Введение в машинное обучение

Омелюсик Владимир Степанович

Национальный Исследовательский Университет
«Высшая школа экономики»
–
Факультатив «Введение в анализ данных и машинное обучение на Python»

23 ноября 2019 г.

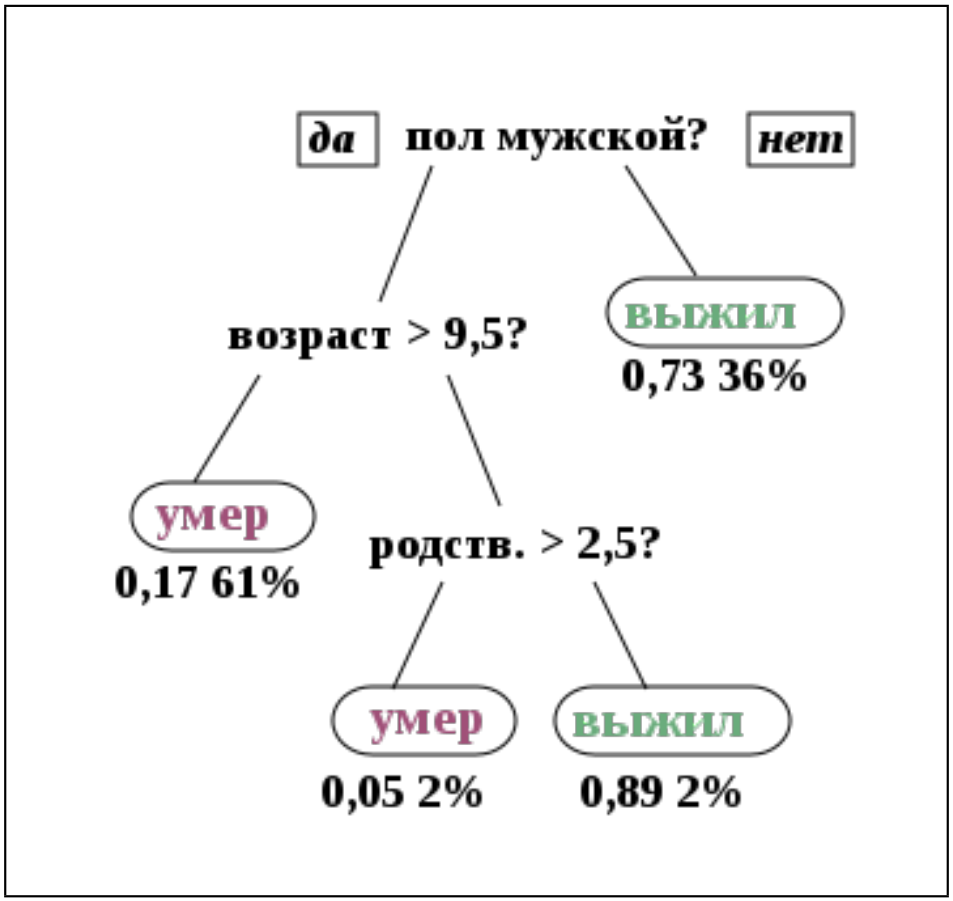
Немного истории

- 1956 год – первый семинар по проблемам искусственного интеллекта.
  Задача: моделирование интеллекта человека математическими методами.

- До 1970-х годов – простейшие системы AI.
  - Доказательство теорем методом дедукции.
  - ELIZA – синтаксический анализатор.

- 1980-е годы – решающие деревья.
  В каждом узле дерева стоит некоторое условие. В зависимости от выполнения данного условия, дерево переходит в следующую ветвь.

Дерево решений: пример



Задача «Титаник» (Источник)

Немного истории

- 1990-е годы – развитие более продвинутых систем машинного обучения:
  - Нейронные сети.
  - Генетические алгоритмы.
- 2000-е годы и современность – Deep Learning.
  Построение моделей с очень высокой точностью распознавания.

Определение

Машинное обучение

Область науки, изучающая построение моделей и алгоритмов, позволяющих компьютерным системам воспроизводить зависимости между различными объектами без их непосредственного программирования.

Неформально:

- «Обучение» специальных моделей на некоторых данных.
- В ходе «обучения» происходит «запоминание» зависимостей, представленных в данных (важна репрезентативность выборки).
- После «обучения» модель способна давать корректные предсказания на новых данных.

Зависимости

- Зависимости позволяют нам давать ответы на интересующие нас вопросы.
- Зависимость можно сформулировать словесно:
  - «Площадь прямоугольника равна произведению его длины и ширины».
  - «Вероятность выжить на Титанике зависит от пола пассажира».
  - «Вероятность того, что данный цветок ириса принадлежит к виду I. kaempferi, зависит от длины его лепестков».
- Но для получения чётких количественных результатов нужно формализовать словесные формулировки.
- Необходимо выразить их математическими функциями.
- Это не всегда просто (иногда – невозможно), так как истинные функции могут быть достаточно сложными.

Пример зависимости: килограммы и тонны

- Вопрос: как перевести массу в тоннах в массу в килограммах?
- Зависимость: 1 тонна = 1000 кг.
- Формализация:

$$f(x) = \frac{x}{1000},$$

где x – масса в тоннах.

Пример зависимости: предсказание погоды

- Вопрос: какая завтра будет погода?
- Зависимость: как погода завтра зависит от ...?
- Формализация:
  Уравнения Навье-Стокса (частично, источник):

$$\frac{\partial \rho}{\partial t} + \vec{\nabla} \cdot (\rho \vec{u}) = 0 \tag{1}$$

$$\frac{\partial (\rho \vec{u})}{\partial t} + \vec{\nabla} \cdot [\rho \overline{\overline{u \otimes u}}] = -\vec{\nabla} p + \vec{\nabla} \cdot \bar{\bar{\tau}} + \rho \vec{f} \tag{2}$$

$$\frac{\partial (\rho e)}{\partial t} + \vec{\nabla} \cdot ((\rho e + p) \vec{u}) = \vec{\nabla} \cdot (\bar{\bar{\tau}} \cdot \vec{u}) + \rho \vec{f} \vec{u} + \vec{\nabla} \cdot (\vec{\dot{q}}) + r \tag{3}$$

- Позволяют найти давление и скорость воздуха в любой точке. Но тяжело решать.

Пример зависимости: анализ тональности текста

«Быть или не быть, вот в чем вопрос. Достойно ль
Смиряться под ударами судьбы...»

У. Шекспир «Гамлет»

- Вопрос: какая тональность у данного фрагмента текста?
- Зависимость: ...?
- Формализация: x – фрагмент текста, $f(x) = ...?$ Непонятно.

Более сложные вопросы

- Какой будет спрос на овощи в продуктовом магазине в следующем месяце?
- Выдать ли клиенту кредит?
- На фотографии кошка или собака?

Найти точные математические функции для ответа на данные вопросы сложно (или невозможно). Но если у нас есть некоторый набор данных, можно попытаться приблизить зависимости некоторыми математическими моделями.

Приближение зависимостей

Цель машинного обучения

Используя только данные, а не теорию, попытаться восстановить истинные зависимости.

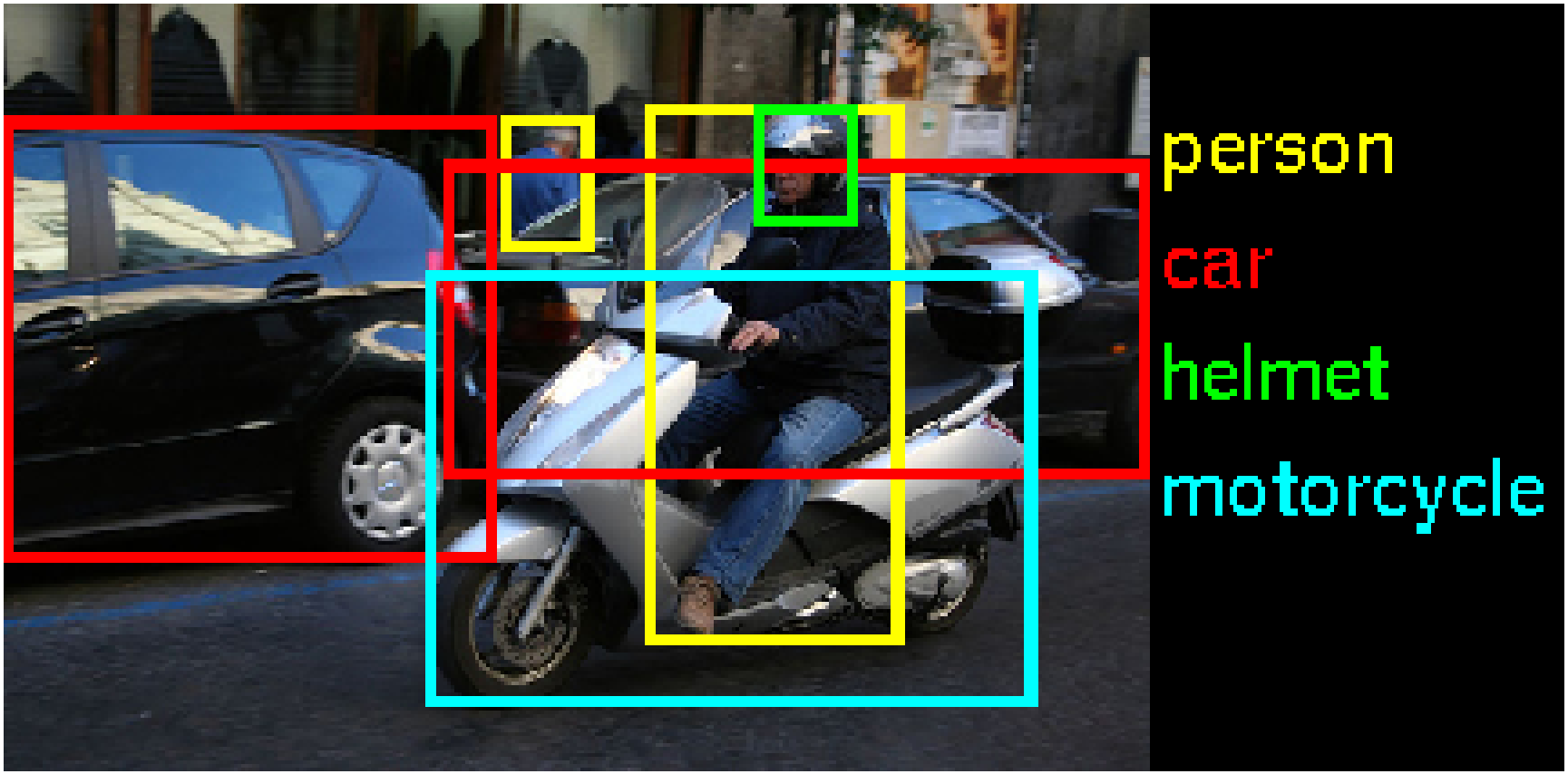
- Пример с монеткой: истинная вероятность того, что выпадет орёл равна p, её оценка равна $\hat{p}$.
- Формально: пусть истинная зависимость: $y(x)$ – и её мы не знаем. Будем пытаться по данным подобрать некоторую функцию $\hat{y}(x)$, которая приближает истинную зависимость.

Применение машинного обучения: AlphaGo

- Нейронная сеть, победившая чемпиона мира в 2016 году.
- Обучалась, играя сама с собой.

Применение машинного обучения: ImageNet

- Соревнование по распознаванию объектов на изображении.
- Решается с помощью нейронных сетей.



Применение машинного обучения: Отдел кадров

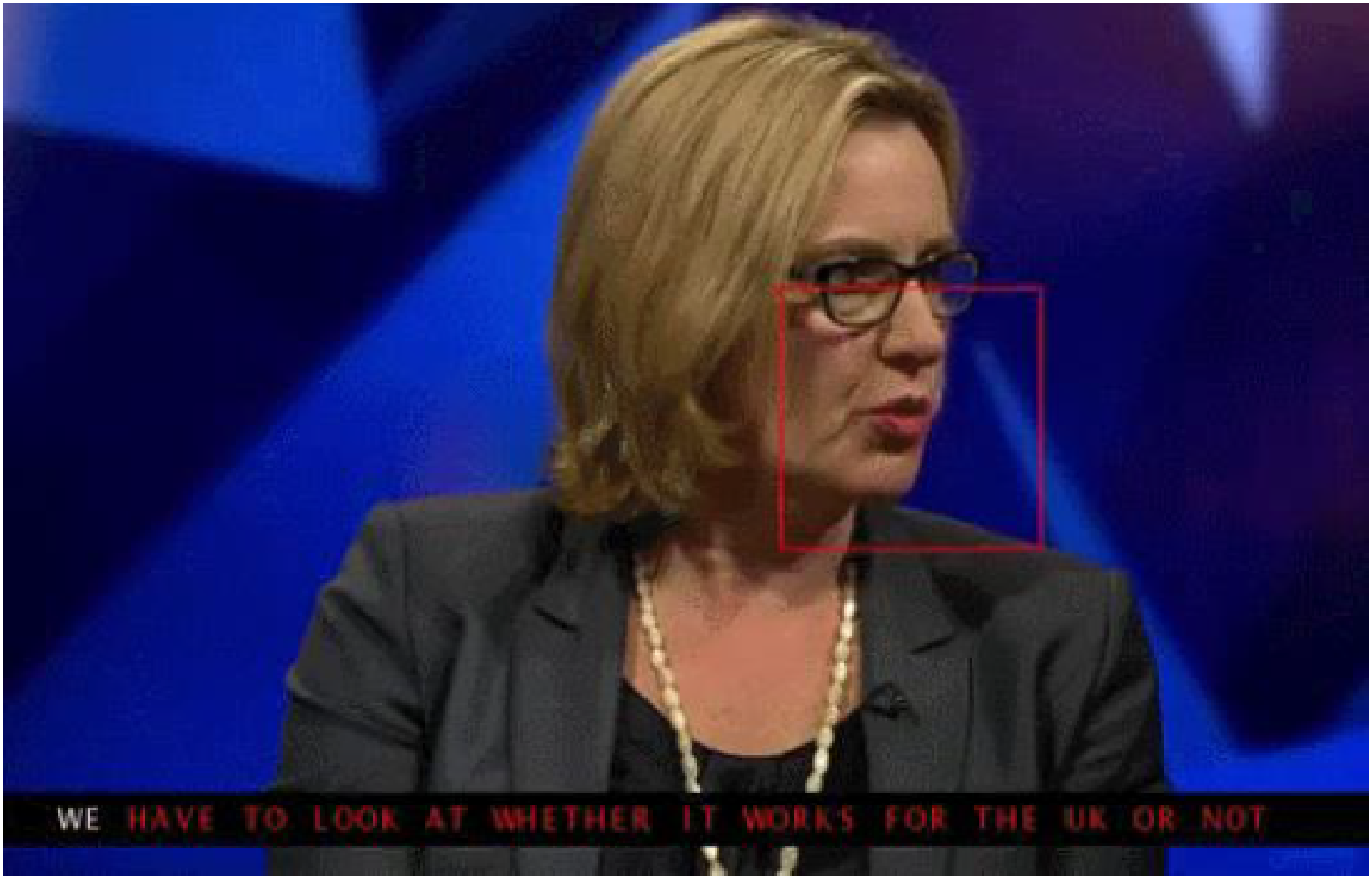
- Поиск кандидатов, предсказание результата собеседования.
- Предсказание ухода сотрудника.
- Анализ внутренних каналов информации, определение жалоб.

Применение машинного обучения: Рекомендательные и поисковые системы

- Рекомендательные системы: Netflix, Amazon, ... – на основании поведения пользователя определяют, какой товар или услугу разумно ему предложить.
- Поисковые системы: Google, Яндекс, ... – на основании запроса пользователя определяют, какие веб-сайты наиболее соответствуют запросу.

Применение машинного обучения: Чтение по губам

- Google Deepmind: модель, которая была способна превзойти профессионального чтеца по губам.



Типы задач машинного обучения

- Мы уже знакомы с основными понятиями машинного обучения: целевая переменная (target) и признаки (features).
- Типы задач машинного обучения (в зависимости от наличия целевой переменной):
  1. Обучение с учителем (supervised learning).
     - ★ Классификация.
     - ★ Регрессия.
     - ★ Ранжирование.
  2. Обучение без учителя (unsupervised learning).
     - ★ Кластеризация.
  3. Обучение с подкреплением (reinforcement learning).

Обучение с учителем

Обучение с учителем

Вид обучения, когда имеется целевая переменная, и модель обучается так, чтобы наиболее правильно предсказывать целевую переменную.

В обучении с учителем выделяют следующие виды задач:

- Регрессия: $Y \in \mathbb{R}$.
- Классификация: $|Y| < \infty$.
- Ранжирование: Y – конечное упорядоченное множество.

Задача регрессии

- $Y \in \mathbb{R}$, то есть зависимая переменная может принимать любые вещественные значения (бесконечное число значений).
- Пример: линейная регрессия

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots \hat{\beta}_k X_{ki},$$

где $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ – оценки истинных коэффициентов.

Задача регрессии: Пример

Предсказание цены на жильё в зависимости о среднего числа комнат (по датасету `boston` в `sklearn`):

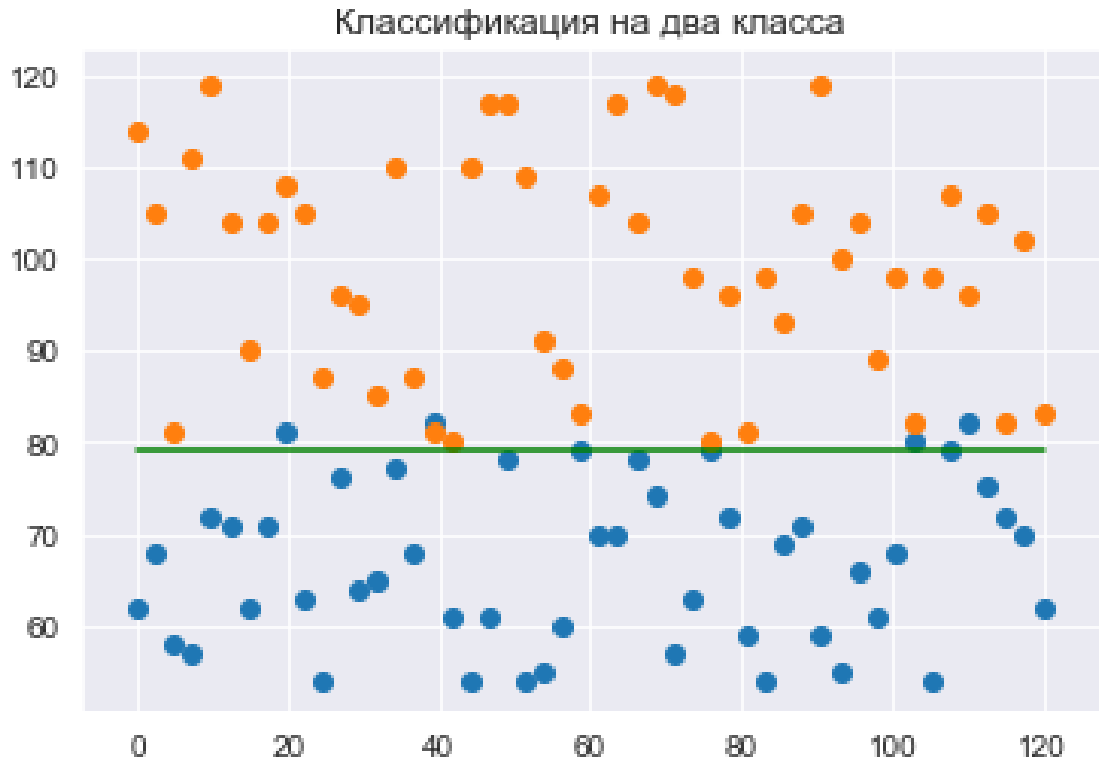
$$\hat{Y}_i = -34.67 + 9.1 X_i^{RM}$$

Задача классификации

- $|Y| < \infty$, то есть зависимая переменная может принимать ограниченное число значений.
- Виды:
  - Бинарная классификация: $Y = 1$ или $Y = 0$.
  - Многоклассовая классификация: $Y = 1$, или $Y = 2$, . . ., или $Y = K$.
  - Классификация с пересекающимися классами: Y может принимать несколько значений из множества: $\{1, 2, \ldots K\}$.

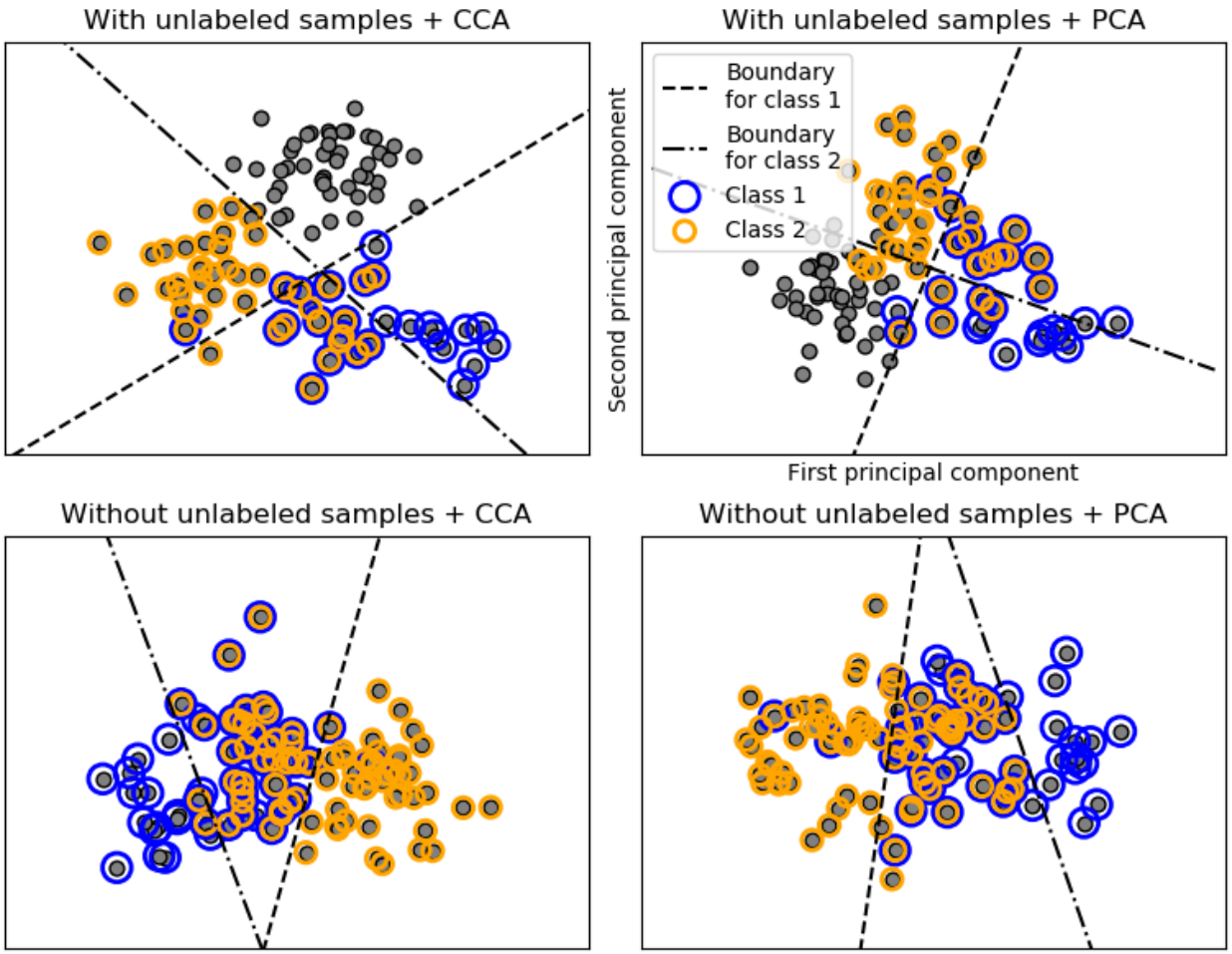
Бинарная классификация: Пример

Задача: провести линию так, чтобы наиболее точно разделить объекты разных классов.



Многоклассовая классификация: Пример

Задача: провести линии так, чтобы наиболее точно разделить объекты разных классов.



Примеры реальных задач классификации и регрессии

1. Медицинская диагностика.
   - Наблюдение: пациент в момент времени t.
   - Предсказание: диагноз.
2. Предсказание оттока клиентов.
   - Наблюдение: клиент в момент времени t.
   - Предсказание: уйдёт или нет в течение трёх месяцев.
3. Прогнозирование времени сна млекопитающих.
   - Наблюдение: млекопитающее в момент времени t.
   - Предсказание: среднее время сна в сутки в секундах.

Примеры реальных задач классификации и регрессии

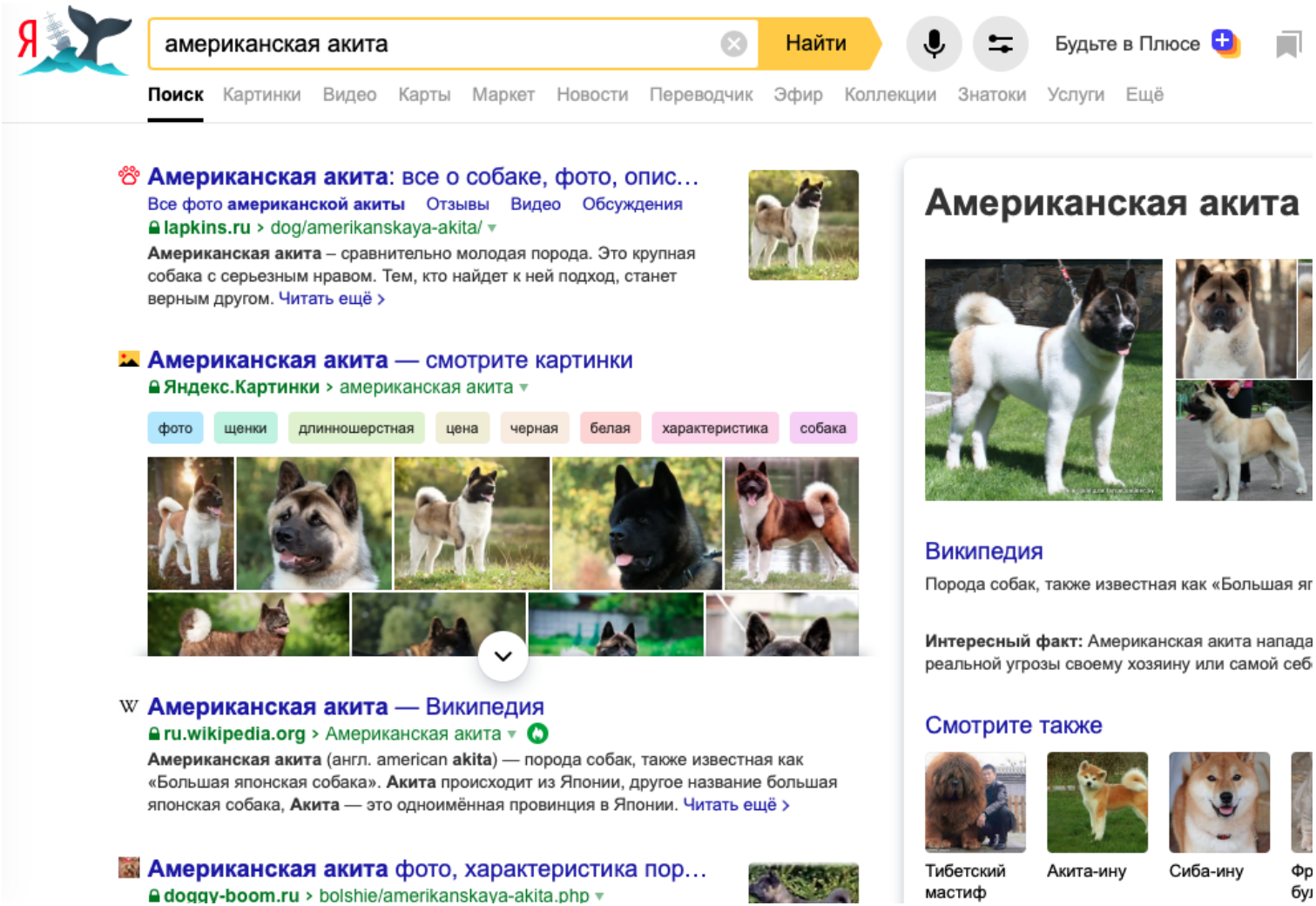
1. Медицинская диагностика.
   - Наблюдение: пациент в момент времени t.
   - Предсказание: диагноз.
   - Многоклассовая классификация.
2. Предсказание оттока клиентов.
   - Наблюдение: клиент в момент времени t.
   - Предсказание: уйдёт или нет в течение трёх месяцев.
   - Бинарная классификация.
3. Прогнозирование времени сна млекопитающих.
   - Наблюдение: млекопитающее в момент времени t.
   - Предсказание: среднее время сна в сутки в секундах.
   - Регрессия.

Задача ранжирования

- Y – конечное упорядоченное множество (например, пользовательские оценки веб-сайтов).
- Дан набор «запросов» $Q = (q_1, q_2, \ldots q_n)$ и набор «документов» $D = (d_1, d_2, \ldots d_m)$.
- Цель: используя Y, построить модель $R(q_i, D)$, которая для запроса q «правильно» бы упорядочивала набор документов D.

Задача ранжирования: Пример

Задача: упорядочить веб-сайты в соответствии с релевантностью запросу.



Обучение без учителя

Обучение без учителя

Вид обучения, когда целевая переменная неизвестна или отсутствует, и модель обучается только по признакам объектов.

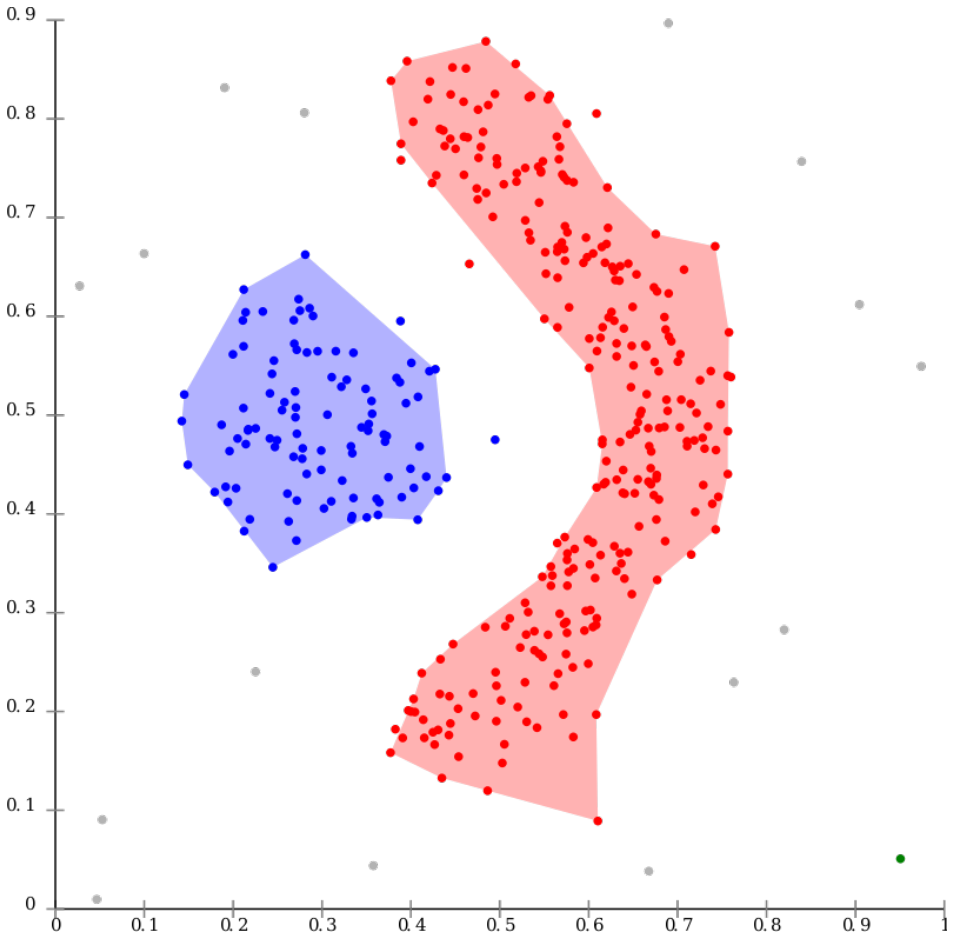
- Задача кластеризации: Y – отсутствует.

  Цель: найти группы похожих объектов (то есть разделить выборку на кластеры).

  Обучение происходит только на основе признаков объектов.

Задача кластеризации: Пример

Задача: разбить представленную выборку на кластеры, руководствуясь только признаковым описанием объектов.

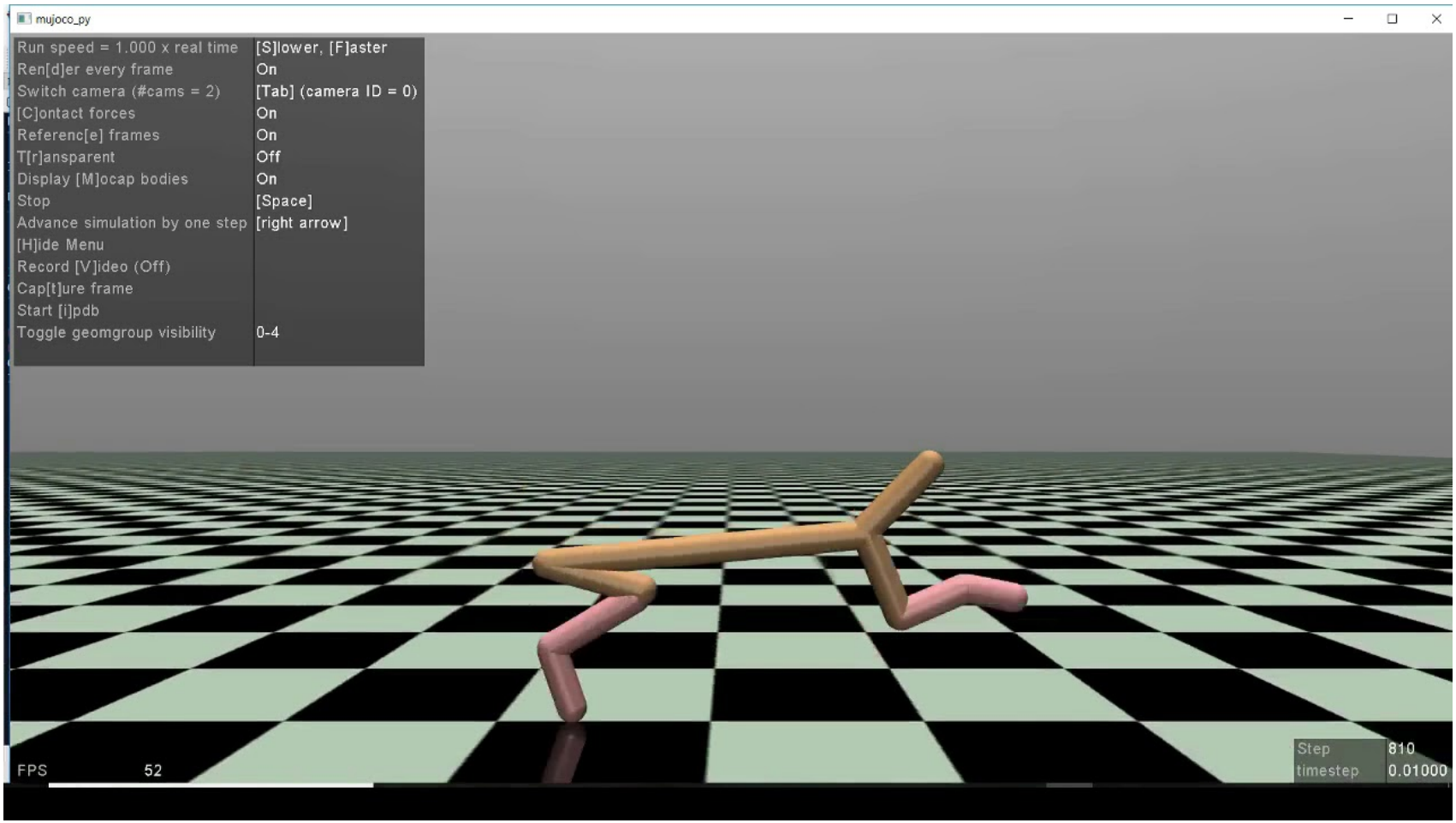


Обучение с подкреплением

- Другой подход к обучению: существует *среда* и отделённый от неё *агент*.
- На каждом шаге агент получает вознаграждение за выполненное им действие (может быть отрицательным).
- Учитель отсутствует: обучение происходит через максимизацию суммарного вознаграждения.
- Примеры:
  - AlphaGo.
  - Контроль движений робота.
  - Управление энергетической станцией.
  - Реализация управления вертолётом.

Обучение с подкреплением: Пример

Задача: научить агента использовать нужные движения, чтобы переместиться на максимальное расстояние.



Другие понятия машинного обучения

- Что знаем теперь:
  - Можем решать разные типы задач: регрессия, классификация, кластеризация.
  - Y – зависимая переменная, X_1, X_2, ... – признаки.
  - Хотим восстановить зависимость $Y(X)$ (для обучения с учителем).
  - Для восстановления зависимости строим модель $\hat{Y}(X)$.
- Как происходит обучение? $\Rightarrow$ Функция потерь.
- Как определить качество модели? $\Rightarrow$ Функционал качества, Обобщающая способность.

Функция потерь

Функция потерь

Функция, измеряющая ошибку алгоритма. Иначе говоря, мера корректности алгоритма.

- Много различных вариантов.
- Алгоритм обучается путём минимизации функции потерь.
- Пример: среднеквадратичная ошибка (MSE, mean squared error):

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Функционал (метрика) качества

Функционал качества

Функция, используемая для оценки качества и сравнения различных моделей.

- Много различных вариантов.
- С помощью функционала качества мы сравниваем различные модели.
- Пример: доля правильных ответов (accuracy) – для задачи классификации:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\hat{Y}_i = Y_i\},$$

где $\mathbb{I}\{\cdot\} = \begin{cases} 1, & \text{если условие в скобках выполнено.} \\ 0, & \text{если условие в скобках не выполнено.} \end{cases}$

Обобщающая способность

Обобщающая способность

Способность модели давать корректные предсказания на новых данных, не участвовавших при её обучении.

Недообучение

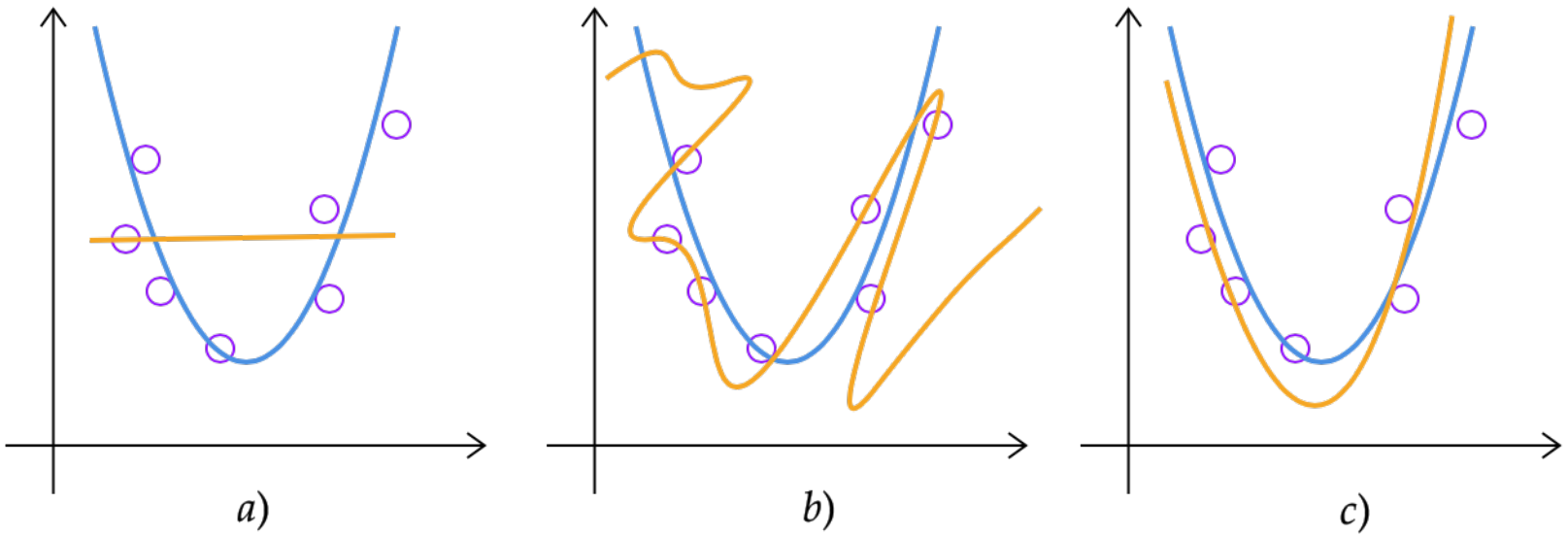
Ситуация, когда модели не удалось правильно «запомнить» зависимости в данных. В этом случае качество будет низким как на обучающей выборке, так и на новых данных.

Переобучение

Ситуация, когда модель идеально «запомнила» соотношения, представленные в обучающей выборке, но не зависимости в данных. В этом случае качество будет высоким на обучающей выборке, но низким на новых данных.

Обобщающая способность

- В случаях недо- и переобучения обобщающая способность модели низкая.
- Пример: синий – истинная зависимость, оранжевый – оценённая зависимость, фиолетовый – выборка.



- a) – недообучение, b) – переобучение, c) – корректно обученная модель.