Машинное обучение для жизни

Курс «Аналитическое мышление»

Алексей Кузьмин
Директор разработки ДомКлик.ру

Алексей Кузьмин

Директор разработки
ДомКлик.ру

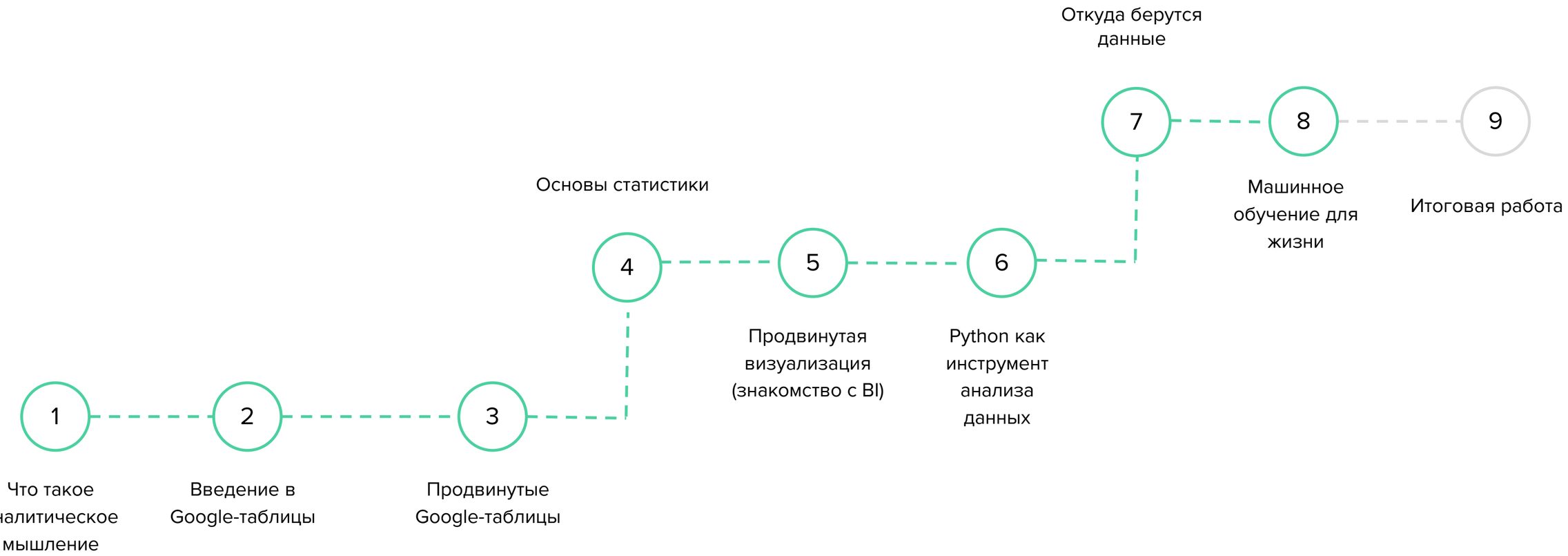
О спикере

- 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

В слаке

@Alexey Kuzmin

Структура курса



План занятия

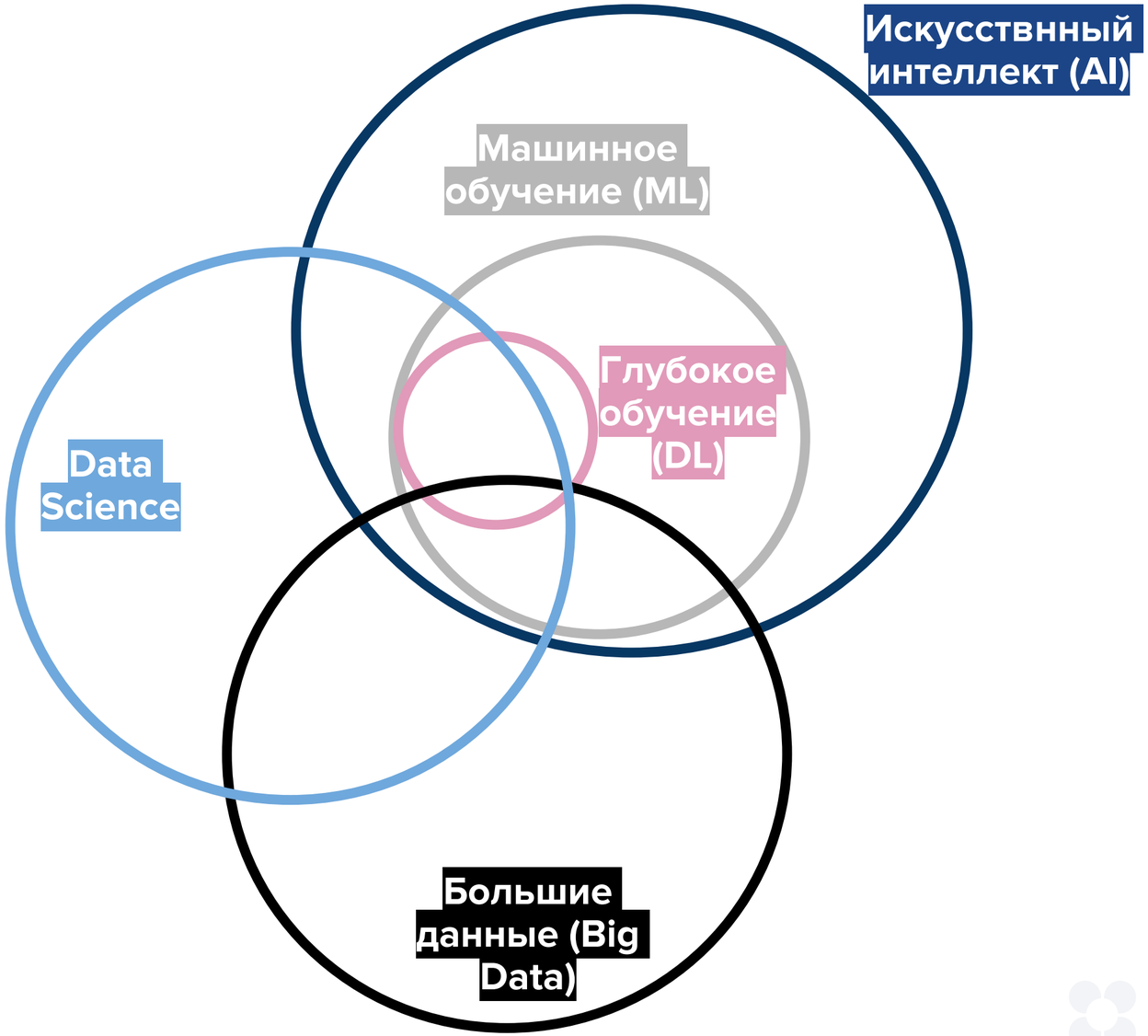
1. Что такое машинное обучение и зачем оно нужно
2. Три вида задач машинного обучения
3. Типовой процесс решения Data Science задачи
4. Пример конкретного алгоритма машинного обучения: в теории и на Python

Машинное обучение

1

Области Data Science

- **Искусственный интеллект:** «научить машины думать»
- **Машинное обучение:** инструменты для извлечения знаний из данных
- **Глубокое обучение:** многослойные нейронные сети
- **Data Science:** понимание и придание смысла данным
- **Большие данные:** совокупность подходов к обработке огромных объемов неструктурированных данных



HP

Компания Hewlett-Packard оценивает риски ухода своих сотрудников, которых более 333 000 по всему миру.

Алгоритм помогает менеджерам заранее найти замену человеку, который планирует покинуть свою должность в компании.

Результат: Система с 80% точностью вычисляла имена тех, кто планирует уйти. Иногда человек ещё сам не знал об этом, а машина «предсказывала» его судьбу на ближайшие полгода.

Dubai Airports

Международный аэропорт Дубая. Большие данные о показателях работы аэропорта, рейсах и перемещении пассажиров широко используются для оптимизации работы аэропорта и повышения удовлетворенности пассажиров.

Что в основе: Аэропорт использует сложные алгоритмы оптимизации для динамического назначения выходов для посадки и прилета.

Пассажиры посещают Дубай для шоппинга, что приводит к частым опозданиям на рейс. Многие не говорят на английском или арабском – языках, на которых делаются объявления.

После внедрения новой программы оповещений в каждом магазине аэропорта сканируются посадочные талоны пассажиров, и они получают оповещения на языке, которым владеют.

Результат: оптимизировано назначение выходов на посадку и прилет, значительно сократилось число опозданий на рейсы.

Сбербанк

2014

Сбербанк разработал систему анализа фотографий для идентификации клиентов и предотвращения мошенничества с документами

Как работает: в основе работы системы лежит сравнение фотографий с помощью технологий компьютерного зрения.

Платформа: биометрическая платформа «Каскад-Поиск» от компании «Техносерв».

Преимущества: система работает очень быстро. Благодаря ряду инновационных решений, таких как In-Memory Processing, сопоставление изображений камеры и изображений в базе занимает несколько секунд.

Результат: потери от мошенничества с документами физических лиц сократились в 10 раз.

https://habr.com/ru/company/newprolab/blog/318208/

Машинное обучение

Основные понятия

1

2

Объекты и признаки

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

Пример:

- Объект: человек
- Признаки: рост, возраст, вес и т.д.

3 классические задачи

Классификация — определять тип (мужчина/женщина)

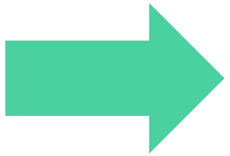
Регрессия — прогнозировать значения для объектов (возраст, доход, рост)

Кластеризация — группировать (школьники, бизнесмены, политики, любители чая)

3 классические задачи

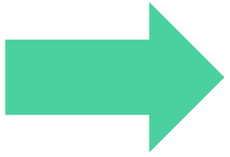
Классификация

Регрессия



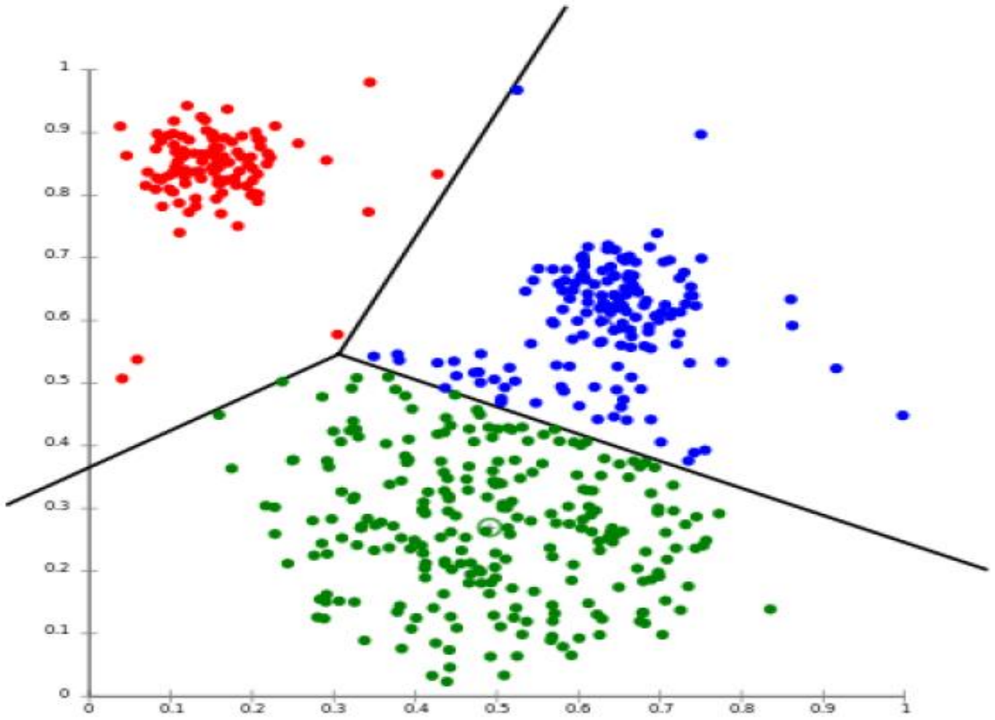
Нужны правильные ответы (обучение с учителем)

Кластеризация



Нужны только объекты и признаки (обучение без учителя)

Классификация



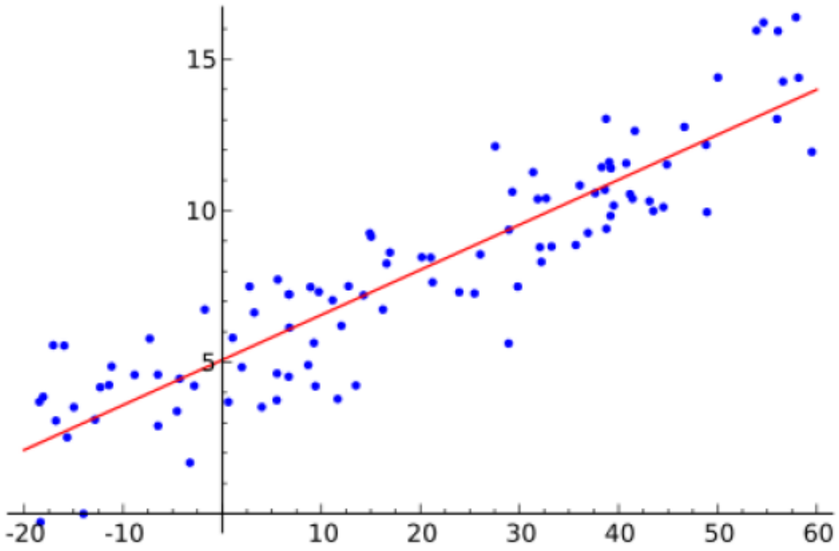
Дано:

– 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



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.

Геометрически алгоритм восстанавливает зависимость между признаками и целевой переменной.

Важно помнить

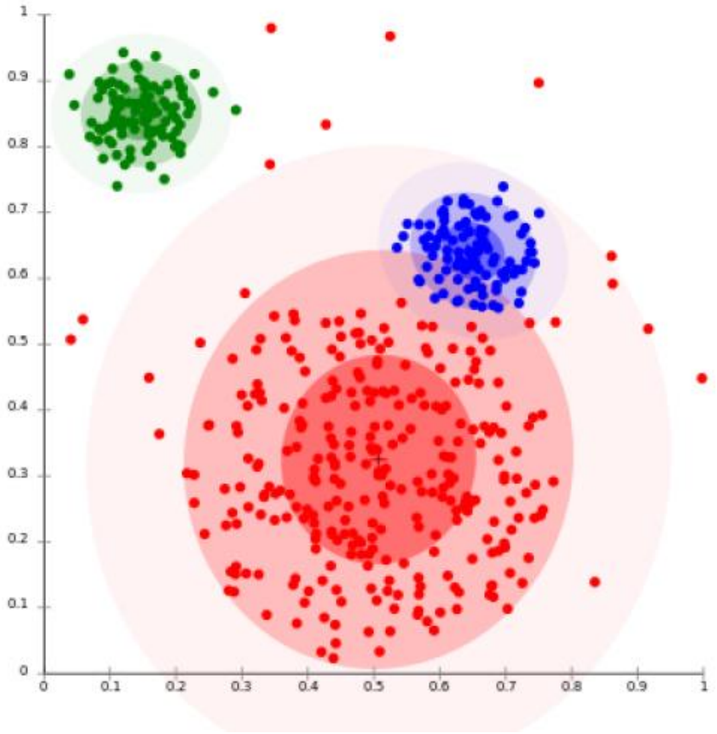
Классификация

- 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



Геометрически алгоритм группирует данные объекты в кластеры наилучшим образом

Дано:

– обучающая выборка, состоящая из признакового описания объектов.

Найти:

– разделение всех объектов на кластеры.

Ответов нет. Есть только объекты и признаки!

Примеры задач

Классификация:

- Возьмет ли клиент кредит
- Что изображено на картинке
- Будет ли отзыв положительным

Регрессия:

- Предсказание погоды
- Прогноз цены акций
- Прогноз спроса

Кластеризация:

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

Процесс решения

Как решить любую DS-задачу

1 2 3

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

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

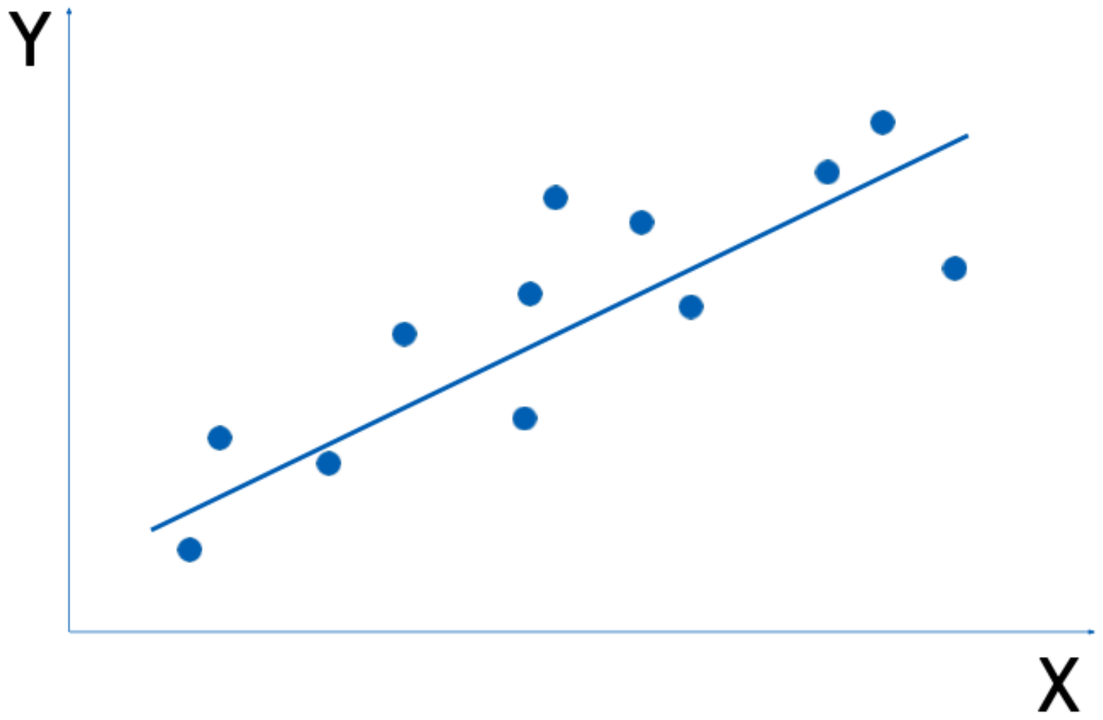
Линейная регрессия

Как пример

1 2 3 4

Линейная регрессия

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$



y^ — предсказание модели

b0, b1, ..., bk — коэффициенты модели

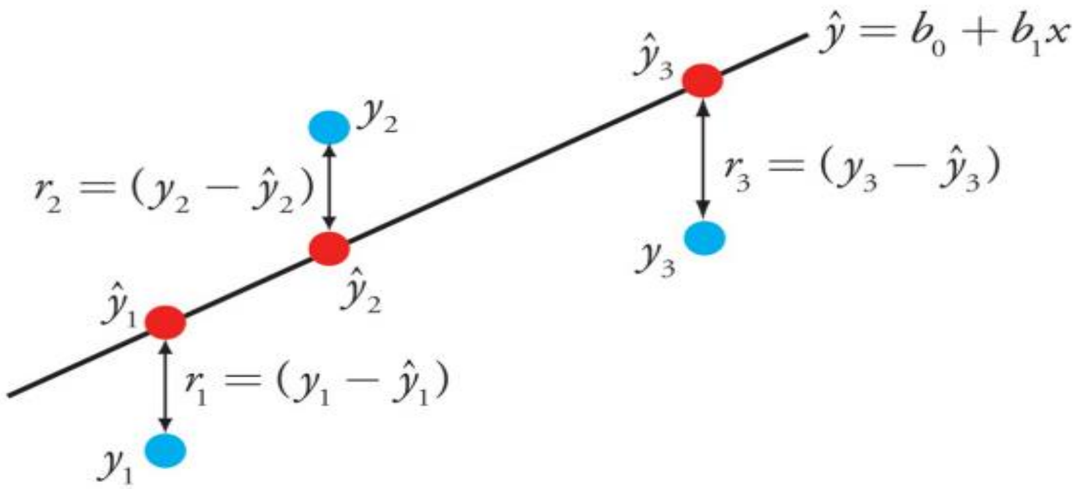
x1, ... , xk — признаки объекта

Как оценить качество модели?

Разные коэффициенты b0, b1, …, bk дают разные по качеству модели. Как понять, какая из них наилучшего качества?

Вводим функцию качества предсказания модели:

$$S^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



Часто функцию S2 дополнительно усредняют по количеству наблюдений. Получается функция MSE — средняя квадратичная ошибка

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Пример

Разберем на примере домов

количество спален	площадь	этажей	реальная цена
2	50	1	10000000
1	40	1	6000000
4	120	2	25000000

Пример

Есть две модели с разными коэффициентами

модель	стартовая стоимость дома	стоимость спальни	стоимость метра	стоимость этажа
1	1000000	1000000	10000	1000000
2	2000000	2000000	5000	5000000

Пример

Для каждой модели и каждого дома можно построить прогноз

Цена дома 1 по версии модели 1 = стартовая стоимость дома (модель 1) + стоимость спальни (модель 1) * количество спален (дом 1) + стоимость метра (модель 1) * количество метров (дом 1) + стоимость этажа (модель 1) * количество этажей (дом 1)

Ошибка прогноза дома 1 для модели 1 = (цена дома 1 по версии модели 1 - цена дома 1 реальная) ** 2

Пример

реальная цена	прогноз цены по модели 1	прогноз цены по модели 2	ошибка модели 1	ошибка модели 2
10000000	4500000	11250000	30250000000000	1562500000000
6000000	3400000	9200000	6760000000000	10240000000000
25000000	8200000	20600000	282240000000000	19360000000000

ср. ошибка модели 1	ср. ошибка модели 2
106416666666667	10387500000000

Средняя ошибка меньше у модели 2 => она лучше, чем модель 1

Линейная регрессия в жизни

- Прогноз продаж
- Предсказание цены товара на основе его характеристик
- Построение трендов на основе предыдущих значений
- ...

Одна из самых простых моделей, которую можно построить

А как на Python?

Пример в коде

1 2 3 4 5

scikit-learn

Библиотека машинного обучения на Python

- Огромный набор инструментов для создания моделей на основе машинного обучения
- Качественная документация
- Высокая скорость работы
- Единообразный API взаимодействия

https://scikit-learn.org

Содержит

Множество моделей

- Включая линейную регрессию
- from sklearn.linear_model import LinearRegression

Методы оценки качества

- Включая mse
- from sklearn.metrics import mean_squared_error

Методы разделения данных

- Включая разделение на выборку для обучения и выборку для валидации
- from sklearn.model_selection import train_test_split

Единый подход для всех моделей

```
from sklearn.linear_model import LinearRegression

X, y = КАКИЕ-ТО ДАННЫЕ

model = LinearRegression(fit_intercept=True)

model.fit(X, y)

prediction = model.predict(X)
```

Практика

- Попробуем предсказать цены на дома в Бостоне при помощи линейной регрессии

Итоги

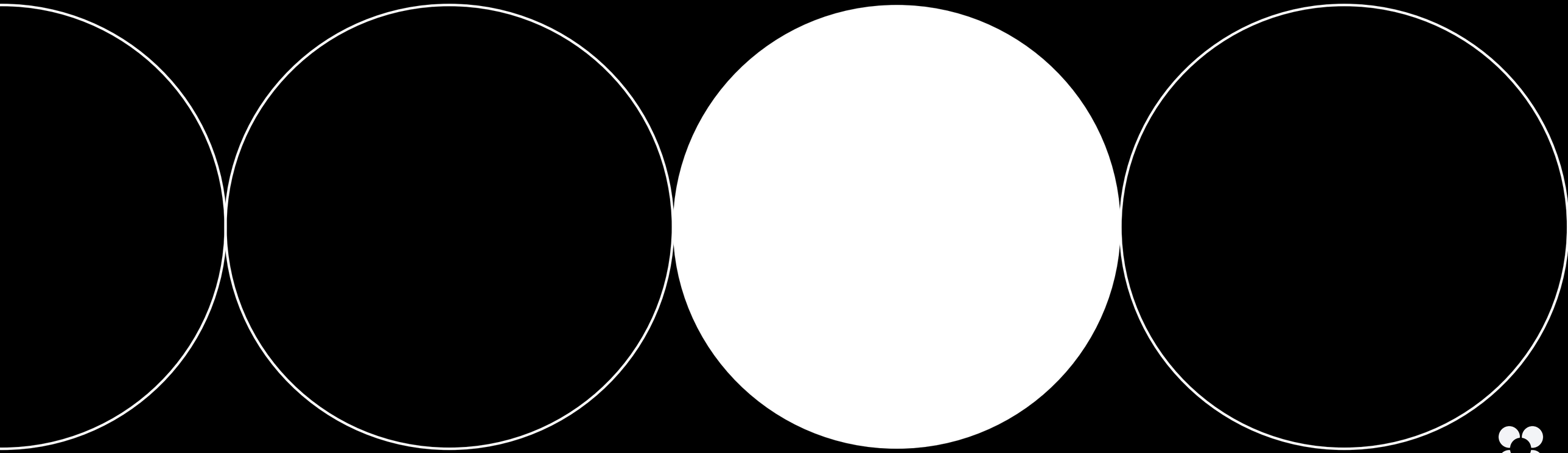
Алексей Кузьмин

Аналитическое мышление

Что мы узнали сегодня

- Что такое машинное обучение и посмотрели кейсы, как оно применяется в реальной жизни
- Что такое классификация, регрессия и кластеризация — три основных вида задач машинного обучения
- Поняли, как обычно решаются задачи машинного обучения
- Познакомились с линейной регрессией и научились предсказывать цены на дома на ее основе

Дополнительные ресурсы



Рекомендуемая литература

- Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных
- Python и машинное обучение
- Профессия «Data Scientist» в Нетологии

Спасибо за внимание!

Алексей
Кузьмин

нетология