Введение в прикладной анализ данных

Анвар Курмуков, 2022

Анализ данных и машинное обучение

1. Задача
2. Данные
3. Решение

Анализ данных и машинное обучение

1. Задача
2. Данные
3. **Решение**

Анализ данных и машинное обучение

1. Задача
2. Данные
3. **Решение**
    a. Настройка инфраструктуры
    b. Прототип (алгоритм)
    c. Поддержка и развитие

Анализ данных и машинное обучение

1. Задача
2. Данные
3. **Решение**
    a. Настройка инфраструктуры
    b. **Прототип (алгоритм)**
    c. Поддержка и развитие

Алгоритмы машинного обучения

Классическая таксономия методов ML

https://vas3k.ru/blog/machine_learning/

Алгоритмы машинного обучения

Классическая таксономия методов ML

https://vas3k.ru/blog/machine_learning/

Обучение “с учителем”

$D = \{(x_1, y_1),\ldots,(x_n, y_n)\}$ - набор данных

$x_i \in R^m$ - наблюдение

$y_i \in R^1$ - задача регрессии

или

$y_i \in \{1,\ldots,c\}$ - задача классификации

Цель: построить $f(x, W)$: $x \rightarrow y$

Пример: выдача кредита

Имя	Образование	…	Выдать кредит
Маруся	2	…	0
Алиса	2	…	1
Олег	1	…	0

$X_{n \times m}$ Y_n

Пример: часть речи

token	Embedding, 512	Часть речи
Однажды	…	нар.
в студеную	…	прил.
зимнюю	…	прил.
пору	…	сущ.
я	…	мест.

$X_{n \times m}$ Y_n

Пример: стоимость недвижимости

ID	Площадь	…	Стоимость
1	120	…	570
2	60	…	320
3	95	…	415
4	80	…	440

$X_{n \times m}$ Y_n

Обучение “с учителем”

Цель: “обучить” функцию (алгоритм ML)

$f(x, W): x \rightarrow y$

такую что:

$\text{Error}(f(x, W, \Theta), y) \rightarrow \min_W$

x и **y** у нас фиксированы, поэтому свобода выбора у нас ограничена

Обучение “с учителем”

$$\text{Error}(f(x, W, \Theta), y) \rightarrow \min_W$$

Error - функция ошибки, например среднеквадратичное отклонение (MSE);

f - модель машинного обучения (ML), например *Линейная регрессия*, *Метод ближайших соседей;*

W - параметры модели ML, подбираются в процессе обучения “автоматически”;

Θ - гипер-параметры модели ML, задаются вручную.

Обучение “с учителем”

Выбор зависит от задачи

$$\text{Error}(f(x, W, \Theta), y) \rightarrow \min_{W}$$

Error - функция ошибки, например среднеквадратичное отклонение (MSE);

f - модель машинного обучения (ML), например *Линейная регрессия*, *Метод ближайших соседей;*

W - параметры модели ML, подбираются в процессе обучения “автоматически”;

Θ - гипер-параметры модели ML, задаются вручную.

Обучение “с учителем”

Выбор зависит от задачи

$$\text{Error}(f(x, W, \Theta), y) \rightarrow \min_W$$

Error - функция ошибки, например среднеквадратичное отклонение (MSE);

f - модель машинного обучения (ML), например *Линейная регрессия*, *Метод ближайших соседей;*

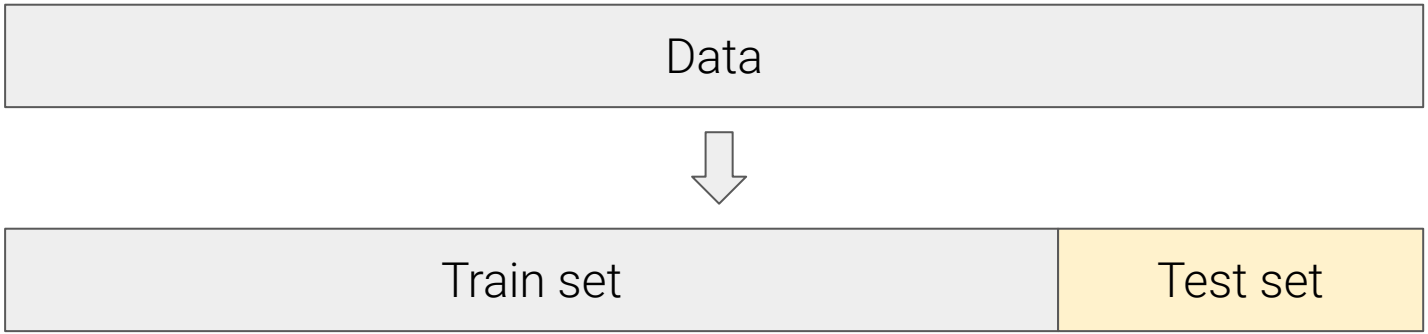
W - параметры модели ML, подбираются в процессе обучения “автоматически”;

Θ - гипер-параметры модели ML, задаются вручную.

Отличаются для разных алгоритмов

Обучение “с учителем”: переобучение

Мы хотим получить такую **f(x)** которая на новых данных сделает “хорошее” предсказание, поэтому во время обучения давайте эмулировать ситуацию “новых данных” и разделять данные на тестовую часть и тренировочную.



Контакты

1. Ссылка на гитхаб: https://github.com/kurmukovai/iitp-ml-ds
2. Почты: kurmukovai@gmail.com (Анвар); rilshok@pm.me (Влад)

Материалы

1. Учебники по ML:
    a. https://ml-handbook.ru Учебник по ML от ШАДа (на русском)
    b. ESL: The elements of statistical learning https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf настольная книга по классическому ML (на английском)
2. Другие курсы по ML (на английском)
    a. Курс от Killian Weinberg из университета Cornell https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/
    b. Курс от Andrew Ng из университета Stanford http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599
3. Материалы по Python
    a. https://youtu.be/fgf57Sa5A-A?list=PLRDzFCPr95fLuusPXwvOPgXzBL3ZTzybY (на русском)
    b. https://www.youtube.com/c/Coreyms/videos (на английском)