Технологии работы с большими данными

Лекция 2. Машинное обучение и Data Science

нетология

Алексей Кузьмин

Директор разработки; ДомКлик.ру

О спикере:

- Руковожу направлением работы с данными и Data Science
- Преподаю в Нетологии
- Работаю в IT с 2010 года (ABBYY, ДомКлик)
- Окончил МехМат МГУ в 2012 году

Я в Слаке:

@Alexey Kuzmin

Работа с данными

Источники данных

Сбор данных
Лекция 8

SQL-БД
Лекция 1

NoSQL
Лекция 4

MapReduce
Лекция 5-7

Мотивация Big Data
Лекция 3

Data Science
Лекция 2

Отчеты
Лекция 1

Модели
Лекция 2

Люди и процессы
Лекция 9

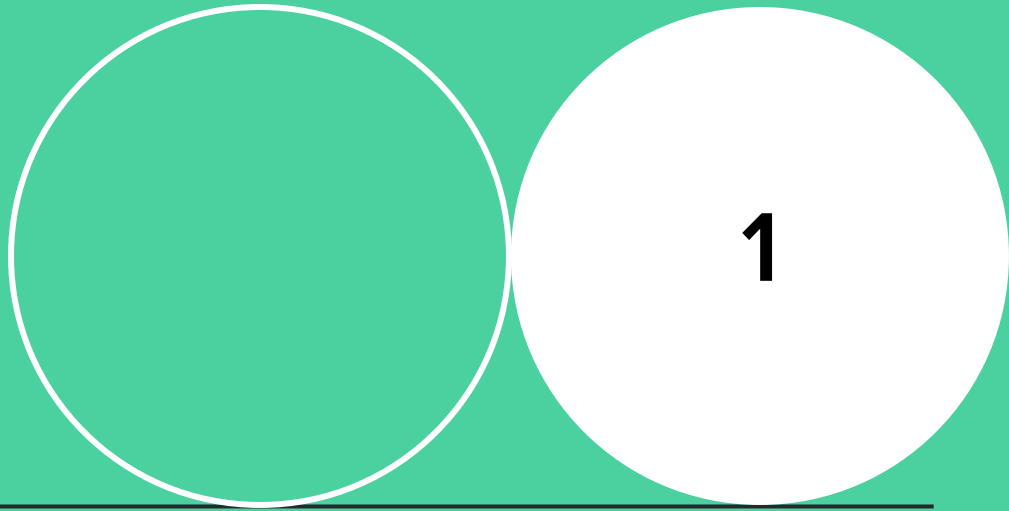
Примеры кейсов
Лекция 10

Что будет сегодня?

- Какие задачи решает машинное обучение?
- Основные инструменты Data Science
- Подробно разберем практический пример

Машинное обучение

Что, где и как?



Алексей Кузьмин

Технологии работы с большими данными

нетология

Объекты и признаки

- Объект - сущность, для которой мы проводим анализ
- Признаки - характеристики объекта

Пример:

- Объект - человек
- Признаки - рост, возраст, вес и тд

3 классические задачи

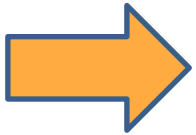
Машинное обучение глядя на выборку объектов с признаками может решать одну из 3-х задач:

- **Классификация** - Определять тип (Мужчина/Женщина)
- **Регрессия** - Прогнозировать значения для объектов (Возраст, Доход, Рост)
- **Кластеризация** - Группировать (Школьники, Бизнесмены, Политики, Любители Чая)

3 классические задачи

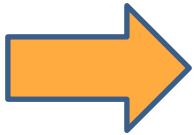
Классификация

Регрессия



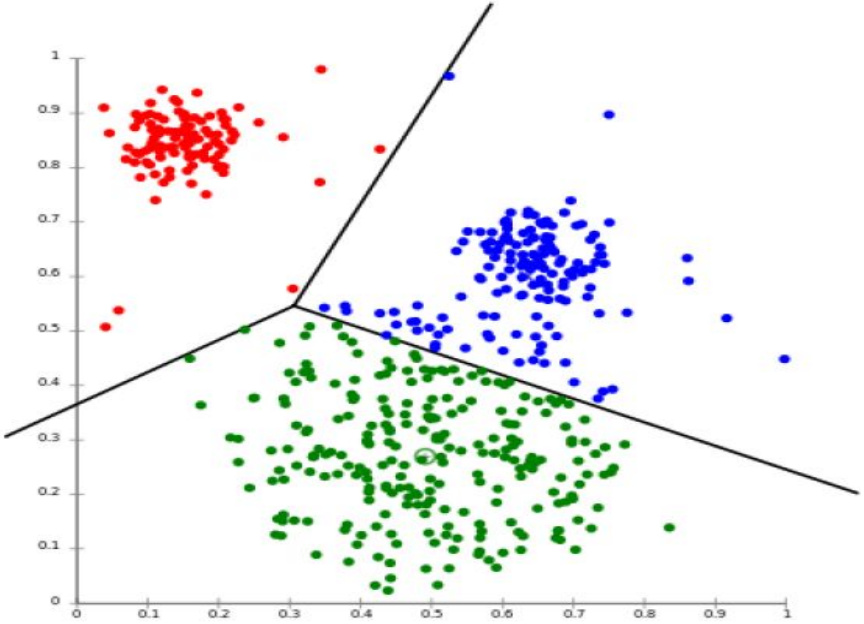
Нужны правильные ответы (обучение с учителем)

Кластеризация



Нужны только объекты и признаки (обучение без учителя)

Классификация



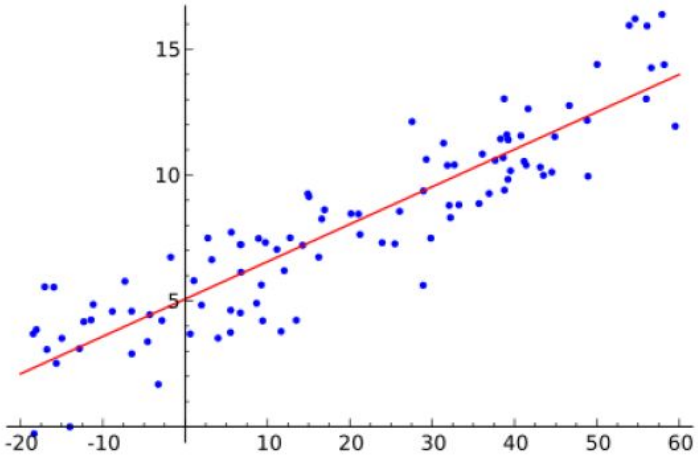
- Дано:

– 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

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

Важно помнить

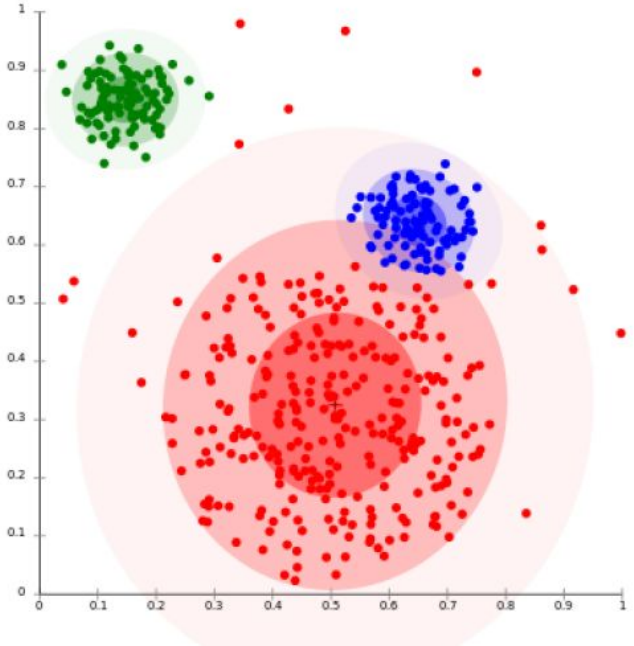
Классификация

- Ответ алгоритма - конечное количество меток
- Нужна “обучающая” выборка - объекты, их признаки и правильные ответы

Регрессия

- Ответом алгоритма может быть любое число
- Нужна “обучающая” выборка - объекты, их признаки и правильные ответы

Кластеризация



- Дано:

– Обучающая выборка, состоящая из признакового описания объектов

- Найти:

– Разделение всех объектов на кластеры

Ответов нет. Есть только объекты и признаки!

Геометрически, алгоритм группирует данные объекты в кластеры наилучшим образом

Примеры задач

Классификация:

- Возьмет клиент кредит или нет
- Что изображено на картинке
- Отзыв положительный или негативный

Регрессия:

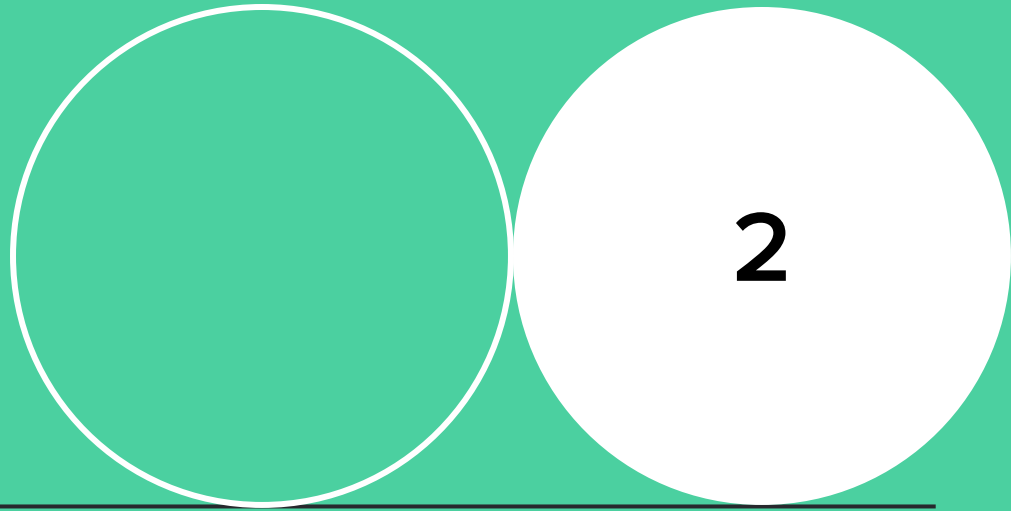
- Предсказание погоды
- Прогноз цены акций
- Прогноз спроса

Кластеризация:

- Какие основные темы обращений клиентов?
- Какие группы пользователей у нас есть?

Инструменты Data Science

Чем будем пользоваться



Алексей Кузьмин

Технологии работы с большими данными

нетология

Python

Python — высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода.

Основной язык для изучения данных и построения моделей машинного обучения.

PANDAS

Модуль Python’а, предназначенный для работы с табличными данными, полученными из различных источников.

SCIKIT-LEARN

Модуль Python’a, предназначенный для работы с алгоритмами машинного обучения

COLAB

Облачная среда для работы с Python, предоставляемая google совершенно бесплатно.

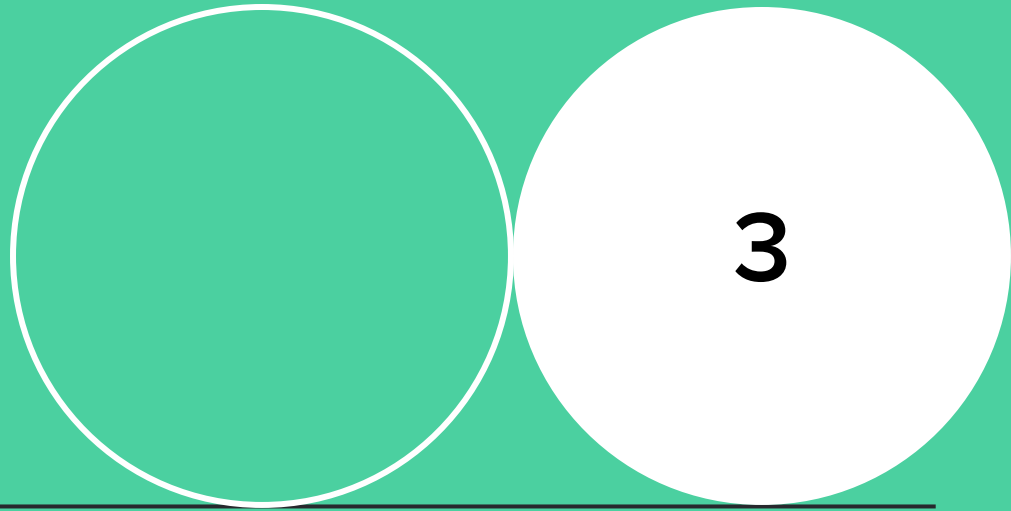
- **Ноутбук** - файл с кодом
- Состоит из **ячеек** (код или текст)
- Код - ячейки с кодом на языке Python. Можно выполнять в произвольном порядке
- Текст - текстовые комментарии (в формате markdown)

Практика 1

- https://colab.research.google.com/ <- основной рабочий инструмент
- Загрузим данные и посмотрим, что у нас с ними
- iris.csv
- Ссылка на ноутбук:
  - https://colab.research.google.com/drive/1ELVgU7aw6Og5elKZqBW7bdM2uKZOnTEM?usp=sharing

Процесс решения

Как решить любую DS-задачу



Алексей Кузьмин

Технологии работы с большими данными

нетология

Общая схема

1. Получить данные
2. Подготовить объекты и признаки
3. Разделить данные на обучающую и тестовую выборку при необходимости
4. Выбрать алгоритм машинного обучения
5. Обучить модель на обучающей выборке
6. Оценить качество на тестовой выборке

1. Получить данные

Много способов. Самый простой - из csv-файла, рассматривали в предыдущей практике

2. Подготовить объекты и признаки

Часто данные бывают некачественными или неподходящими для машинного обучения:

- Есть строковые значения (математика же работает только на цифрах)
- Есть пропуски
- Есть выбросы и шумы

Перед применением алгоритма - данные нужно привести в порядок.

3. Разделить данные

Машина, как и человек, может понять закономерность, а может просто “зазубрить” обучающую выборку.

Нужно уметь честно оценивать качество работы алгоритма на данных, которые он не видел.

Для этого имеющееся множество делят на 2 группы:

- **Обучающая выборка -** используется при обучении алгоритма
- **Тестовая выборка -** скрыта от алгоритма и используется только для оценки качества

4. Выбрать алгоритм

- Алгоритм зависит от:
  - Задачи (классификация/регрессия/кластеризация)
  - Структуры и особенностей данных
  - … <- работа data science

5. Обучить модель

Выбрав алгоритм, ему надо подать на вход обучающуюу выборку, чтобы он на ее основе вывел основные закономерности в данных (Обучился)

6. Оценить качество

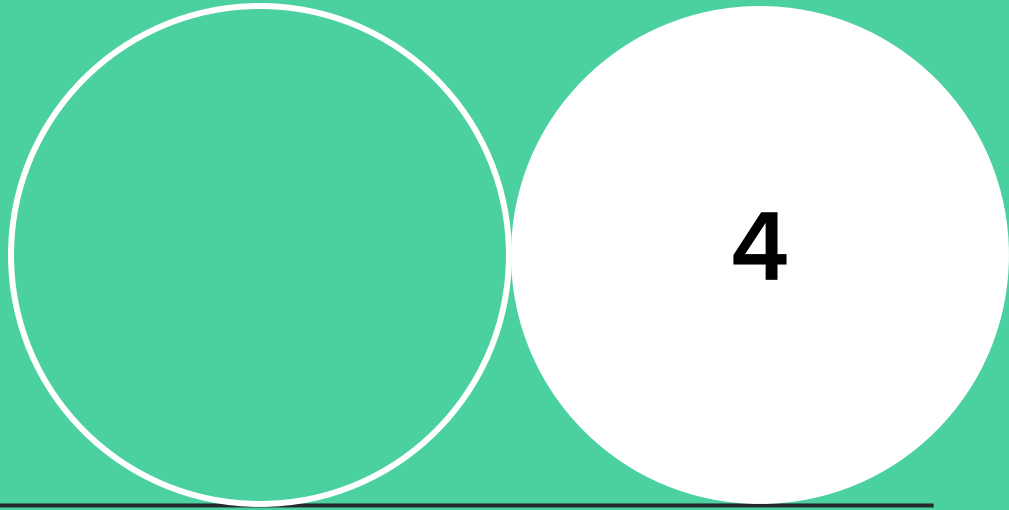
Чтобы оценить качество алгоритма нужно:

1. Выбрать меру качества (в зависимости от задачи они бывают разные)
2. Сделать предсказания для тестовой выборки
3. Оценить насколько они похожи на правильные ответы

Качество алгоритма оцениваем на тестовой выборке!

Алгоритмы

Некоторые основные модели МО



Алексей Кузьмин

Технологии работы с большими данными

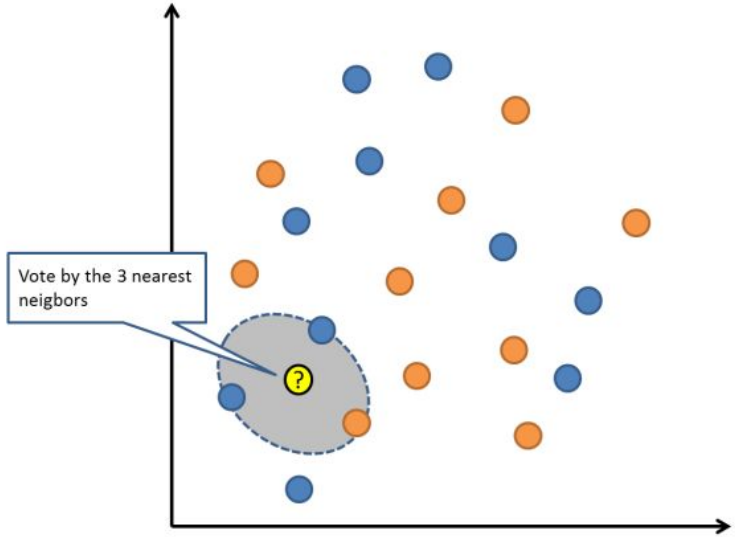
нетология

Классификация: деревья решений

Пытаемся оптимальным образом построить дерево так, чтобы объекты обучающей выборки классифицировались максимально правильно

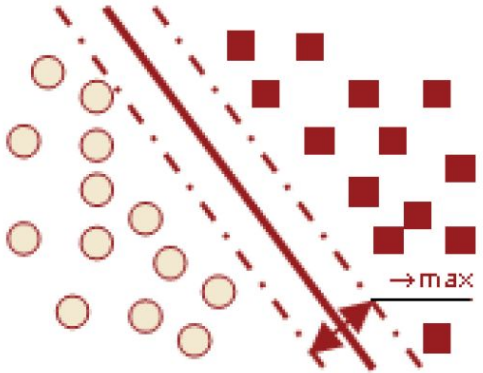
Классификация: метод ближайшего соседа

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



Классификация: метод опорных векторов

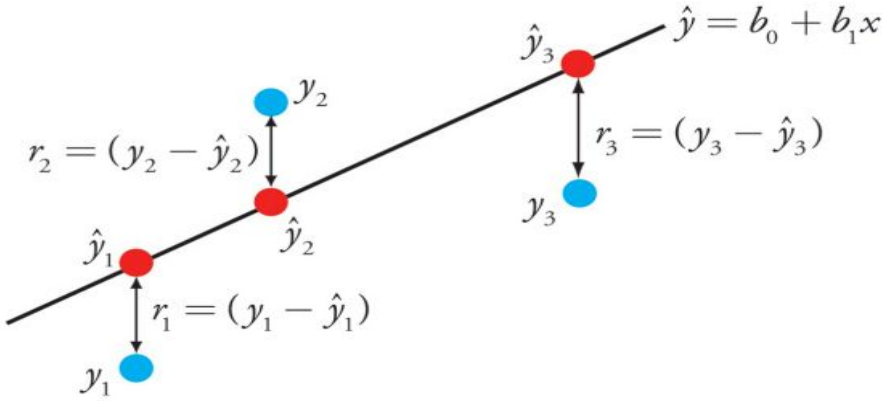
Пытаемся провести разделяющую поверхность так, чтобы максимизировать зазор между объектами обучающей выборки разных классов



Регрессия: линейная регрессия

– Метод наименьших квадратов, известный со школы

– Ищем значение целевой переменной в виде линейной комбинации признаков



Кластеризация: KMeans

Идея:

– Задаем количество кластеров

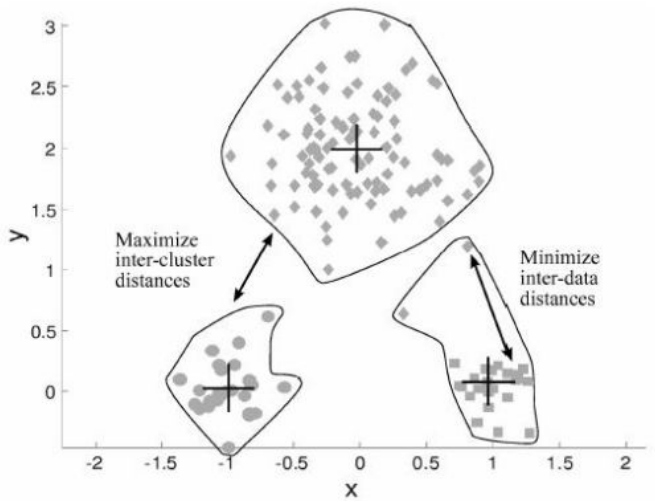
– Задаем центры кластеров

– Каждый объект принадлежит к тому кластеру, центр которого ближе

– Уточняя центры кластеров нахфодим оптимальное разбиение

Результат:

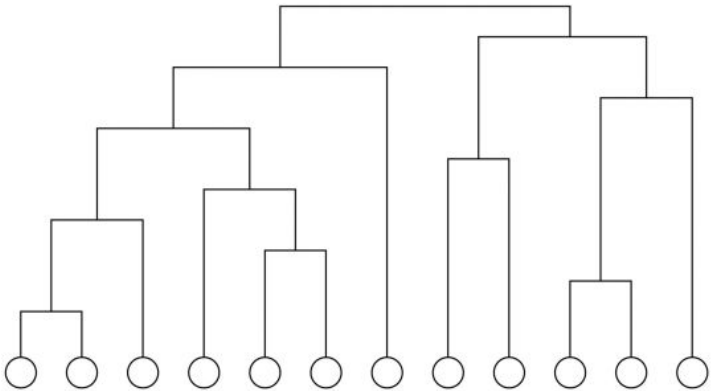
– Наиболее оптимальное разбиение данных объектов на кластеры



Кластеризация: иерархическая кластеризация

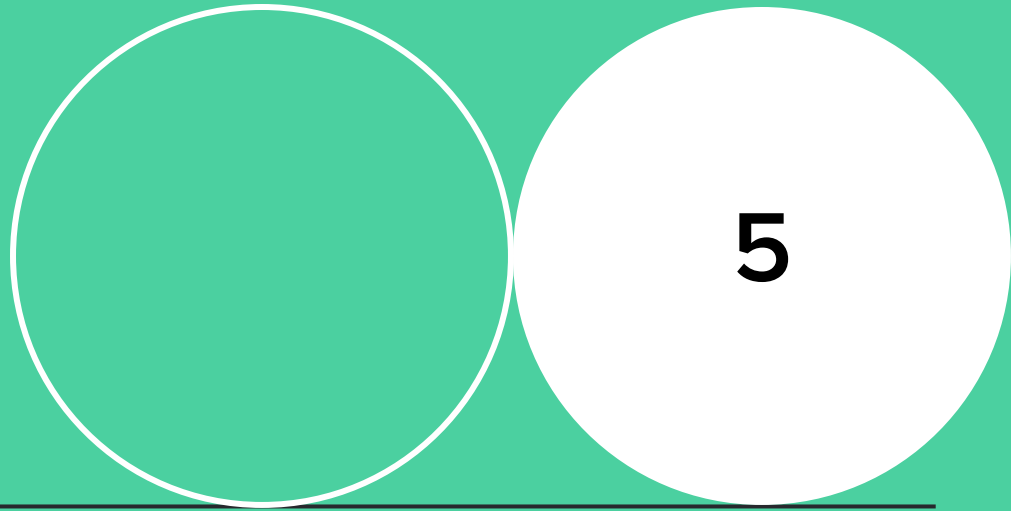
– Изначально каждый объект – отдельный кластер

– Постепенно объединяем похожие кластеры между собой на основе метрики схожести



Метрики

Как понять как обучили?



Алексей Кузьмин

Accuracy

Для задачи классификации.

Точность предсказания (не путать с precision).

Количество правильно классифицированных / общее количество примеров

MSE

Для задачи регрессии.

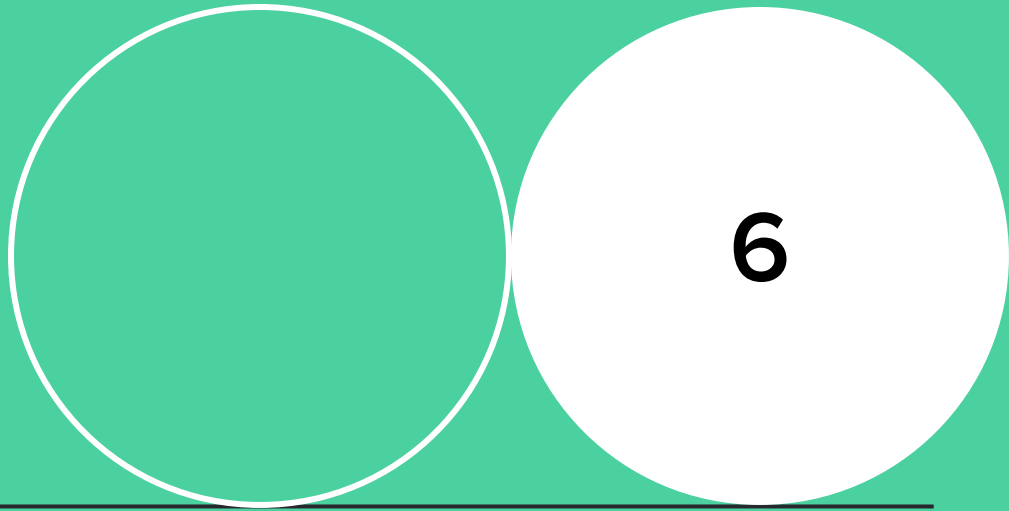
Среднее отклонение предсказания от правильного значения

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Практика 2

Применим машинное обучение к нашей задаче

Домашнее задание



Алексей Кузьмин

Технологии работы с большими данными

нетология

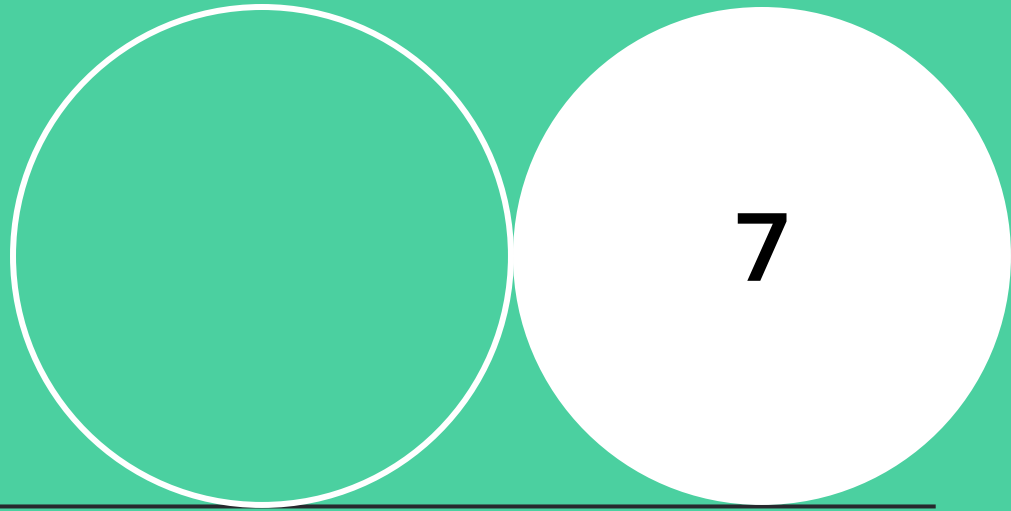
Домашнее задание

- Взять датасет homework.csv
- Описание датасета доступно тут - https://www.kaggle.com/c/boston-housing/overview
- Предсказываем значение столбца MEDV на основе других признаков
- Решить задачу регрессии, используя алгоритм линейной регрессии:
  - https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- Оценить качество регрессии при помощи метрики MSE:
  - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html

Шаблон для выполнения домашнего задания:

https://colab.research.google.com/drive/1x4tenHozBvzWG-I_TcKcdCS12V0xhZ8f?usp=sharing

Полезные материалы



Алексей Кузьмин

Технологии работы с большими данными

нетология

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
- https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f

Спасибо за внимание

Алексей
Кузьмин

нетология