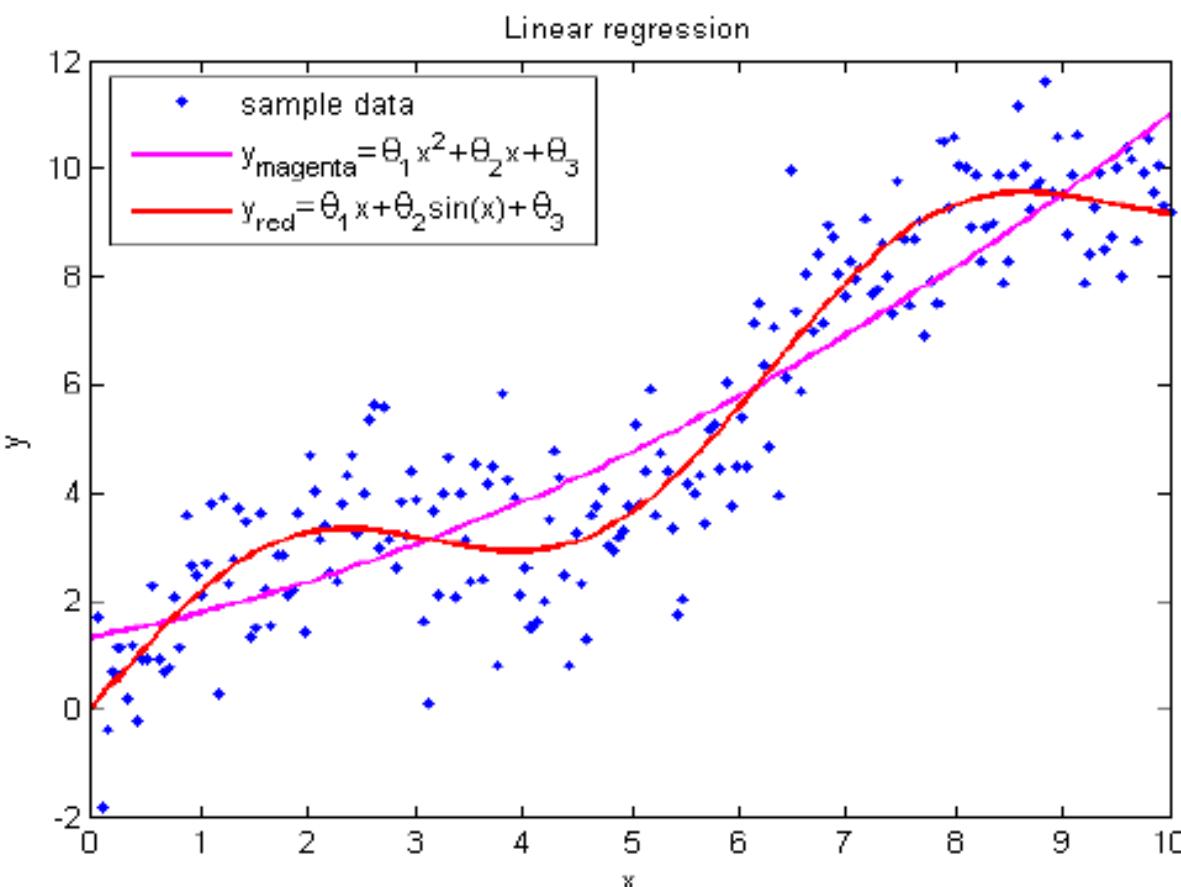
- $Y = R$ или $Y = R^n$

ПРИМЕРЫ ЗАДАЧ РЕГРЕССИИ

- Предсказание стоимости недвижимости (стоимость квартиры в Москве)
- Предсказание прибыли ресторана
- Предсказание поведения временного ряда в будущем (стоимость акций)
- Предсказание зарплаты выпускника вуза по его оценкам

ЗАДАЧА РЕГРЕССИИ

$X = Y = \mathbb{R}$, $\ell = 200$, $n = 3$ признака: $\{x, x^2, 1\}$ или $\{x, \sin x, 1\}$



ТИПЫ ЗАДАЧ В ЗАВИСИМОСТИ ОТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

Классификация

- $Y = \{0, 1\}$ – классификация на 2 класса
- $Y = \{1, \dots, M\}$ – классификация на M непересекающихся классов
- $Y = \{0, 1\}^M$ - классификация на M классов, которые могут пересекаться

Регрессия

- $Y = R$ или $Y = R^n$

Ранжирование

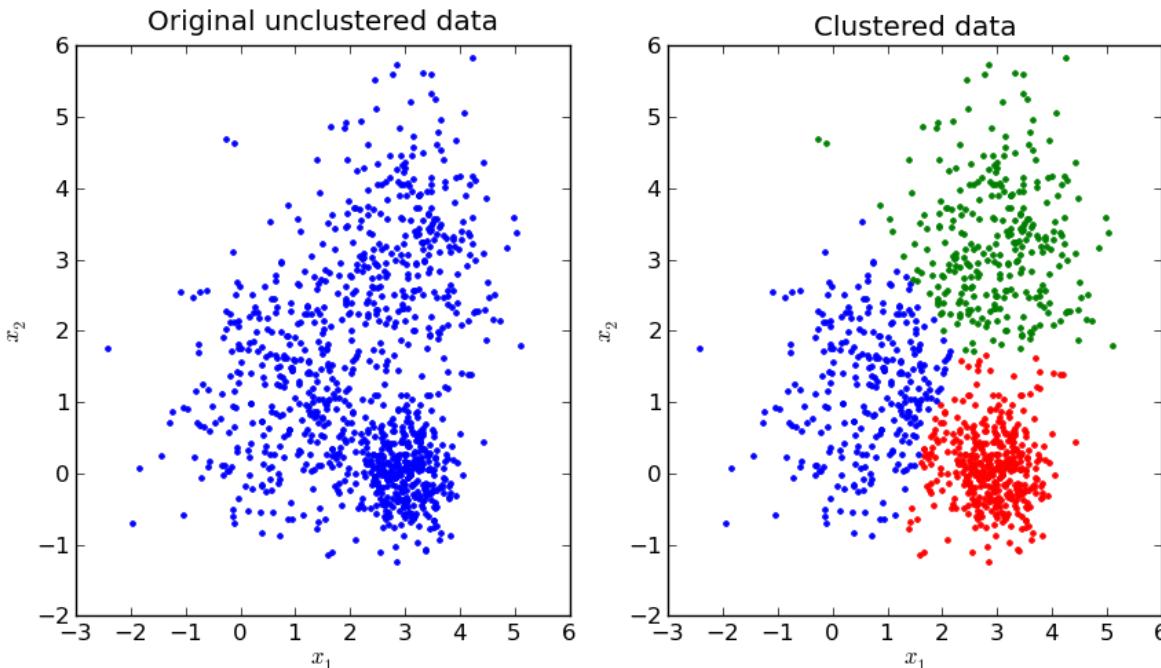
- Y – конечное упорядоченное множество

ПРИМЕРЫ ЗАДАЧ РАНЖИРОВАНИЯ

- Вывести подходящие запросу документы в порядке уменьшения релевантности
- Вывести кандидатов на должность в порядке уменьшения релевантности

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.



ПРИМЕРЫ ЗАДАЧ КЛАСТЕРИЗАЦИИ

- Разбить пользователей на группы, внутри каждой из которых будут похожие пользователи
- Разбить текстовые документы на группы по похожести документов

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

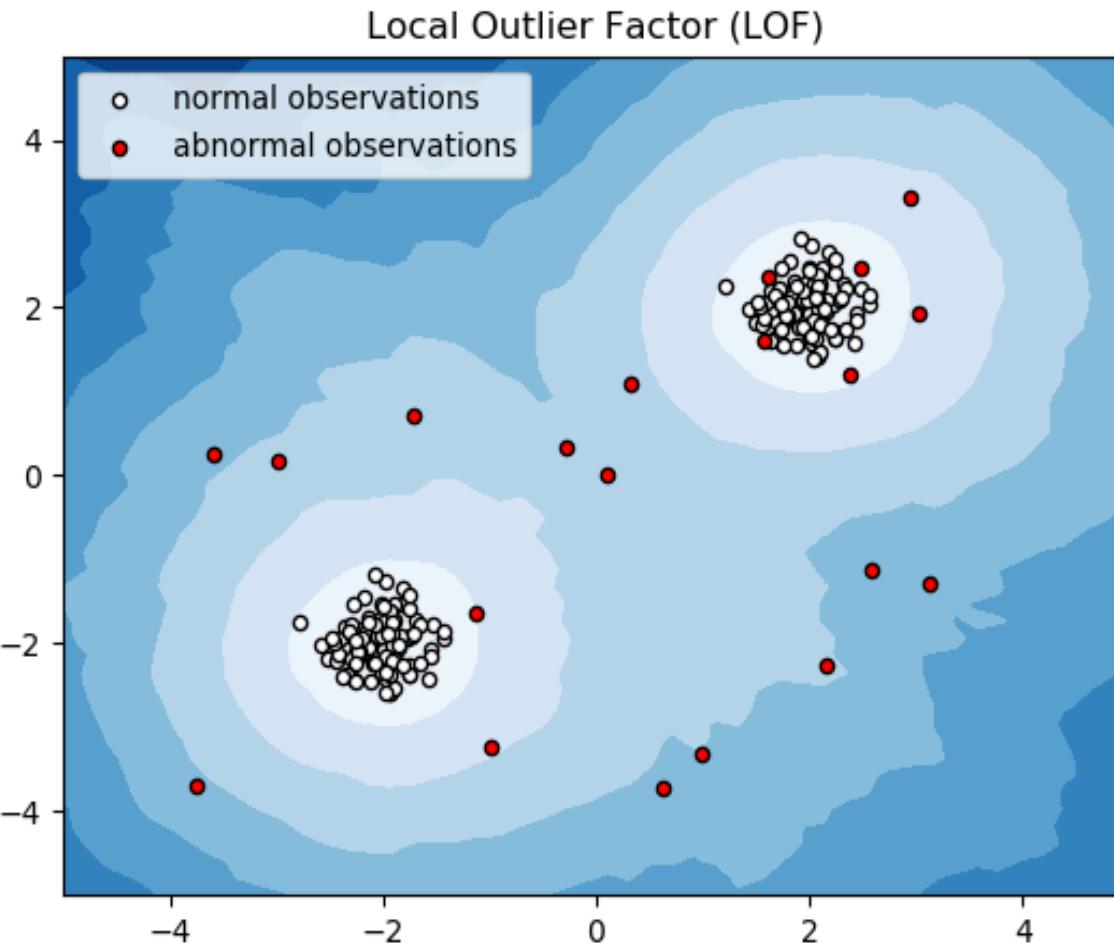
- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.

ПРИМЕР ОЦЕНИВАНИЯ ПЛОТНОСТИ

- Поиск аномалий с помощью оценивания плотностей



ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.
- **Визуализация** – задача изображения многомерных объектов в 2х или 3хмерном пространстве с сохранением зависимостей между ними.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

ОЦЕНКА СКРЫТОГО СОСТОЯНИЯ МОДЕЛИ

Объекты в задачах машинного обучения характеризуются:

- Наблюдаемыми переменными (признаками)
- Скрытыми переменными (например, целевая переменная в задаче обучения с учителем)

ОЦЕНКА СКРЫТОГО СОСТОЯНИЯ МОДЕЛИ

Объекты в задачах машинного обучения характеризуются:

- Наблюдаемыми переменными (признаками)
- Скрытыми переменными (например, целевая переменная в задаче обучения с учителем)

Задача машинного обучения:

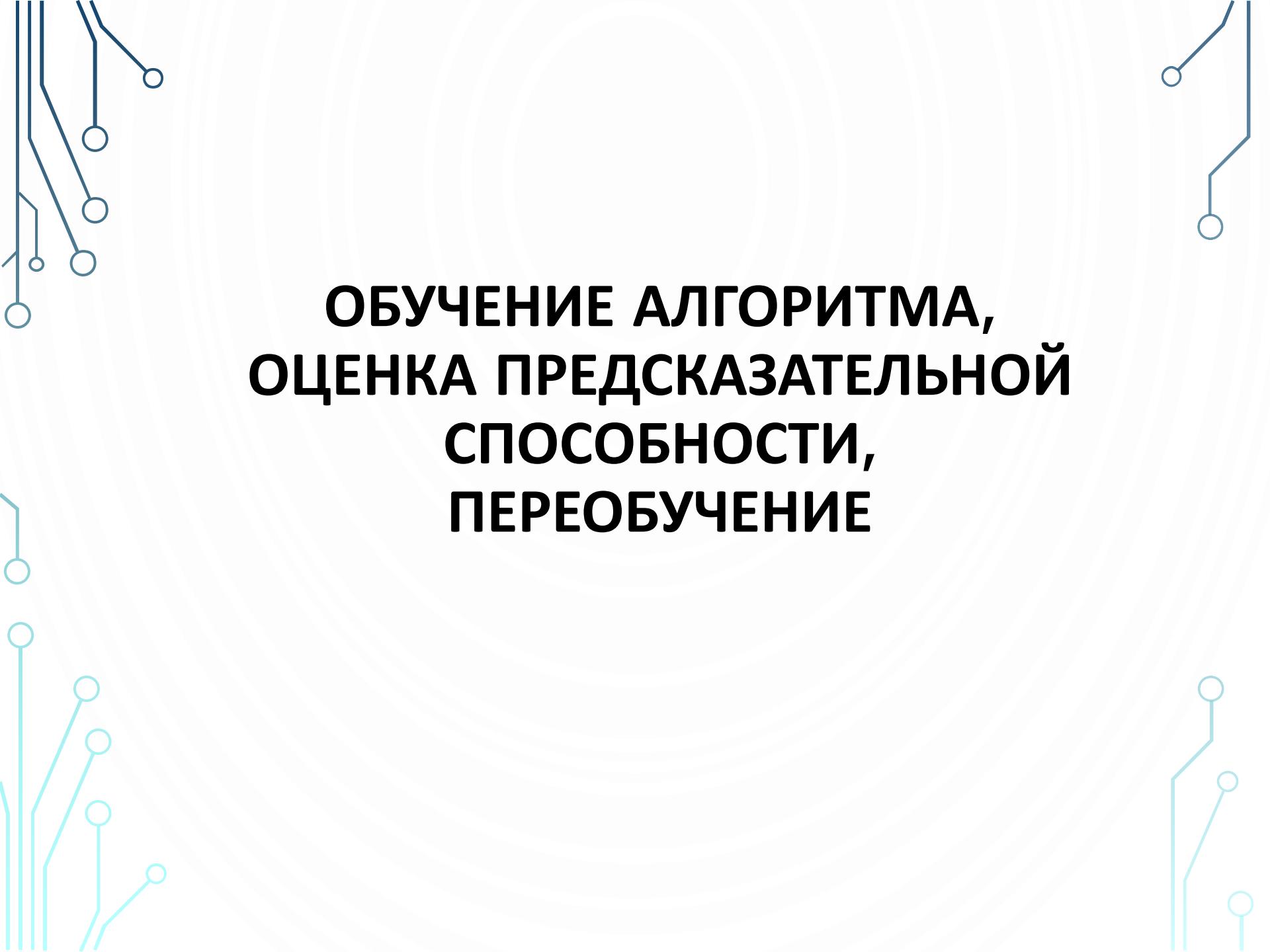
определение зависимостей между наблюдаемыми переменными и скрытыми переменными

ОЦЕНКА СКРЫТОГО СОСТОЯНИЯ МОДЕЛИ

Задача машинного обучения:

определение зависимостей между наблюдаемыми переменными и скрытыми переменными

- Как правило, зависимости задаются параметрическими решающими правилами с параметрами w (весами)
- В ходе обучения определяются значения весов w .



ОБУЧЕНИЕ АЛГОРИТМА, ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ, ПЕРЕОБУЧЕНИЕ

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):

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

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 *количество комната (x_2)*.



ОБУЧЕНИЕ АЛГОРИТМА

Предположим, что мы хотим предсказать *стоимость дома* у по его *площади (x_1)* и *количество комната (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 *количество комната (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 *количество 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

Пример (семейство линейных моделей):

$$A = \{a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d \mid w_0, w_1, \dots, w_d \in \mathbb{R}\}$$

Функционал ошибки:

$$Q(a, X) = \frac{1}{l} \sum_{i=1}^l (\textcolor{blue}{a}(x_i) - \textcolor{red}{y}_i)^2$$

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*.

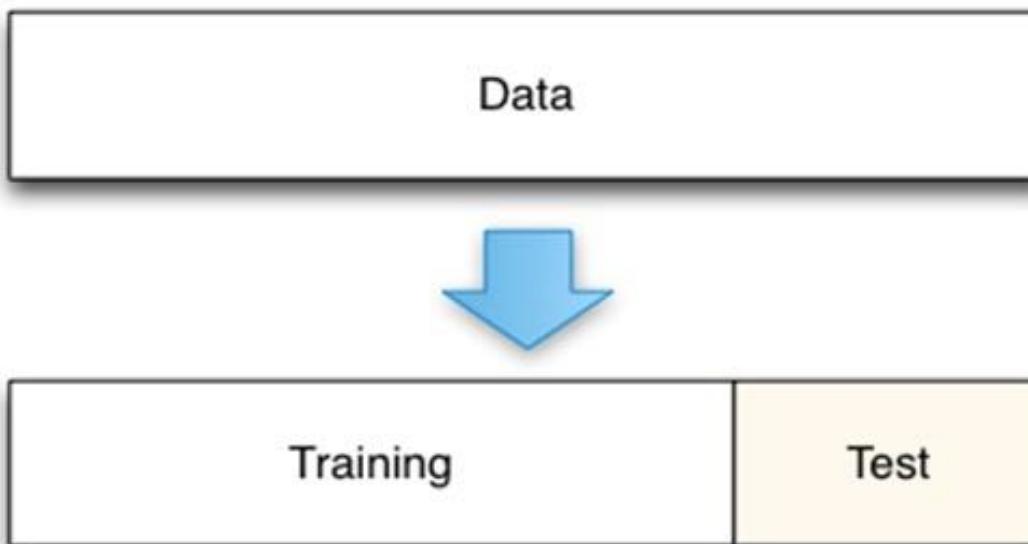
Примеры:

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

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

ОЦЕНКА ПРЕДСКАЗАТЕЛЬНОЙ СПОСОБНОСТИ АЛГОРИТМА

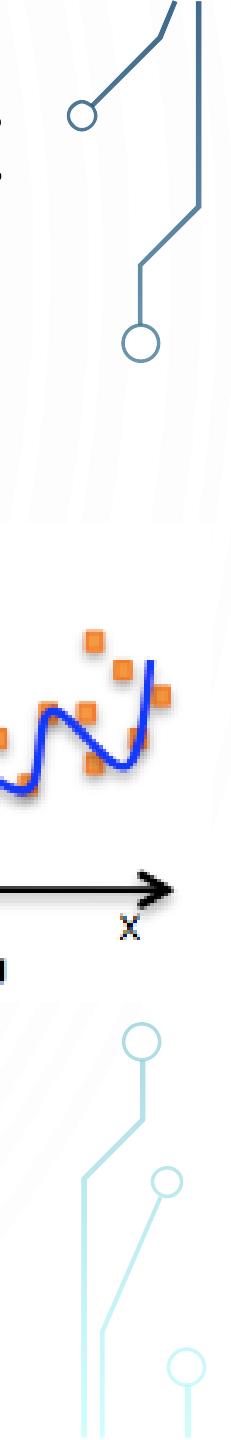
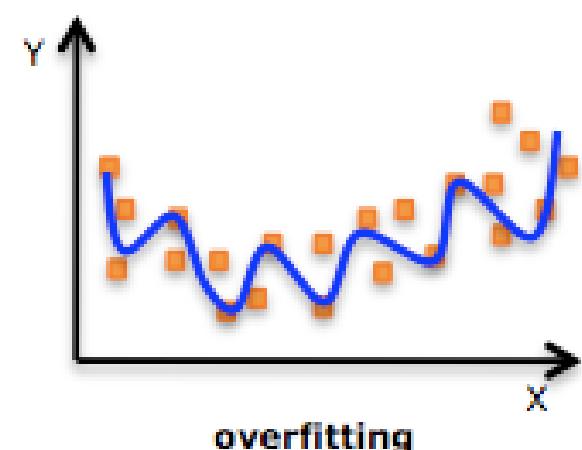
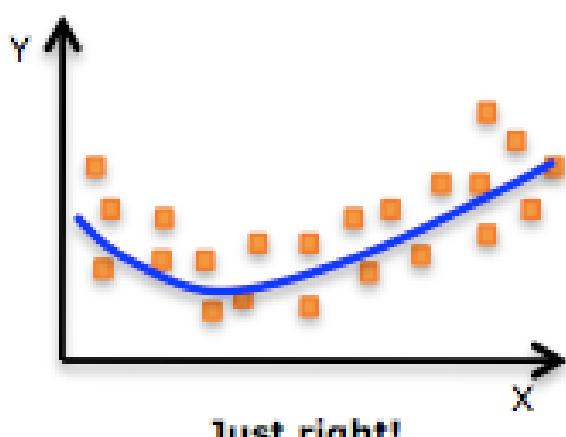
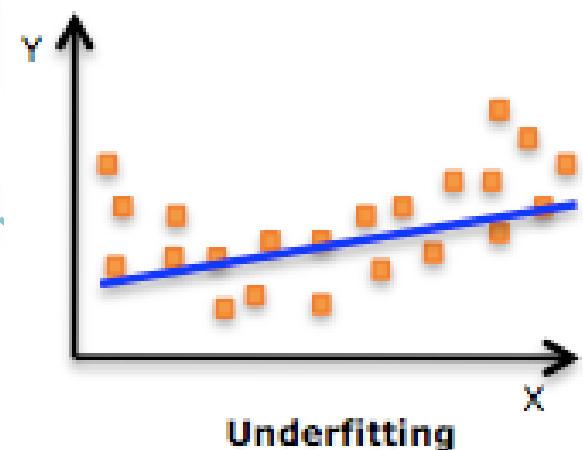
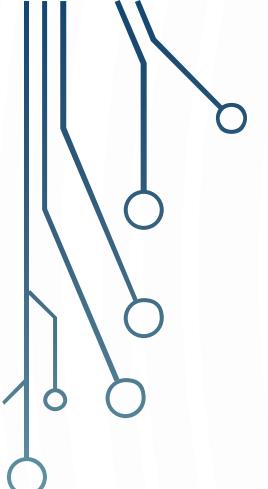
- Перед началом обучения отложим часть обучающих объектов и не будем использовать их для построения модели (отложенная выборка).



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.

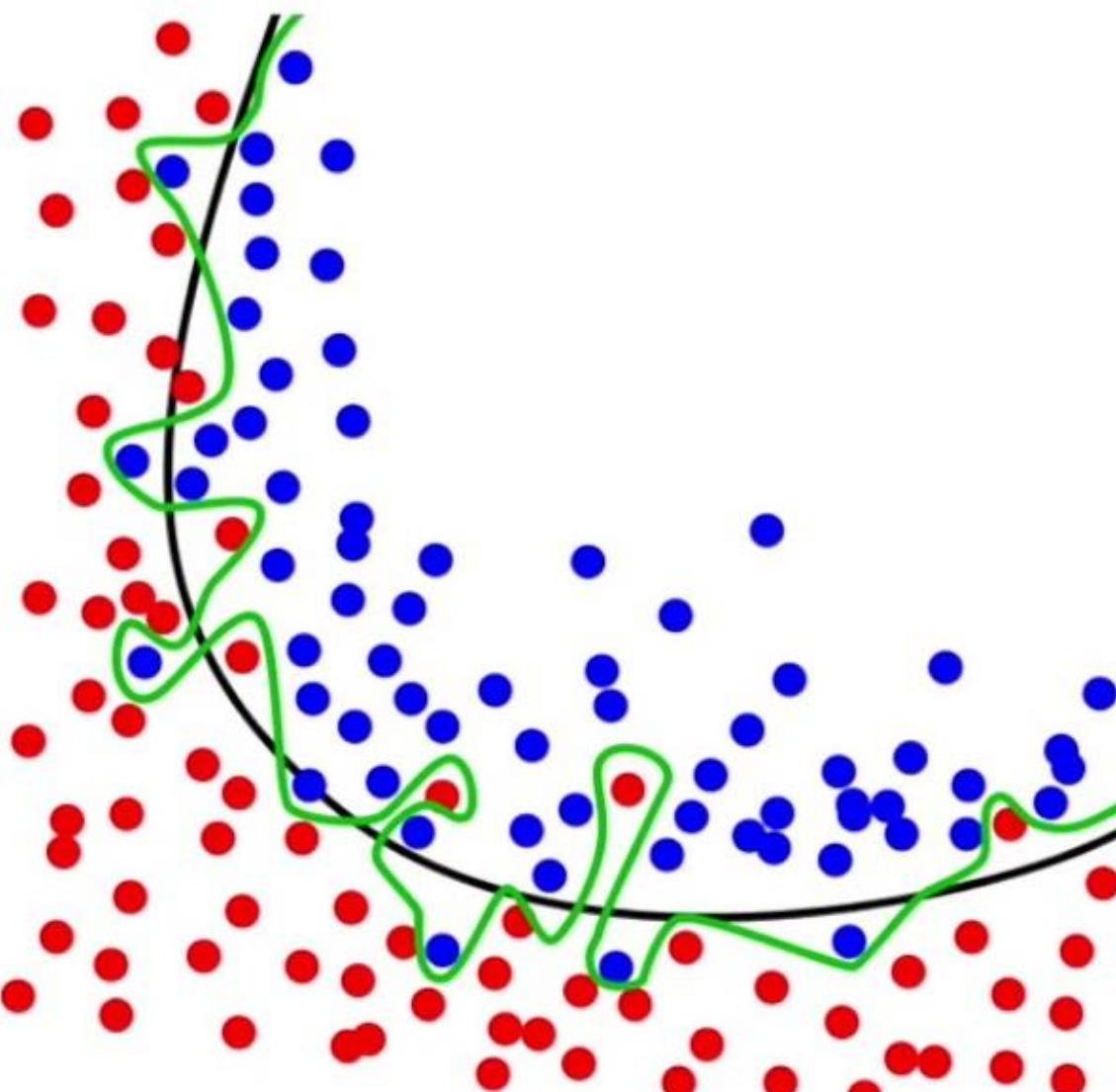
ПЕРЕОБУЧЕНИЕ И НЕДООБУЧЕНИЕ



ИЗ-ЗА ЧЕГО ВОЗНИКАЕТ ПЕРЕОБУЧЕНИЕ

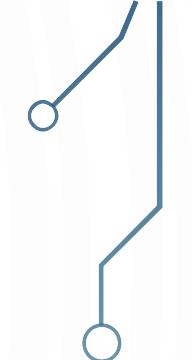
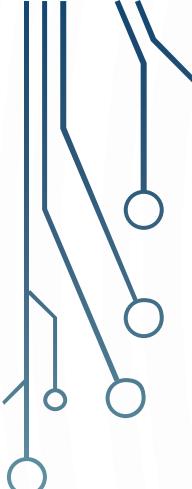
- Избыточная сложность пространства параметров Ω , лишние степени свободы в модели $a(x, w)$ “тратятся” на чрезмерно точную подгонку под обучающую выборку.
- Переобучение есть всегда, когда есть оптимизация параметров по конечной (заведомо неполной) выборке.

ПРИМЕР ПЕРЕОБУЧЕНИЯ В ЗАДАЧЕ КЛАССИФИКАЦИИ



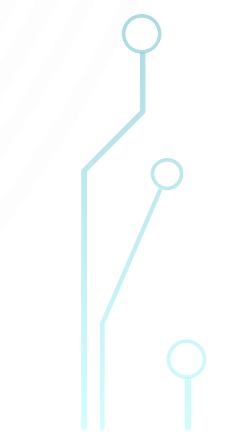
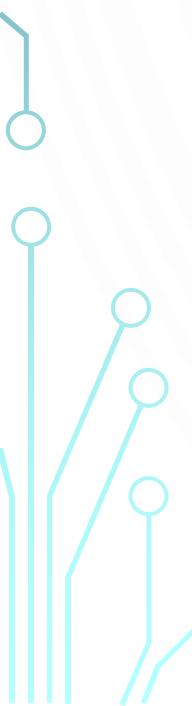
ПРИЗНАК ПЕРЕОБУЧЕНИЯ

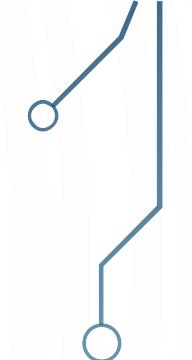
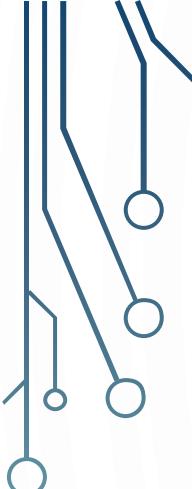
- *Если качество на отложенной выборке сильно ниже качества на обучающих данных, то происходит переобучение*



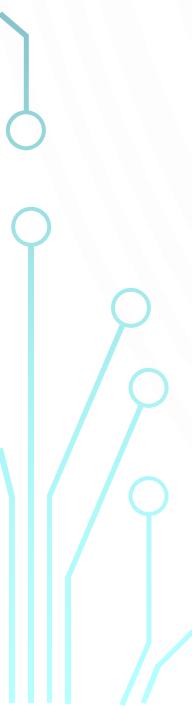
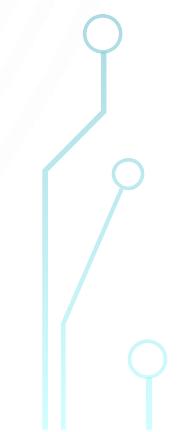
АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

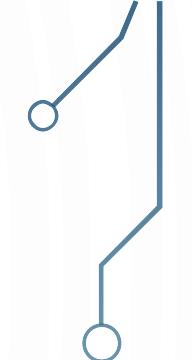
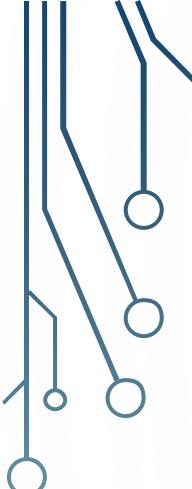
1. Постановка задачи



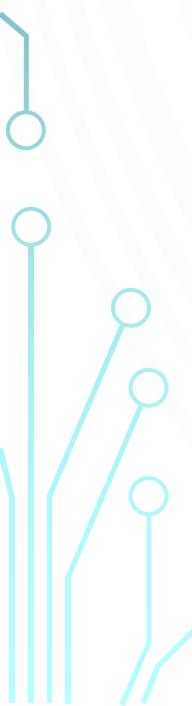
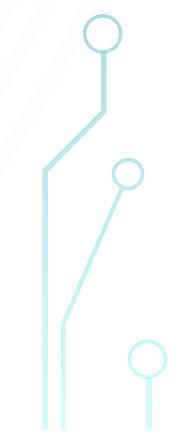


АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

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

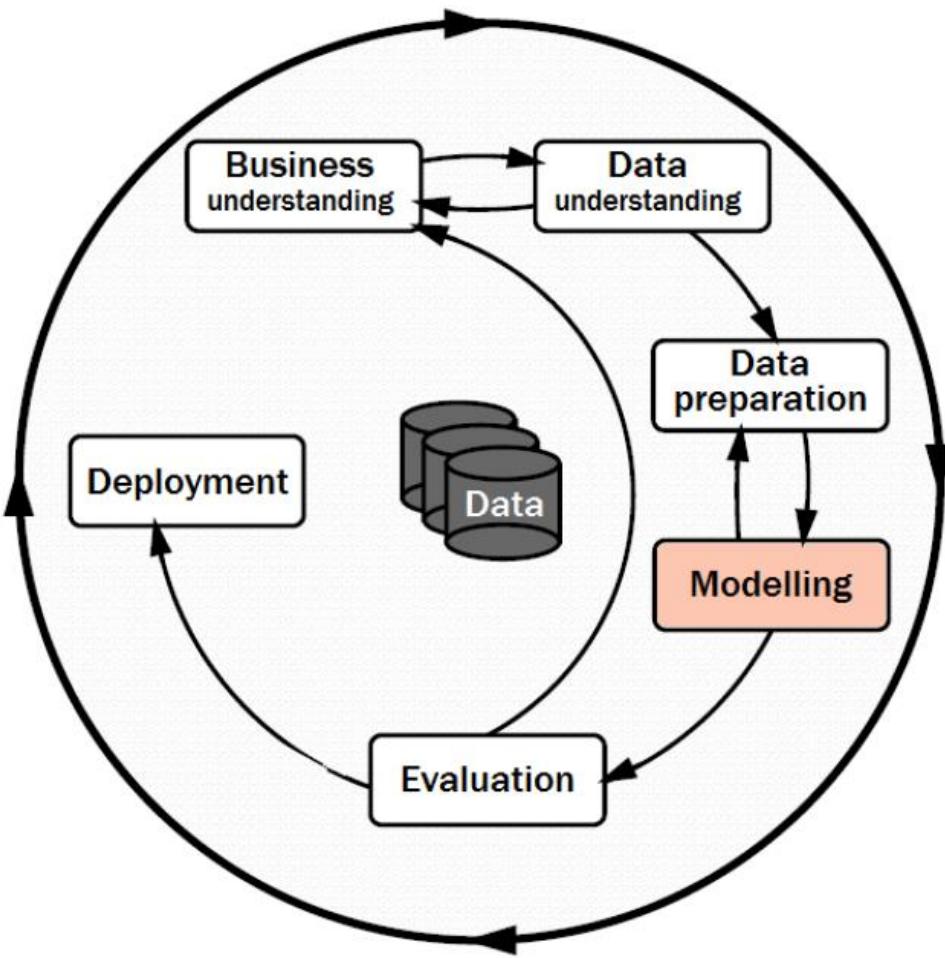
АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ АНАЛИЗА ДАННЫХ

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



[http://wiki.cs.hse.ru/Машинное обучение \(ФЭН\) - 2020](http://wiki.cs.hse.ru/Машинное_обучение_(ФЭН) - 2020)