INFO 204 Assignment 1

Worth: 10%, Due: 5pm 27th August 2021

The goal of this assignment is to complete a small supervised learning project. Briefly, you will:

1. Load the supplied data, perform some basic exploratory data analysis, and produce a transformed data set (3 marks)
2. Perform a hyperparameter analysis of k-nearest neighbour and CART decision trees (3 marks)
3. Fit and compare the following models using repeated cross validation:
    a. Linear regression using the original data
    b. Linear regression using the transformed data from Step 1
    c. k-Nearest Neighbour regression using the optimal neighbourhood size determined in step 2
    d. CART decision tree using the optimal minimum split size determined in step 2

    Include a suitable visualisation of the results and a brief commentary on the performance of the four approaches. (3 marks)
4. Build and visualise a CART decision tree using a minimum split size that reasonably balances tree size. Describe at least one prediction pathway in the tree (1 mark)

More detailed instructions for each step are outlined below. You will complete your work in a Jupyter notebook, and upload your resulting document to Blackboard.

Every Python (and related library) routine that you will need to call upon to complete this assignment has an exemplar that has been discussed or presented in lectures/labs. All examples up to and including Lab 5 and up to and including Lecture 10 will help with completing this assignment.

There will be time in Lab 7 to work on this assignment, but you should expect to spend a reasonable amount of your own time to complete this assignment.

You are expected to work on and submit this assignment on your own. You must adhere to all the usual requirements of academic integrity for this assignment (refer to the University's Academic Integrity pages for more information).

Step 1: Exploratory Data Analysis (EDA) (3 marks)

The provided data is a modified version of the Boston Housing data used in earlier labs. The response variable, medv, has been modified and so its relationship to the other features of the data set should be unknown even if you have worked with this data before. More details about this data set (in its unmodified form) can be found at:

https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

In this step, you are expected to load the provided data, and perform an analysis of the data set in terms of visual and descriptive statistics, similar to that done in Labs 4 and 5 and discussed in Lecture 9. In the process of your analysis, you should identify two features that could be removed from the data set, and two features that could be transformed to present a stronger linear relationship with the target (the medv variable).

Be sure to include appropriate visualisations and justify your decisions. This step is intentionally loosely defined to test your ability to conduct an EDA. However, you should feel free to discuss your ideas with staff.

Step 2: Hyperparameter Analysis (3 marks)

This step is similar in nature to the work you perform in Lab 5. In particular, you take two methods, k-Nearest Neighbours and CART decision trees, and use cross validation to analyse their behaviour in response to changes to a key hyperparameter setting (specifically the neighbourhood size for kNN, and the minimum split size for CART. Ultimately, your goal is to find the "best" hyperparameter setting for these algorithms that you will use in the next step.

You are free to specify your own parameter grid for these exercises, although a suitable set for the neighbourhood size might be [1, 2, 5, 10, 20, 40, 80, 120, 360], while an acceptable set for minimum split would be [2, 5, 10, 20, 40, 80, 120, 360]. Be warned that larger sets will present a more thorough examination of the hyperparameter at the expense of more time require to process.

You are also free to use the helper routines provided by scikit-learn to ease this process (in other words, you are not expected to replicate the laborious manual code presented in Lecture 10).

Finally, be sure to incorporate feature standardisation for kNN if your EDA process in the previous step suggests that it is warranted. A pipeline for this may help (see Lab 5).

Step 3: Final Comparison (3 marks)

Once you have tuned kNN and CART in the previous step, you must now compare their performance on the modified Boston Housing problem to that of linear regression. As in the previous step, cross validation (3 rounds of 10 folds) will be used to estimate the generalisation performance of the methods.

When examining the performance of linear regression, ensure that you consider both the "original" features, and also the effect of EDA (i.e., the performance of linear regression using the transformed data identified through EDA).

An appropriate visualisation to compare the results is useful here, and be sure to provide a description of the relative performance of the methods.

Step 4: Knowledge Generation (1 mark)

Similar to the process that you performed in Lab 4, generate a CART decision tree and visualise it using scikit-learn's built-in functions. Describe at least one prediction pathway in the tree – be sure to use an appropriately large value for min_samples_split so that the tree is not too detailed!