

Билет 13

1. Качество классификации.

Доля правильно классифицированных объектов (accuracy)

Accuracy — самая простая оценка классификации:

$$Accuracy(a) = \frac{\sum_{i=1}^N \mathbb{I}[a(x_i) = y_i]}{N} = \frac{TP + TN}{TP + TN + FP + FN}$$

По сути это вероятность того, что класс будет предсказан правильно.

Например, если мы ловим сумасшедших, accuracy показывает долю правильных диагнозов.

- Работает для многоклассовой классификации.
- Плохо работает при высокой априорной вероятности у одного из классов. В таком случае константное предсказание может давать высокое значение **accuracy** (равное этой априорной вероятности).

Точность (Precision)

$$Precision(a) = \frac{TP}{TP + FP}$$

Точность показывает какую долю объектов, **распознанных** как объекты положительного класса, мы предсказали верно.

На примере: точность — это сколько из пойманных нами и посаженных в психушку людей реально сумасшедшие.

- Только бинарная классификация.
- Не зависит от априорной вероятности положительного класса.

Полнота (Recall)

$$Recall(a) = \frac{TP}{TP + FN}$$

Полнота показывает, какую долю объектов, **реально** относящихся к положительному классу, мы предсказали верно.

На примере: полнота — это сколько из сумасшедших людей, которых мы проверили, мы посадили в психушку.

- Только бинарная классификация

- Не зависит от априорной вероятности положительного класса.

F-мера

Точность и полнота хорошо оценивают качество классификатора для задач со смещенной априорной вероятностью, но если мы обучили модель с высокой точностью, то может случиться так, что полнота у такого классификатора низкая и наоборот. Чтобы связать точность с полнотой вводят F-меру как среднее гармоническое точности и полноты:

$$F_{measure} = \frac{2Precision \cdot Recall}{Precision + Recall}$$

В некоторых задачах одна метрика важнее другой (например при выдаче поисковых запросов полнота важнее точности, неинтересные страницы мы можем сами пропустить, а вот если поисковик пропустит несколько страниц то мы можем остаться без каких то важных деталей). Для установления важности конкретной метрики мы рассматриваем параметрическую F-меру:

$$F_{measure}_{\beta} = \frac{(1 + \beta^2)Precision \cdot Recall}{\beta^2 Precision + Recall}$$

Где $\beta \in [0, \infty)$ при $\beta = 0$ получаем точность, при $\beta = 1$ — непараметрическую F-меру, при $\beta = \infty$ — полноту.

2. Правильно классифицированные алгоритмом объекты. Неправильно классифицированные алгоритмом объекты.

Общая правильность модели — это количество правильных предсказаний, деленное на общее количество предсказаний. Оценка правильности дает значение от 0 до 1, где 1 — идеальная модель.

Правильность = количество правильных прогнозов / общее количество прогнозов

Этот показатель редко следует использовать отдельно, поскольку на материале несбалансированных данных, где один класс значительно превосходит другой, правильность может быть очень обманчивой.

Вернемся к примеру с раком. Представьте, что у вас есть набор данных, в котором только 1% образцов являются раковыми. Классификатор, который предскажет все результаты как доброкачественные, достигнет 99% правильности. Однако на практике такая модель была бы не просто бесполезной, но и опасной, поскольку она никогда не обнаружила бы раковые образцы.

Практика

```
class DataAnalysis:
    ...
```

Дана непустая последовательность целых чисел, оканчивающаяся нулем.

Найти:

- а) сумму всех чисел последовательности;
- б) количество всех чисел последовательности.

Решить задачу используя циклическую конструкцию `while`.

...

```
def __init__(self):
    self._array = []
    self._inputArray()

def _inputArray(self):
    number = None
    while number != 0:
        number = int(input('-> '))
        self._array.append(number)

@property
def sum(self):
    return sum(self._array)

@property
def len(self):
    return len(self._array)

array = DataAnalysis()
print(f'Сумма: {array.sum}, длина {array.len}')
```