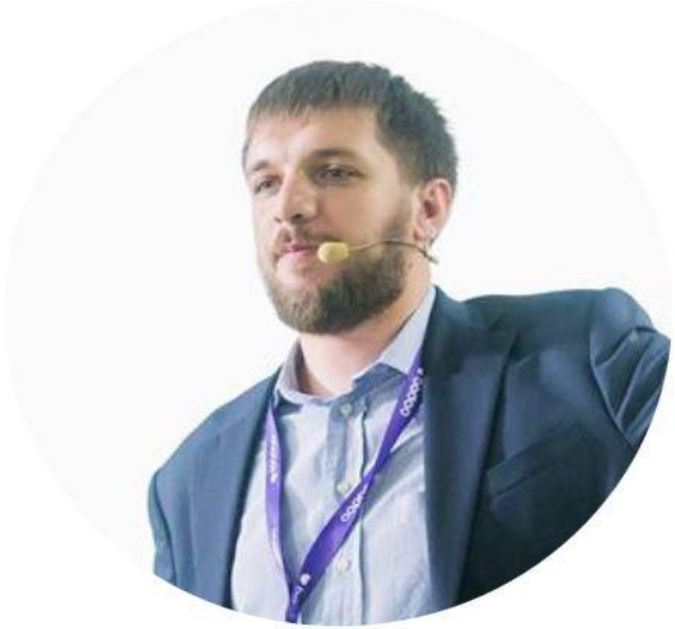
НЕТОЛОГИЯ
групп

Машинное обучение в DataScience



Алексей Кузьмин

Директор разработки; Data Scientist

ДомКлик.ру

aleksej.kyzmin@gmail.com

Работа с данными

Источники данных

Сбор данных
Лекция 7

SQL-БД
Лекция 1

NoSQL
Лекция 4

MapReduce
Лекция 5-6

Мотивация Big Data
Лекция 3

Data Science
Лекция 2

Отчеты
Лекция 1

Модели
Лекция 2

Люди и процессы
Лекция 8

Примеры кейсов
Лекция 9

Что будет сегодня?

- Понятия объекта и признаков
- 3 классические задачи машинного обучения
- Извлечение, отбор и преобразование признаков
- Алгоритмы машинного обучения
- Метрики качества алгоритмов классификации

Объекты и признаки

Объекты и признаки зависят от рассматриваемого контекста (задачи)

- B2C абонент сотового оператора
  - ARPU, модель телефона, количество финансовых блокировок, количество входящих звонков, средняя сумма пополнения баланса
- Пара сим-карт
  - Количество общих контактов, модели телефонов, количество совместных регистраций на базовой станции
- Электронное сообщение
  - Длина сообщения, наличие цифр, количество слов, предложений, сами слова
- Ресторан
  - Средний чек, количество посетителей за месяц, район, количество официантов

Практика 1

- https://colab.research.google.com/ <- основной рабочий инструмент
- Загрузим данные и посмотрим, что у нас с ними
- data.csv

3 классические задачи

Машинное обучение умеет:

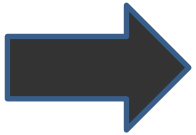
– Извлечь признаковое описание объектов (рост, цвет волос, размер одежды, количество детей, образование, наличие смартфона)

– Посмотрев на объекты, научиться:
  - Классифицировать (Мужчина/Женщина)
  - Прогнозировать значения для объектов (Возраст, Доход, Рост)
  - Группировать (Школьники, Бизнесмены, Политики, Любители Чая)

3 классические задачи

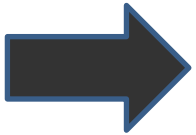
Классификация

Регрессия



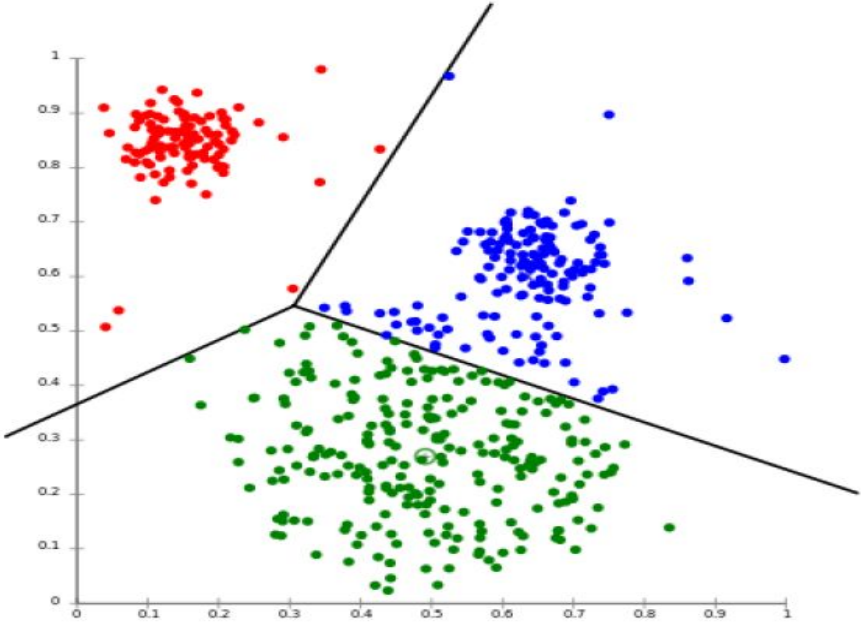
Обучение с учителем

Кластеризация



Обучение без учителя

Классификация



- Дано:

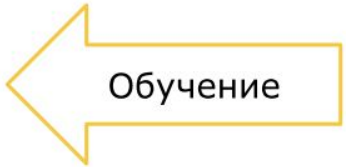
– Обучающая выборка, состоящая из признакового описания объектов и метки класса для каждого объекта

- Найти:

– Алгоритм, который бы для каждого нового объекта по его признаковому описанию прогнозировал класс этого объекта

Классификация

ID	ARPU	Кол-во блокировок	Кол-во входящих звонков	ОТТОК
9034948911	123.5	5	15	1
9034948912	245.6	10	124	0
9034948913	890.4	0	23	1
9034948914	50.3	101	0	0



АЛГОРИТМ: **ВХОД** - [ID, ARPU, Кол-во блокировок, Кол-во входящих звонков], **ВЫХОД** – [ОТТОК]

903494895	12.6	8	12	?
9034948916	1012.2	10	256	?
9034948917	132.9	112	10	?

Классификация

- Важно помнить:

– Классификация – это обучение с учителем (supervised learning), в роли учителя выступает обучающая выборка

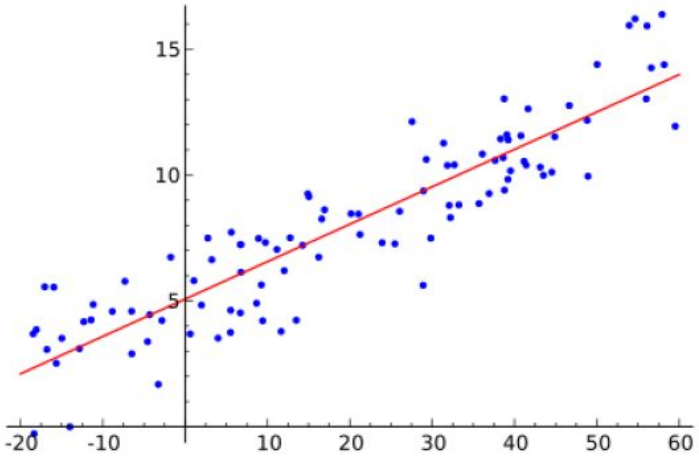
– Классификация прогнозирует метку (класс) для объекта, который может принимать набор дискретных значений

– В результате получается алгоритм, который на вход принимает признаковое описание объекта, а на выходе выдает его класс

- Примеры задач:

– Классификация абонентов по полу, классификация спама, классификация абонентов на наличие второго устройства

Регрессия



Геометрически, алгоритм восстанавливает зависимость между признаками и целевой переменной

- Дано:
  - Обучающая выборка, состоящая из признакового описания объектов и значения целевой переменной для каждого объекта
- Найти:
  - Алгоритм, который бы для каждого нового объекта по его признаковому описанию прогнозировал целевую переменную этого объекта

Регрессия

ID	ARPU	Модель телефона	Интернет-трафик	ДОХОД
9034948911	123.5	Samsung	1500.4	10000
9034948912	245.6	iPhone 6	1124.7	25000
9034948913	890.4	Nokia	2312.6	135000
9034948914	50.3	Samsung	1321.3	90000

АЛГОРИТМ: **ВХОД** - [ID, ARPU, Модель телефона, Интернет-трафик], **ВЫХОД** - [ДОХОД]

903494895	12.6	iPhone 5S	12123.6	?
9034948916	1012.2	HTC	13256.9	?
9034948917	132.9	Samsung	101333.1	?

Регрессия

- Важно помнить:

– Регрессия – это обучение с учителем (supervised learning), в роли учителя выступает обучающая выборка

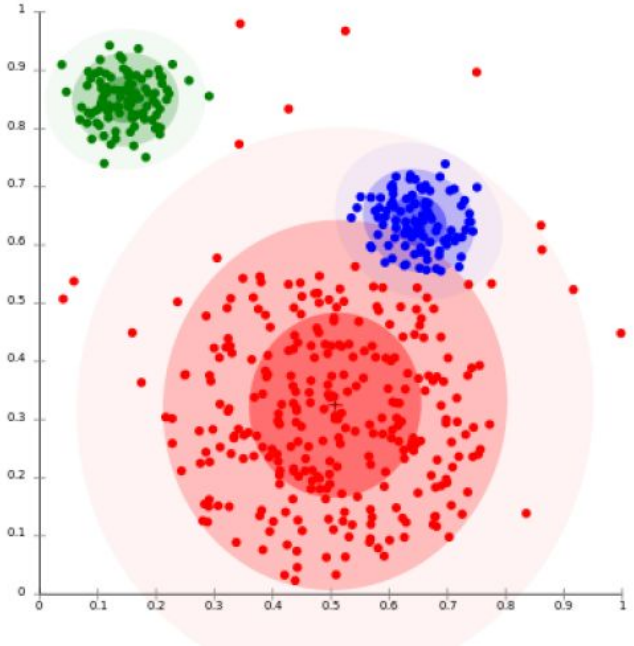
– Регрессия прогнозирует значение целевой переменной для объекта, которая может принимать любое действительное значение

– В результате получается алгоритм, который на вход принимает признаковое описание объекта, а на выходе выдает значение целевой переменной

- Примеры задач:

– Прогнозирование дохода абонента, прогнозирование нагрузки на колл-центр, прогнозирование прибыли ресторана

Кластеризация



- Дано:

– Обучающая выборка, состоящая из признакового описания объектов

- Найти:

– Разделение всех объектов на кластеры

Геометрически, алгоритм группирует данные объекты в кластеры наилучшим образом

Кластеризация

ID	ARPU	Модель телефона	Интернет-трафик	Кол-во блокировок
9034948911	123.5	Samsung	1500.4	5
9034948912	245.6	iPhone 6	1124.7	10
9034948913	890.4	Nokia	2312.6	0
9034948914	50.3	Samsung	1321.3	101
9034948915	12.6	iPhone 5S	12123.6	8
9034948916	1012.2	HTC	13256.9	10
9034948917	132.9	Samsung	101333.1	2
9034948918	152.0	Nokia	1498.2	76
9034948919	14.6	Samsung	4135.7	54

Кластеризация

- Важно помнить:

– Кластеризация – это обучение без учителя (unsupervised learning), размеченная (обучающая) выборка не нужна

– Кластеризация группирует данное множество объектов на кластеры наилучшим образом

– В результате получается алгоритм, который на вход принимает признаковое описание набора объектов и на выходе выдает разбиение объектов на группы

- Примеры задач:

– Выделение домохозяйств среди абонентской базы, выделение сообществ, определение архетипа абонента

Извлечение, отбор и преобразование признаков

Человеко-читаемые, извлекаются сразу

Признаки для простых объектов (человек, сим-карта) берутся на основе целевой переменной:

– Опытным путем (наверное, на доход влияет ARPU)

– Из статей (если задача ранее решалась)

НЕ человеко-читаемые, извлекаются с помощью алгоритмов

Для сложных объектов (лицо на изображении, слова в тексте, номер на видео) признаки извлечь очень тяжело:

– Либо из статей (то, что ученые придумали)

– Либо извлекать автоматически (Deep Learning подход)

Извлечение, отбор и преобразование признаков

- Машина – не человек:
  - Если в качестве признака есть дата, то машина не понимает время суток
  - Если дано имя – машина не понимает, что оно женское
  - Если дан числовой признак – машина не понимает, много это или мало
  - Машина не может группировать признаки
  - Машина не различает «много» или «мало»

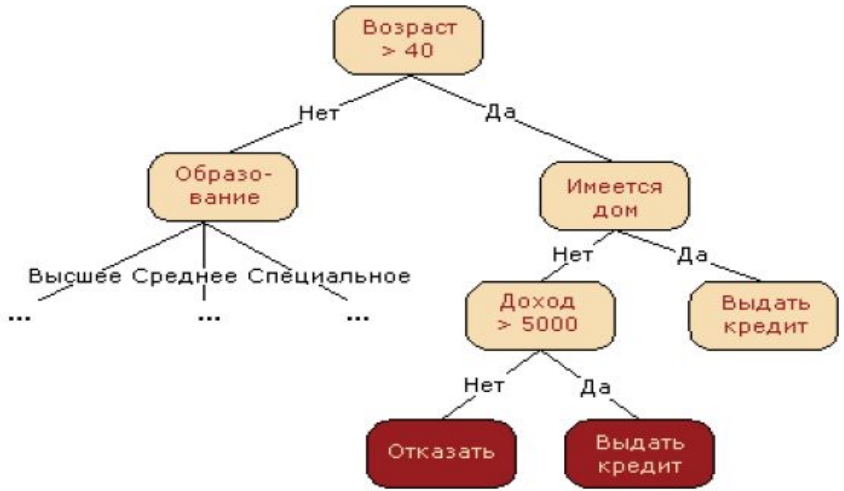
- Примеры преобразования признаков:
  - При прогнозировании спроса на вело прокат дату можно преобразовать в признаки - «утро», «день», «вечер»
  - При прогнозировании цены квартиры «длину» и «ширину» нужно преобразовать в площадь

Алгоритмы машинного обучения

Наиболее простые подходы к задачам машинного обучения:

- Классификация
  - Деревья решений (Decision Trees), метод ближайшего соседа (kNN), метод опорных векторов (SVM)
- Регрессия
  - Линейная регрессия (Linear Regression)
- Кластеризация
  - KMeans, иерархическая кластеризация (Hierarchical Clustering)

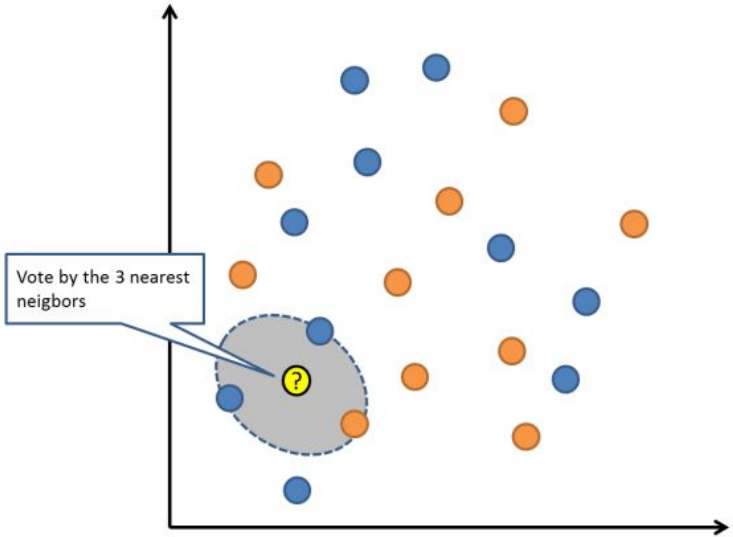
Классификация: деревья решений



- Идея:
  - Пытаемся оптимальным образом построить дерево так, чтобы объекты обучающей выборки классифицировались максимально правильно
- Результат:
  - Для каждого нового объекта сможем пройтись по дереву и классифицировать объект, выдав при этом причину классификации

Наиболее часто используемый алгоритм в медицине и банковском скоринге ввиду человеко-читаемости

Классификация: метод ближайшего соседа



- Идея:

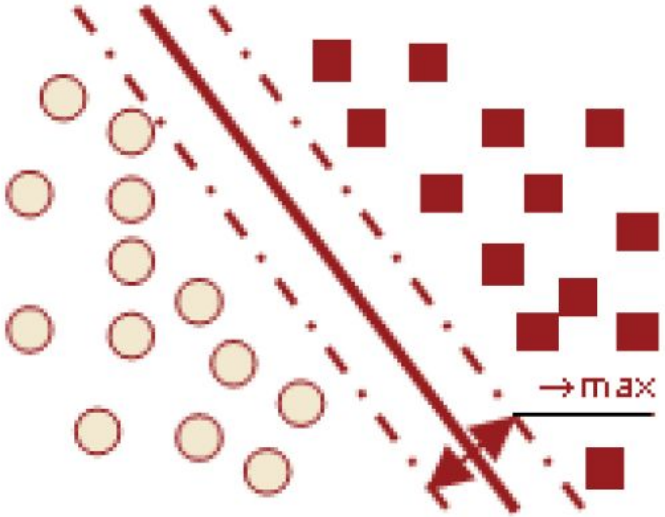
– Наверное, новый объект такого же класса, как и его окружение

- Результат:

– Для каждого нового объекта смотрим его окружение и говорим, на кого он больше похож

Алгоритм не используется в продуктивных задачах, т.к. для каждого нового объекта мы должны искать ближайших – это долго

Классификация: метод опорных векторов



- Идея:

– Пытаемся провести разделяющую поверхность так, чтобы максимизировать зазор между объектами обучающей выборки разных классов

- Результат:

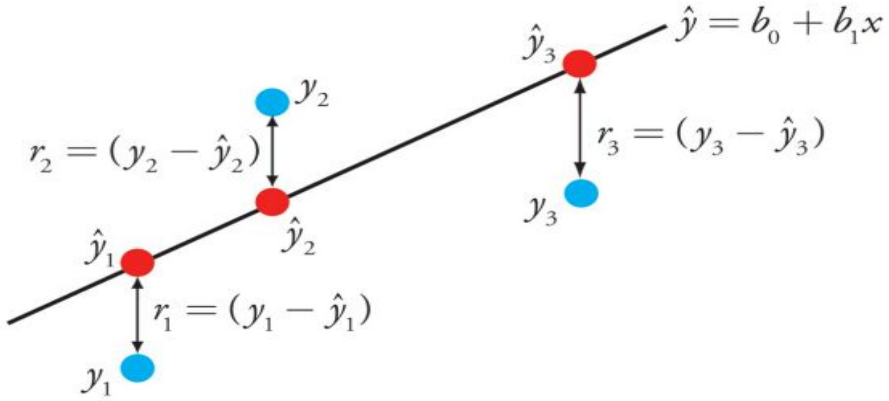
– Для каждого нового объекта смотрим, с какой стороны от разделяющей поверхности он лежит, тем самым, классифицирую объект

Один из самых распространенных алгоритмов классификации, ввиду своей гибкости

Практика 2

- Попробуем решить задачу классификации несколькими различными алгоритмами
- Посмотрим, что у нас получится

Регрессия:
линейная регрессия

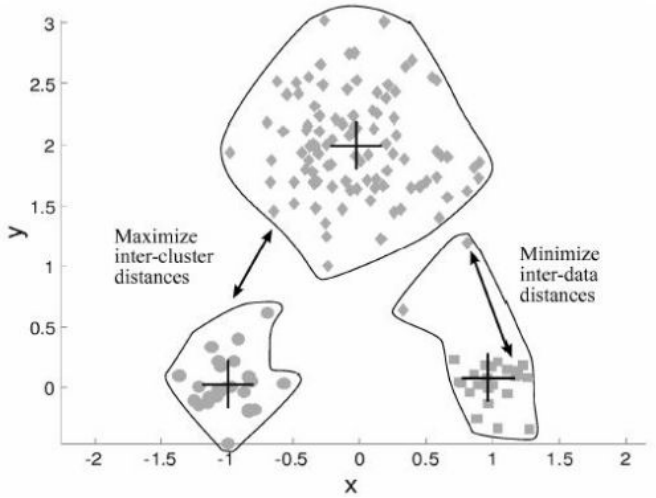


- Идея:
  - Метод наименьших квадратов, известный со школы
  - Ищем значение целевой переменной в виде линейной комбинации признаков

- Результат:
  - Для каждого нового объекта смотрим по восстановленной зависимости (формуле) считаем значение целевой переменной

Линейные модели допускают гибкую настройку и большое количество эвристик. Настраивать сложно, но алгоритмы наиболее подходят для продуктивных решений ввиду своей простоты

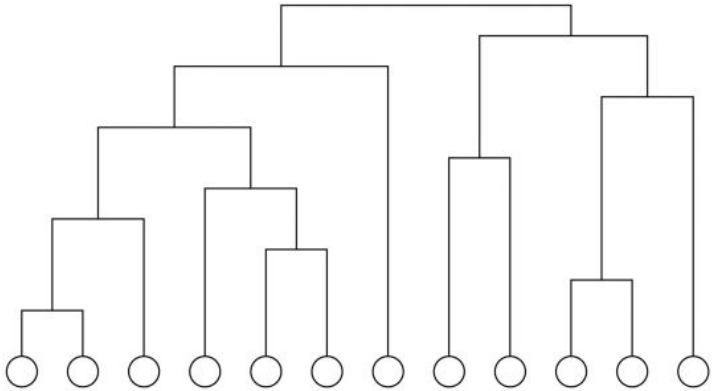
Кластеризация: KMeans



KMeans является наиболее распространенным методом кластеризации,
однако важно правильно определить

- Идея:
  - Задаем количество кластеров
  - Задаем центры кластеров
  - Каждый объект принадлежит к тому кластеру, центр которого ближе
  - Уточняя центры кластеров нахфодим оптимальное разбиение
- Результат:
  - Наиболее оптимальное разбиение данных объектов на кластеры

Кластеризация: иерархическая кластеризация



- Идея:
  - Изначально каждый объект – отдельный кластер
  - Постепенно объединяем похожие кластеры между собой на основе метрики схожести
- Результат:
  - Дендрограмма – иерархическое древовидное представление кластеризации

Иерархическая кластеризация более гибкая,
чем KMeans в бизнес приложениях, но и более чувствительна к настройкам

Метрики качества алгоритмов классификации

Бинарная классификация		Истинные значения	
		1	0
Результат алгоритма	1	TP	FP
	0	FN	TN

Доля

$$\text{Accuracy} = \frac{TP + TN}{TP + FP}$$

Точность

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

Полнота

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

F-мера

$$F = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- True Positive
  - верно угадали 1
- True Negative
  - верно угадали 0
- False Positive
  - ошибка первого рода
- False Negative
  - ошибка второго рода

Практика 3

- Попробуем оценить качество моделей из прошлой практики
- Посмотрим, что у нас получится

Примеры прикладных задач анализа данных: сферы

- Финансовые организации (Bank of America, Citigroup, Сбербанк, HomeCredit)
- Ритейл (Amazon, Target, Metro, Лента)
- Телеком (Vodafone, China Mobile, Вымпелком, МТС, Мегафон)
- Социальные сети (Facebook, Baidu, ВКонтакте, Одноклассники)
- Медицина (Enlitic, Lumiata, Numerate)
- Урбанистика (Uber, ГенПлан, РЖД, ДИТ Правительства Москвы)
- Интернет-компании (Google, Facebook, Яндекс, Mail.ru,)

Примеры прикладных задач анализа данных

- Обработка естественного языка (Natural Language Processing)

– Машинный перевод, анализ отзывов, выделение названий, логические выводы

- Анализ социальных сетей (Social Network Analysis)

– Рекомендация друзей, поиск сообществ, выделение лидеров мнения

- Анализ изображений и видео (Computer Vision)

– Выделение лиц на изображениях, извлечение номеров, названий с камер

- Анализ аудио сигналов (Signal Processing)

– Распознавание речи, классификация музыки, рекомендация плейлиста

- Рекомендательные системы (Recommended Systems)

– Рекомендация товаров, друзей, прогнозирование оценок к фильмам

- Поиск ассоциативных правил (Association Rule Learning)

– Построение логических правил, анализ чеков

Примеры прикладных задач анализа данных

- Поиск спама (Spam Detection)
- Gmail, Mail.ru, Яндекс.Почта, …
- Рекомендательные системы (Product Recommendation)
- Netflix, Amazon, Ozon, RetailRocket, Facebook, …
- Сегментация потребителей (Customer Segmentations)
- Facebook, Google, Яндекс, Вымпелком, …
- Выявление фрода (Fraud Detection)
- Google, Facebook, Вымпелком....
- Прогнозирование оттока (Churn Prediction)
- Amazon, Netflix, Вымпелком, МТС, Мегафон, МГТС, …
- Распознавание речи (Speech Understanding)
- Apple (Siri), Amazon
- Классификация изображений (Image Understanding)
- Facebook, Google, Instagram, Яндекс, Mail.ru, …

Домашнее задание

Домашнее задание

- Взять датасет homework.csv
- Описание датасета доступно тут - https://www.kaggle.com/c/boston-housing/overview
- Решить задачу регрессии (как минимум один из):
  - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html
- Оценить качество регрессии при помощи метрик:
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html

Полезные материалы

Ссылки

- https://scikit-learn.org/stable/index.html
- https://pandas.pydata.org/
- https://habr.com/ru/company/ods/blog/322626/
- https://pandas.pydata.org/pandas-docs/stable/visualization.html
- https://matplotlib.org
- https://netology.ru/blog/03-2019-python-knigi-novichkam

НЕТОЛОГИЯ
групп

Спасибо за внимание!

Алексей Кузьмин  aleksej.kyzmin@gmail.com