Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления» – ИУ5
Факультет «Радиотехнический» – РТ5

Отчёт по лабораторной работе №1 по курсу Технологии машинного обучения

7
(количество листов)

Исполнитель

студент группы РТ5-61б _____ Нижаметдинов М. Ш.

“___” _____ 2023 г.

Проверил

Преподаватель кафедры ИУ5 _____ Гапанюк Ю. Е.

“___” _____ 2023 г.

Москва, 2023 г.

Задание

- Выбрать набор данных (датасет).
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Дополнительно примеры решения задач, содержащие визуализацию, можно посмотреть в репозитории курса mlcourse.ai - https://github.com/Yorko/mlcourse.ai/wiki/Individual-projects-and-tutorials-(in-Russian)

Набор данных

https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-recognition-dataset

Исходный текст проекта

1) Текстовое описание набора данных

В качестве набора данных мы будем использовать набор данных по результатам химического анализа вин - https://scikit-learn.org/stable/datasets/toy_dataset.html#wine-recognition-dataset

Данные являются результатом химического анализа вин, выращенных в одном и том же регионе Италии тремя разными культиваторами. Для различных компонентов, содержащихся в трех типах вина, проводится тринадцать различных измерений.

Датасет состоит из одного файла, подключаемого через функцию библиотеки sklearn.

Каждый файл содержит следующие колонки:
- Алкоголь
- Яблочная кислота
- Пепел
- Щелочность золы
- Магний
- Всего фенолов
- Флавоноиды
- Нефлаваноидные фенолы
- Проантоцианы
- Интенсивность цвета
- Оттенок
- OD280/OD315 разбавленных вин

- Пролин

Импорт библиотек

Импортируем библиотеки с помощью команды import. Как правило, все команды import размещают в первой ячейке ноутбука, но мы в этом примере будем подключать все библиотеки последовательно, по мере их использования.

```
import numpy as np
import pandas as pd
from sklearn.datasets import *
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")

# Загрузка данных (вины Фишера)

wine = load_wine()

type(wine)

# Датасет возвращается в виде словаря со следующими ключами
for x in wine:
    print(x)

wine['target_names']

wine['feature_names']

# Размерность данных
wine['data'].shape

# Размерность целевого признака
wine['target'].shape

data = pd.DataFrame(data= np.c_[wine['data'], wine['target']],
                    columns= wine['feature_names'] + ['target'])

data

## 2) Основные характеристики датасета

data.head()

data.shape
```

```
total_count = data.shape[0]
print('Всего строк: {}'.format(total_count))

# Список колонок
data.columns

# Список колонок с типами данных
data.dtypes

# Проверим наличие пустых значений
# Цикл по колонкам датасета
for col in data.columns:
    # Количество пустых значений - все значения заполнены
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{} - {}'.format(col, temp_null_count))

# Основные статистические характеристки набора данных
data.describe()

# Определим уникальное значения для целевого признака
data['target'].unique()
```

3) Визуальное исследование датасета

Диаграмма рассеяния
Позволяет построить распределение двух колонок данных и визуально обнаружить наличие зависимости. Не предполагается, что значения упорядочены (например, по времени).

```
fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='proline', y='alcohol', data=data)

fig, ax = plt.subplots(figsize=(10,10))
sns.scatterplot(ax=ax, x='proline', y='alcohol', data=data, hue='target')
```

Гистограмма
Позволяет оценить плотность вероятности распределения данных.

```
fig, ax = plt.subplots(figsize=(10,10))
sns.distplot(data['proline'])
```

Jointplot
Комбинация гистограмм и диаграмм рассеивания.

```
sns.jointplot(x='proline', y='alcohol', data=data)
```

```
sns.jointplot(x='proline', y='alcohol', data=data, kind="hex")

sns.jointplot(x='proline', y='alcohol', data=data, kind="kde")
```

"Парные диаграммы"
Комбинация гистограмм и диаграмм рассеивания для всего набора данных.

Выводится матрица графиков. На пересечении строки и столбца, которые соответствуют двум показателям, строится диаграмма рассеивания. В главной диагонали матрицы строятся гистограммы распределения соответствующих показателей.

```
sns.pairplot(data)

sns.pairplot(data, hue="target")
```

Ящик с усами
Отображает одномерное распределение вероятности.

```
sns.boxplot(x=data['proline'])

# По вертикали
sns.boxplot(y=data['proline'])

# Распределение параметра target сгруппированные по propile.
sns.boxplot(x='target', y='proline', data=data)
```

Violin plot
Похоже на предыдущую диаграмму, но по краям отображаются распределения плотности - https://en.wikipedia.org/wiki/Kernel_density_estimation

```
sns.violinplot(x=data['proline'])

fig, ax = plt.subplots(2, 1, figsize=(10,10))
sns.violinplot(ax=ax[0], x=data['proline'])
sns.distplot(data['proline'], ax=ax[1])

# Распределение параметра Humidity сгруппированные по Occupancy.
sns.violinplot(x='target', y='proline', data=data)

sns.catplot(y='proline', x='target', data=data, kind="violin", split=True)
```

4) Информация о корреляции признаков
Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки датасета) наиболее сильно коррелируют с целевым признаком (в нашем примере это колонка "target"). Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

data.corr()

Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков.

Корреляционная матрица симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой).

На основе корреляционной матрицы можно сделать следующие выводы:

- Целевой признак наиболее сильно коррелирует с пропиленом (-0.63), OD280/OD315 разбавленных вин (-0.79), общим кол-во фенолов (-0,71) и флавоноидами (-0,84). Эти признаки обязательно следует оставить в модели.
- Целевой признак отчасти коррелирует с алкоголем (-0.33). Этот признак стоит также оставить в модели.
- Целевой признак слабо коррелирует с золой (-0.05) и магнием (-0.2). Скорее всего эти признаки стоит исключить из модели, возможно они только ухудшат качество модели.

```
sns.heatmap(data.corr())

# Вывод значений в ячейках
fig, ax = plt.subplots(figsize=(20,20))
sns.heatmap(data.corr(), annot=True, fmt='.3f', ax=ax)

# Изменение цветовой гаммы
fig, ax = plt.subplots(figsize=(20,20))
sns.heatmap(data.corr(), cmap='YlGnBu', annot=True, fmt='.3f', ax=ax)

# Треугольный вариант матрицы
fig, ax = plt.subplots(figsize=(40,40))
mask = np.zeros_like(data.corr(), dtype=np.bool)
# чтобы оставить нижнюю часть матрицы
# mask[np.triu_indices_from(mask)] = True
# чтобы оставить верхнюю часть матрицы
```

```
mask[np.tril_indices_from(mask)] = True
sns.heatmap(data.corr(), mask=mask, annot=True, fmt='.3f', ax=ax)

fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(30,10))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```

Тепловая карта с указание размера

```
fig, ax = plt.subplots(1, 1, sharex='col', sharey='row', figsize=(15,10))
fig.suptitle('Корреляционная матрица')
sns.heatmap(data.corr(), ax=ax, annot=True, fmt='.3f')
```