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 ИУ5-22М
Лун Сыхань

Москва — 2024г.

1. Цель лабораторной работы

Изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

2. Задание

1. 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.

2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

1). устранение пропусков в данных;

2). кодирование категориальных признаков;

3). нормализация числовых признаков.

3. Ход выполнения работы

```
!pip install scikit-learn
import pandas as pd
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
```

```
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.10/dist-packages (1.2.2)
Requirement already satisfied: numpy>=1.17.3 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.25.2)
Requirement already satisfied: scipy>=1.3.2 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.11.4)
Requirement already satisfied: joblib>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn) (3.5.0)
```

```
# 加载数据集
url = 'https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv'
df = pd.read_csv(url)
df.head()
```

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250
3	1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35.0	1	0	53.1000
4	0	3	Mr. William Henry Allen	male	35.0	0	0	8.0500

устранение пропусков в данных:

```
# 查看缺失值情况
df.isnull().sum()

# 使用中位数填充数值型特征的缺失值
num_features = df.select_dtypes(include=['int64', 'float64']).columns
imputer = SimpleImputer(strategy='median')
df[num_features] = imputer.fit_transform(df[num_features])

# 使用最频繁值填充类别型特征的缺失值
cat_features = df.select_dtypes(include=['object']).columns
imputer = SimpleImputer(strategy='most_frequent')
df[cat_features] = imputer.fit_transform(df[cat_features])

# 查看处理后的缺失值情况
df.isnull().sum()
```

```
Survived                  0
Pclass                    0
Name                      0
Sex                       0
Age                       0
Siblings/Spouses Aboard   0
Parents/Children Aboard   0
Fare                      0
dtype: int64
```

кодирование категориальных признаков:

```
# 使用独热编码处理类别型特征
encoder = OneHotEncoder(sparse=False, drop='first')  # drop='first'避免虚拟变量陷阱
encoded_features = pd.DataFrame(encoder.fit_transform(df[cat_features]), columns=encoder.get_feature_names_out(cat_features))

# 删除原来的类别型特征，并将编码后的特征加入数据集中
df = df.drop(cat_features, axis=1)
df = pd.concat([df, encoded_features], axis=1)

# 查看处理后的数据集
df.head()
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/preprocessing/_encoders.py:868: FutureWarning: `sparse` was renamed to `sparse_output` in version 1.2
  warnings.warn(
```

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	Name_Col. John Weir	Name_Col. Oberst Alfons Simonius-Blumer	Name_Don. Manuel E Uruchurtu	Name_Dr. Alfred Pain	...	Name_Ms. Encarnacion Reynaldo	Nam Cl L Ki
0	0.0	3.0	22.0	1.0	0.0	7.2500	0.0	0.0	0.0	0.0	...	0.0	
1	1.0	1.0	38.0	1.0	0.0	71.2833	0.0	0.0	0.0	0.0	...	0.0	
2	1.0	3.0	26.0	0.0	0.0	7.9250	0.0	0.0	0.0	0.0	...	0.0	
3	1.0	1.0	35.0	1.0	0.0	53.1000	0.0	0.0	0.0	0.0	...	0.0	
4	0.0	3.0	35.0	0.0	0.0	8.0500	0.0	0.0	0.0	0.0	...	0.0	

5 rows × 893 columns

нормализация числовых признаков:

```
# 对数值型特征进行标准化处理
scaler = StandardScaler()
df[num_features] = scaler.fit_transform(df[num_features])

# 查看处理后的数据集
df.head()
```

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	Name_Col. John Weir	Name_Col. Oberst Alfons Simonius-Blumer	Name_Don. Manuel E Uruchurtu	Name_Dr. Alfred Pain	...	Name_M Encarnaci Reynal
0	-0.792163	0.830524	-0.529366	0.429904	-0.474981	-0.503586	0.0	0.0	0.0	0.0	...	
1	1.262366	-1.561277	0.604265	0.429904	-0.474981	0.783412	0.0	0.0	0.0	0.0	...	
2	1.262366	0.830524	-0.245958	-0.475856	-0.474981	-0.490020	0.0	0.0	0.0	0.0	...	
3	1.262366	-1.561277	0.391709	0.429904	-0.474981	0.417948	0.0	0.0	0.0	0.0	...	
4	-0.792163	0.830524	0.391709	-0.475856	-0.474981	-0.487507	0.0	0.0	0.0	0.0	...	

5 rows × 893 columns