Основы машинного обучения

Преподаватель – аспирант института физико – математических наук и информационных технологий направления Информатика и вычислительная техника(Computer Science)

Сотрудник лаборатории “Интеллектуальная робототехника”

Ширкин Александр Евгеньевич

Рекомендации к курсу

Machine learning это –

1. 35 % линейная алгебра
2. 25 % математическая статистика и теория вероятностей
3. 15 % математический анализ
4. 15 % алгоритмы
5. 10 % подготовка данных

Рекомендую параллельно повторять математику

https://proglib.io/p/ml-3months/ Машинное обучение за 3 месяца

Обзор курса

- 8 лекций
- Основные алгоритмы и их использование
- Домашние задания и практики
- Соревнования – (но это не точно)
- Индивидуальные проекты

Особенности курса

- Обилие практики. Задания на каждом занятии и после него
- Теоретическое понимание алгоритмов
- Знакомство с соревнованиями по анализу данных
- Собственный проект

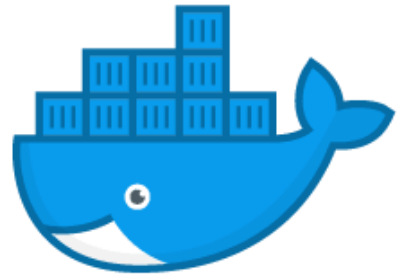
Логистика

- Все общение – в Slack Kant FU Data Science
- За домашние задания max 10 баллов
- За проекты, соревнования – max 40 баллов
- Текущий рейтинг – ‘ссылка появится позже’
- Все материалы курса будут опубликовываться на GitHub по ходу прохождения курса

Инструменты

- Язык Python
- Jupyter notebooks
- GitHub
- Docker (опц)

Notebook



docker

Занятие 1

- Знакомство с Python
- Анализ данных Pandas
- Практика на знакомство с данными

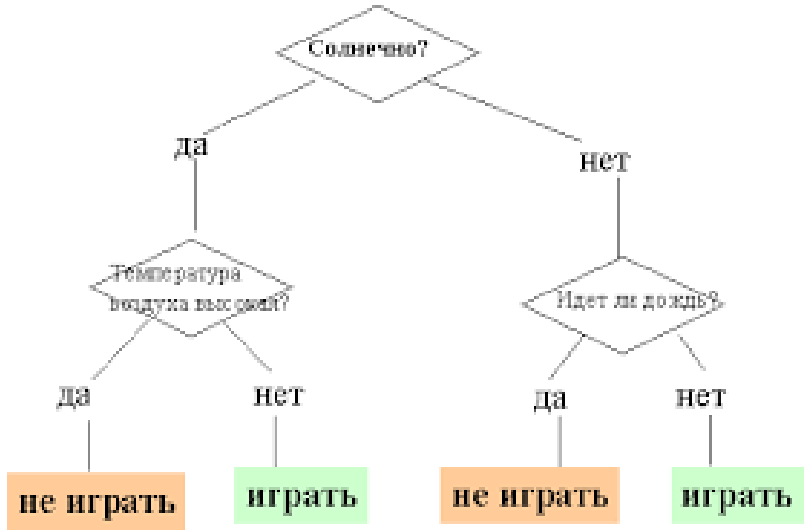
In [5]:
```
complete_excel = pd.read_excel('gapminder.xlsx')
complete_excel.head(5)
```

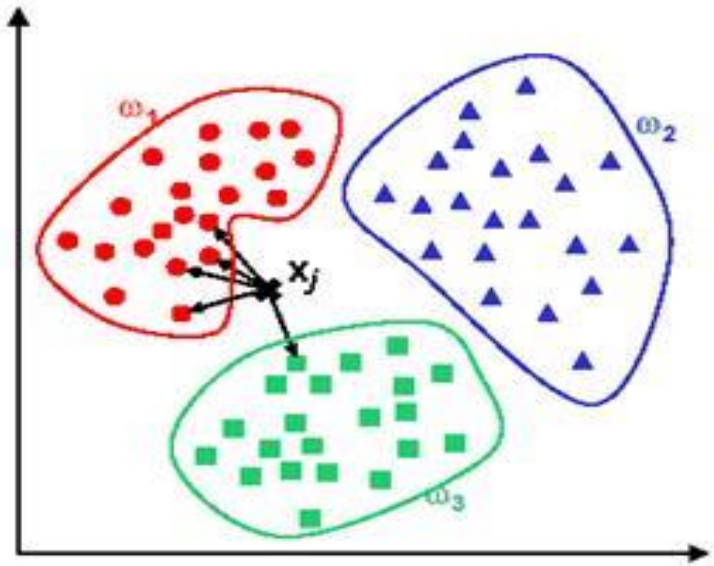
Out[5]:

	gdp pc 2011 ppp	1800	1801	1802	1803
0	Afghanistan	634.400014	634.400014	634.400014	634.400014
1	Albania	793.136557	793.960291	794.784880	795.610326
2	Algeria	1520.025973	1519.988511	1519.951050	1519.913589
3	Angola	650.000000	NaN	NaN	NaN
4	Antigua and Barbuda	771.878735	771.878735	771.878735	771.878735

Занятие 2

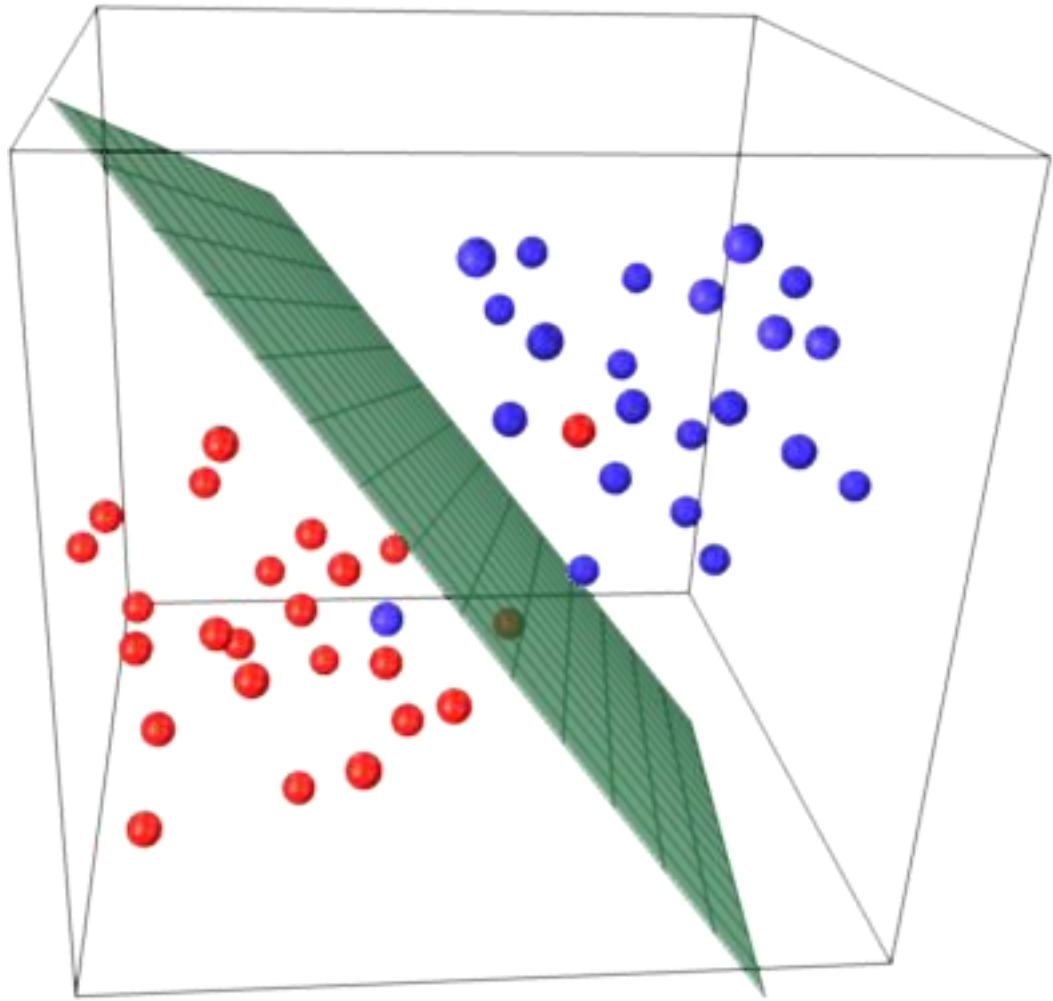
- Деревья решений
- Метод ближайших соседей
- Практика на знакомство с библиотекой Scikit-learn





Занятие 4

- Линейные модели классификации
- Регуляризация
- Кросс-валидация
- Практика на логистическую регрессию



Занятие 5

- Композиции алгоритмов, случайный лес
- Практика на применение случайного леса и оценке важности признаков

Занятие 6

- Обучение без учителя
- Principal Component Analysis
- Кластеризация
- Практика на кластеризацию данных

Занятие 7

- Современный взгляд на бустинг
- Теоретические основы
- Лучшие на сегодня реализации

Занятие 8 – экзамен?

- Защиты проектов
- Теоретический экзамен – вы будете собеседоваться на работу, никаких билетов, только хардкор.
- Практика –
  набрать нужное количество баллов за все активности курса.
  Те кто попадут в 5 % лучших (среди двух групп) получают автомат на экзамене.
  Необходимый порог, который нужно пройти по баллам, для зачета, будет уточнен позже.

Семинары

- До и после каждого занятия будут проводиться семинары
- Семинар заключается в решении различных задач в Jupyter Notebook

Индивидуальный проект

- В течение всего курса
- Четкий план
- Лучше свои данные
- Отличный опыт

Чихуахуа или маффин?

Что такое машинное обучение

- Анализ данных или машинное обучение – наука, изучающая способы извлечения закономерностей из ограниченного числа признаков

Пример - ресторан

- Пространство объектов - Множество всех возможных точек размещения ресторана (X)
- Прибыль ресторана в течение одного года (Целевая переменная, Y) – Множество ее значений – пространство ответов Y
- Y - множество вещественных чисел R
- Каждый ранее открытый ресторан - обучающий пример, а их множество – обучающая выборка
- $X = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ – обучающая выборка, где x_i - обучающий пример, y_i – ответ, n – количество обучающих примеров.

Признаки

- Объекты – абстрактные сущности, которые нужно формально описать с помощью некоторого набора характеристик называемых признаками
- Вектор всех признаков объекта x называется признаковым описанием объекта.
- Признаки бывают разные: бинарные, вещественные, категориальные, …
- Признаки с внутренней структурой - фотографии

Feature engineering

- В нашей задаче полезными могут оказаться признаки, связанные с демографией (
- средний возраст жителей ближайших кварталов,
- динамика изменения количества жителей
- недвижимостью (например, средняя стоимость квадратного метра в окрестности, количество школ, магазинов, заправок, торговых центров, банков поблизости).
- Разработка признаков (feature engineering) для любой задачи является одним из самых сложных и самых важных этапов анализа данных.

Обучение с учителем

- Описанная задача является примером задачи обучения с учителем (supervised learning), а более конкретно задачей регрессии — именно так называются задачи с вещественной целевой переменной.

Виды задач машинного обучения

- 1. Y = {0, 1} — бинарная классификация. Например, мы можем предсказывать, кликнет ли пользователь по рекламному объявлению, вернет ли клиент кредит в установленный срок, сдаст ли студент сессию, случится ли определенное заболевание с пациентом (на основе, скажем, его генома).
- 2. Y = {1, . . . , K} — многоклассовая (multi-class) классификация. Примером может служить определение предметной области для научной статьи (математика, биология, психология и т.д.).
- 3. Y = {0, 1} K — многоклассовая классификация с пересекающимися классами (multi-label classification). Примером может служить задача автоматического проставления тегов для ресторанов (логично, что ресторан может одновременно иметь несколько тегов).
- 4. Частичное обучение (semi-supervised learning) — задача, в которой для одной части объектов обучающей выборки известны и признаки, и ответы, а для другой только признаки. Такие ситуации возникают, например, в медицинских задачах, где получение ответа является крайне сложным (например, требует проведения дорогостоящего анализа).

Обучение без учителя

- 1. Кластеризация — задача разделения объектов на группы, обладающие некоторыми свойствами. Примером может служить кластеризация документов из электронной библиотеки или кластеризация абонентов мобильного оператора.
- 2. Оценивание плотности — задача приближения распределения объектов. Примером может служить задача обнаружения аномалий, в которой на этапе обучения известны лишь примеры «правильного» поведения оборудования (или, скажем, игроков на бирже), а в дальнейшем требуется обнаруживать случаи некорректной работы (соответственно, незаконного поведения игроков).
- 3. Визуализация — задача изображения многомерных объектов в двумерном или трехмерном пространстве таким образом, что сохранялось как можно больше зависимостей и отношений между ними.
- 4. Понижение размерности — задача генерации таких новых признаков, что их меньше, чем исходных, но при этом с их помощью задача решается не хуже (или с небольшими потерями качества, или лучше — зависит от постановки).

И снова ресторан

- Предположим, что мы собрали обучающую выборку и изобрели некоторое количество признаков. Результатом будет матрица «объекты-признаки»
- $X \in R^{l*d}$ (ℓ — число объектов, d — число признаков), в которой каждая строка содержит признаковое описание одного из обучающих объектов.
- строки в матрице объекты,
- столбцы — признаки.

Функционал качества

- $a : X \Rightarrow Y$
- Такая функция называется алгоритмом или моделью
- Чтобы оценить правильность нашего алгоритма, нужно ввести функционал качества
- $Q(a, x) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2$ - MSE – Средне квадратичная ошибка
- Функционал качества – это как правило сумма ошибок каждого примера
- Функция потерь - $(y - z)^2$ – ошибка для конкретного примера

- Заметим, что именно функционал качества будет определять во всех дальнейших рассуждениях, какой алгоритм является лучшим. Если метрика выбрана неудачно и не соответствует бизнес-требованиям или особенностям данных, то все дальнейшие действия обречены на провал. Именно поэтому выбор базовой метрики является крайне важным этапом в решении любой задачи анализа данных. Она не обязательно должна обладать хорошими математическими свойствами (непрерывность, выпуклость, дифференцируемость и т.д.), но обязана отражать все важные требования к решению задачи.

Самый изученный алгоритм

- линейная модель
- $\frac{1}{l}\sum_{i=1}^{l}\left(w_0 + \sum_{j=1}^{d}(w_j x_{ij}) - y_i\right)^2 \to min$