Exploratory Data Analysis

1 Introduction

Learning Goals

- Visualizing data (unlabeled and labeled)
- Interpreting diagrams and graphs

Plotting in Python

There are a lot of different plotting libraries available in Python. Not every diagram type is available in each library. The basic library is `matplotlib`, which takes care of correctly rendering the graphs. Most other libraries are built on top of `matplotlib` and add high-level functionality for changing the plot's appearance or for advanced or combined diagram types.

- matplotlib: basic rendering utilities
- pandas: straightforward plotting of `DataFrame` objects.
- seaborn: pretty graphs, easy integration of `DataFrames`
- plotly: interactive graphs

2 Lab Tasks

Reading the Dataset

```
from sklearn.datasets import load_wine
# configure load_wine to split data into features X and labels y
X, y = load_wine(as_frame=True, return_X_y=True)
# deep copy the features
df = X.copy()
# deliberately set targets (labels) to all zeros
df['target'] = 0
# we'll use the true labels in y in the subsequent task!
```

To summarize:

- `X` contains features only
- `y` contains labels only
- `df` contains total information (features for all samples and their associated labels)
- You may call `df.head()` and `df.describe()` to get a first impression of the data

Exploration of Unlabeled Data

1. Display all features in a single plot by using a parallel lines plot or a heatmap.
2. Decide whether the data needs scaling.
3. Detect any outliers in the features' distribution through means of a boxplot.
4. Find out which data points have outliers in any feature.
5. Detect correlations between features (e.g. by plotting a scatter matrix, computing correlation values, . . .).

```
import matplotlib.pyplot as plt
from pandas.plotting import parallel_coordinates, scatter_matrix
from sklearn.preprocessing import StandardScaler
```

Exploration of Labeled Data

Redo the above steps for the labeled dataset by producing grouped plots according to the labeling information.

```
# set true labels in df to use them in plotting
df['target'] = y
```

3 Homework

Plot Upload

Redo the tasks for the Iris Dataset. Information on the dataset can be found in the sklearn documentation. Save your favorite **(iris-related) plot and upload** it to Moodle. Make sure that the plot is self-contained:

- concise yet descriptive title
- fitting and well formatted axes labels
- appropriate color scheme
- legend (if necessary)

In Moodle, make sure to also **submit a short description** of your interpretation of what can be seen in the plot. This is an **individual assignment**, therefore you are graded individually. Collaboration in discussing the approach or some technical details is encouraged, yet your own assignment needs to be prepared and uploaded individually. Make sure that essential original content (in structure, reasoning and interpretation) makes individual grading possible on the instructor's side.

Quiz

Take the quiz in Moodle. As some questions refer to the work discussed in this lab session, please make sure to have your have your code up and running for both datasets!

Deadlines

Deadlines for plot upload and quiz are set in Moodle - make sure to stick to these guidelines as late submissions are not enabled! You have time until midnight before our next lab session to complete your work.

Further Reading

Refer to ITSM1DCEIL's lecture part and the literature linked within its slides. As an additional input on why data visualization matters, have a look at Anscombe's Quartet.