Exercise Basics Data Mining

Prof. Dr.-Ing. Steffen Schober

1 Exercise 1: Statistical description of data

Let there be a small data set with one feature: {8, 2, 4, 5, 1, 2, 6}

1. Calculate the following statistical features by hand! (Please do NOT use Python or any calculator):
   - mean value
   - median
   - quantile $Q_{0.25}$
   - quantile $Q_{0.75}$
2. Now use Python to calculate the statistical features. Do you get the same values?
3. Manually draw (by hand!) a histogram with bins of width 2 (bins: (0,2] , (2,4] ,. . .)
4. Now use Python to plot the histogram. Do you get the same result?

2 Exercise 2: Project understanding and Data understanding

First download the wine dataset from Moodle (wine.csv). You can read a CSV with pandas using `pandas.from_csv`. Importing pandas is achieved with

```
import pandas as pd
# df = pd.read_csv(...)
```

Using `pd` as alias is a convention.

2.1 Project and data understanding

Project Goal: Using chemical analysis to determine the origin of wines using the „wine“ data set.

Your data to solve the task:

- 3 different types of Italian wine
  - number of instances: 180
  - number of features: 13
  - number of classes: 3

- features:
  - Alcohol
  - Malic acid
  - Ash
  - Alcalinity of ash
  - Magnesium
  - Total phenols
  - Flavanoids
  - Nonflavanoid phenols
  - Proanthocyanins
  - Color intensity
  - Hue
  - OD280/OD315 of diluted wines
  - Proline
- one column „class“: with the types of wine $1, 2, 3$

Read the csv-file with the wine data set in a Pandas data frame.

1. Check if all data objects and features are available, compare the number of lines with the description above.
2. Check the types of your attributes (there is one column where it does not make sense),
3. also check for duplicates and missing values.

If you find duplicates or missing values remove the corresponding objects.

Hints:

- There is one column with a non-sense value in it.
- If a file is read, the types of each column are determined automatically (if possible). It might happen that there are different types in one column. As the documentation tells us: **Columns with mixed types are stored with the object dtype**.
- Duplicates can be removed with the method `DataFrame.duplicated()`.
- Missing values can be found with `DataFrame.isnull()`.
- Missing values can dropped with `DataFrame.dropna()`.

2.2 Data understanding and preparation, visualization

There are outliers in the data set (hint: 4 obvious outliers in one column, which you will find without having any background in chemistry).

- Find the outliers and remove the entire instances (the entire rows). You can use Python commands and visualization (e.g. histograms or box plots). Which outliers did you find?

Hints:

- The function `DataFrame.describe()` is useful, check out the argument `percentiles`.
- Make a boxplot of the suspicious column with the member function (`.plot.box()`).

3 Exercise 3: Use simple grouping to understand and classify data

There are many features for each class. A useful feature to classify wine is such that it behaves differently for different classes. First, let us check the mean of each class. Using the pandas `groupby` function (member function of a DataFrame), you can compute aggregate functions of groups. Use this to compute the mean of each feature for each group. If you found an interesting column, the following command vizualizes the distribution for the different classes.

```python
import seaborn as sns
# sns.displot(data=df, x=column, hue='class', kind='kde')
```