

Лекция 1

Введение в анализ данных

Анализ данных
Андрей Фильченков

08.09.2021

План лекции

- Организационные вопросы
 - Терминологические вопросы
 - Как устроены данные
 - Примеры задач
-
- Слайды доступны: shorturl.at/txMW7
 - Видео доступны: shorturl.at/asL08

План лекции

- **Организационные вопросы**
- Терминологические вопросы
- Как устроены данные
- Примеры задач

Курсы про машинное обучение и анализ данных

- Анализ данных (осень 2021)
- Машинное обучение (весна 2021)
- Дополнительные главы машинного обучения (осень 2022)

Wanna know more?

Магистерская программа по современному глубокому обучению на ФИТиП с осени 2022.

Лаборатория машинного обучения

- Часть Центра Компьютерных Технологий
- Области исследований:
 - автоматическое машинное обучение
 - обработка и генерация изображений
 - профилирование пользователей и анализ социальных сетей
 - выбор признаков
 - маршрутизация
 - фундаментальные исследования
 - применение (медицина, анализ кода, финансы, производство)
 - ...

Подписывайтесь на нас!

- Telegram: t.me/itmo_mllab
- Instagram: [instagram.com/itmo.mllab](https://www.instagram.com/itmo.mllab)
- Facebook: [facebook.com/itmo.mllab](https://www.facebook.com/itmo.mllab)
- Twitter: twitter.com/itmo_mllab
- Medium: medium.com/@itmo.mllab
- YouTube: [youtube.com/c/mllabitmo](https://www.youtube.com/c/mllabitmo)
- Twitch: www.twitch.tv/itmo_ml_lab

План курса

- Python (2 лекции)
- Модели данных (3 лекции)
- Работа с разными типами данных (5 лекций)
- Введение в статистику, лучшие практики работы с данными и полный цикл анализа данных

Как получить оценку?

- Практика — сдача лабораторных работ
- Теория — сдача экзамена
- Бонусные баллы

План лекции

- Организационные вопросы
- **Терминологические вопросы**
- Как устроены данные
- Примеры задач

Знания и данные

Знания \neq данные

А в чем отличие?

Знания и данные

Знания \neq данные

Знания это закономерности в некоторой области (принципы, ограничения, отношения, правила, законы), получаемые в ходе профессиональной деятельности, которые позволяют формулировать и решать проблемы в этой области.

Анализ данных часто путают с:

- Большие данные (Big Data)
- Бизнес-информатика (Business Intelligence)
- Информационный поиск (information retrieval)
- Машинное обучение (machine learning)
- Глубокое обучение (deep learning)
- Искусственный интеллект (artificial intelligence)

Data Mining

Формально, DM является одним из шагов в извлечении знаний из баз данных (**knowledge discovery in databases**) и включает в себя:

1. Сбор данных
2. Выделение признаков
3. Применение алгоритмов машинного обучения

Фактически, синонимично data analysis.

Data Science

1. Сбор данных
2. Интеграция данных (data integration)
3. Хранение данных (data warehousing)
4. Анализ данных
5. Высокопроизводительные вычисления (high-performance computing)

Data Analysis

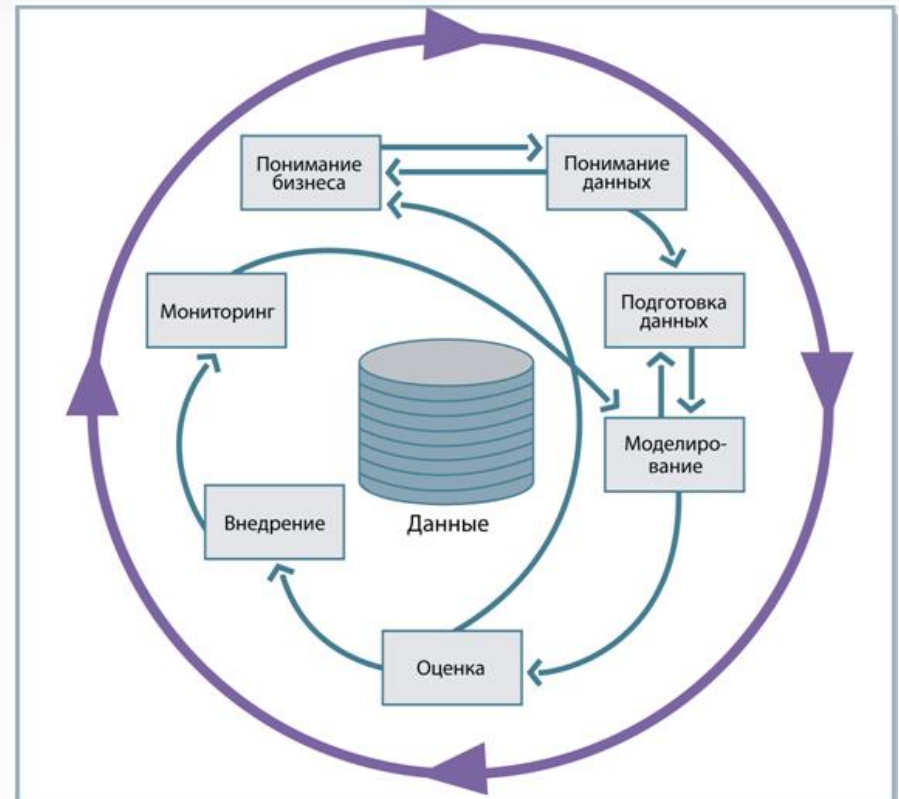
(Интеллектуальный) анализ данных

1. Эксплораторный анализ данных (exploratory DA)
2. Конфирмационный анализ данных (confirmatory DA)
3. Предсказательный анализ данных
4. Визуализация данных

Методология CRISP-DM

Шаги:

- Понимание бизнеса
- Понимание данных
- Подготовка данных
- Моделирование
- Оценка
- Внедрение
- Мониторинг



Кто тут работает

- Data analyst
- Data scientist
- Business analyst
- Data engineer
- Machine learning engineer
- Machine learning researcher

План лекции

- Организационные вопросы
- Терминологические вопросы
- **Как устроены данные**
- Примеры задач

Размеченный набор данных

X — множество объектов;

Y — множество меток (ответов);

$y : X \rightarrow Y$ неизвестная целевая функция (зависимость).

$\mathcal{D} = \{(x_i, y_i)\}$ — размеченный набор данных,
где $\{x_1, \dots, x_{|\mathcal{D}|}\} \subset X$ — объекты, а $y_i = y(x_i)$ — известные метки (значения целевой функции).

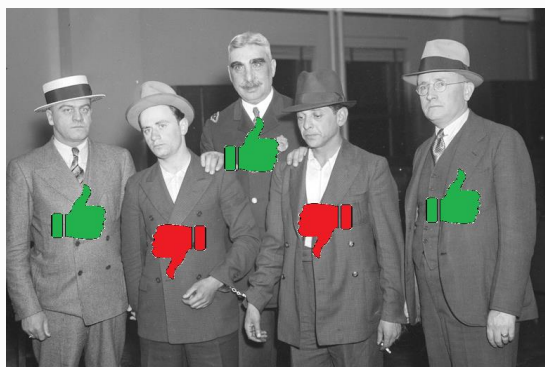
Но метки бывают не всегда.

Если они есть, то это **задача обучения с учителем**.

Если их нет, то это **задача обучения без учителя**.

Предсказание

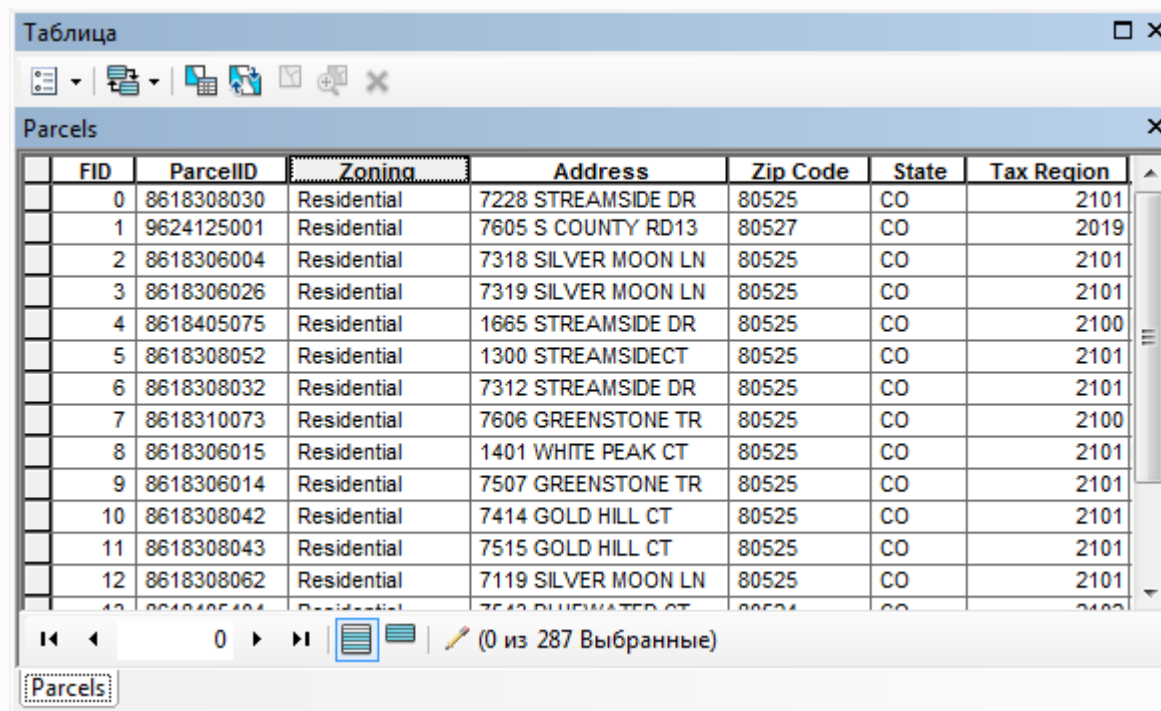
- Наука занимается осуществлением предсказаний
- Предсказание это самая популярная (но не единственная) задача в анализе данных



Разные типы данных

Данные собираются из разных источников.

Будем говорить про табличные данные:



Таблица

Parcels

FID	ParcelID	Zoning	Address	Zip Code	State	Tax Region
0	8618308030	Residential	7228 STREAMSIDE DR	80525	CO	2101
1	9624125001	Residential	7605 S COUNTY RD13	80527	CO	2019
2	8618306004	Residential	7318 SILVER MOON LN	80525	CO	2101
3	8618306026	Residential	7319 SILVER MOON LN	80525	CO	2101
4	8618405075	Residential	1665 STREAMSIDE DR	80525	CO	2100
5	8618308052	Residential	1300 STREAMSIDE CT	80525	CO	2101
6	8618308032	Residential	7312 STREAMSIDE DR	80525	CO	2101
7	8618310073	Residential	7606 GREENSTONE TR	80525	CO	2100
8	8618306015	Residential	1401 WHITE PEAK CT	80525	CO	2101
9	8618306014	Residential	7507 GREENSTONE TR	80525	CO	2101
10	8618308042	Residential	7414 GOLD HILL CT	80525	CO	2101
11	8618308043	Residential	7515 GOLD HILL CT	80525	CO	2101
12	8618308062	Residential	7119 SILVER MOON LN	80525	CO	2101
13	8618308064	Residential	7519 SILVER MOON LN	80525	CO	2101

0 (0 из 287 Выбранные)

Parcels

Сведение к табличному типу

Все типы данных можно сводить к табличному типу.

Но не все сводятся:

- текст
- изображение
- видео
- звук
- сигналы (в целом)

Что представляют собой объекты?

$f_j : X \rightarrow D_j, j = 1, \dots, n$ — признаки (features, attributes) объектов.

Типы признаков:

- **бинарный**: $D_j = \{0, 1\}$ (гендер в XVIII веке);
- **категориальный (номинальный)**:
 D_j конечно (цвет);
- **порядковый (ординальный)**:
 D_j конечно и упорядочено (сорт муки);
- **численный (количественный)**: $D_j = \mathbb{R}$ (длина).

Табличные данные

$(f_1(x), \dots, f_n(x))$ — признаковое описание объекта x . Объект отождествляется с его признаковым описанием.

Данные часто представляются в табличном виде (матрица «объекты — признаки») :

$$F = \|f_j(x_i)\|_{|\mathcal{D}| \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_{|\mathcal{D}|}) & \dots & f_n(x_{|\mathcal{D}|}) \end{pmatrix}.$$

Что представляют собой ответы?

Для классификации:

- $Y = \{-1, +1\}$ — бинарная классификация (родился ли человек в СССР);
- $Y = \{1, \dots, M\}$, M непересекающихся классов (в какой стране человек родился);
- $Y = \{0, 1\}^M$, M пересекающихся классов (гражданином каких стран человек является).

Для ранжирования:

- Y — конечно (частично) упорядоченное множество (ранжирование стран по предпочтительности посещения).

Для регрессии:

- $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$ (с какой вероятностью человек посетит Ирак / каждую из стран-членов ООН).

План лекции

- Организационные вопросы
- Терминологические вопросы
- Как устроены данные
- **Примеры задач**

Примеры 1/3

1. Медицинские диагнозы

Предсказание диагноза для пациента и потенциального лечения

2. Кредитный скоринг / HR

Определение того, стоит ли или не стоит давать клиенту кредит или рабочее место

3. Фильтрация спама и обнаружение вредоносов

Определение того, является ли письмо / файл, соответственно, спамом / вредоносом или нет

4. Категоризация документов и

Определение категорий документов

5. Сегментация пользователей

Объединение пользователей в группы по их поведению

Примеры 2/3

6. Предсказание стоимости жилья

Предсказание на основе разнородных данных того, сколько будет стоить то или иное жилье

7. Предсказание биржевых индексов

Предсказание того, сколько будут стоить акции

8. Коллаборативная фильтрация / рекомендательные системы

Предсказание предпочтений пользователей по данным их поведения

9. Определение мнения клиентов о продукте

Определение мнений за счет работы с отзывами

10. Поиск проблем в работе оборудования

Обнаружение поломок или опасных ситуаций в работе приборов, оборудования, механизмов и роботов

Примеры 3/3 (не анализ данных)

1. Обучение агента играть в Minecraft / WoW
2. Диалоговые системы и чат-боты
3. Создание изображений несуществующих людей
4. Разработка поискового движка