Introduction to Machine Learning / Labwork 0

Installing the environment

Maxime Ossonce maxime.ossonce@esme.fr

The purpose of this labwork is to make sure all needed packages are properly installed and the dataset is downloaded for next session. This labwork has to be fulfilled before the first session. No assistance concerning *this* labwork will be provided during the first session. Any assistance has to be sought before the first session.

At the end of the labwork, you will save the last image produced and upload it on moodle.

You will use the Spyder IDE that is surely already installed on your computer.

It is recommended (but not obligatory) to create a virtual environment for the course. For instance, call it `ml`. You will activate the environment when working for the class labworks.

Packages installation

Once you have (optionally) setup and activated your environment, you can install the needed packages (with the package manager of your choice, `conda` or `pip`). Install the follwing packages:

- `pandas`
- `matplotlib`
- `scikitlearn`
- `seaborn`

Packages and dataset loading

- Download the `csv` file containing the data (click here or see url[1]) in your project directory.

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv

- Create a python file in your project directory named wine_db_viz.py in which you will load the library mentioned (see code below).

wine_db_viz.py

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import sklearn
from socket import gethostname as gh
```

Once everything is on track, you can load the dataset with the pandas library.

wine_db_viz.py

```
# download the csv file first!
csv_file = 'winequality-white.csv'
df = pd.read_csv(csv_file, header='infer', delimiter=';')
```

Correlation matrix

Now that the dataset is uploaded, we visualize some properties. For instance the mean and standard deviations of the different variables:

wine_db_viz.py

```
print(df.head())

X = df.drop('quality', axis=1)  # we drop the column "quality"

print('Average')
print(X.mean())

print('\nStandard deviation')
print(X.std())
```

Then, plot the correlation matrix:

wine_db_viz.py

```
corr = X.corr()
plt.title(gh())
sns.heatmap(corr)
plt.show()
```

You should be able to visualize the correlation matrix now! All you have to do is to save it as PNG and upload it on moodle.