MSc in Data Science

Machine Learning

Academic Year: 2022-2023

Exercise 1: Applying the Project Template

Delivery Date: **23/12/2022**

You are provided with two datasets: The breast cancer and the Diabetes datasets. Both datasets are included in Python's scikit-learn package[1]. The objective of the exercise is to apply the project template from lecture 6 on both datasets.

Classification: The Breast Cancer dataset

Using this dataset, you are requested to apply the project template on the dataset. You are expected to provide (among other things) the following:

- The dimensions of the dataset
- A peek at the data
- Statistical summary of all attributes
- The class distribution (number of instances per class)
- Univariate plots to better understand each attribute
- Multivariate plots to better understand relationships between attributes
- Apply a set of algorithms and select the best model
- Split the dataset into training/test sets (with test set being the 20% of the dataset) or use cross-validation and evaluate accuracy, as well as other metrics of the winning algorithm
- Report the confusion matrix

Regression: The Diabetes dataset

Using this dataset, you are requested to apply the project template on the dataset. You are expected to provide (among other things) the following:

- The dimensions of the dataset
- A peek at the data
- Statistical summary of all attributes
- The class distribution (number of instances per class)
- Univariate plots to better understand each attribute (histograms, density plots, whisker plots)
- Multivariate plots to better understand relationships between attributes (scatter plot matrix, correlations)
- Do you have any ideas for feature engineering?
    - Remove the most correlated attributes?
    - Normalising the dataset to reduce the effect of differing scales of attributes?
    - Standardising the dataset to reduce the effects of differing distributions?
- Evaluate algorithms also with normalisation/standardisation (along with the baseline)
- Improve results with tuning for the winning algorithm

[1] sklearn.datasets.load_diabetes, sklearn.datasets.load_breast_cancer

The exercise has the following deliverables:

Two Jupyter notebooks (one for each dataset). Explanations on the various steps and comments regarding the obtained results should be included.