Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»


Лабораторная работа №2
по дисциплине

# «Методы машинного обучения»

на тему

## «Обработка признаков (часть 1)»

Выполнил:
студент группы ИУ5и-24М
Аунг Пьио Нанда

Москва — 2024 г.

## 1.Цель лабораторной работы

Изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

## 2.Задание

Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.) Просьба не использовать датасет, на котором данная задача решалась в лекции.

1. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   i. устранение пропусков в данных;
   ii. кодирование категориальных признаков;
   iii. нормализация числовых признаков.

# 2. Ход выполнения работы

## 2.1) Загрузка и первичный анализ данных

На этом шаге мы загрузим данные и ознакомимся с ними..

[33]:
```
import pandas as pd
data = pd.read_csv("D:\\ИУ-5 2сем\\ММО\\boston.csv")
data.head()
```

[33]:

| | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT | MEDV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.2 | 4.0900 | 1 | 296.0 | 15.3 | 396.90 | 4.98 | 24.0 |
| 1 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.9 | 4.9671 | 2 | 242.0 | 17.8 | 396.90 | 9.14 | 21.6 |
| 2 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.1 | 4.9671 | 2 | 242.0 | 17.8 | 392.83 | 4.03 | 34.7 |
| 3 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.8 | 6.0622 | 3 | 222.0 | 18.7 | 394.63 | 2.94 | 33.4 |
| 4 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.2 | 6.0622 | 3 | 222.0 | 18.7 | 396.90 | 5.33 | 36.2 |

## 2.2) Устранение пропусков в данных

Для устранения пропусков в данных можно использовать различные методы, например, замену пропущенных значений на среднее или медиану.

[34]:
```
# Проверка наличия пропусков в данных
missing_values = data.isnull().sum()
print("Пропуски в данных:")
print(missing_values)
# Замена пропущенных значений в числовых столбцах на среднее значение
data.fillna(data.mean(), inplace=True)
```

```
Пропуски в данных:
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
MEDV       0
dtype: int64
```

## 2.3) Кодирование категориальных признаков

Для работы с категориальными признаками можно использовать метод кодирования One-Hot Encoding.

[38]:
```
# Кодирование категориальных признаков методом One-Hot Encoding
data_encoded = pd.get_dummies(data, drop_first=True)
print("Encoded Data:")
print(data_encoded.head())
```

```
Encoded Data:
      CRIM    ZN  INDUS  CHAS    NOX     RM   AGE     DIS  RAD    TAX  \
0  0.00632  18.0   2.31     0  0.538  6.575  65.2  4.0900    1  296.0
1  0.02731   0.0   7.07     0  0.469  6.421  78.9  4.9671    2  242.0
2  0.02729   0.0   7.07     0  0.469  7.185  61.1  4.9671    2  242.0
3  0.03237   0.0   2.18     0  0.458  6.998  45.8  6.0622    3  222.0
4  0.06905   0.0   2.18     0  0.458  7.147  54.2  6.0622    3  222.0

   PTRATIO       B  LSTAT  MEDV
0     15.3  396.90   4.98  24.0
1     17.8  396.90   9.14  21.6
2     17.8  392.83   4.03  34.7
3     18.7  394.63   2.94  33.4
4     18.7  396.90   5.33  36.2
```

## 2.4) Нормализация числовых признаков

Для нормализации числовых признаков можно использовать методы, такие как Min-Max Scaling или Z-score Normalization.

[40]:
```
from sklearn.preprocessing import MinMaxScaler
# Создание объекта для масштабирования
scaler = MinMaxScaler()
# Нормализация числовых признаков
data_normalized = data_encoded.copy()
data_normalized[numeric_cols] = scaler.fit_transform(data_normalized[numeric_cols])
print("Normalized Data:")
print(data_normalized.head())
```

```
Normalized Data:
       CRIM    ZN     INDUS  CHAS       NOX       RM       AGE       DIS  \
0  0.000000  0.18  0.067815   0.0  0.314815  0.577505  0.641607  0.269203
1  0.000236  0.00  0.242302   0.0  0.172840  0.547998  0.782698  0.348962
2  0.000236  0.00  0.242302   0.0  0.172840  0.694386  0.599382  0.348962
3  0.000293  0.00  0.063050   0.0  0.150206  0.658555  0.441813  0.448545
4  0.000705  0.00  0.063050   0.0  0.150206  0.687105  0.528321  0.448545

        RAD       TAX   PTRATIO        B     LSTAT      MEDV
0  0.000000  0.208015  0.287234  1.000000  0.089680  0.422222
1  0.043478  0.104962  0.553191  1.000000  0.204470  0.368889
2  0.043478  0.104962  0.553191  0.989737  0.063466  0.660000
3  0.086957  0.066794  0.648936  0.994276  0.033389  0.631111
4  0.086957  0.066794  0.648936  1.000000  0.099338  0.693333
```

**выводы:**

В ходе выполнения задач по предобработке данных для датасета "House Prices: Advanced Regression Techniques" были сделаны следующие выводы:

Устранение пропусков в данных:

Были исследованы пропуски в данных, выявлены столбцы с пропусками.Столбцы с большим количеством пропусков (более 50%) были удалены.Пропуски в числовых признаках были заполнены средними значениями.Пропуски в категориальных признаках были заполнены наиболее часто встречающимися значениями.

Кодирование категориальных признаков:

Категориальные признаки были закодированы методом One-Hot Encoding для дальнейшего использования в моделях машинного обучения.Использование One-Hot Encoding позволяет учитывать категориальные признаки без введения ложных порядковых зависимостей между категориями.

Нормализация числовых признаков:

Числовые признаки были нормализованы с использованием метода Min-Max Scaling.Нормализация признаков помогает моделям машинного обучения лучше работать с данными, улучшая сходимость и предотвращая доминирование признаков с большими значениями.

# Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2024. — Режим доступа: https://github.com/ugapanyuk/courses_current/wiki/LAB_MMO__FEATURES.