Лекция 1
Основные задачи машинного обучения и анализа данных

Габдуллин Р.А., Макаренко В.А.

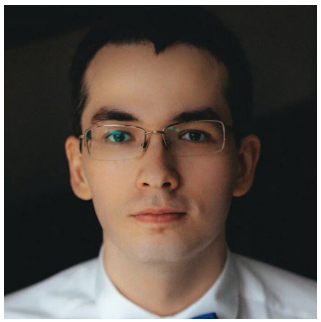
МГУ им. М.В. Ломоносова

10 января 2022

Преподаватели

- **Габдуллин Руслан Айдарович**

  rgabdullin@cs.msu.ru



- **Макаренко Владимир Александрович**

  vlamakarenko@mail.ru

Содержание курса

- Основные задачи машинного обучения и анализа данных
- Регрессия и классификация
- Линейные модели в задаче регрессии
- Линейные модели в задаче классификации
- Оценки качества моделей в задачах регрессии и классификации
- Выбор модели. Кросс-валидация. Отбор признаков
- Регуляризация. Метод опорных векторов
- Деревья решений
- Ансамбли моделей
- Введение в нейронные сети
- Анализ временных рядов

Литература

- Christopher M. Bishop. Pattern Recognition and Machine Learning
- T. Hastie, R. Tibshirani, J. Friedman. Elements of statistical learning
- Kevin P. Murphy. Machine Learning: A Probablistic Perspective
- Петер Флах. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных
- В.Н.Вапник. Восстановление зависимостей по эмпирическим данным.
- В.Н.Вапник, А.Я.Червоненкис. Теория распознавания образов.
- Vladimir N. Vapnik. The Nature of Statistical Learning Theory.

Ресурсы

- Сайт machinelearning.ru
- Школа Анализа Данных (видеолекции, сайт)
- Академия MADE
- Stanford (видеолекции, курс)
- Курсы на Coursera
- Сайт towardsdatascience.com
- Платформа Kaggle
- Блог А.Г.Дьяконова
- Курс от Open Data Science

Формат занятий и отчетность

Аудиторные часы:

- Лекция и семинар в неделю
- Консультации по необходимости

Самостоятельная работа:

- Две домашние работы (200 баллов)
- Проект (100 баллов)

Оценка	От	До
Отлично	255	300
Хорошо	195	254.9
Удовлетворительно	120	194.9
Неудовлетворительно	60	119.9

Технические рекомендации

- Компьютер с установленной средой Anaconda
- **Камера и микрофон (!)**

Что такое машинное обучение?

Машинное обучение (Machine learning)

Обширный раздел прикладной математики, находящийся на стыке математической статистики, оптимизации, искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться по эмпирическим данным.

Анализ данных (Data mining)

Процесс извлечения знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т. д. Полученные знания должны быть достоверными, полезными и интерпретируемыми.

Основные направления I

С точки зрения данных

- Классический ML (табличные данные)
- Компьютерное зрение
  - Распознавание людей на фотографиях
  - Анализ видеоконтента
  - Обработка изображений (фильтры, стиль, inpainting, . . .)
  - Генерация изображений
- Обработка естественного языка (NLP)
  - Машинный перевод
  - Генерация текстов
  - Распознавание именованных сущностей (NER)
  - Суммаризация (Summary)
  - Чат-боты
- Анализ временных рядов
- Рекомендательные системы

Основные направления II

По типу задач

- Обучение с учителем (supervised learning)
- Обучение без учителя (unsupervised learning)
- Обучение с частичным привлечением учителя (semi-supervised learning)
- Обучение с подкреплением (reinforcement learning)

Основные направления II

По типу задач

- **Обучение с учителем (supervised learning)**
- Обучение без учителя (unsupervised learning)
- Обучение с частичным привлечением учителя (semi-supervised learning)
- Обучение с подкреплением (reinforcement learning)

Примеры задач.Распознавание лиц

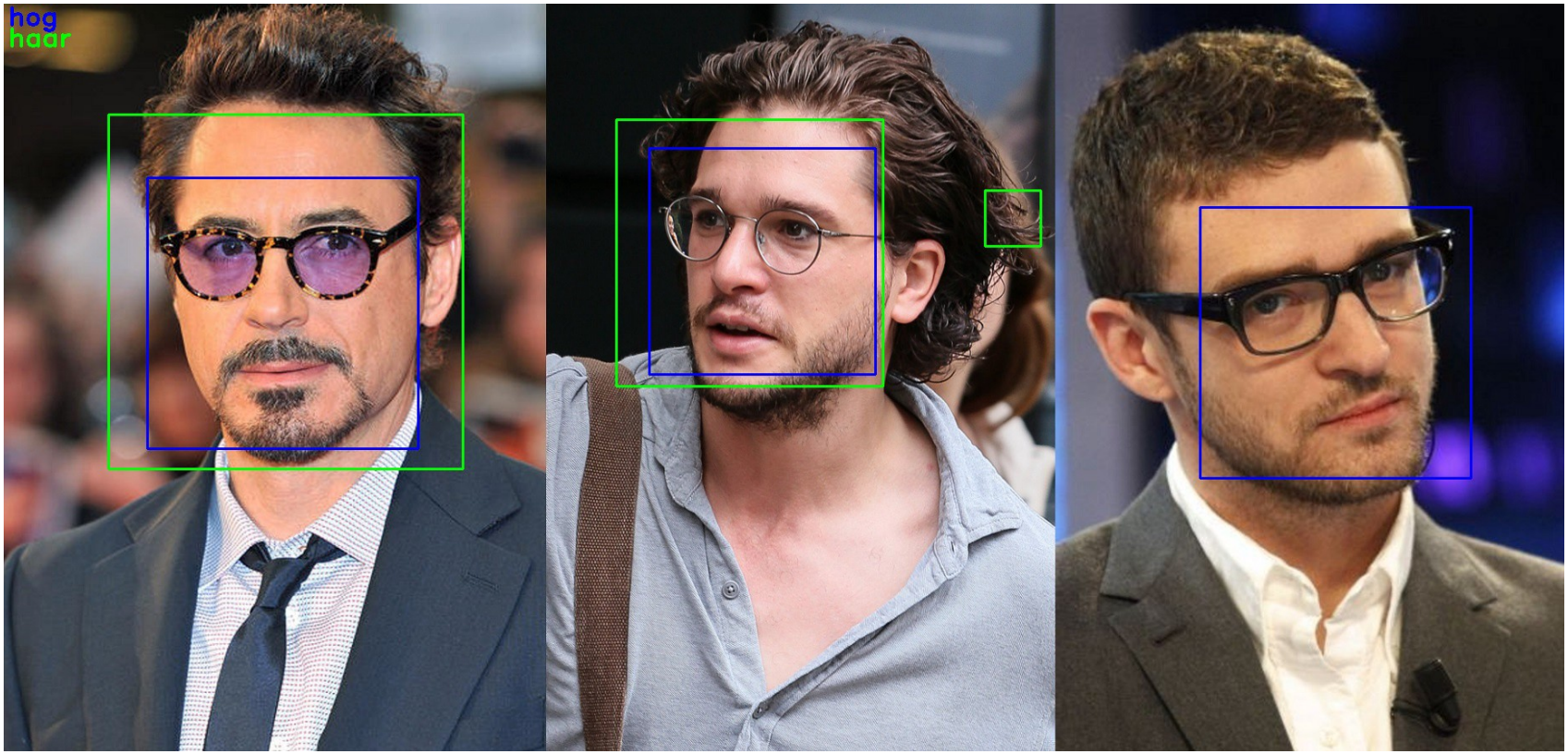


Рис. 1: Источник: medium.com

Примеры задач. Генерация изображений

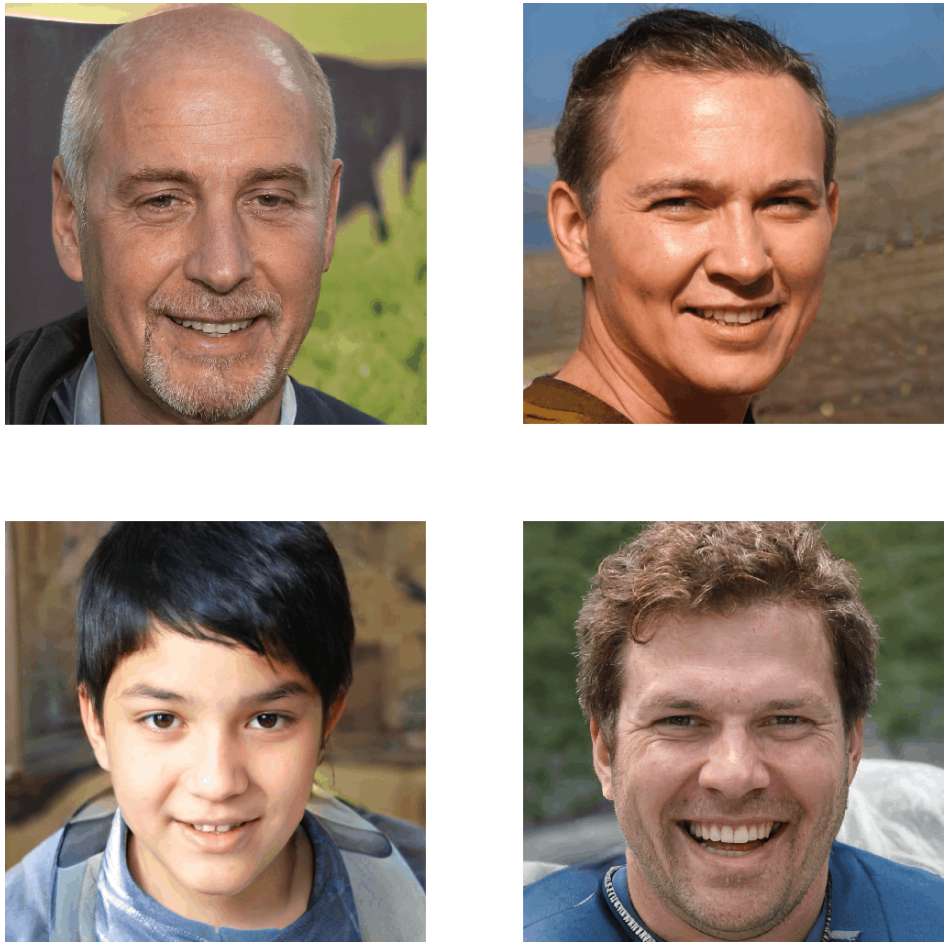


Рис. 2: Источник: thispersondoesnotexist.com

Примеры задач. Машинный перевод

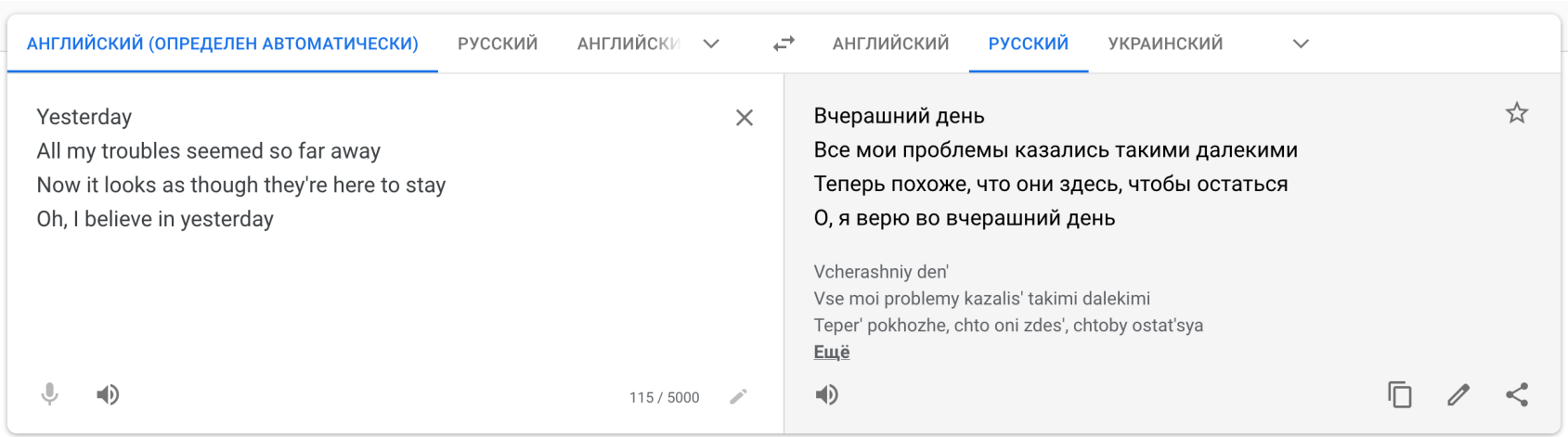


Рис. 3: Источник: translate.google.com

Примеры задач. Распознавание именованных сущностей

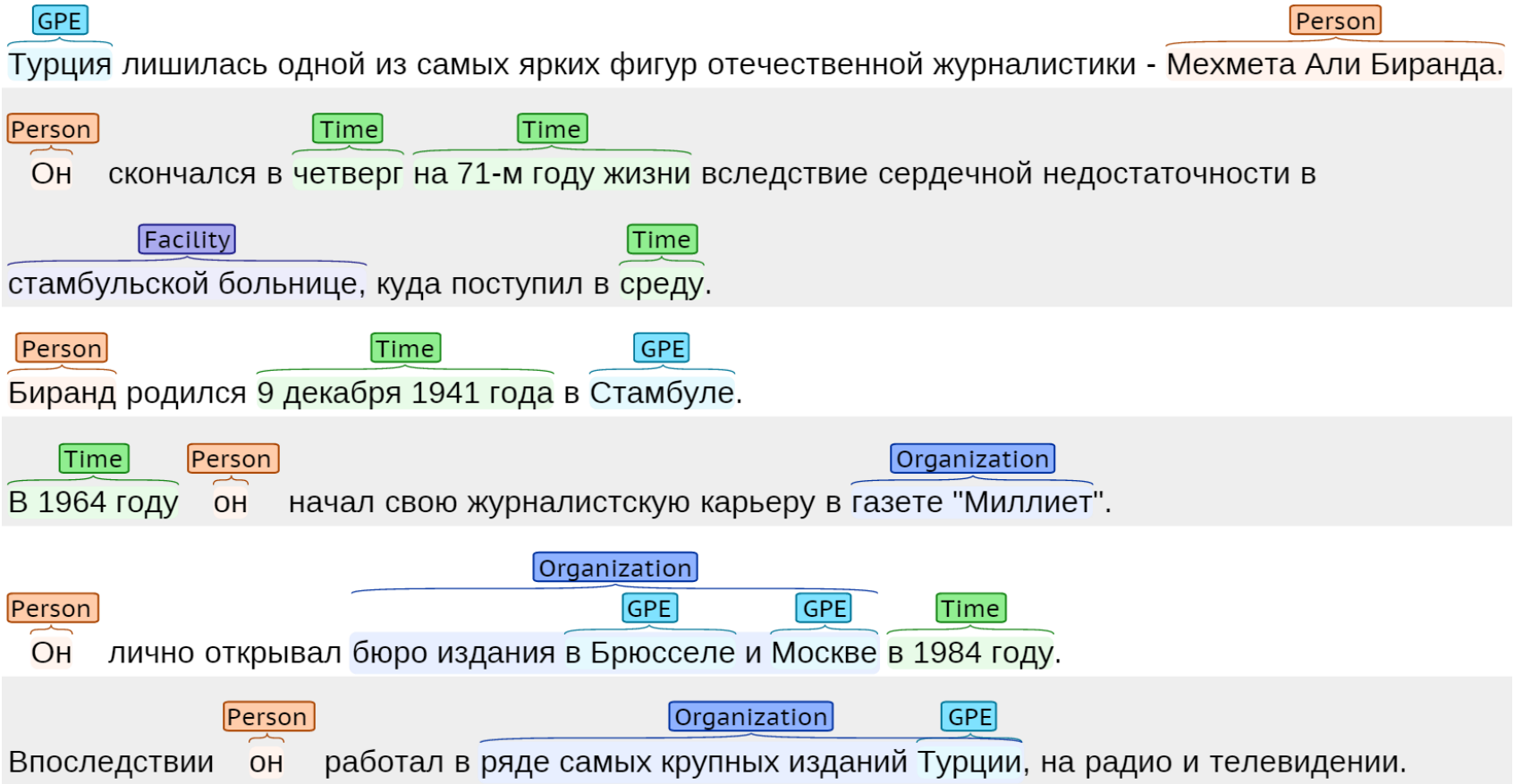


Рис. 4: Источник: habr.com

Примеры задач. Анализ временных рядов

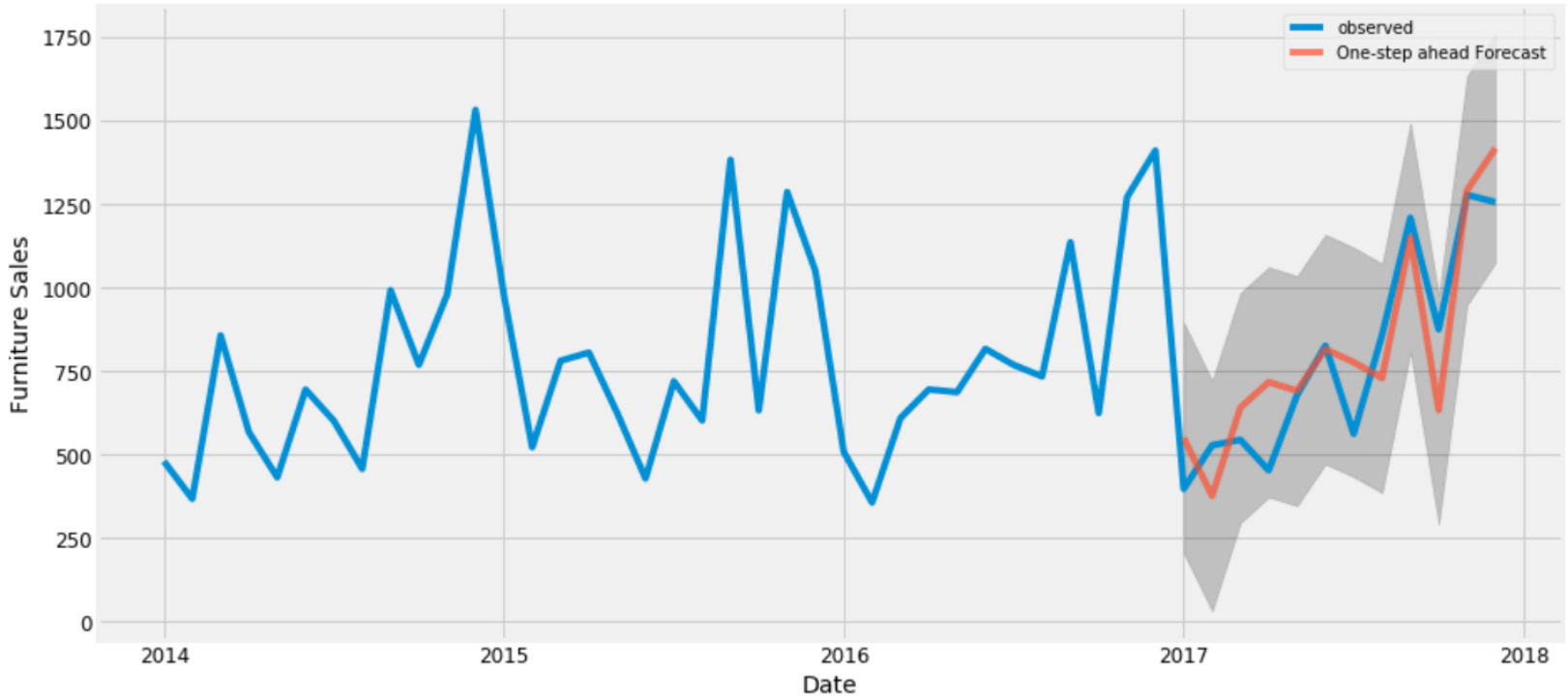


Рис. 5: Источник: becominghuman.ai

Примеры задач. Рекомендательные системы

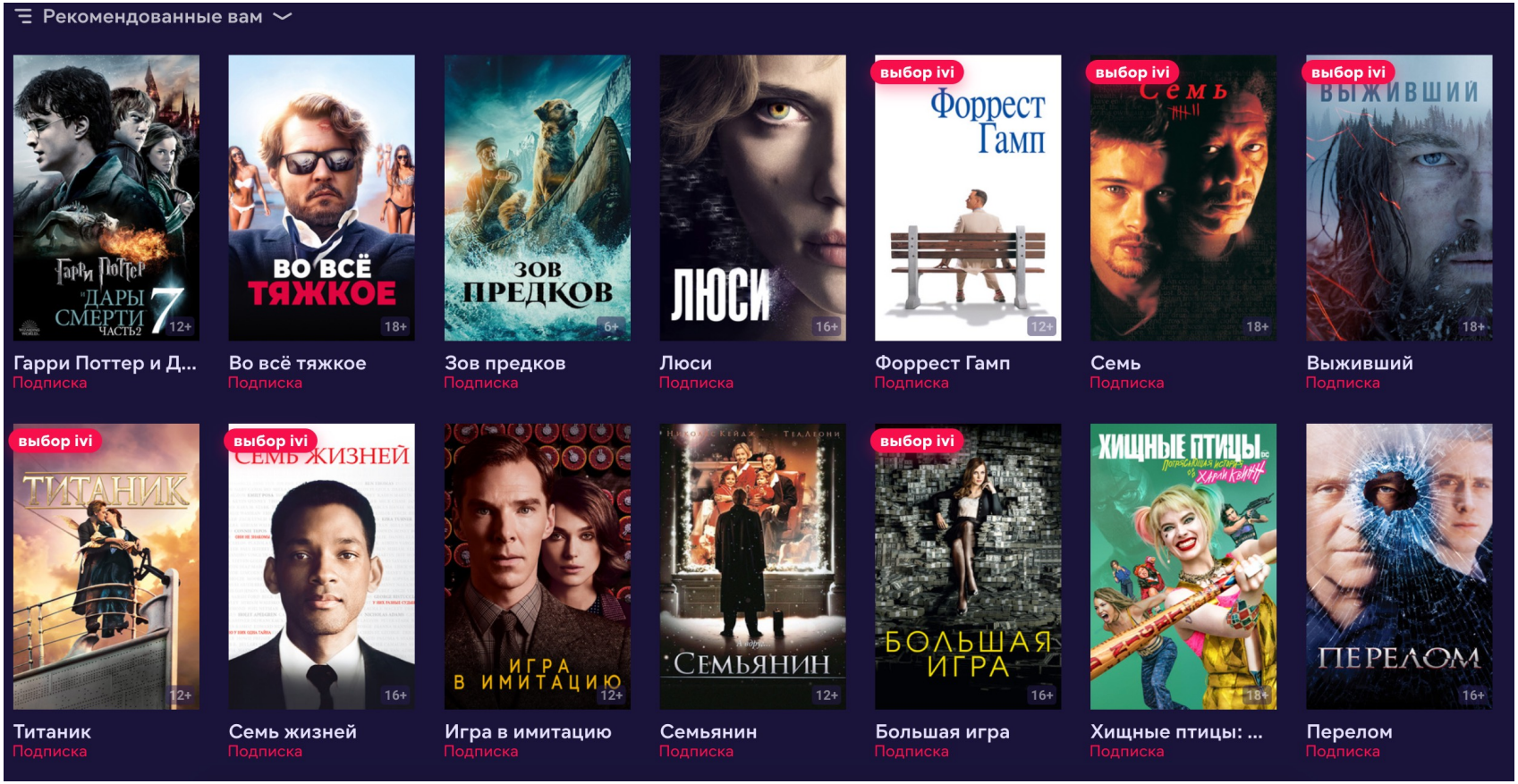


Рис. 6: Источник: ivi.ru

Обучение с учителем

X – множество объектов,
Y – множество ответов,
$y : X \to Y$ – неизвестная зависимость.

Дано:
$\{x_1, x_2, \ldots, x_\ell\} \subset X$ – обучающая выборка,
$y_i = y(x_i),\ i = 1, \ldots, \ell$ – известные ответы.

Найти:
$a : X \to Y$ – решающая функция, приближающая y на всём X.

Описание объектов. Признаки

X – множество объектов,
$f_j : X \to F_j, \quad j = 1, \ldots, n$ – признаки объектов (features),

Типы признаков:

Бинарные	Binary	$F_j = \{\text{true, false}\}$
Номинальные	Categorical	F_j – конечное мн-во
Порядковые	Ordinal	F_j – конечное упорядоченное мн-во
Количественные	Numerical	$F_j = \mathbb{R}$

$(f_1(x), f_2(x), \ldots, f_n(x))$ – признаковое описание объекта $x \in X$.
Матрица «объекты-признаки» (feature data)

$$F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \ldots & \ldots & \ldots \\ f_1(x_\ell) & \ldots & f_n(x_\ell) \end{pmatrix}$$

Примеры признаков

Пример: прогнозирование стоимости недвижимости.
Признаки:

Бинарные	Номинальные	Порядковые	Количественные
Наличие газа Наличие электричества Наличие балкона Наличие подвала	Регион расположения	Класс (эконом, средний, премиум)	Удаленность от общественного транспорта Удаленность от центра Число владельцев Число комнат Число этажей Площадь

Как задаются ответы

Задача восстановления регрессии:

- Вещественный ответ: $Y = \mathbb{R}$ или $Y = \mathbb{R}^m$

Задача классификации:

- Два класса: $Y = \{0, 1\}$
- Несколько классов: $Y = \{1, 2, 3, \ldots, m\}$
- Несколько пересекающихся классов: $Y = \{0, 1\}^m$

Примеры задач. Классификация

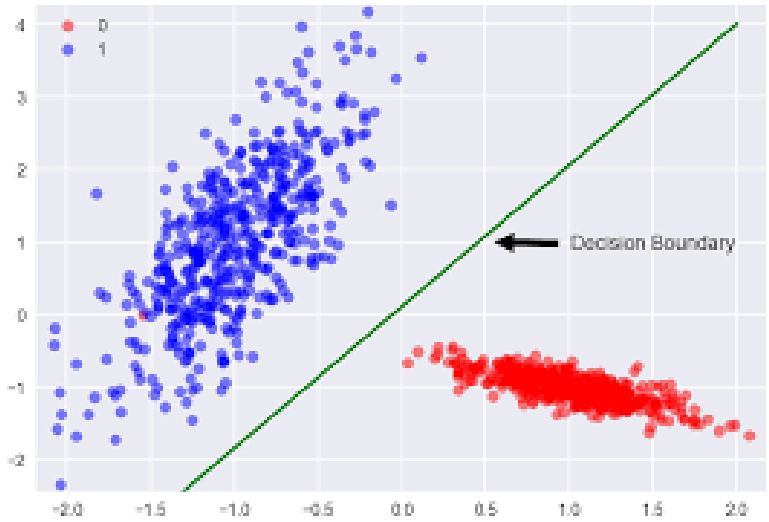


Рис. 7: Источник: kaggle.com

- Выявление email-спама: spam/ham
- Распознавание рукописных цифр: десять классов
- Задача кредитного скоринга: выдать кредит/отказать в выдаче кредита

Примеры задач. Регрессия

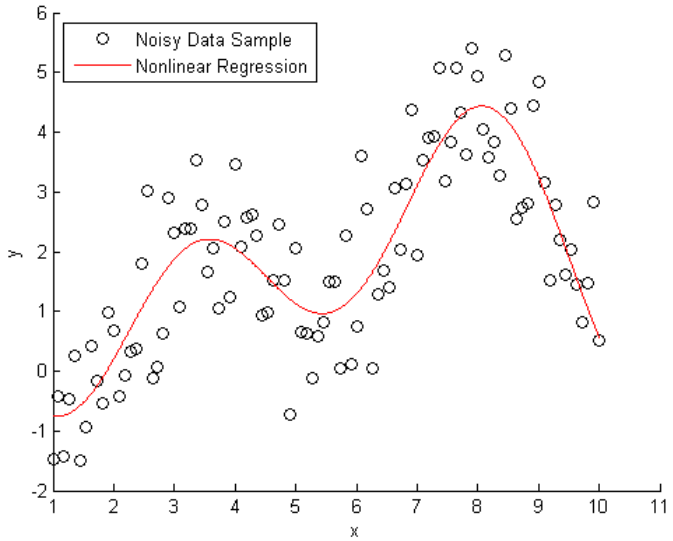


Рис. 8: Источник: datascience.stackexchange.com

- Оценка стомости недвижимости
- Кредитный скоринг: по анкете заемщика оценить величину кредитного лимита
- Предсказание объема продаж товара в магазине

Обучение без учителя

Дано:

- Признаки объектов без ответов

Требуется:

- Найти зависимости между объектами

Типичные задачи:

- Кластеризация объектов
- Снижение размерности данных
- Поиск аномальных объектов
- Восстановление плотности распределения

Обучение без учителя. Кластеризация

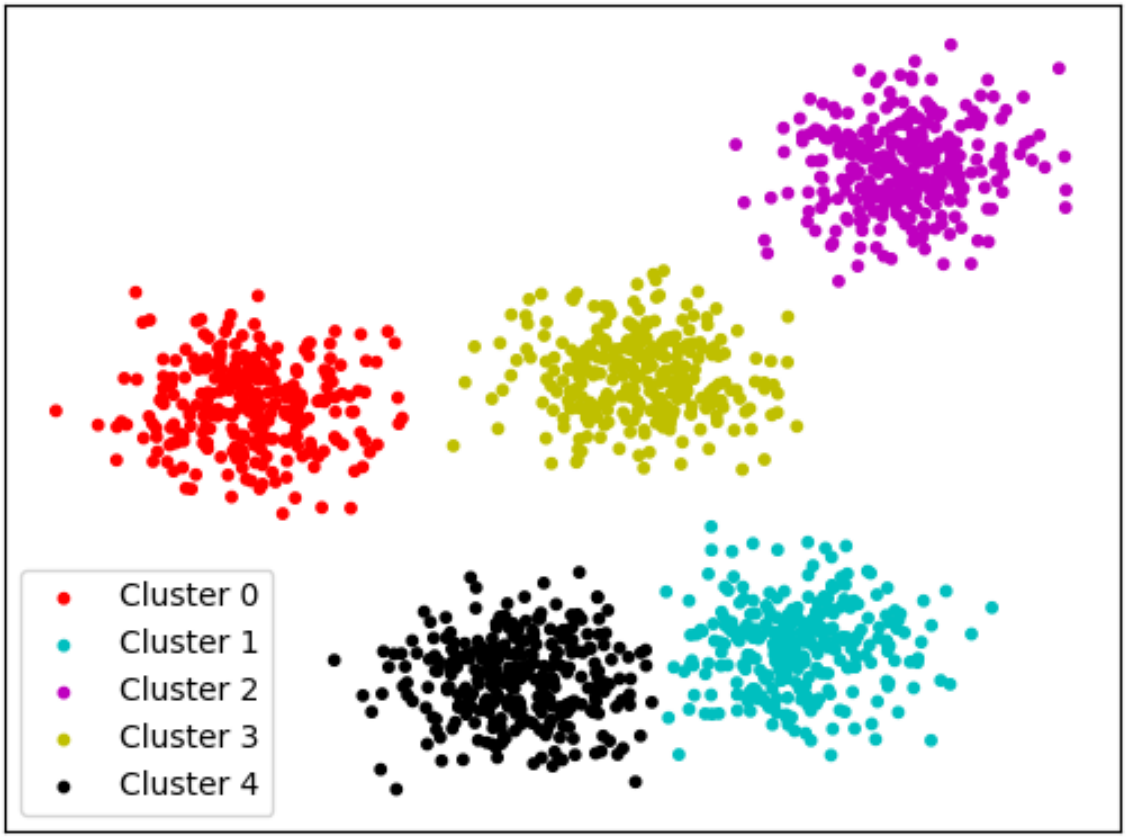


Рис. 9: Источник: medium.com

Обучение без учителя. Снижение размерности данных

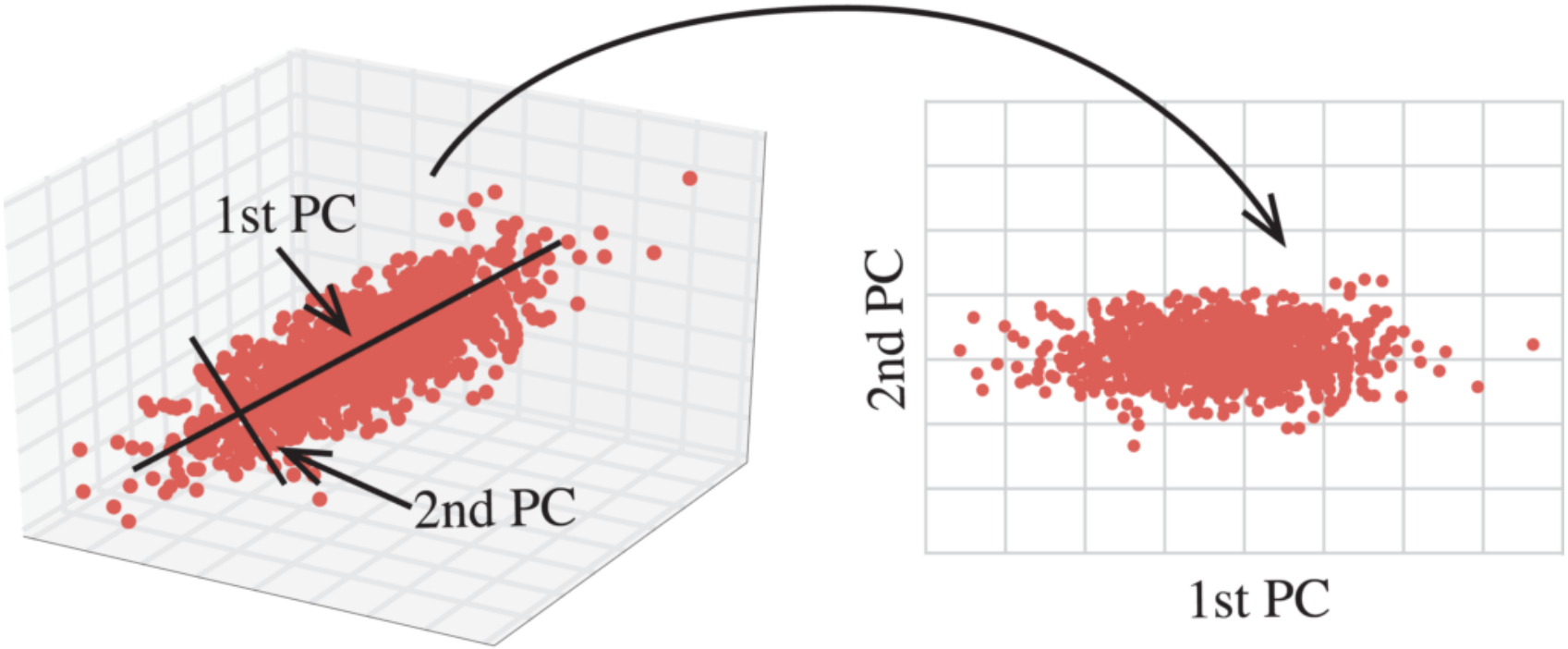
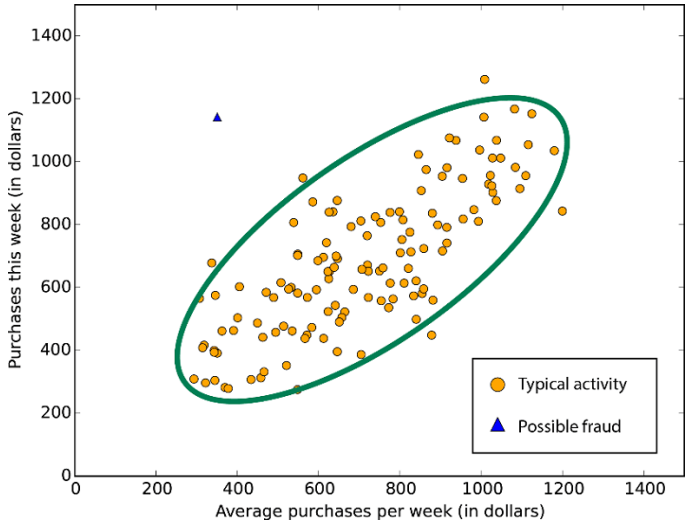
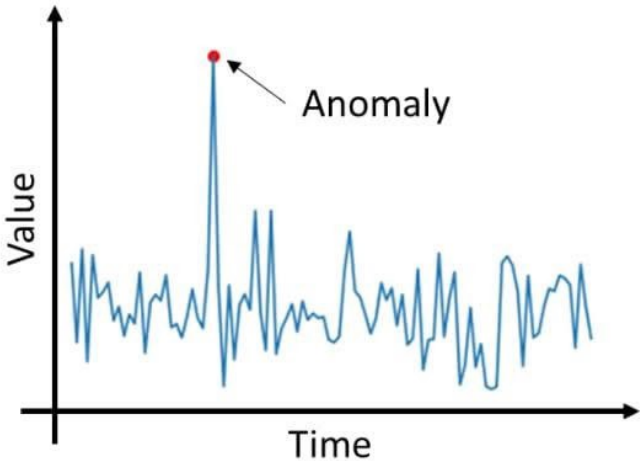


Рис. 10: Источник: medium.com

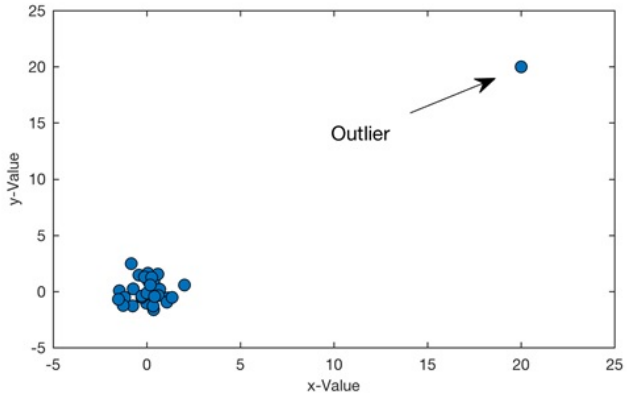
Обучение без учителя. Поиск аномалий



(a) Источник: researchgate.net



(b) Источник: medium.com



(c) Источник: datasciencecentral.com

Обучение без учителя. Восстановление плотности распределения

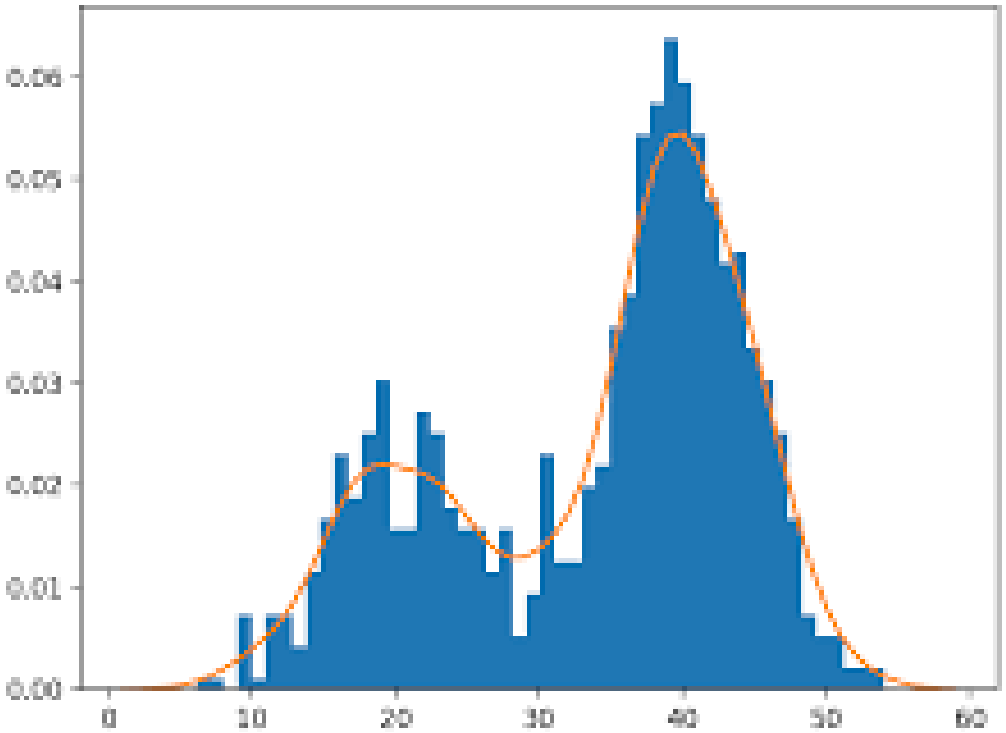


Рис. 12: Источник: machinelearningmastery.com

Обучение с частичным привлечением учителя

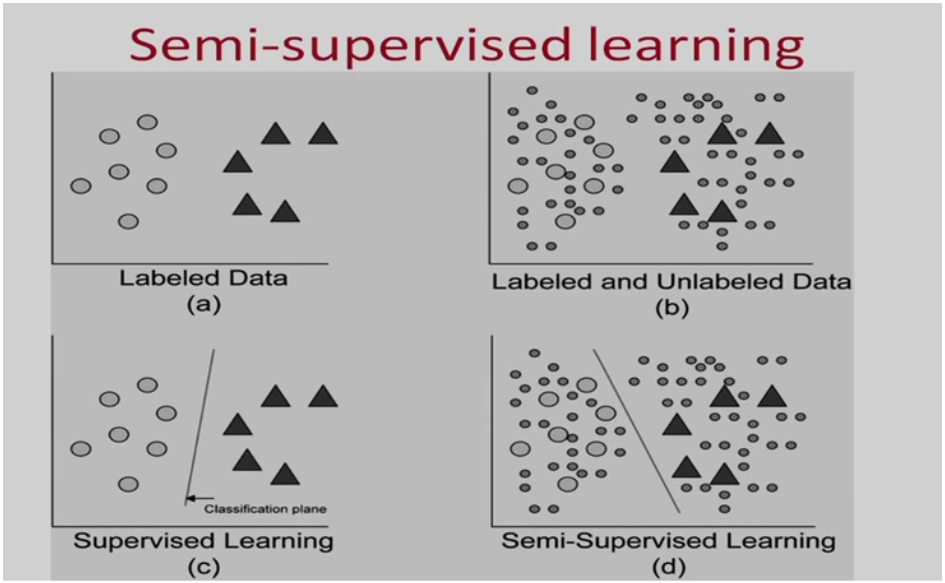


Рис. 13: Источник: medium.com

Дано:

- Признаки объектов
- У небольшой части объектов известны ответы

Требуется:

- Найти решающую функцию

Обучение с подкреплением

Рис. 14: Источник: kdnuggets.com

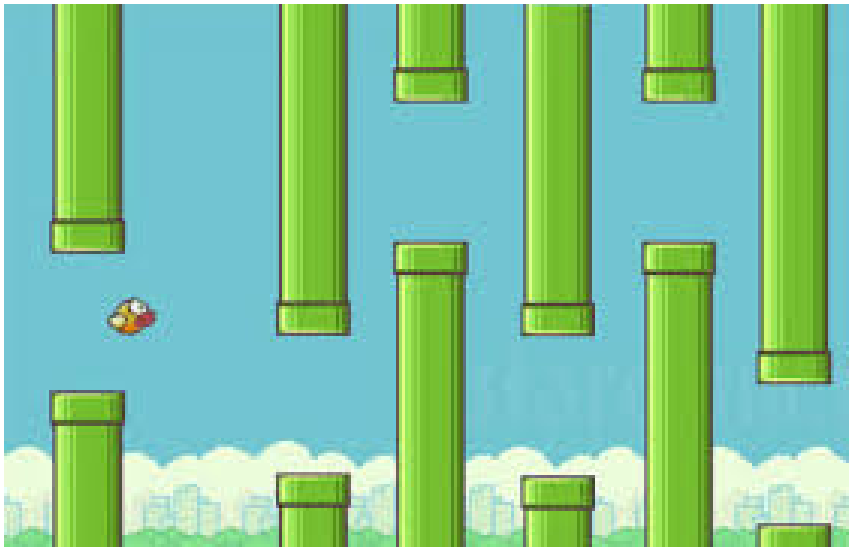
- Модели доступен ограниченный набор действий
- Модель (агент) действует в динамической среде
- Модель получает награду или штраф за выбранное действие
- Чаще всего получение обратной связи затруднено: дорого, долго, вычислительно затратно
- Модель ограничена в получении обратной связи в единицу времени
- Цель: максимизировать награду

Обучение с подкреплением. Игры

(a) Источник: Википедия

(b) Источник: Википедия

(c) Источник: dendyemulator.ru



(d) Источник: psmag.com

Обучение с подкреплением. Робототехника

Рис. 16: Источник: bostondynamics.com

Обучение с подкреплением. Беспилотные автомобили

Рис. 17: Источник: computerra.ru

Модель алгоритмов

Зачастую семейство решающих правил (алгоритмов) задается в виде семейства параметрических функций

$$A = \{g(x, \theta) | \theta \in \Theta\},$$

где $g : X \times \Theta \to Y$ – фиксированная функция,
Θ – множество допустимых значений параметра θ.

Пример.

Линейная модель:

$$\theta = (\theta_1, \ldots, \theta_n), \quad \Theta = \mathbb{R}^n.$$

Для задачи регрессии:

$$g(x, \theta) = \sum_{k=1}^{n} \theta_k f_k(x).$$

Для задачи классификации:

$$g(x, \theta) = \text{sign} \sum_{k=1}^{n} \theta_k f_k(x).$$

Обучение и применение модели

- Обучение: по объектам и ответам подобрать $\theta \in \Theta$.
  Метод обучения:
  $$\mu : (X, Y)^\ell \to A$$
  По выборке $\mathbb{X} = (x_i, y_i)_{i=1}^\ell$ получаем алгоритм $a = \mu(\mathbb{X})$.

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix} \xrightarrow{\mu} a$$

- Применение модели: получение ответов на новых данных.

$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_\ell) & \dots & f_n(x'_\ell) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_\ell) \end{pmatrix}$$

Функционалы качества

- Функция потерь $L(x, a)$ – величина ошибки алгоритма $a \in A$ на объекте x.

  Функция потерь для классификации:

  $$L(x, a) = [a(x) \neq y(x)]$$

  Функции потерь для задачи регрессии:

  $$L(x, a) = (a(x) - y(x))^2, \quad L(x, a) = |a(x) - y(x)|$$

- Теоретический риск:

  $$Q(a) = \mathbb{E}L(x, a)$$

- Эмпирический риск:

  $$Q(a, \mathbb{X}) = \frac{1}{\ell} \cdot \sum_{i=1}^{\ell} L(x_i, a)$$

Как обучать модель?

- Наилучший алгоритм:

$$a^* = \min_{a \in A} Q(a).$$

- Проблема: имеем лишь конечное число объектов, информации о всей генеральной совокупности нет.
- Решение: будем минимизировать эмпирический риск

$$a^* = \min_{a \in A} Q(a, \mathbb{X}).$$

Проблема переобучения (overfitting)

Если алгоритм показывает хорошее качество на обучающей выборке, то совсем не факт, что он будет хорошо работать на новых данных. Например, алгоритм

$$a(x) = \sum_{i=1}^{\ell} y(x_i)[x = x_i]$$

имеет нулевую ошибку на тренировочной выборке, но совершенно бесполезен.

Если качество модели на тренировочной выборке сильно лучше, чем на новых данных, то говорят, что модель переобучена.

Проблема переобучения (overfitting)

- Причина переобучения: семейство алгоритмов слишком гибкое («лишние» степени свободы тратятся на запоминание шума в данных)
- Как обнаружить переобучение: разделить обучающую выборку на две части: train и test. Обучить модель на train, оценить качество на test.

Резюме лекции

- Объекты, признаки, ответы
- Направления машинного обучения
- Типы задач
  - Обучением с учителем
    - Классификация
    - Регрессия
  - Обучение без учителя
    - Кластеризация
    - Снижение размерности
    - Поиск аномалий
  - Обучение с частичным привлечением учителя
  - Обучение с подкреплением
- Модель алгоритмов
- Функционалы качества
  - Теоретический риск
  - Эмпирический риск
- Проблема переобучения