**Семинар 4. Метрики качества классификации**

Для начала подгружаем библиотеки:

**from** sklearn.model\_selection **import** train\_test\_split # функция, чтобы разбить данные на трейн и тест

**from** sklearn.linear\_model **import** LogisticRegression # наша модель для классификации

Воспользуемся встроенным датасетом, который содержит информацию об опухолях груди:

**from** sklearn.datasets **import** load\_breast\_cancer # подгружаем датасет

breast\_cancer = load\_breast\_cancer()

Теперь зададим зависимую и независимые переменные:

Y = breast\_cancer.target ## Наша целевая переменная, 0 — если рака нет, 1 — если есть

X = breast\_cancer.data # X - признаки, по которым мы будем предсказывать рак

Разбиваем выборку на обучающую и тестовую и обучаем нашу модель:

X\_train, X\_val, Y\_train, Y\_val = train\_test\_split(X, Y, test\_size = 0.3)

model = LogisticRegression()

model.fit(X\_train, Y\_train)

Готово! Теперь осталось только вычислить необходимые метрики:

**from** sklearn.metrics **import** accuracy\_score, precision\_score, recall\_score, f1\_score

Y\_predicted = model.predict(X\_val)

print(accuracy\_score(Y\_val,Y\_predicted))

print(precision\_score(Y\_val,Y\_predicted))

print(recall\_score(Y\_val,Y\_predicted))

print(f1\_score(Y\_val,Y\_predicted))

**Задание 3B2.2**

Загрузите встроенный в библиотеку *sklearn* датасет про ирисы с помощью функции load\_iris. Обучите модель логистической регрессии (random\_state=50, размер тестовой выборки 0.3) и укажите полученное значение метрики Accuracy.

**Задание 3B2.3**

У вас есть датасет с параметрами мобильных телефонов. Переменная price\_range отвечает за то, к какой категории относится телефон: 1 — дорогие, 0 — дешевые.

Ваша задача состоит в том, чтобы наиболее точно научиться классифицировать телефоны по этим двум категориям на основании других параметров.

## **Шаг 1**

Для начала нам надо отобрать признаки, с помощью которых мы будем предсказывать категорию телефона.

**Задание 3B.3.1 Отбор признаков**

Выберите пять признаков, у которых наибольшая взаимосвязь с целевой переменной (с помощью корреляции).

## **ШАГ 2**

Теперь необходимо обучить алгоритм.  Для начала разбейте выборку на тестовую и обучающую, размер тестовой задайте 0.2. Параметр random\_state=31. В качестве модели возьмите логистическую регрессию. В качестве предикторов возьмите пять ранее отобранных признаков.

Рассчитайте метрику, которая покажет, какая доля телефонов, обозначенных классификатором как дорогие, действительно относится к этой категории.

Что это за метрика?

**Задание 3B2.4**

**Изображение выглядит как текст, Шрифт, линия, снимок экрана

Автоматически созданное описание**

**Если имеется несколько классов то logloss вычисляется так:**

**Изображение выглядит как текст, снимок экрана, Шрифт, алгебра

Автоматически созданное описание**

Посчитайте  logloss для данных в таблице (без нормализации). Укажите число с точностью до сотых:

Изображение выглядит как текст, снимок экрана, линия, Шрифт

Автоматически созданное описание

Посчитайте  logloss для данных в таблице. Необходимо найти среднюю ошибку. Классификация на три класса:

Изображение выглядит как текст, снимок экрана, линия, число

Автоматически созданное описание