Московский государственный технический университет им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

Задания для РК №1.

по дисциплине «Методы машинного обучения»

Выполнила:

студентка группы ИУ5И-23М

Цзян Юхуэй

Москва — 2024 г.

Номер варианта: 15 + 4 = 19

Номер задачи №1: 19

Номер задачи №2: 39

Задача №19

Для набора данных проведите масштабирование данных для одного (произвольного) числового признака с использованием метода "Mean Normalisation".

Задача №39

Для набора данных проведите процедуру отбора признаков (feature selection). Используйте класс SelectPercentile для 10% лучших признаков, и метод, основанный на взаимной информации.

Часть 1. Задача №19

В этом задании мы используем набор данных "Bike Sharing Dataset".

```
import pandas as pd

import matplotlib.pyplot as plt

from io import BytesIO

import requests

import zipfile

# 下载并解压数据集

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip"

response = requests.get(url)

zip_file = zipfile.ZipFile(BytesIO(response.content))

bike_data = pd.read_csv(zip_file.open('day.csv'))

# 选择数值特征 'temp' (温度)

feature = 'temp'

original_values = bike_data[feature]

# 计算均值、最大值和最小值

mean = original_values.mean()

max_val = original_values.max()

min_val = original_values.min()

# 应用 Mean Normalisation

normalized_values = (original_values - mean) / (max_val - min_val)

# 绘制原始值和缩放后的值的盒须图

plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)

plt.boxplot(original_values, vert=False)

plt.title(f'Boxplot of Original {feature}')

plt.xlabel('Value')

plt.subplot(1, 2, 2)

plt.boxplot(normalized_values, vert=False)
```

```
plt.title(f'Boxplot of Normalized {feature}')

plt.xlabel('Value')

plt.show()
```

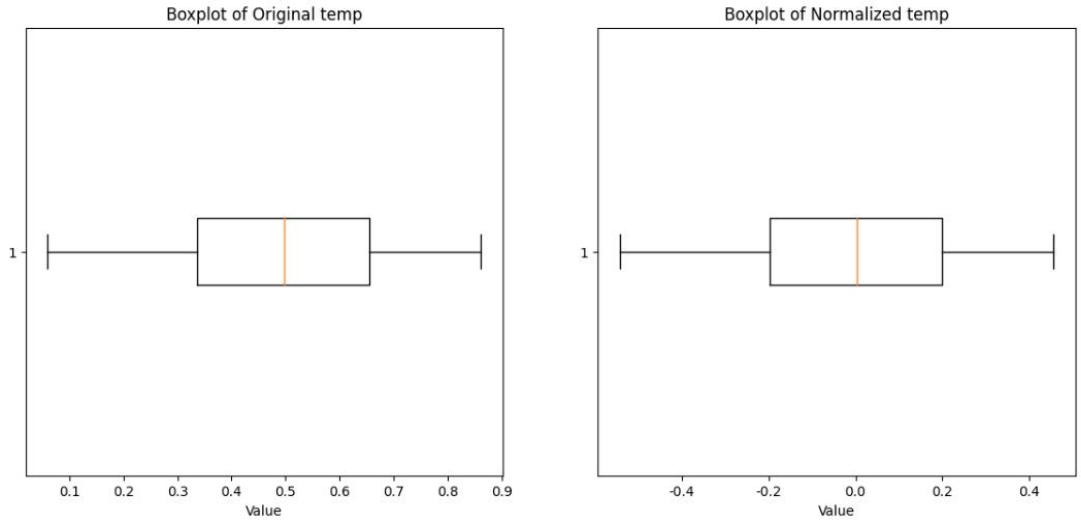

Рис 1. Боксплот (boxplot) исходного и нормализованного темпа.

Часть 2. Задача №39

```
from sklearn.feature_selection import SelectPercentile, mutual_info_regression

# 特征矩阵 X 和目标变量 y

X = bike_data.drop(columns=['cnt', 'casual', 'registered', 'dteday'])

y = bike_data['cnt']

# 使用 SelectPercentile 和互信息方法选择 10%的最佳特征

selector = SelectPercentile(mutual_info_regression, percentile=10)

X_selected = selector.fit_transform(X, y)

# 选择的特征

selected_features = X.columns[selector.get_support()]

# 显示选择的特征

print(f"Selected features:\n{selected_features}")

# 绘制选择的特征的盒须图

plt.figure(figsize=(14, 6))

for i, feature in enumerate(selected_features):

    plt.subplot(1, len(selected_features), i + 1)

    plt.boxplot(bike_data[feature], vert=False)

    plt.title(f'Boxplot of {feature}')

    plt.xlabel('Value')

plt.tight_layout()

plt.show()
```

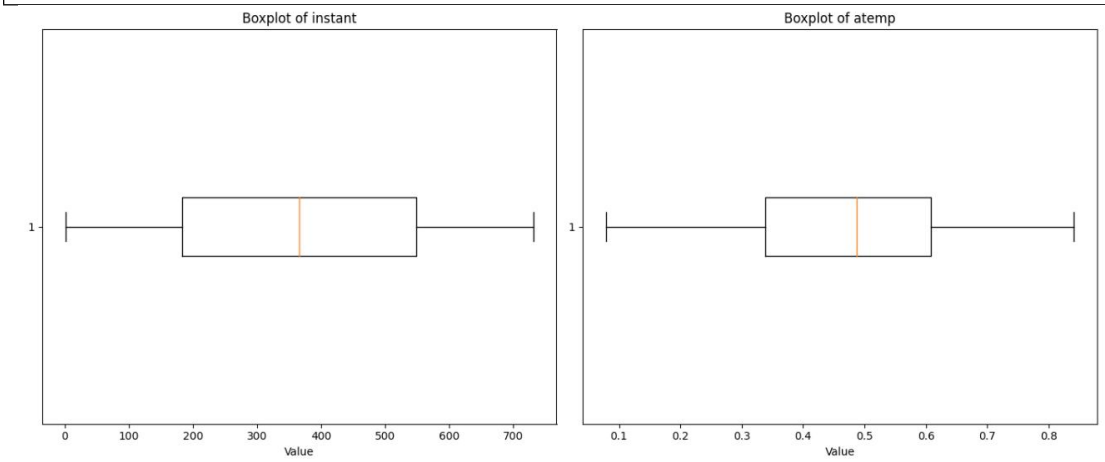

Рис 2. Боксплот для instant и atemp.