MSc in Data Science

Machine Learning

Academic Year: 2017-2018

Exercise 2: Applying the Project Template

Delivery Date: **19/12/2017**

You are provided with two datasets: The Iris dataset, and the Boston Housing Data. Both datasets are included in Python's Scikit-learn package, and have been used as examples in lectures (The Iris dataset has been used in lecture 3 – Decision Trees – and the Boston Housing Data has been used in lecture 10 – Applied Machine Learning 2). Both datasets are also available in Lesson's Github repository at:

https://github.com/MSc-in-Data-Science/class_material/tree/master/semester_1/Machine_Learning/datasets

The objective of the exercise is to apply the project template from lecture 10 on both datasets.

Classification: The Iris 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) and evaluate accuracy of the winning algorithm
- Report the confusion matrix

Regression: The Boston Housing Data 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