



Data Mining in Action: Introduction

vk Data Mining in Action | ВКонтакте[vk.com > data_mining_in_action](#) ▼

Москва, Россия Денис Семенов. Администратор сообщества. Data Mining in Action. So it begins. Местоположение: Москва, Россия. . Data Mining in Action запись закреплена. 6 мая в 23:04.

Нашлось 8 млн результатов

[Дать объявление](#) [Показать все](#)**h** Process Mining: знакомство / Хабрахабр[habrahabr.ru > post/244879/](#) ▼

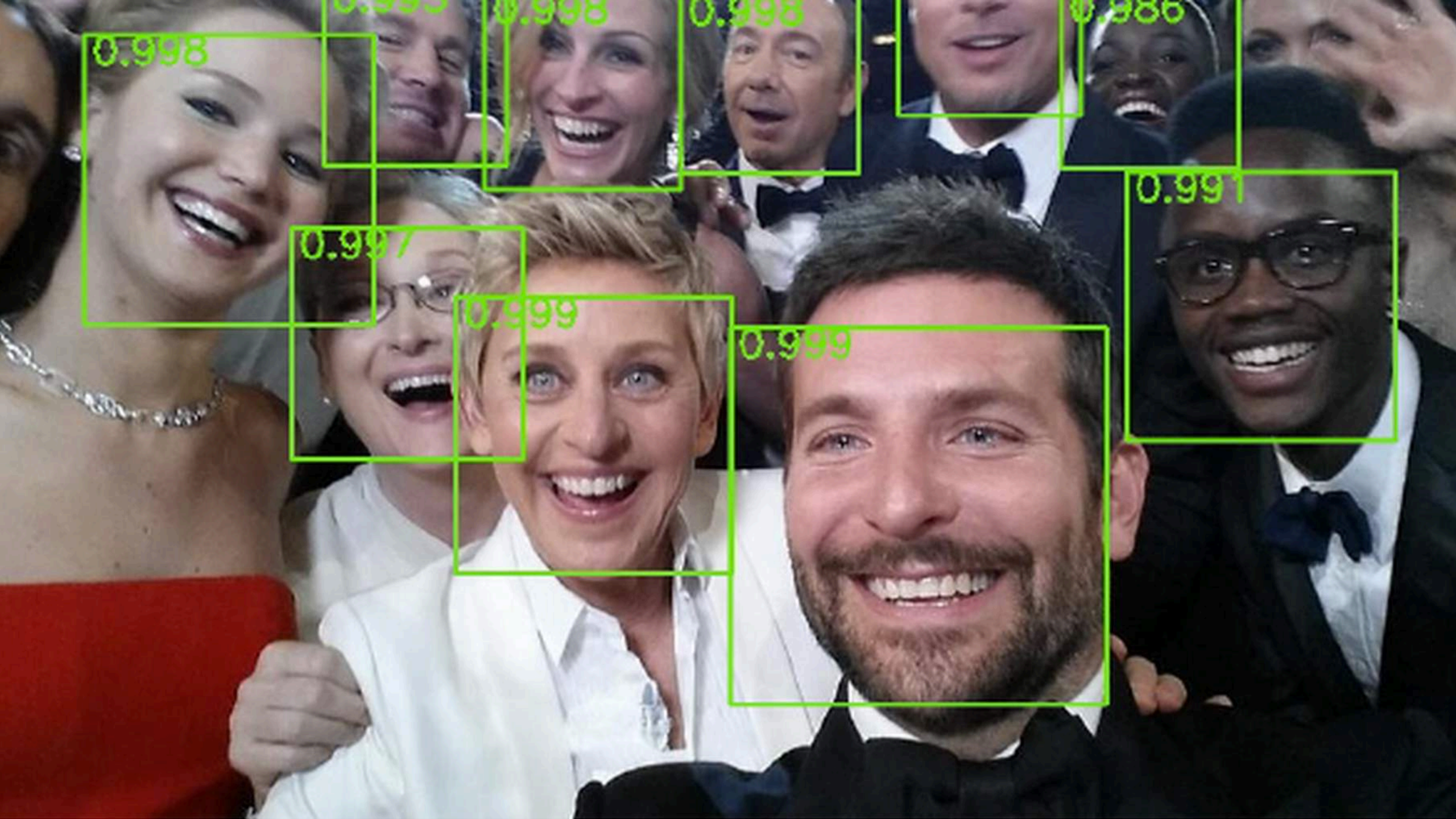
Статья подготовлена на основе материалов онлайн курса Process **Mining: Data Science in Action**, являющихся собственностью Технического университета Эйндховена.

∞ Process Mining: Data science in Action... | Coursera[coursera.org > learn/process-mining](#) ▼









0.998

0.998

0.998

0.998

0.986

0.997

0.999

0.999

0.991

Направления



Ирхин Илья

Руководитель направления
«Индустрия»

Lead Data Scientist

Яндекс.Такси

Что будет на «Индустрии»

- Постановка задач
- Оценка качества в задачах ML
- Часто используемые на практике ML методы
- Инструменты для анализа данных
- Создание прототипов

Постановка задач

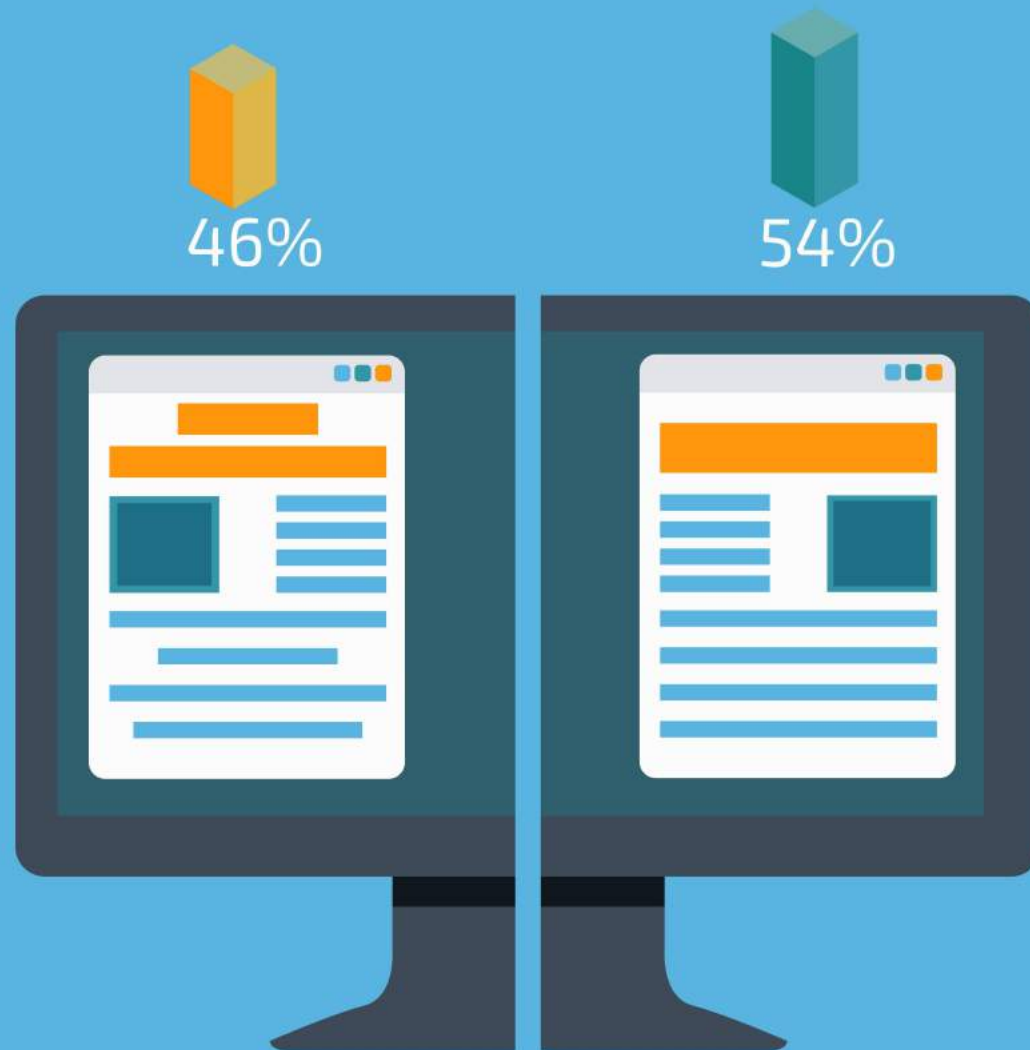
FRAME
ORDER



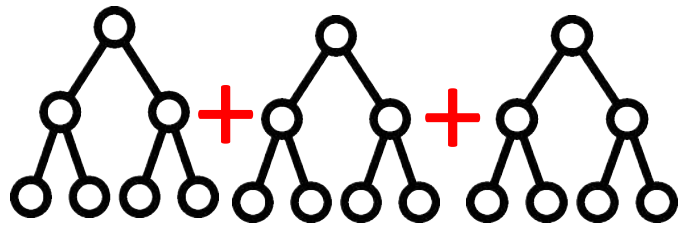
Оценка качества

- Обзор существующих метрик качества
- Переход от задачи бизнеса к метрикам качества
- Оценка потенциального экономического эффекта
- Оценка качества в продакшене

A/B-тестирование



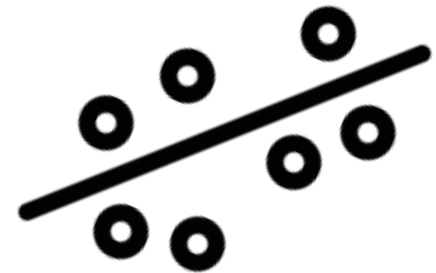
Основные алгоритмы ML



Градиентный бустинг




Случайный лес



Линейные модели

Инструменты для анализа данных

1. *XGBoost*
2. LigthGBM
3. Vawpal Wabbit
4. *Spark* 

И многие другие библиотеки

Инструменты для анализа данных

1. *XGBoost*
2. LigthGBM
3. Vowpal Wabbit
4. *Spark* 

И многие другие библиотеки

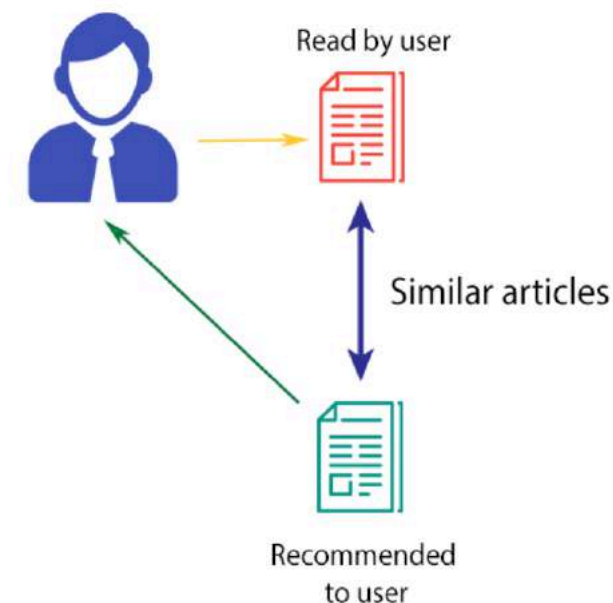
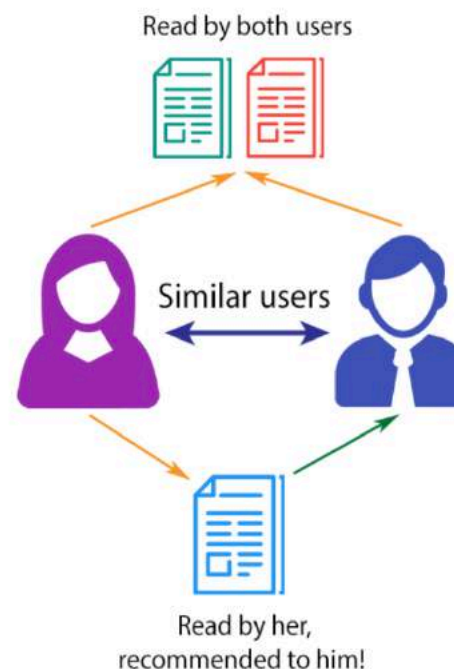
Создание прототипов



Практические кейсы

- Рекомендательные системы
- Анализ тональности отзывов
- Оптимизация техподдержки
- Прогнозирование спроса
- Предсказание оттока
- Оптимизация рекламы

и другие примеры



Вы научитесь:

- Делать правильные с точки зрения бизнеса постановки задач
- Оценивать качество ваших решений как в оффлайн, так и в онлайн экспериментах
- Разбираться в часто используемых на практике методах
- Владеть инструментами для анализа данных
- Создавать прототипы продукта с ML
- Решать реальные практические кейсы, применяя полученные знания

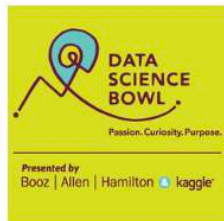
Гущин Александр

Руководитель
направления
«Спорт»

Senior DS Я.Такси



15 Active Competitions



2018 Data Science Bowl

Find the nuclei in divergent images to advance medical discovery

Featured · 2 months to go · biology

\$100,000
1,223 teams



Mercari Price Suggestion Challenge

Can you automatically suggest product prices to online sellers?

Featured · 12 days to go ·

\$100,000
2,236 teams



Toxic Comment Classification Challenge

Identify and classify toxic online comments

Featured · a month to go · arguments, text data

\$35,000
1,780 teams



Nomad2018 Predicting Transparent Conductors

Predict the key properties of novel transparent semiconductors

Research · 6 days to go · chemistry, semiconductors

€5,000
851 teams

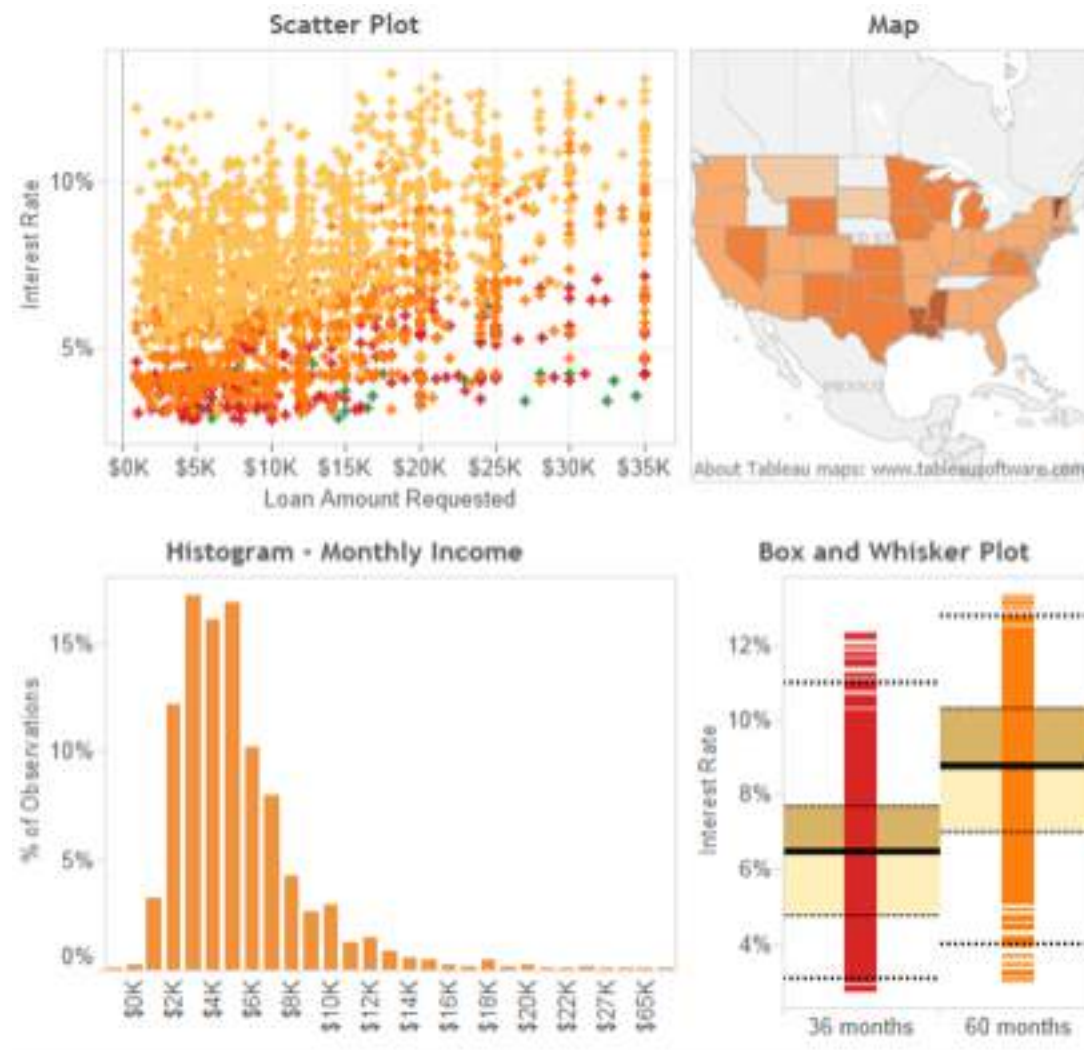
kaggle.com

Что будет на “Соревнованиях”

- Exploratory Data Analysis
- Извлечение и генерация признаков
- Валидация качества решения
- Утечки в соревнованиях
- Ансамбли моделей

Exploratory Data Analysis

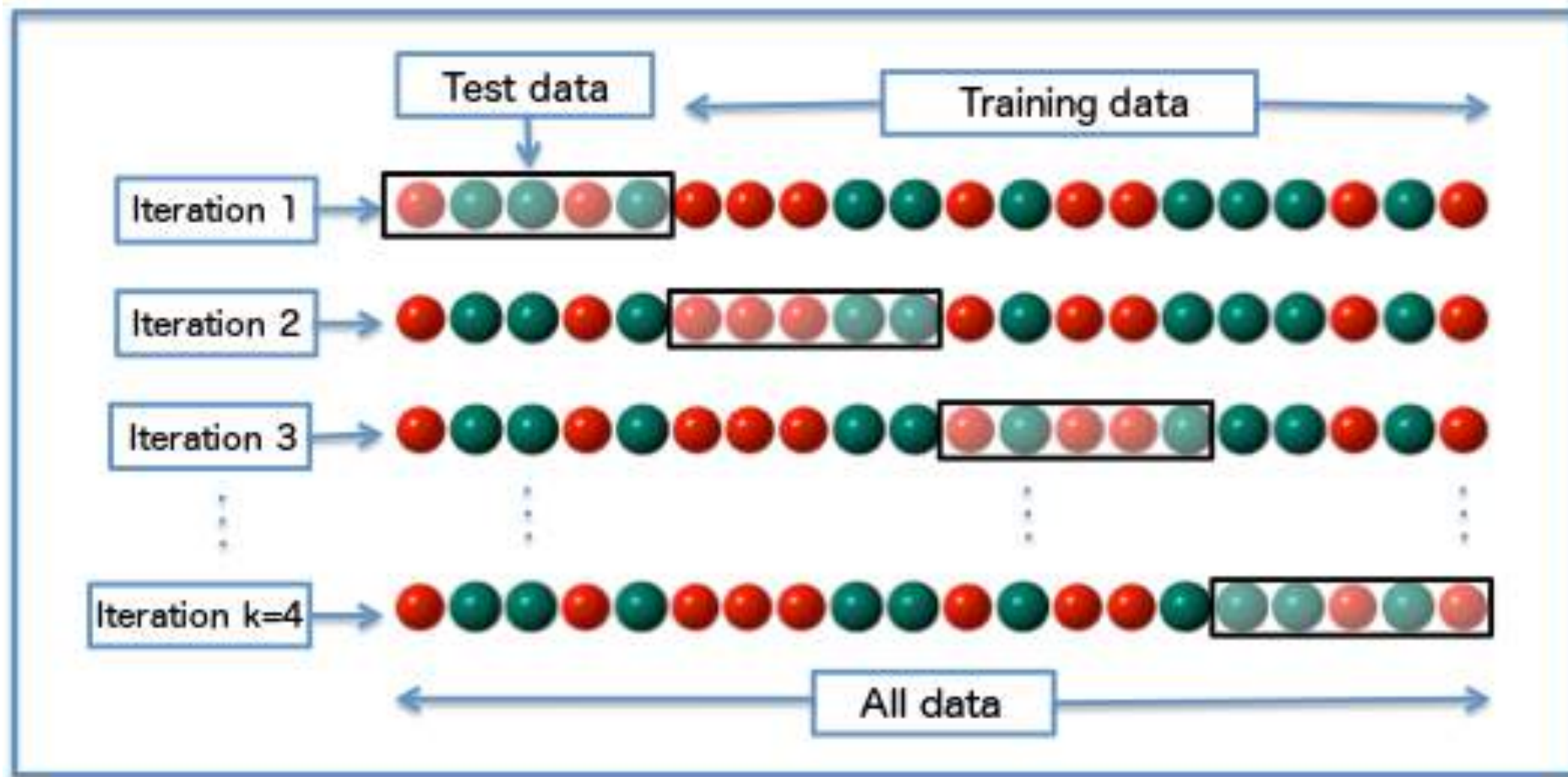
- Визуализировать данные
- Посчитать статистики
- Понять данные и найти новые гипотезы



Генерация признаков

- Время
- Координаты
- Картинки
- Кодирование категориальных признаков средним
- Использование ближайших соседей

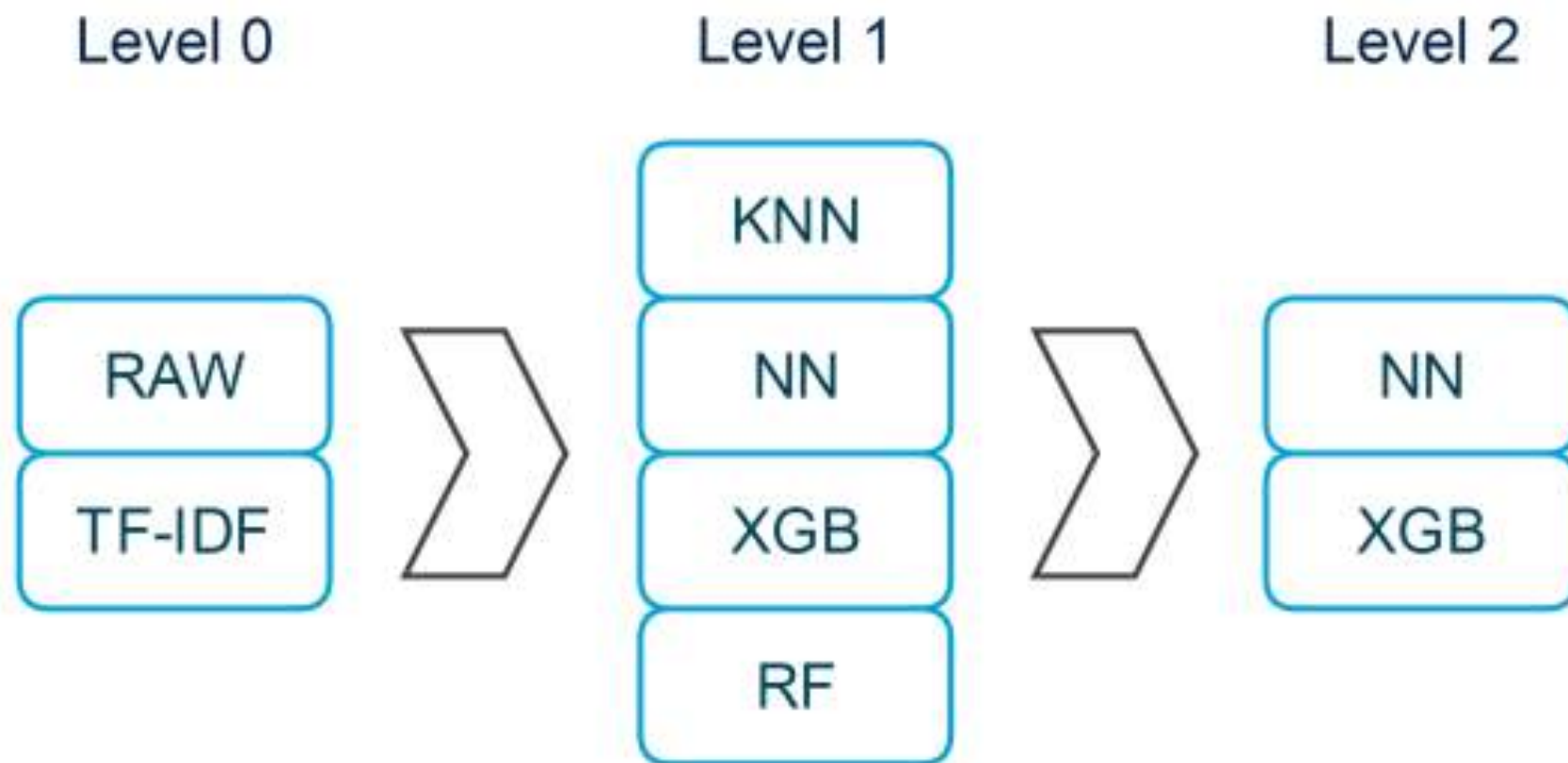
Валидация



Утечки в соревнованиях

- Повторяющиеся строки в train и test
- Порядок строк имеет значение
- Временные ряды с признаками “из будущего”
- etc

Ансамбли



A man with dark hair and glasses, wearing a dark polo shirt, stands with his arms crossed in front of a large window. The window shows a cityscape with buildings and a blue sky with clouds. The man is looking directly at the camera.

Панкратов Антон*

Руководитель направления
«Тренды»

Lead DS Я.Такси

*Антон не любит фотографироваться,
поэтому на фото Ян ЛеКун



Google

Курс Deep Learning
Январь 2016

Andrew Ng

Специализация по
Deep Learning на
Курсере

Август 2017



DMIA: Trends

Object detection

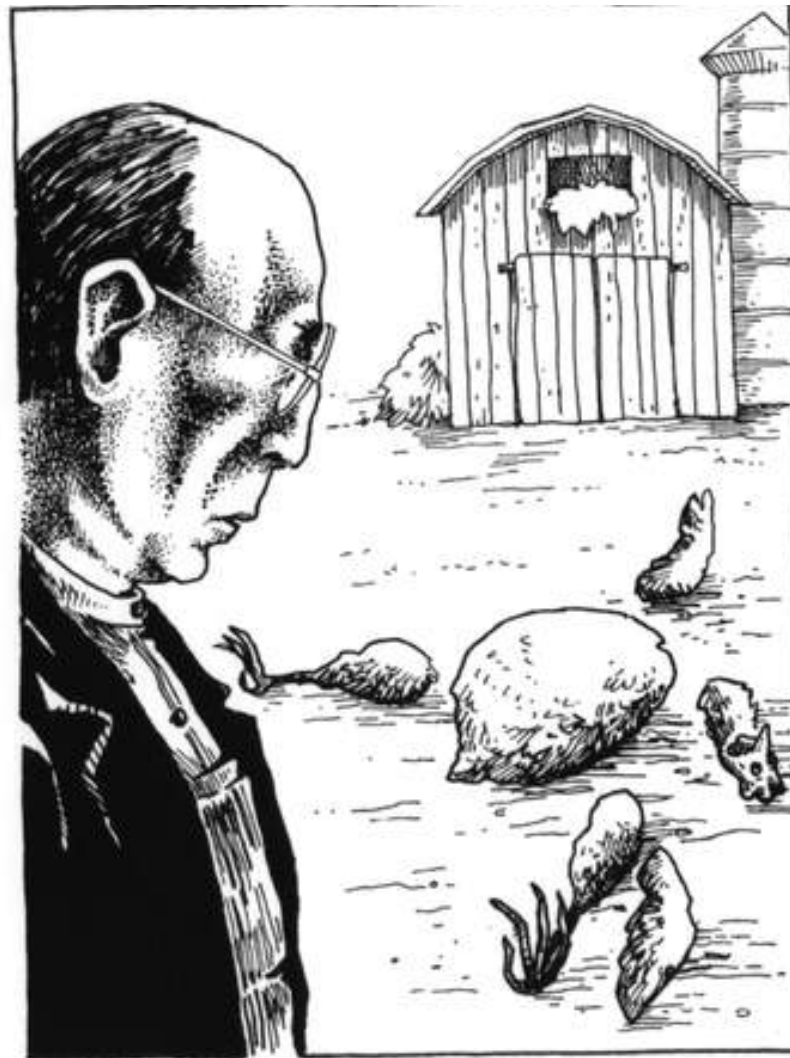
Object segmentation

Instance segmentation

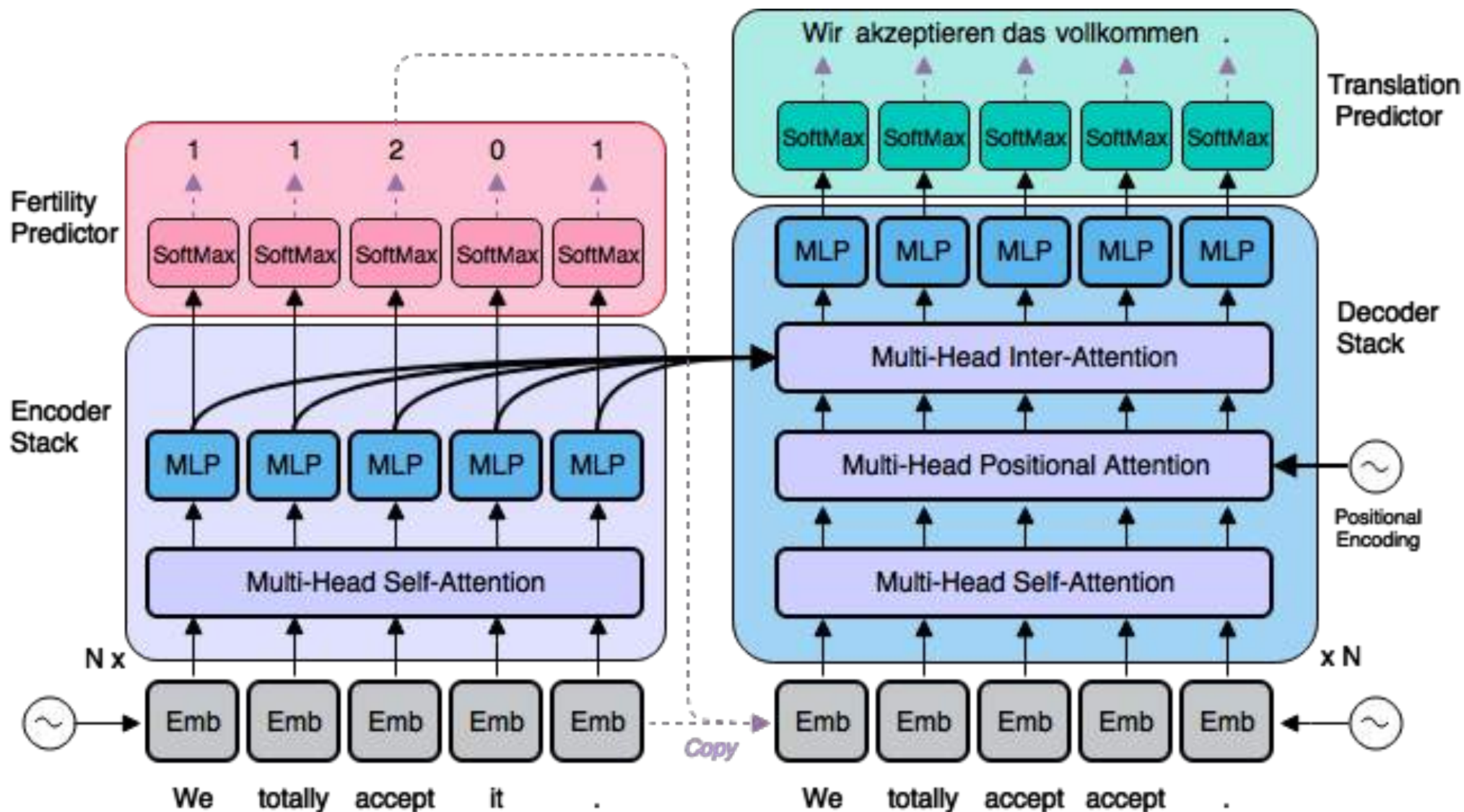
Neural machine translation

Generative Adversarial Networks

Segmentation & Detection



Neural Machine Translation





GANs

Лекции

Что мы обсудим на лекциях

1. Введение: стандартные задачи и методы, настройка алгоритмов
2. Supervised learning: линейные модели, решающие деревья и ансамбли
3. Оценка качества в оффлайне и онлайн
4. Unsupervised learning
5. Рекомендательные системы
6. Предиктивная аналитика
7. Анализ текстов
8. Нейронные сети

На этой лекции

- I. Стандартные задачи и методы машинного обучения
- II. Настройка параметров алгоритмов
- III. Пример проекта по ML
- IV. Инструменты

I. Стандартные задачи и методы машинного обучения

Классификация



Iris setosa



Iris versicolor



Iris virginica

Вход (обучающая выборка):

Признаки N объектов с известными классами

Выход:

Классификатор (алгоритм, прогнозирующий классы новых объектов по их признакам)

Классификация: обучающая выборка

Fisher's Iris Data

Sepal length ↕	Sepal width ▲	Petal length ↕	Petal width ↕	Species ↕
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
5.0	2.3	3.3	1.0	<i>I. versicolor</i>
4.5	2.3	1.3	0.3	<i>I. setosa</i>
5.5	2.4	3.8	1.1	<i>I. versicolor</i>
5.5	2.4	3.7	1.0	<i>I. versicolor</i>
4.9	2.4	3.3	1.0	<i>I. versicolor</i>
6.7	2.5	5.8	1.8	<i>I. virginica</i>
5.7	2.5	5.0	2.0	<i>I. virginica</i>
6.3	2.5	5.0	1.9	<i>I. virginica</i>
6.3	2.5	4.9	1.5	<i>I. versicolor</i>
4.9	2.5	4.5	1.7	<i>I. virginica</i>

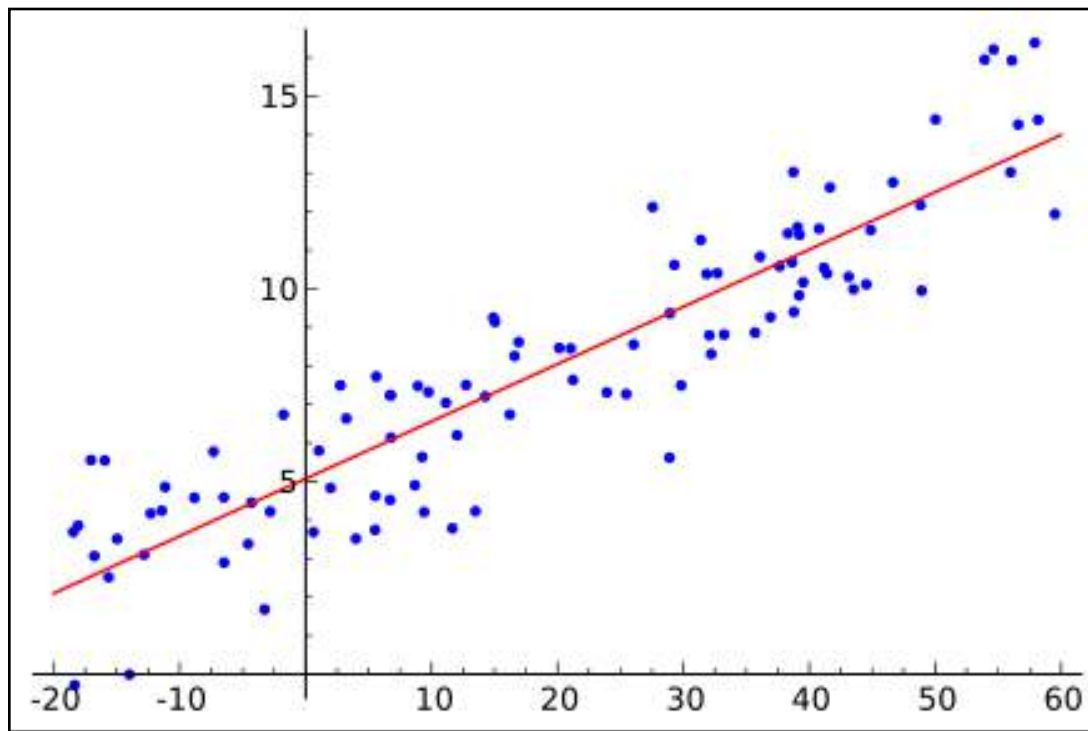
Регрессия

Вход (обучающая выборка):

Признаки N объектов с известными значениями прогнозируемого вещественного параметра объекта

Выход:

Алгоритм, прогнозирующий значение вещественной величины по признакам объекта



Кластеризация

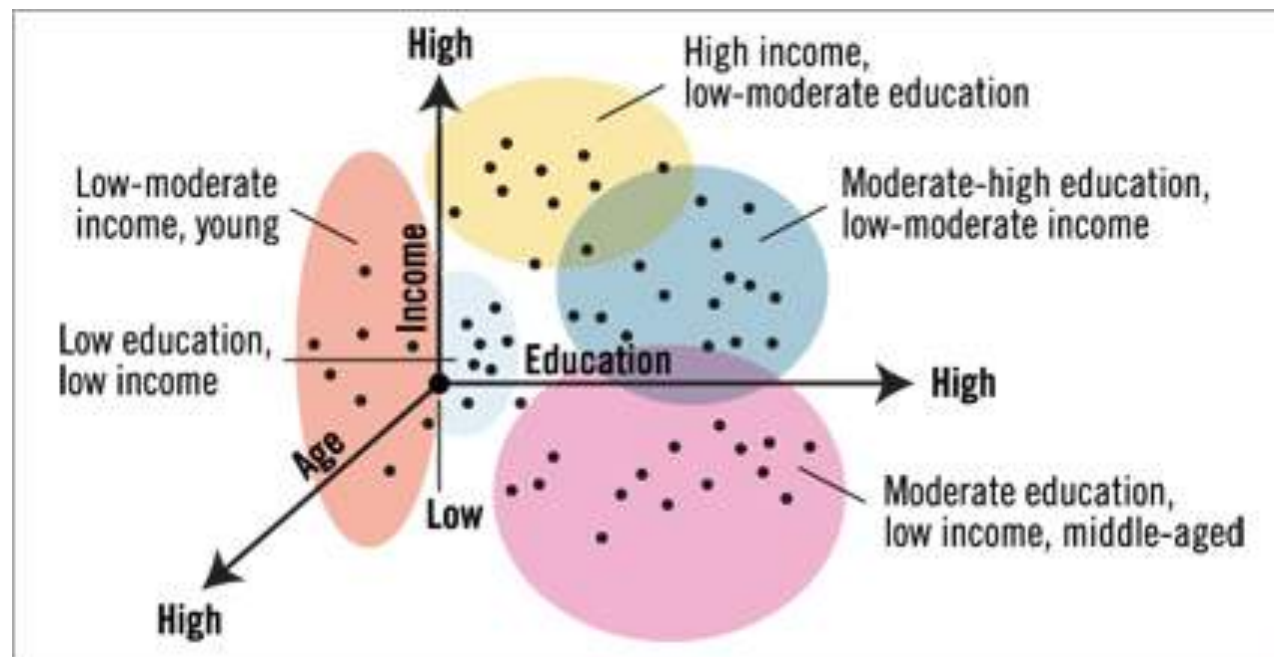
Вход (обучающая выборка):

Признаки N объектов

Выход:

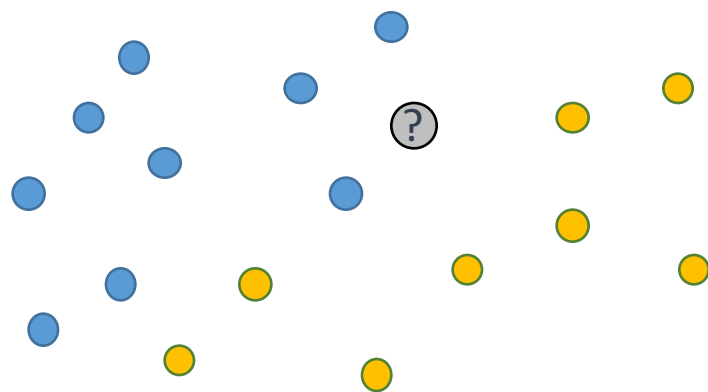
Найденные в выборке классы (кластеры), метки кластеров для объектов из обучающей выборки и алгоритм отнесения новых объектов к кластеру

Пример: сегментация рынка



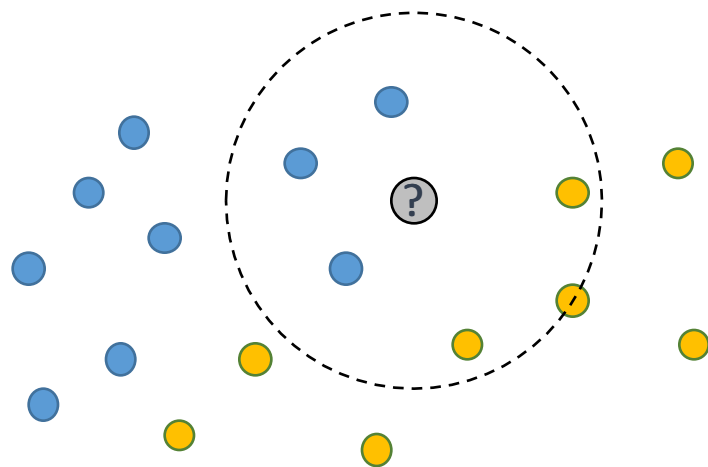
Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):



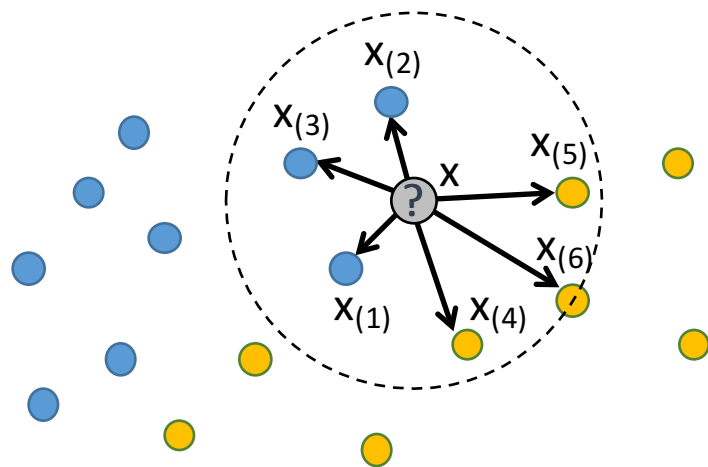
Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):



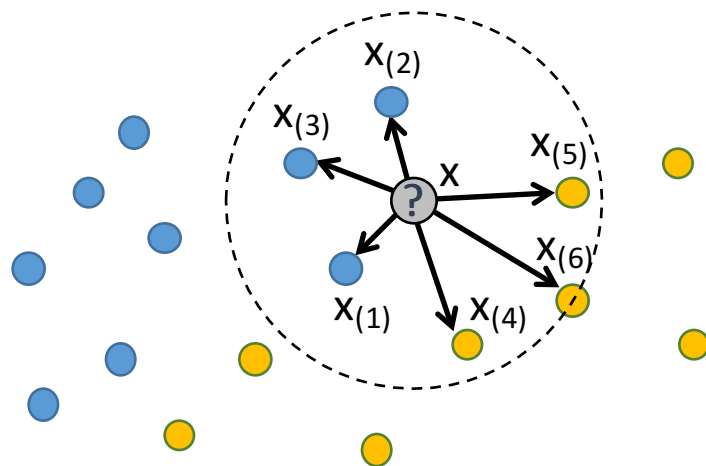
Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):



Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):

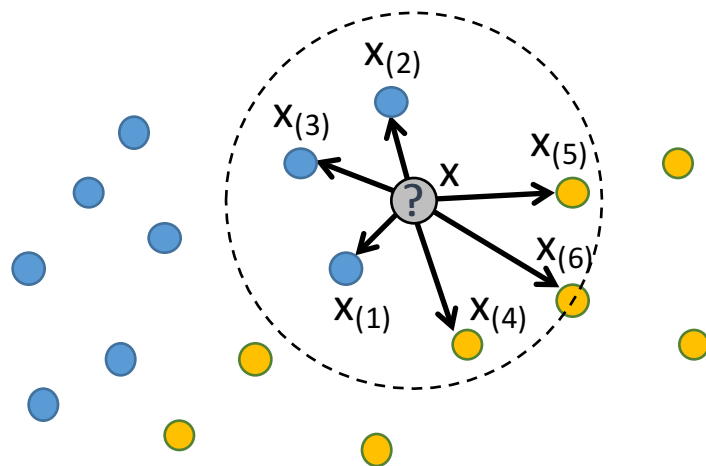


Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

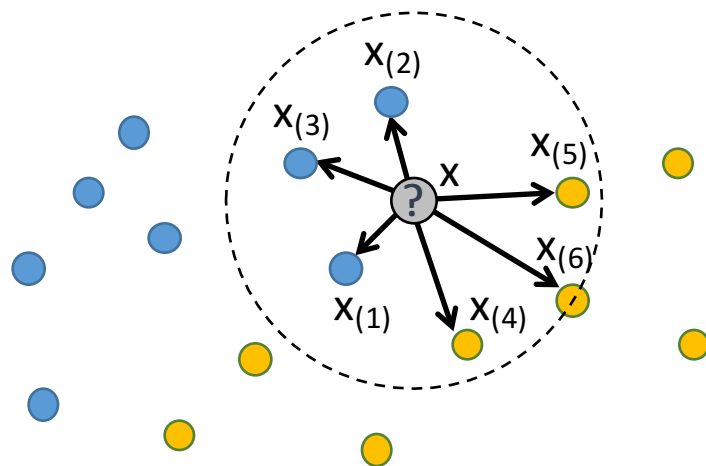
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

Метод k ближайших соседей (kNN)

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

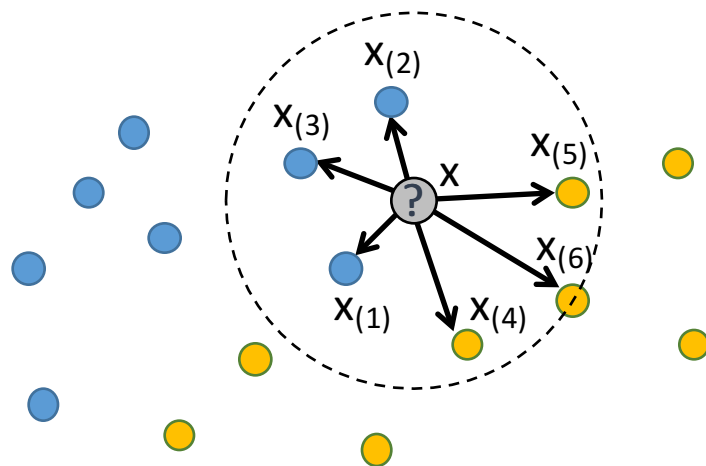
или как функцию расстояния:

$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

$$Z_{\bullet} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Метод k ближайших соседей (kNN)

Пример классификации ($k = 6$):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

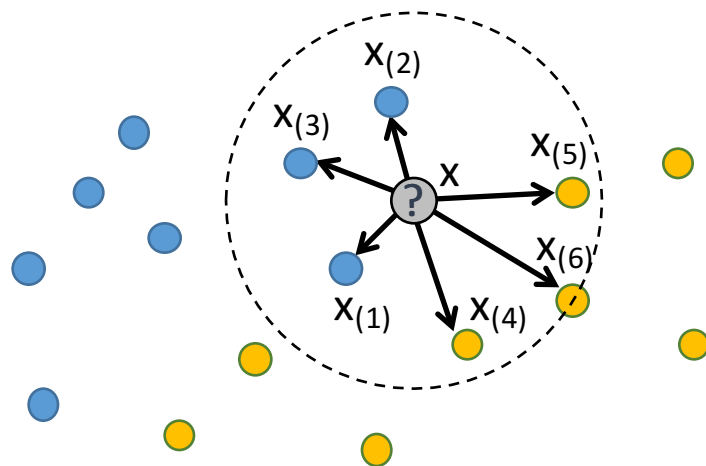
$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\text{yellow}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Метод k ближайших соседей (kNN)

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

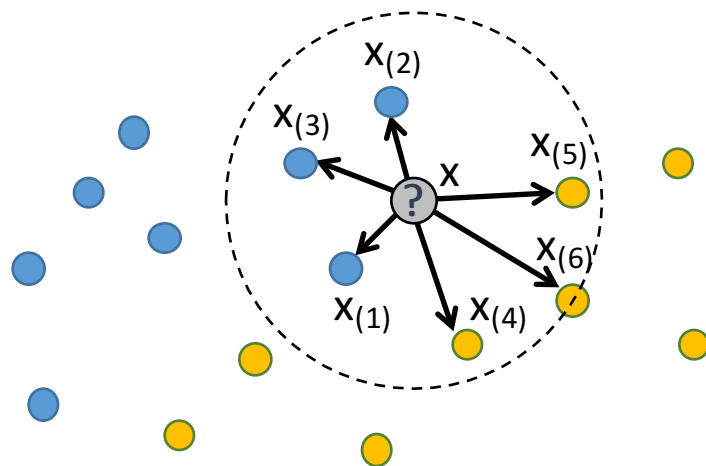
$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$Z_{\text{yellow}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$\text{?} = \underset{\bullet}{\operatorname{argmax}} Z_{\bullet}$$

Метод k ближайших соседей (kNN)

Пример классификации (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

$$Z_{\text{blue}} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

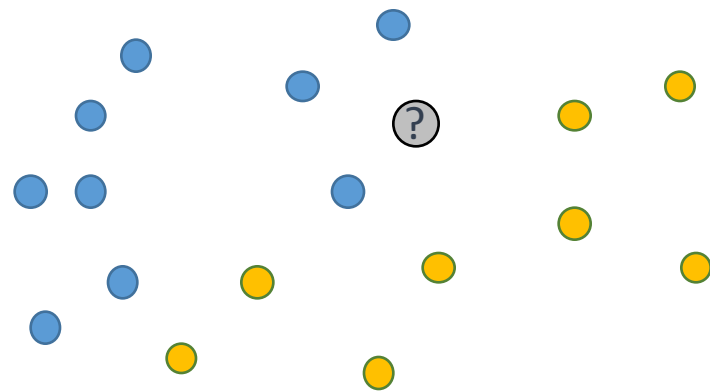
$$Z_{\text{yellow}} = \frac{w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

$$\text{?} = \underset{\text{?}}{\operatorname{argmax}} Z_{\text{?}}$$

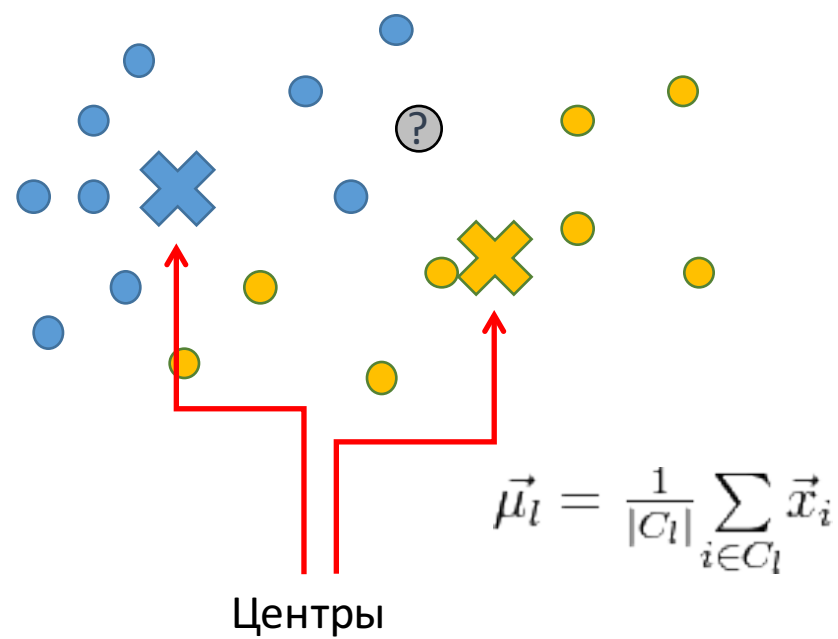
$$\text{if } Z_{\text{yellow}} > Z_{\text{blue}} : \quad \text{?} = \text{yellow}$$

$$\text{if } Z_{\text{yellow}} < Z_{\text{blue}} : \quad \text{?} = \text{blue}$$

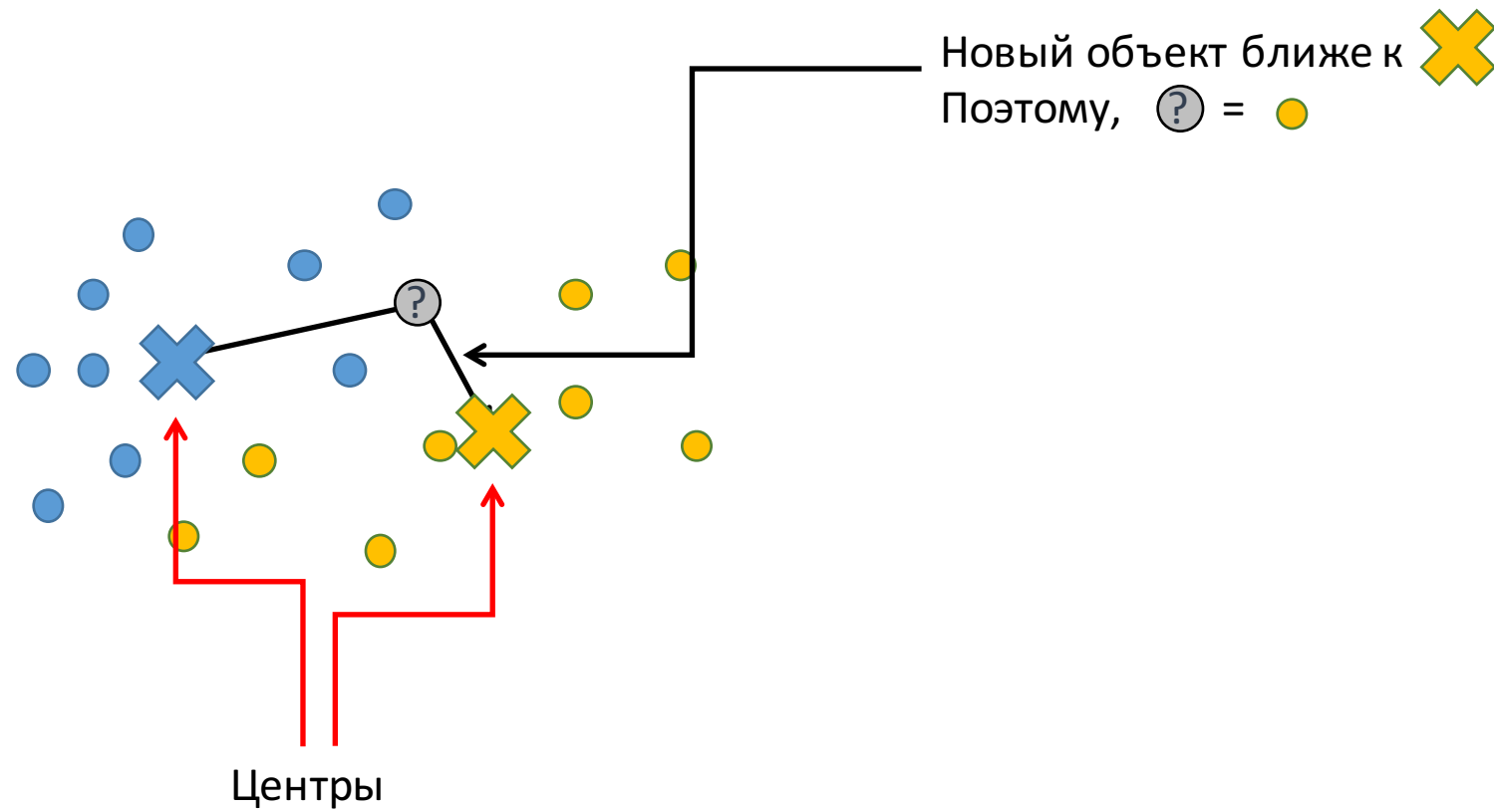
Центроидный классификатор



Центроидный классификатор

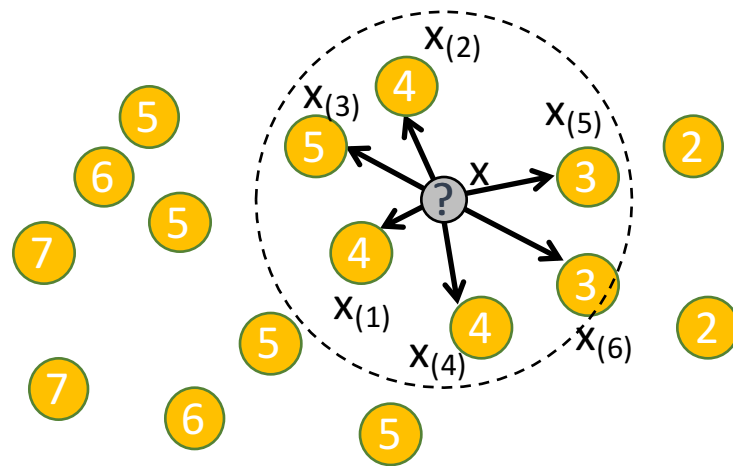


Центроидный классификатор



Взвешенный kNN для регрессии

Пример (k = 6):



Веса можно определить как функцию от соседа или его номера:

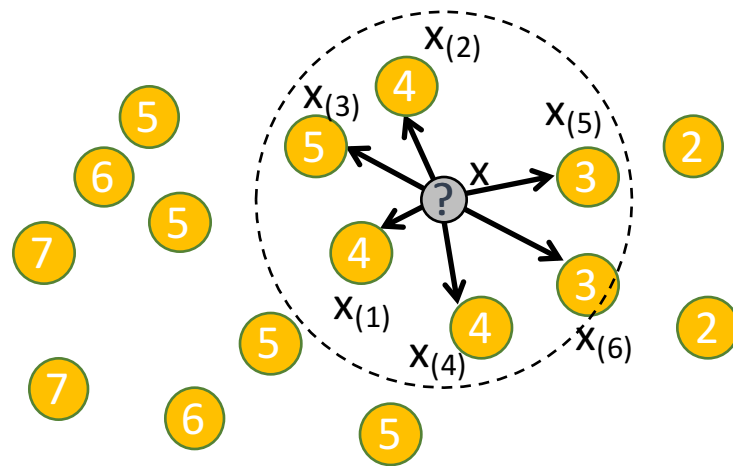
$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

kNN для задачи регрессии

Пример (k = 6):



Веса можно определить как функцию от соседа или его номера:

$$w(x_{(i)}) = w(i)$$

или как функцию расстояния:

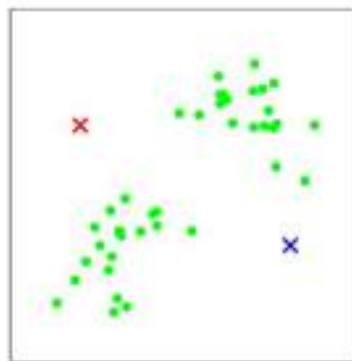
$$w(x_{(i)}) = w(d(x, x_{(i)}))$$

$$\textcircled{?} = \frac{4 \cdot w(x_{(1)}) + 4 \cdot w(x_{(2)}) + 5 \cdot w(x_{(3)}) + 4 \cdot w(x_{(4)}) + 3 \cdot w(x_{(5)}) + 3 \cdot w(x_{(6)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

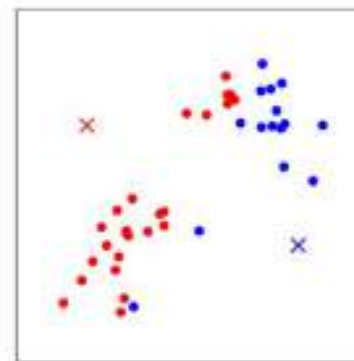
Простой алгоритм кластеризации: kMeans



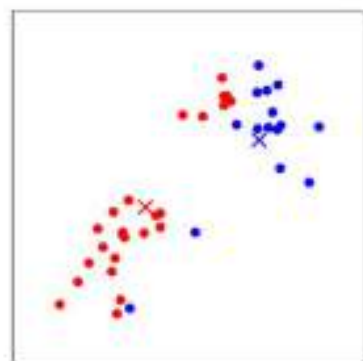
(a)



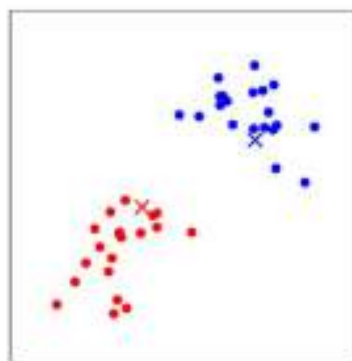
(b)



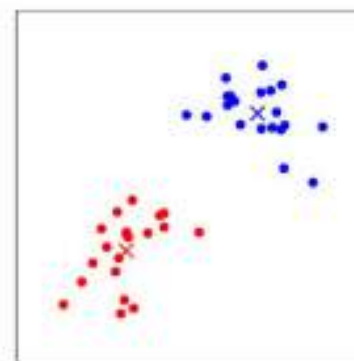
(c)



(d)



(e)



(f)

Часто используемые алгоритмы

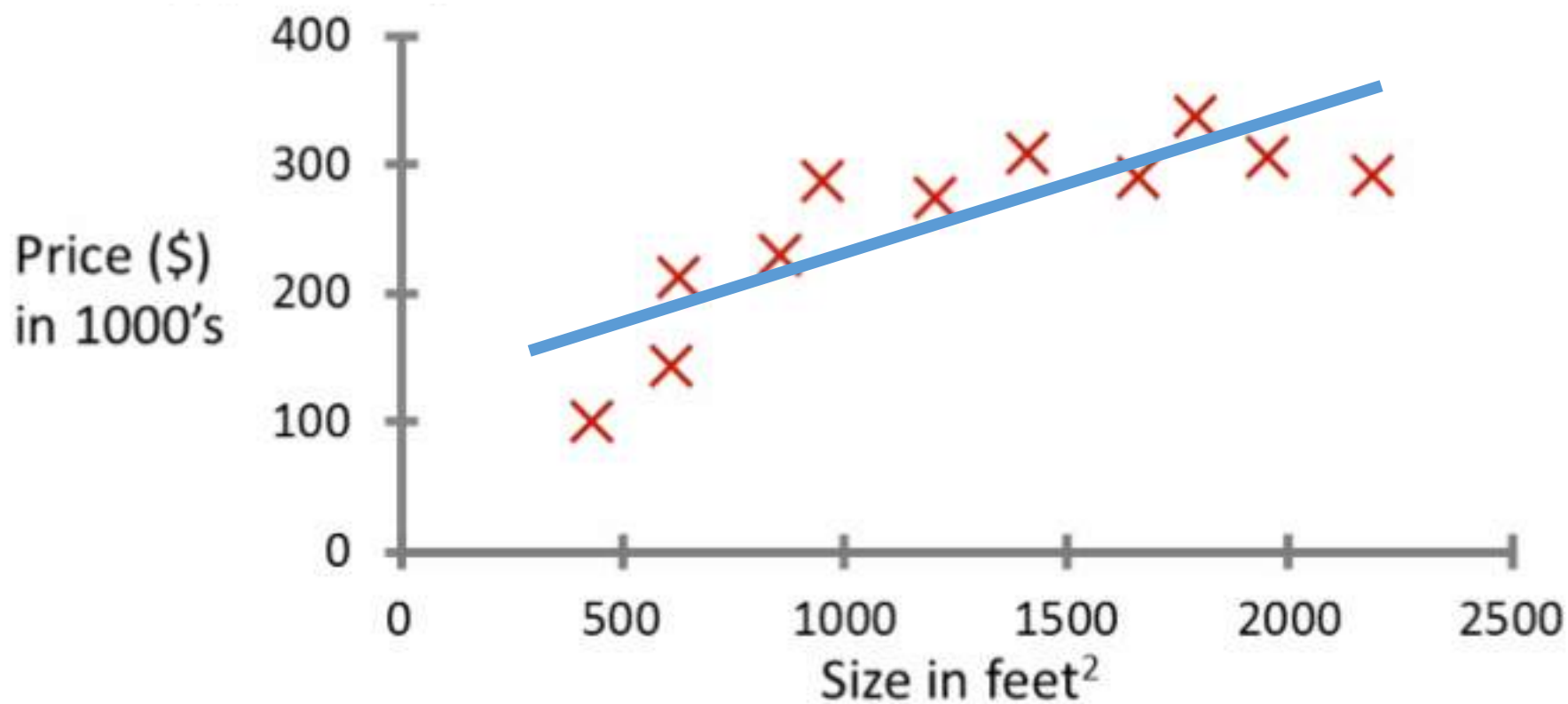
- Линейные модели
- Решающие деревья
- Ансамбли решающих деревьев
- Нейронные сети*

Часто используемые алгоритмы

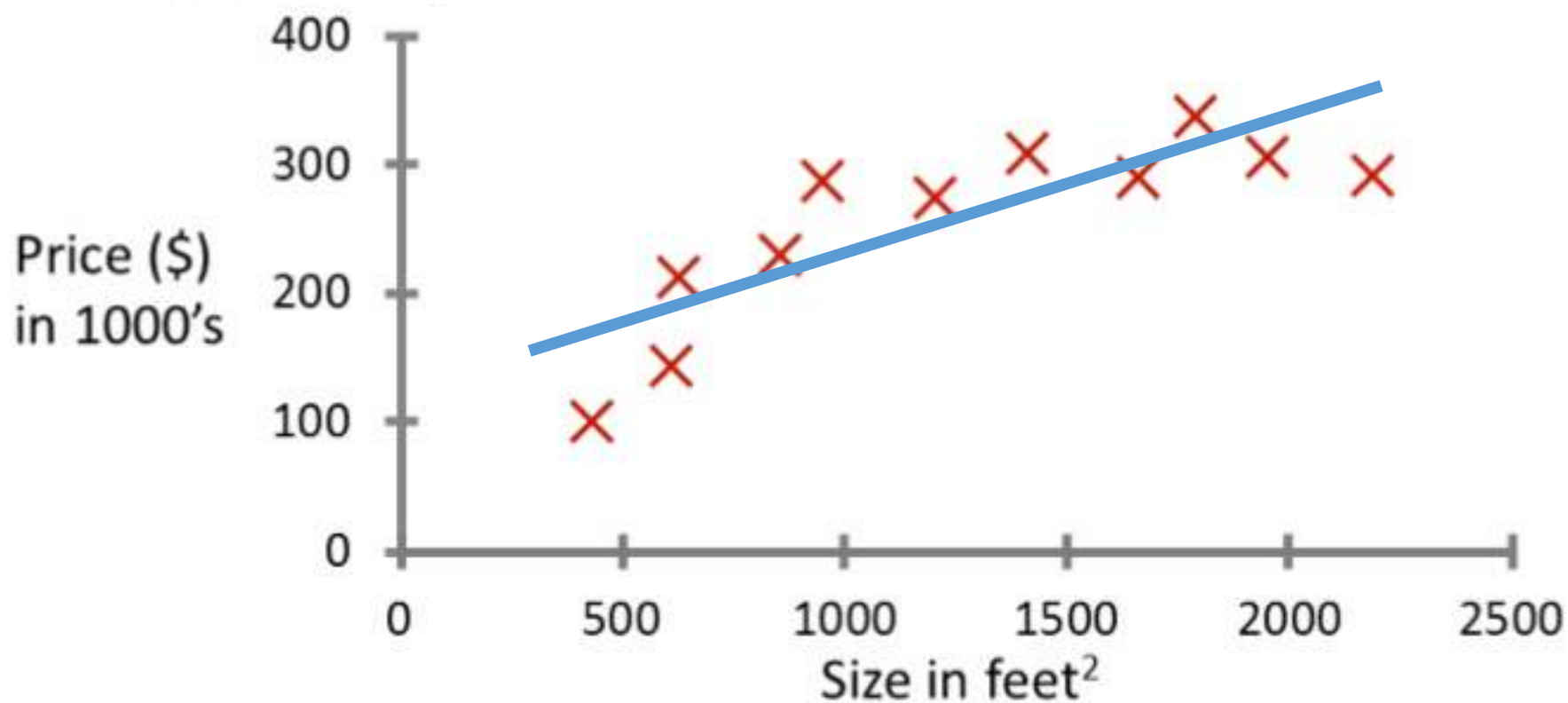
- Линейные модели
- Решающие деревья
- Ансамбли решающих деревьев
- Нейронные сети*

* В задачах анализа изображений, звука, текста и прогнозировании временных рядов

Линейные модели

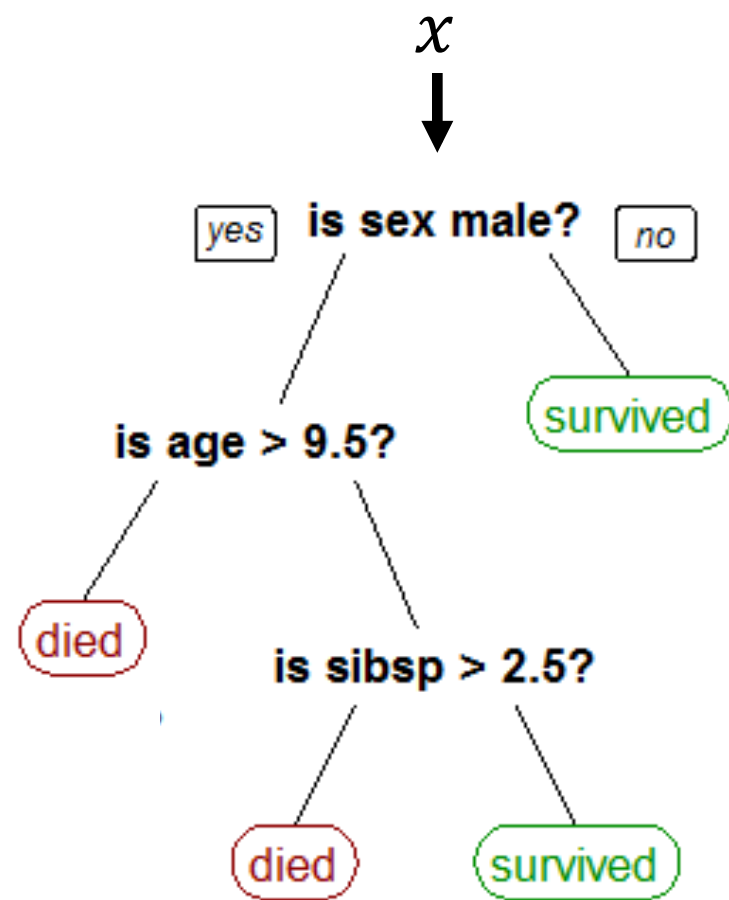


Линейные модели

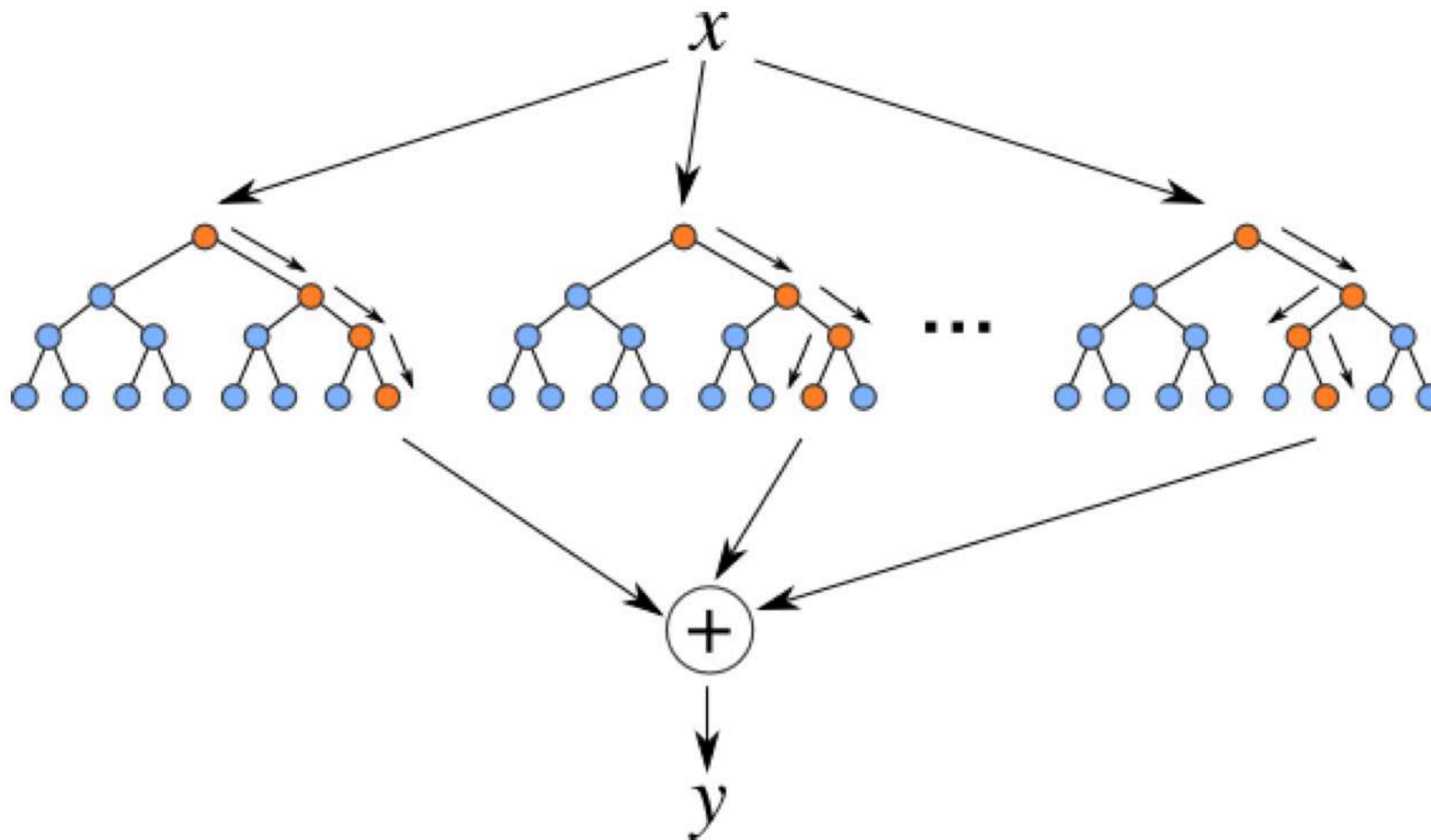


Особенно хороши на разреженных признаках

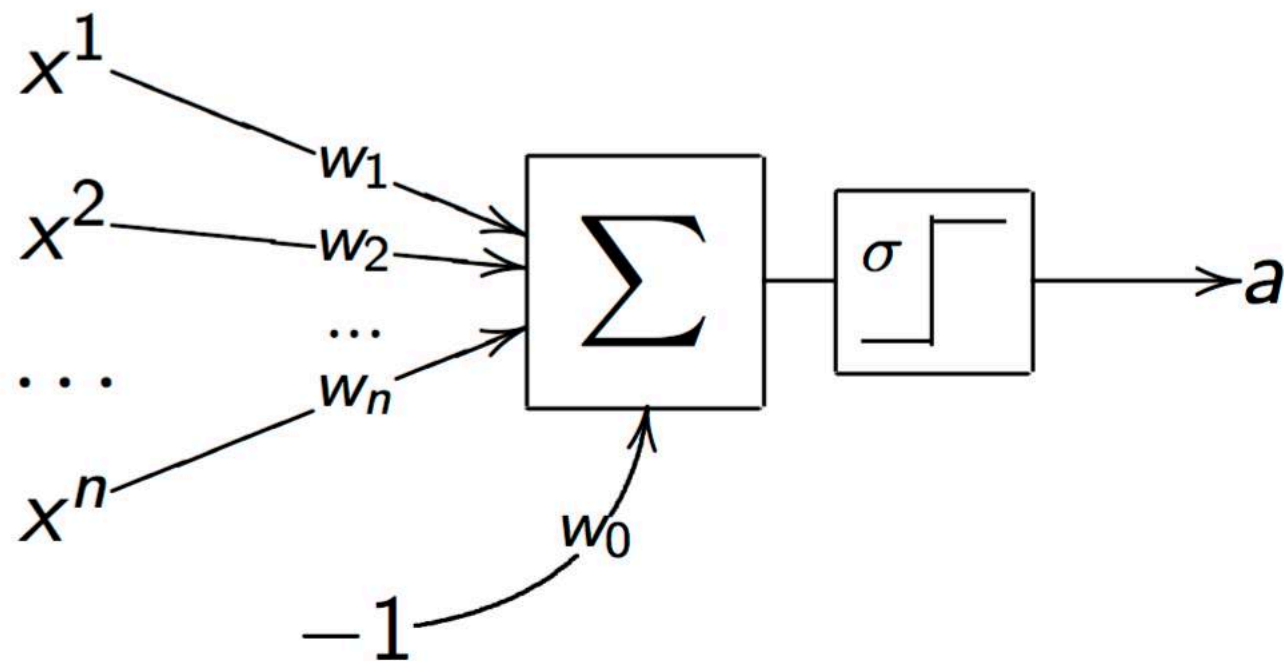
Решающие деревья



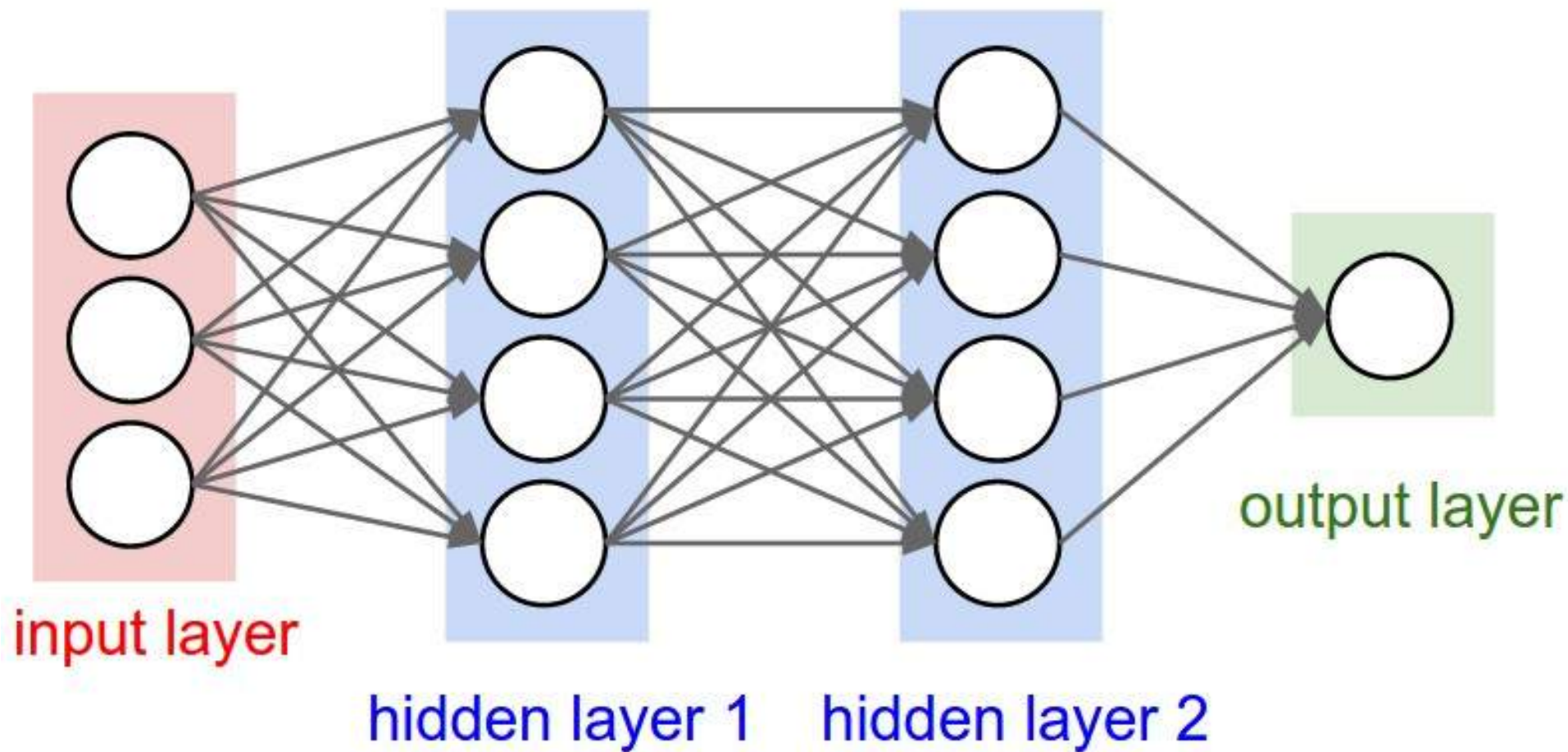
Ансамбли решающих деревьев



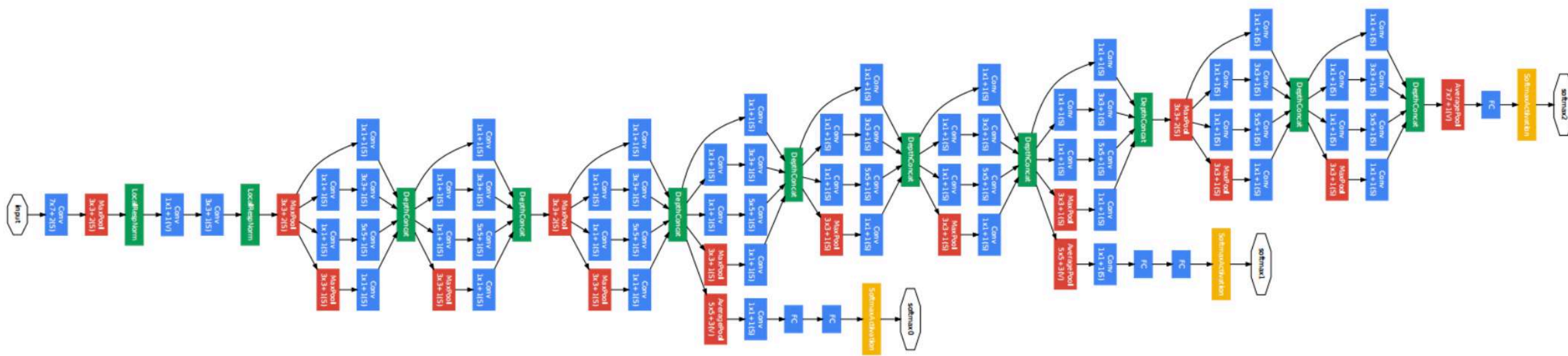
Нейронные сети



Нейронные сети



Нейронные сети



GoogLeNet

Поможем Даше сопоставить алгоритмы и задачи

kNN

KMeans

Решающее дерево

Классификация

Регрессия

Кластеризация



Поможем Даше сопоставить алгоритмы и задачи

kNN

KMeans

Решающее дерево

Классификация

Регрессия

Кластеризация



Поможем Даше сопоставить алгоритмы и задачи

kNN

KMeans

Решающее дерево

Классификация

Регрессия

Кластеризация



Поможем Даше сопоставить алгоритмы и задачи

kNN

Классификация

KMeans

Регрессия

Решающее дерево

Кластеризация



Поможем Даше сопоставить алгоритмы и задачи

kNN

KMeans

Решающее дерево

Классификация

Регрессия

Кластеризация



II. Настройка параметров алгоритмов

Оптимизационные задачи в ML

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Задача регрессии

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Например, это:

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

Оптимизационные задачи в ML

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

Например, это:

$$\sum_{i=1}^l |y_i - a(x_i)| \rightarrow \min$$

Оптимизационные задачи в ML

x_1, x_2, \dots, x_l - точки, в которых известны значения некоторой величины:

$$y_1, y_2, \dots, y_l$$

Мы строим прогнозирующий алгоритм:

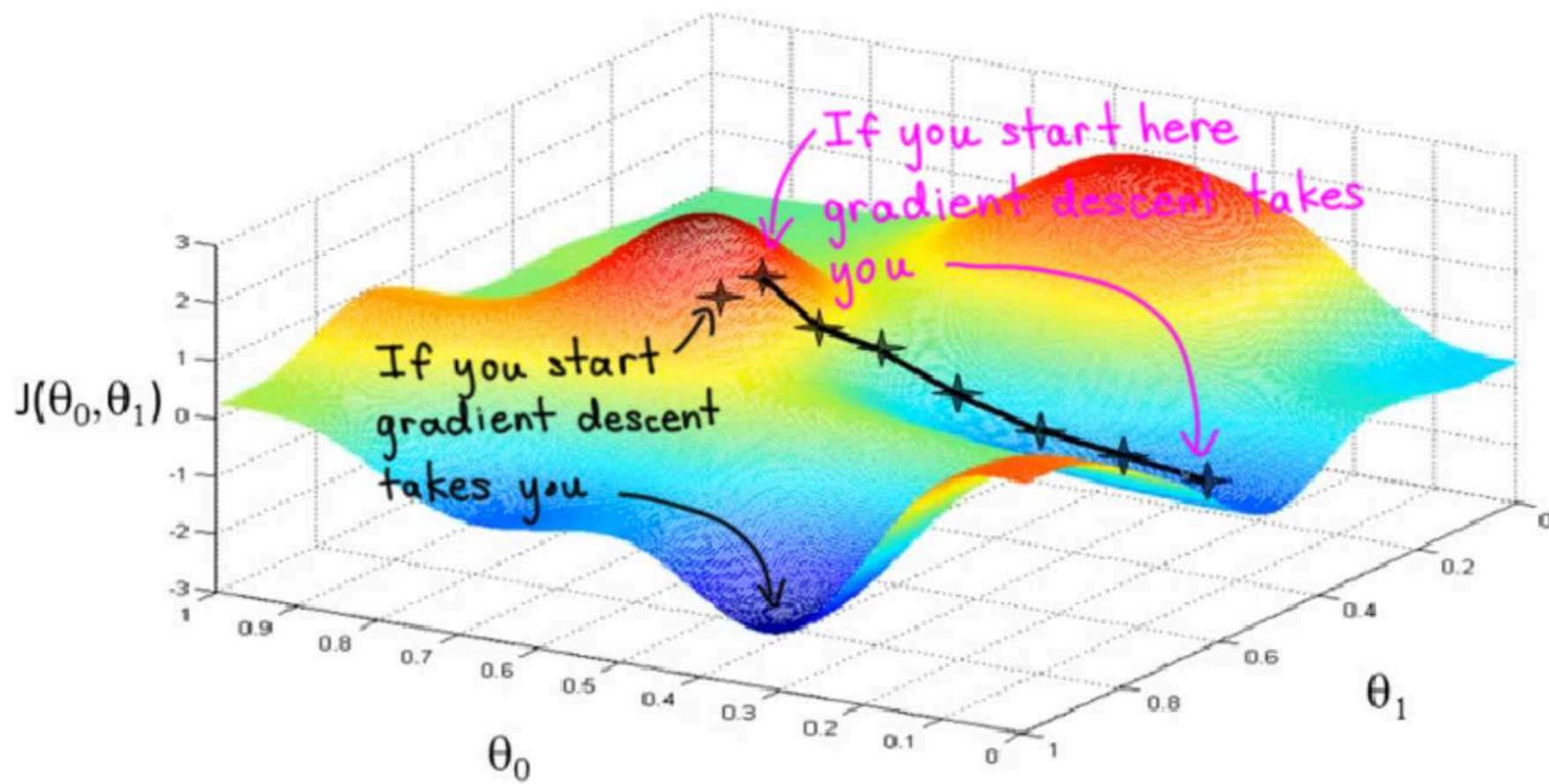
$$y_i \approx a(x_i)$$

Но что значит «примерно равно»?

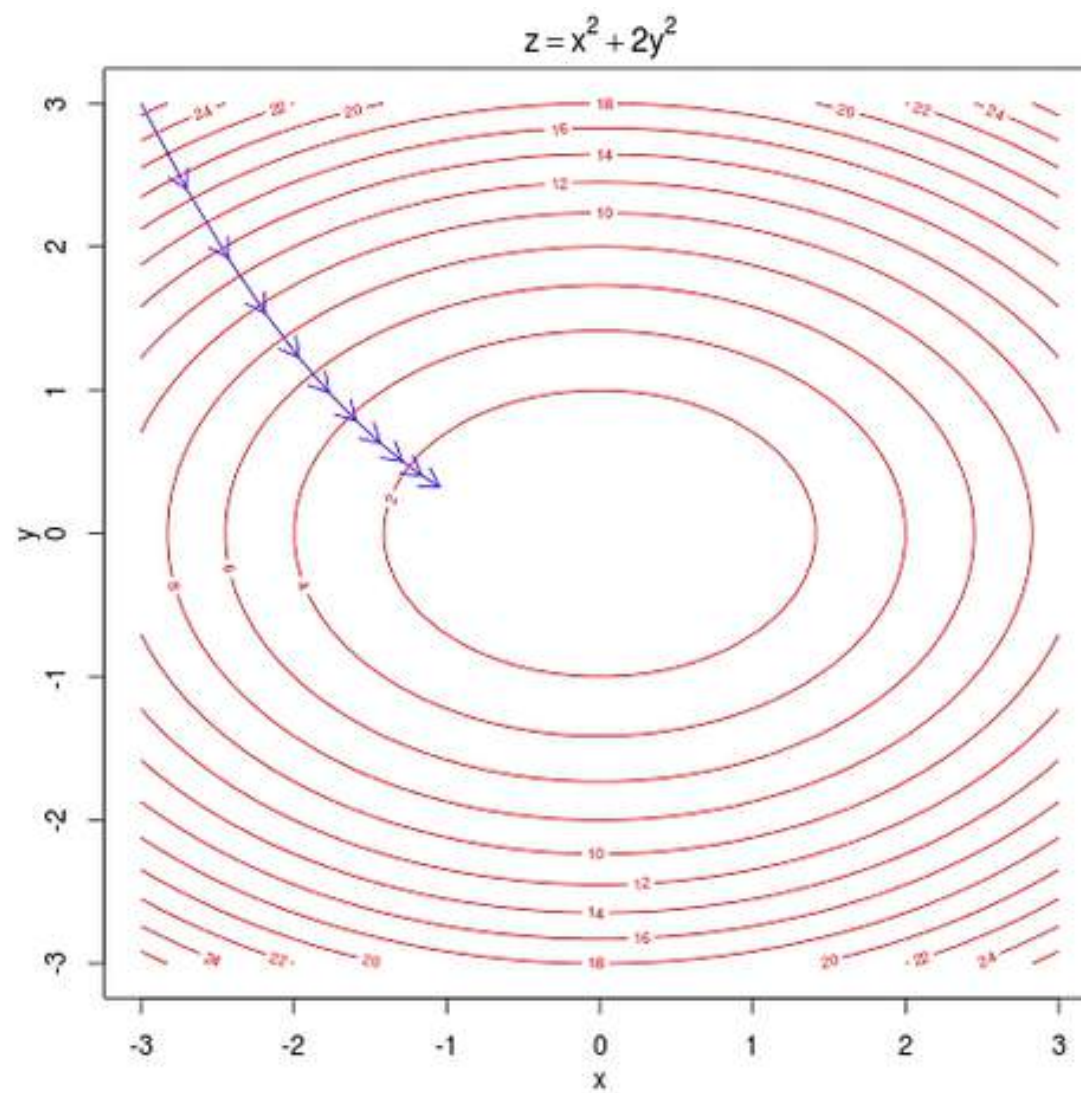
В общем случае:

$$\sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min$$

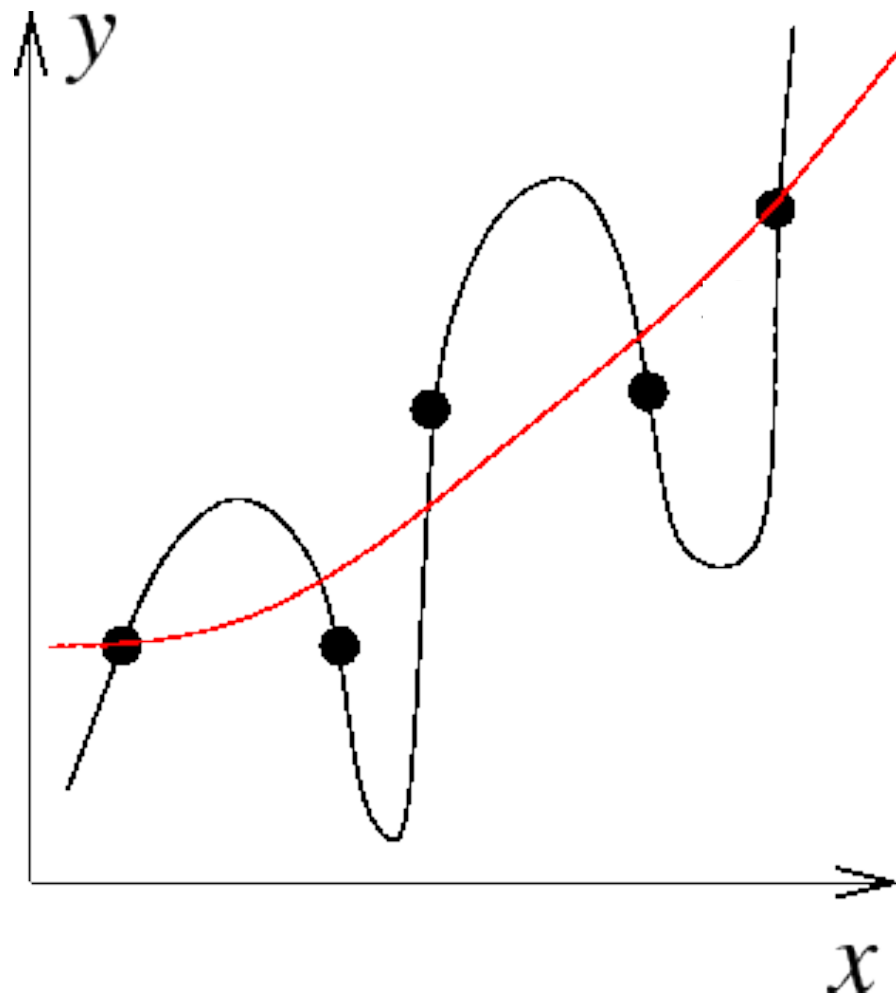
Градиентный спуск



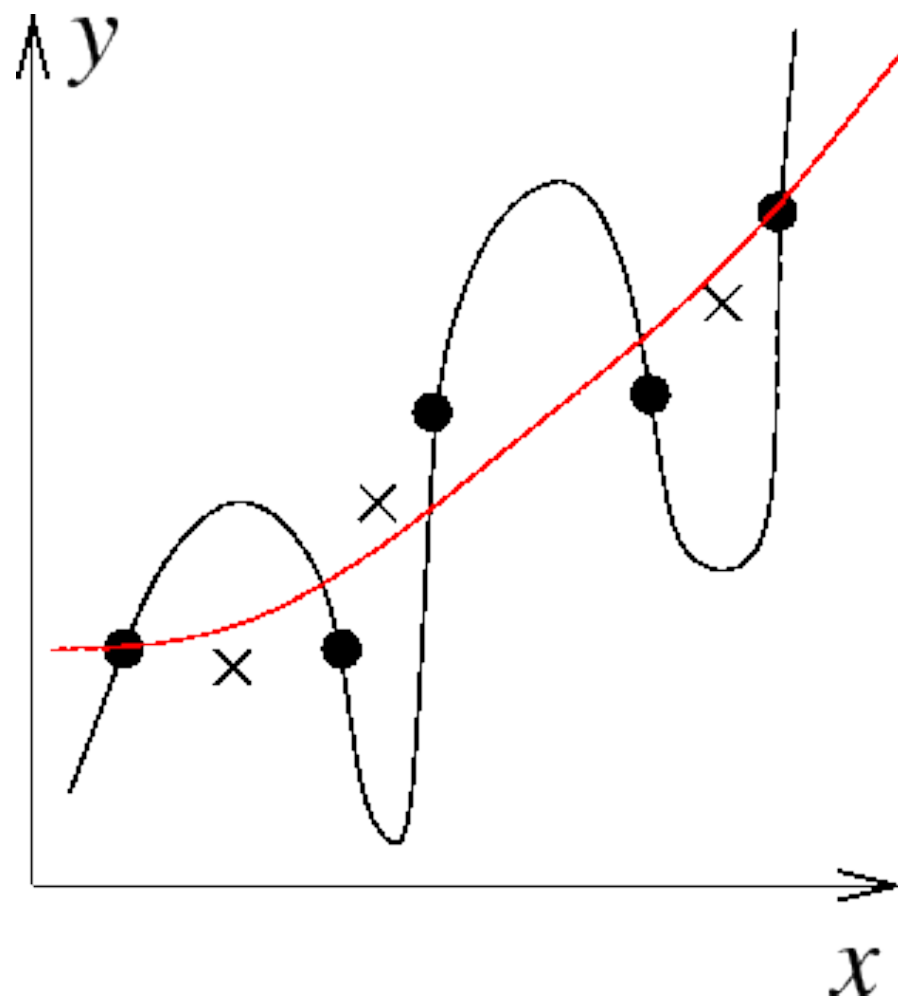
Градиентный спуск



Переобучение (overfitting)



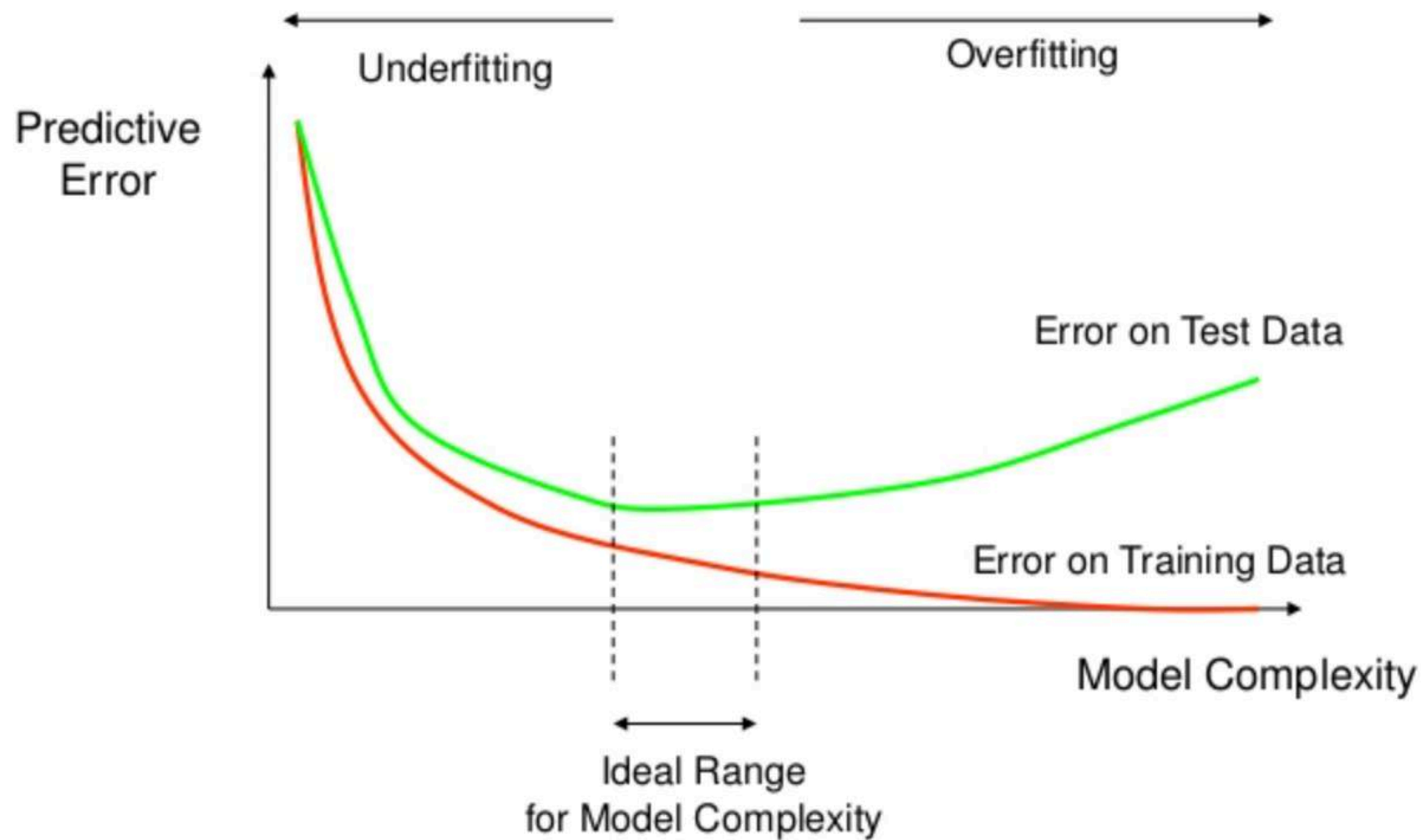
Переобучение (overfitting)



Оценка качества

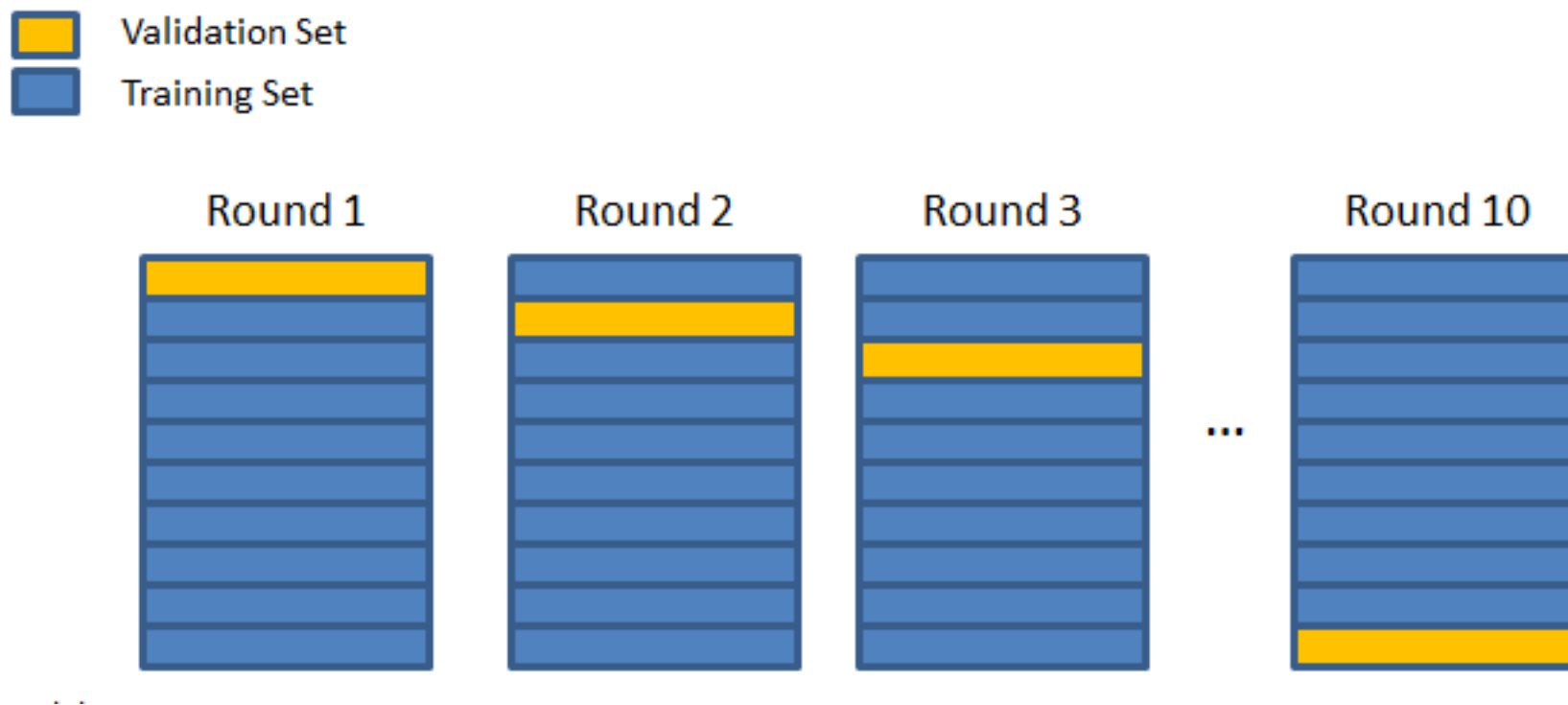


Кривые обучения



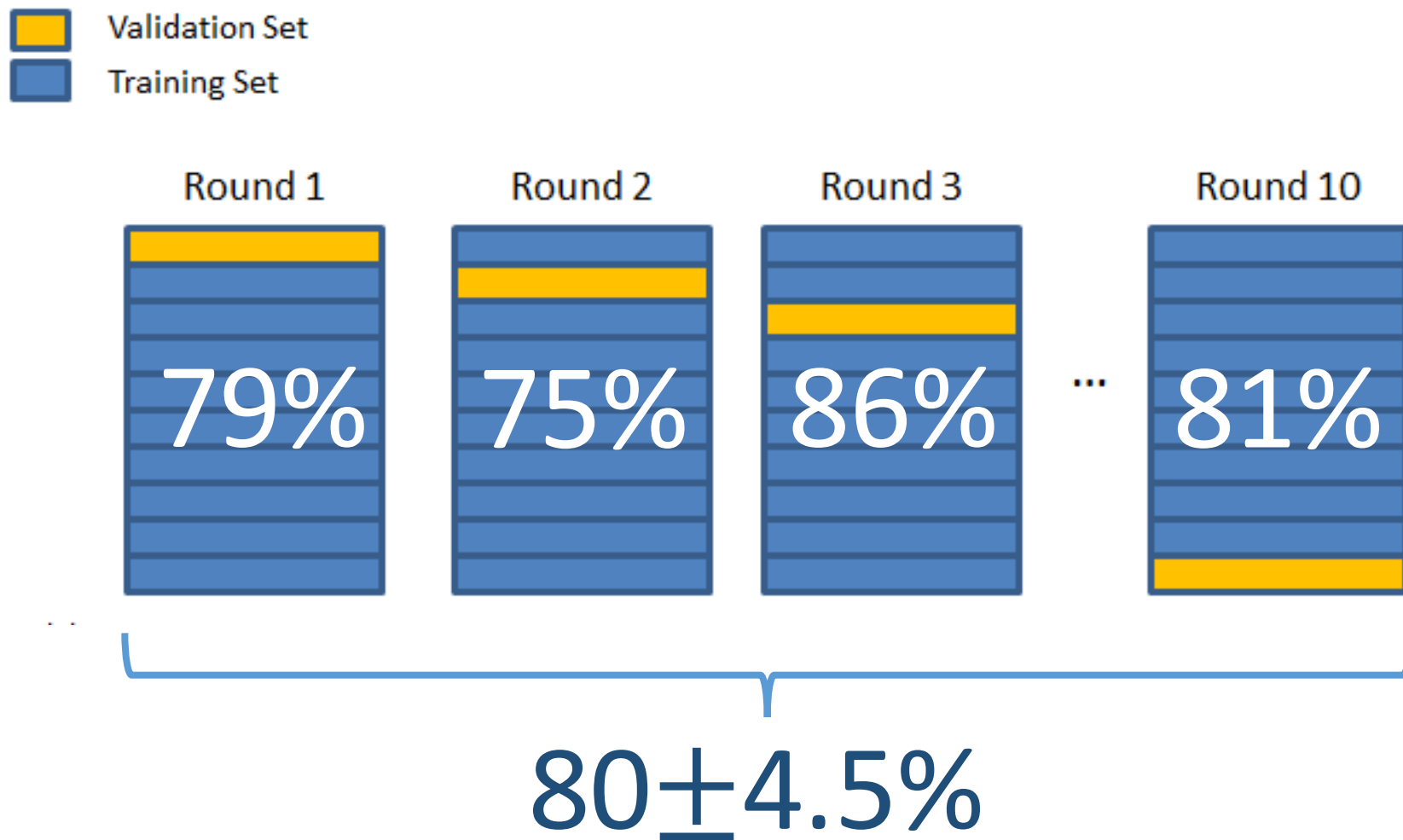
Кросс-валидация

K-Fold cross validation:

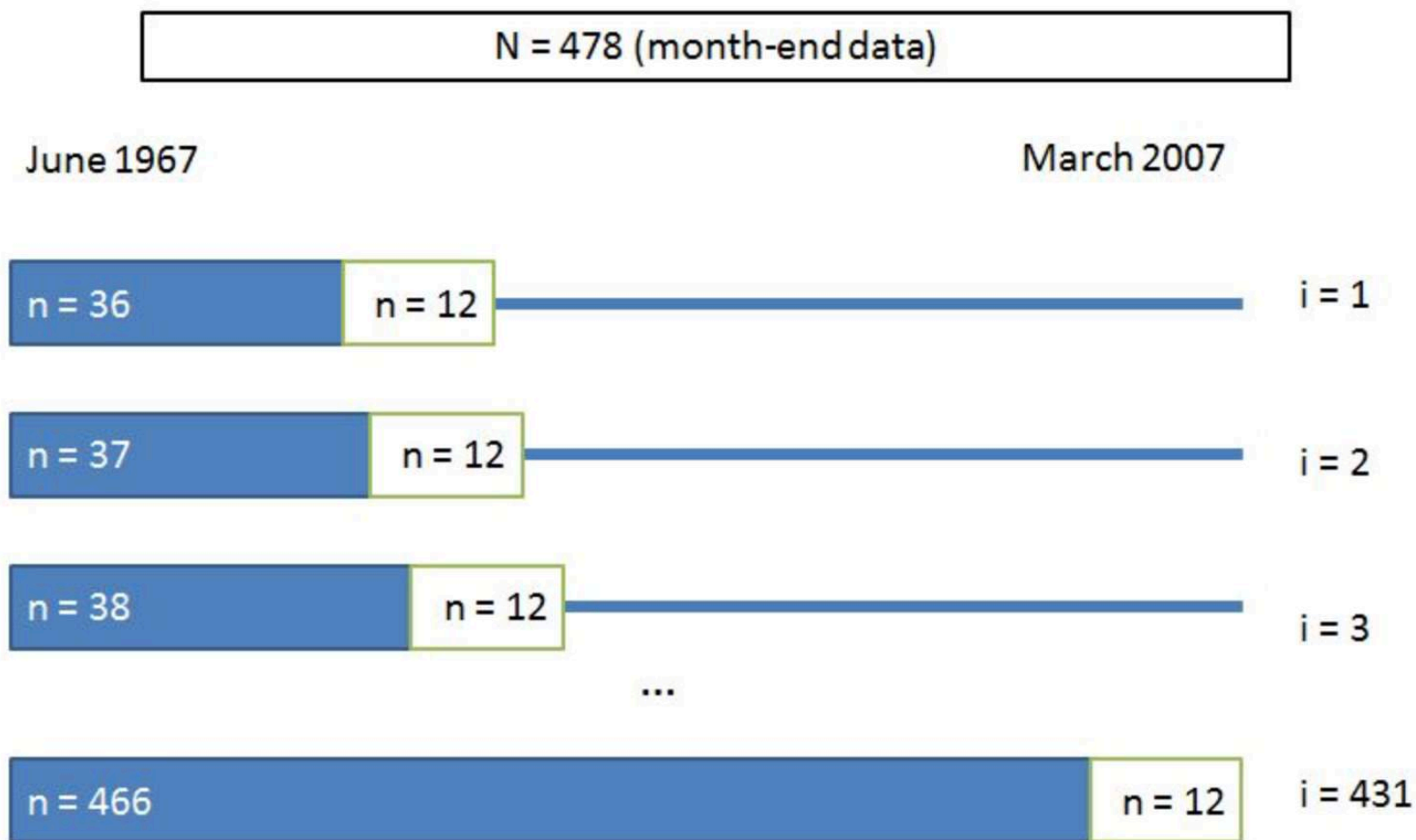


На картинке $k = 10$. Другие частые варианты – 3 и 5.

Учет разброса в CV



Предупреждение: будьте осторожны с CV



История про танки



Классификатор: есть танки на снимке или нет

История про танки



Классификатор: есть танки на снимке или нет

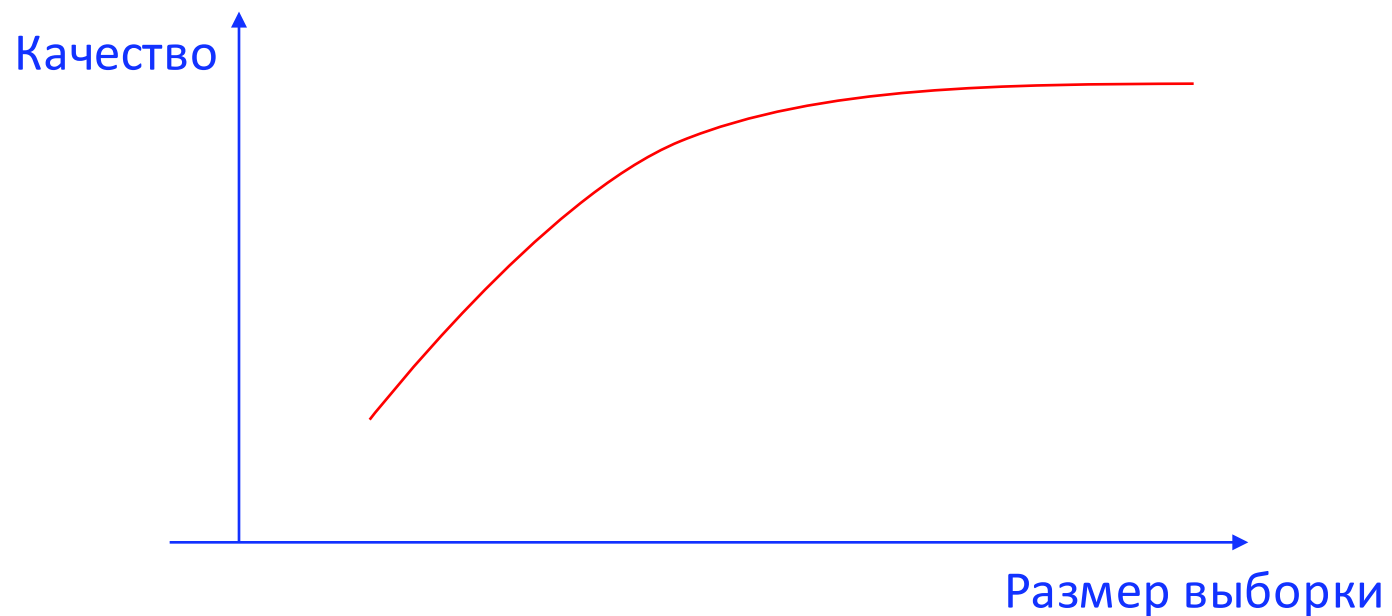
Задача

Для некоторой задачи построили алгоритм обучения с учителем и он работает очень плохо.

- a) Как понять, проблема в недостаточном размере обучающей выборки или в чем-то еще?
- b) В чем еще может быть проблема?

Ответ

Увеличить обучающую выборку – дорого, зато легко можем уменьшить и посмотреть, вышло ли качество на «плато»



III. Пример проекта

Блок рекомендаций

Товар 1	Товар 2	Товар 3	Товар 4
----------------	----------------	----------------	----------------

Возможный вариант заполнения



Puma
Ветровка
3 490 руб.



Crocs
Сланцы
1 990 руб.



Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.



Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Использование прогноза: пример про UX

Сопутствующие товары		Похожие товары	
Товар 1	Товар 2	Товар 3	Товар 4

История про одинаковое качество

- Интегрировали чужое решение, чтобы сравнить качество со своим
- Оценили качество у обоих
- Совпало до тысячных долей
- Не стали использовать чужое решение
- Позже – выяснили, в чем дело

Поставим диагноз

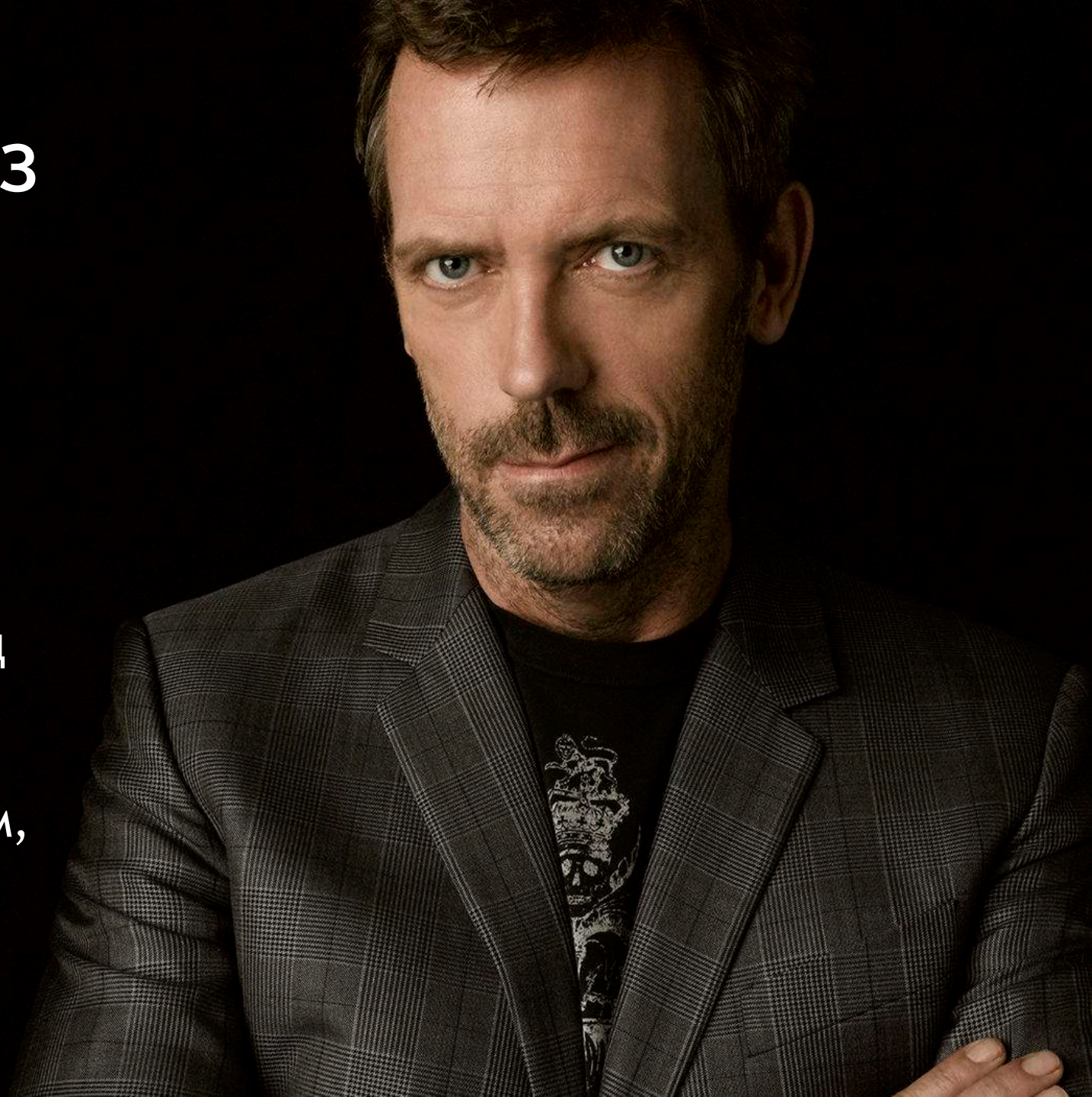
Пациент:

Система, рекомендующая
пользователям товары

Что знаем:

Качество решений двух команд
совпало

Какие еще «анализы» назначим,
чтобы понять, в чем дело?



История про статзначимость



История про статзначимость



Одна кривая отличается от других на 10%
Но разбиение на самом деле – случайное

Рекомендации товаров: вопросы

1. Какой экономический эффект может дать модель в этой задаче? Как он связан с качеством модели? (и как его измерять)
2. Будет ли оценка ожидаемого экономического эффекта на исторических данных совпадать с реальным экономическим эффектом? Как можно измерить его?
3. Какие данные нужны для построения модели?

История о постановке задач

На входе:

- Туроператор хочет персонализировать рассылки своим клиентам
- Данных мало
- Заказчику интересно увеличить конверсию

История о постановке задач

На входе:

- Туроператор хочет персонализировать рассылки своим клиентам
- Данных мало
- Заказчику интересно увеличить конверсию

На выходе:

- Не стали надеяться на какое-то умное обучение
- Попытались кластеризовать и выделить топ рекомендаций по кластерам
- Отдали заказчику в надежде, что в любом случае лучше, чем всем рассылать одно и то же
- Про A/B тесты вообще не слышали

IV. Инструменты

Python

На чем будут примеры

- Почему Python? Потому что можно всего в 5 - 30 строк очень простого кода продемонстрировать интересные явления
- Библиотеки: numpy, scipy, sklearn, matplotlib, pandas и др.
- Что использовать на практике – ваш выбор
- Под Windows проще всего установить Anaconda Python



Scikit-learn

Scikit-learn



[Home](#) [Installation](#) [Documentation](#) [Examples](#)

Google™ Custom Search



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Машинное обучение в несколько строк

```
from sklearn.linear_model import LogisticRegression
```

```
model = LogisticRegression()
```

```
model.fit(X_train, y_train)
```

```
predictions = model.predict(X_test)
```

Резюме

- I. Стандартные задачи и методы машинного обучения
- II. Настройка параметров алгоритмов
- III. Пример проекта по ML
- IV. Инструменты

Конкурс

Найти ошибку на этой картинке и написать комментарий в соответствующей записи в группе DMIA в ВК

Победитель (первый, кто ответит верно) получит приз

Machine Learning



what society thinks I do



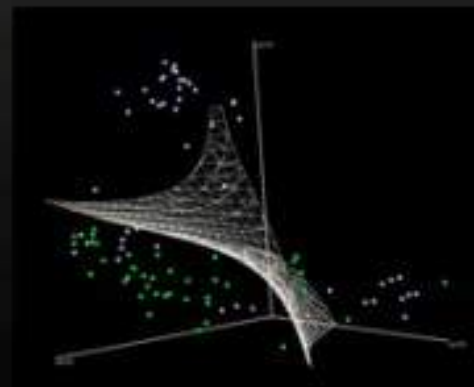
what my friends think I do



what my parents think I do

$$\begin{aligned} L_p &= ||\mathbf{w}'||^2 - \sum_i \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_i \alpha_i \\ \alpha_i &\geq 0, \forall i \\ \mathbf{w} &= \sum_i \alpha_i y_i \mathbf{x}_i, \sum_i \alpha_i y_i = 0 \\ \nabla \hat{g}(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t) \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{(t)}, y_{(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\ \mathbb{E}_{(t)}[\ell(x_{(t)}, y_{(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t) \end{aligned}$$

what other programmers think I do



what I think I do

```
>>> from sklearn import svm
```

what I really do

Контакты



dmia@applieddatascience.ru



<https://t.me/joinchat/B1OITk74nRV56Dp1TDJGNA>



<https://goo.gl/forms/1k17ALSW2urgM91m2>