

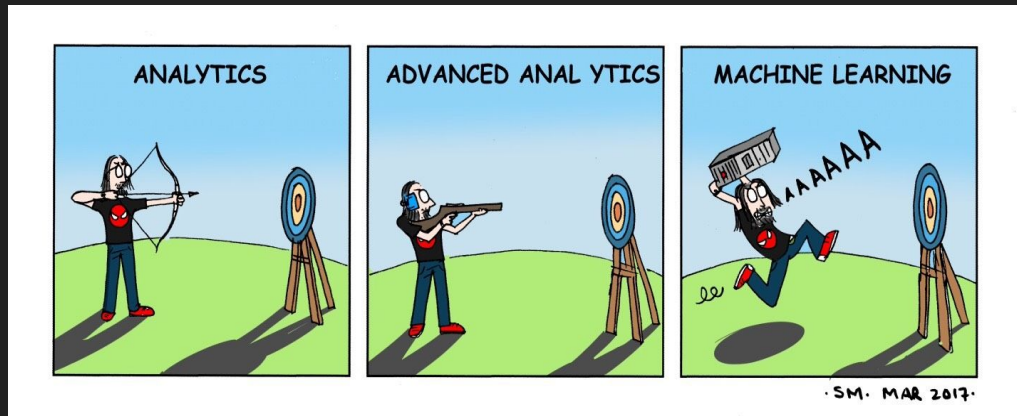
Задачи в ML.

Данные: кто виноват и что делать?



Задачи в ML

- Классификация
- Регрессия
- Кластеризация
- Детекция
- Сегментация
- Перевод голоса в текст и наоборот
- Генерация текста и изображений
- Распознавание текста с изображений
- По сути всё, что вы придумаете



Красным выделены задачи, которые мы будем рассматривать в этом курсе

Метрики vs Loss функции (функции потерь)

Метрика:

- То что имеет смысл в реальном мире

Loss функция:

- Может быть довольно абстрактна

Метрики vs Loss функции (функции потерь)

Метрика:

- То что имеет смысл в реальном мире
- Может быть не дифференцируема, а может и быть => мы не всегда можем её оптимизировать на прямую

Loss функция:

- Может быть довольно абстрактна
- **ОБЯЗАТЕЛЬНО** дифференцируема => мы подбираем лучшую функцию, которая будет помогать нам оптимизировать нашу модель под текущую задачу

Примеры метрик и loss функций

Метрика:

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

Loss функция:

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

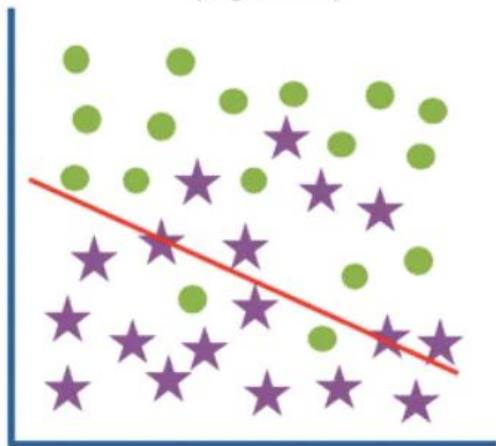
$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Cross Entropy Loss:

$$L(\Theta) = - \sum_{i=1}^k y_i \log(\hat{y}_i)$$

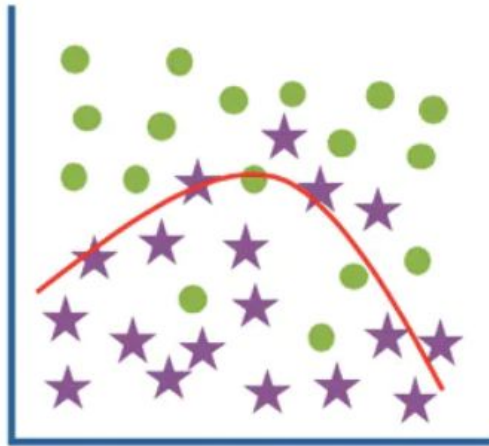
Overfitting and underfitting

Underfit
(high bias)



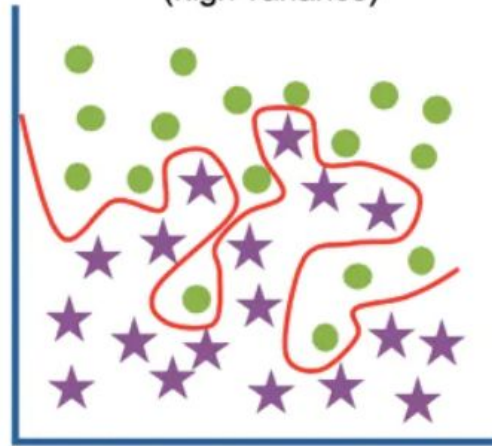
High training error
High test error

Optimum



Low training error
Low test error

Overfit
(high variance)



Low training error
High test error

Данные

Какие могут быть:

- Табличные
- Изображения
- Регрессионные значения
- Текст
- Аудио
- Видео



Что с ними делать?

Простой, но рабочий алгоритм:

1. Смотрим на задачу, которую нужно решить
2. Смотрим на данные, которые у нас есть
3. На пальцах придумываем наш алгоритм решения (мы не думаем о моделях, loss функциях и тд). Задумываемся о критерии успеха (метрика)
4. Думаем о средствах решения (вот здесь перебираем модели и тд)
5. Составляем список рабочих (на первый взгляд) подходов
6. Тестим их
7. Мы молодцы

Некоторые инструменты для работы с данными

- Pandas (табличные)
- Torchvision, alumentations (изображений)
- Деревья (неожиданно, но мы можем выкидывать некоторые фишки, которые не несут какой-то ценности. Условно брать топ n)
- Это не все, но достаточно для нашего курса



Борьба с переобучением за счет данных

1. Validation (blending)
2. Cross-validation
3. Добавить разнообразие в данные
4. Выкинуть выбросы или ненужные фичи

Validation VS Cross validation

