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-23М
Ван Тяньшо

Москва — 2024 г.

1. Цель лабораторной работы

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.

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

устранение пропусков в данных;

кодирование категориальных признаков;

нормализация числовых признаков.

Выбранный мной набор данных: набор данных Titanic

3. Текст программы

```
# 在Colab中运行此代码
# 首先，安装必要的库
!pip install pandas matplotlib seaborn scikit-learn

# 导入必要的库
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.impute import SimpleImputer

# 选择Titanic数据集
url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
data = pd.read_csv(url)

# 显示部分数据
print("原始数据集部分数据：")
print(data.head())

# 1. 处理数据中的缺失值
# 计算缺失值数量
missing_values = data.isnull().sum()
print("\n数据集中缺失值数量：")
print(missing_values)
```

```
# 使用均值填充数值型缺失值，使用众数填充类别型缺失值
numeric_features = data.select_dtypes(include=[np.number]).columns
categorical_features = data.select_dtypes(include=[object]).columns

imputer_numeric = SimpleImputer(strategy='mean')
imputer_categorical = SimpleImputer(strategy='most_frequent')

data[numeric_features] = imputer_numeric.fit_transform(data[numeric_features])
data[categorical_features] = imputer_categorical.fit_transform(data[categorical_features])

print("\n缺失值处理后的数据集：")
print(data.head())

# 2. 对类别型特征进行编码
encoder = OneHotEncoder(sparse=False, drop='first')
encoded_categorical = pd.DataFrame(encoder.fit_transform(data[categorical_features]), columns=encoder.get_feature_names_out(categorical_features))

# 合并编码后的数据
data = data.drop(categorical_features, axis=1)
data = pd.concat([data, encoded_categorical], axis=1)

print("\n编码后的数据集：")
print(data.head())
```

```
# 3. 对数值型特征进行归一化
scaler = StandardScaler()
scaled_numeric = pd.DataFrame(scaler.fit_transform(data[numeric_features]), columns=numeric_features)

# 合并归一化后的数据
data[numeric_features] = scaled_numeric

print("\n归一化后的数据集：")
print(data.head())

# 生成图片结果
# 绘制数值型特征的分布图
plt.figure(figsize=(12, 6))
sns.boxplot(data=data[numeric_features])
plt.title("归一化后的数值型特征分布")
plt.show()
```

4. экранные формы с примерами выполнения программы

4.1 Частичные данные из исходного набора данных：

```
原始数据集部分数据：
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

4.2 устранение пропусков в данных：

```
数据集中缺失值数量：
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64

缺失值处理后的数据集：
   PassengerId  Survived  Pclass  \
0          1.0       0.0     3.0
1          2.0       1.0     1.0
2          3.0       1.0     3.0
3          4.0       1.0     1.0
4          5.0       0.0     3.0

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0    1.0
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0    1.0
2                             Heikkinen, Miss. Laina  female  26.0    0.0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0    1.0
4                           Allen, Mr. William Henry    male  35.0    0.0

   Parch            Ticket     Fare    Cabin Embarked
0    0.0         A/5 21171   7.2500  B96 B98        S
1    0.0          PC 17599  71.2833      C85        C
2    0.0  STON/O2. 3101282   7.9250  B96 B98        S
3    0.0            113803  53.1000     C123        S
4    0.0            373450   8.0500  B96 B98        S
/usr/local/lib/python3.10/dist-packages/sklearn/preprocessing/_encoders.py:868: FutureWarning: `sparse` was renamed to `sparse_output` in version 1.2 and will be removed in 1.4. `sparse_output` is ignored unless you
  warnings.warn(
```

4.3 кодирование категориальных признаков：

编码后的数据集：

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	\
0	1.0	0.0	3.0	22.0	1.0	0.0	7.2500	
1	2.0	1.0	1.0	38.0	1.0	0.0	71.2833	
2	3.0	1.0	3.0	26.0	0.0	0.0	7.9250	
3	4.0	1.0	1.0	35.0	1.0	0.0	53.1000	
4	5.0	0.0	3.0	35.0	0.0	0.0	8.0500	

	Name_Abbott, Mr. Rossmore Edward	Name_Abbott, Mrs. Stanton (Rosa Hunt)	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	

	Name_Abelson, Mr. Samuel	...	Cabin_F G63	Cabin_F G73	Cabin_F2	\
0	0.0	...	0.0	0.0	0.0	
1	0.0	...	0.0	0.0	0.0	
2	0.0	...	0.0	0.0	0.0	
3	0.0	...	0.0	0.0	0.0	
4	0.0	...	0.0	0.0	0.0	

	Cabin_F33	Cabin_F38	Cabin_F4	Cabin_G6	Cabin_T	Embarked_Q	Embarked_S
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0

[5 rows x 1726 columns]

4.4 нормализация числовых признаков：

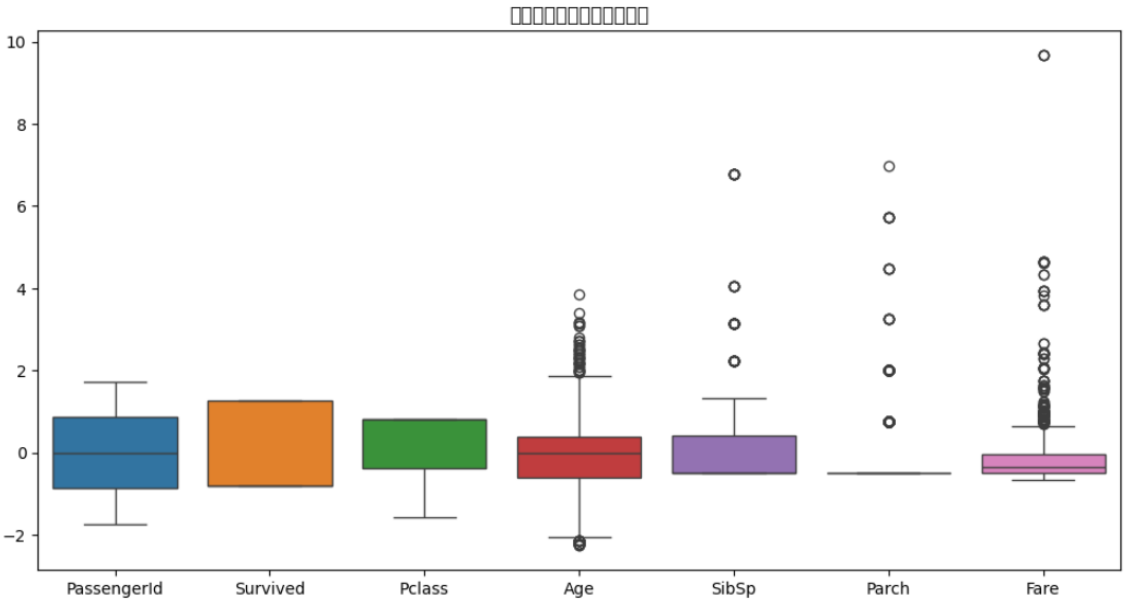
归一化后的数据集：

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	\
0	-1.730108	-0.789272	0.827377	-0.592481	0.432793	-0.473674	-0.502445	
1	-1.726220	1.266990	-1.566107	0.638789	0.432793	-0.473674	0.786845	
2	-1.722332	1.266990	0.827377	-0.284663	-0.474545	-0.473674	-0.488854	
3	-1.718444	1.266990	-1.566107	0.407926	0.432793	-0.473674	0.420730	
4	-1.714556	-0.789272	0.827377	0.407926	-0.474545	-0.473674	-0.486337	

	Name_Abbott, Mr. Rossmore Edward	Name_Abbott, Mrs. Stanton (Rosa Hunt)	\
0	0.0	0.0	
1	0.0	0.0	
2	0.0	0.0	
3	0.0	0.0	
4	0.0	0.0	

	Name_Abelson, Mr. Samuel	...	Cabin_F G63	Cabin_F G73	Cabin_F2	\
0	0.0	...	0.0	0.0	0.0	
1	0.0	...	0.0	0.0	0.0	
2	0.0	...	0.0	0.0	0.0	
3	0.0	...	0.0	0.0	0.0	
4	0.0	...	0.0	0.0	0.0	

	Cabin_F33	Cabin_F38	Cabin_F4	Cabin_G6	Cabin_T	Embarked_Q	Embarked_S
0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	0.0	0.0	0.0	1.0
4	0.0	0.0	0.0	0.0	0.0	0.0	1.0


□□□□□□□□□□□□□□□
10
8
6
4
2
0
−2
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare

Список литературы

[1] Гапанюк Ю. Е. Лабораторная работа «Разведочный анализ данных. Исследование и визуализация данных» [Электронный ресурс] // GitHub. — 2019. — Режим доступа: https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION (дата обращения: 13.02.2019)

[2] https://www.kaggle.com/datasets