Machine Learning for Trading

Project 3: Assess Learners

Tharun Saranga (tsaranga3)

Experimental Methodology:

The purpose of the experiments is to evaluate the performance of various decision tree learners implemented in different forms. The experiment is specifically conducted upon Istanbul data provided for the project. The data is extracted from the provided csv file with header and data column stripped. It is then cut into training data and test data. The insample root mean square error is the error with the training data. The outsample root mean square error is the error with the testing data. The experiment can be run using the following command "*python testlearners.py Data/Istanbul.csv"*

Experiment 1:

Purpose: To check if overfitting occurs with respect to leaf size on Classical Decision Tree Learner. Identify the leaf size at which the overfitting occurs.

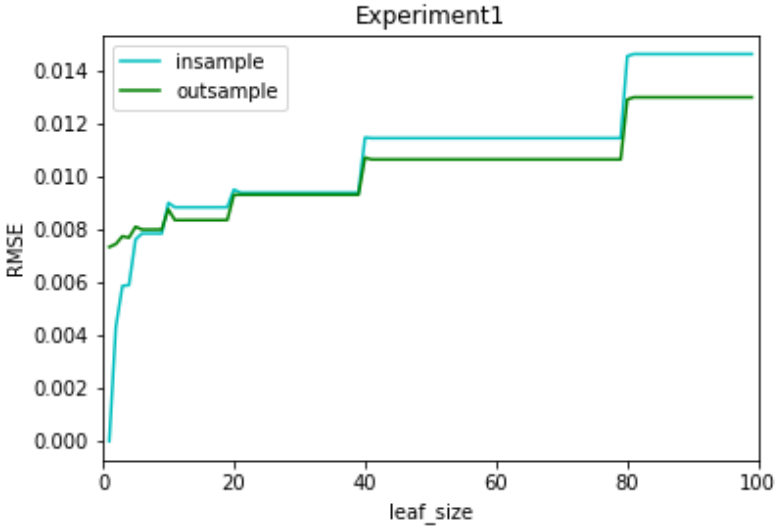

Figure 1: leaf_size vs RMSE for Classical Decision Tree

The graph shows insample and out sample rmse errors for various leaf sizes. Overfitting occurs when the model fits perfectly to the given data resulting in a very small insample rmse and large outsample rmse. We can see this in the graph where the insample error is way less than the outsample error. From the graph we can see this occurs below leaf size of 10. As the leaf size increases the overfitting decreases with overall increase in the error.

Experiment 2:

Purpose: To check if bagging can reduce or eliminate overfitting with respect with leaf size on Classical Decision Tree Learner.

Figure 2: leaf_size vs RMSE DTLearner with Bagging

The above graph shows the rmse errors from insample and outsample data for a classical decision tree with bagging. The number of bags is constant at 20 and leaf size is varied. The bagging has reduced the overfitting effect but not eliminated completely. Overfitting still occurs for leaf size under 10, but now the effect is reduced with a low rmse error than without bagging.

Experiment 3:

Purpose: To compare the performance of Classical Decision Tree and random Decision Tree Quantitatively.

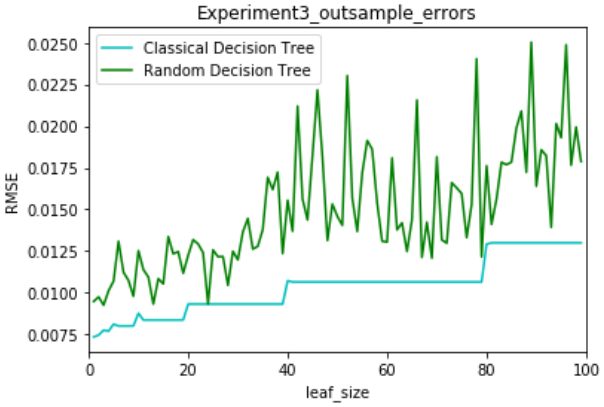

Figure 4: leaf_size vs RMSE of Classical and Random Decision Trees

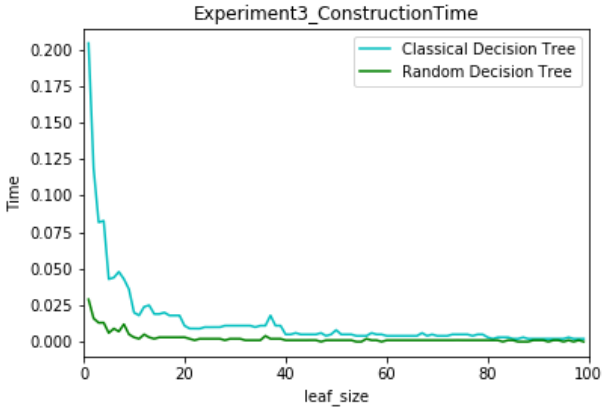

Figure 3: leaf_size vs Construction Time for Classical and Random Decision Trees

Figure 3 shows the plot of the outsample rms errors of both Classical Decision Tree and Random Decision Tree. The Classical Decision Tree has low error rate across all the leaf sizes. The Random Decision Tree sometimes reaches the same error level as that of the Classical Decision Tree when the randomly selected value to split is of more information. The Random Decision Tree is more unstable in the output because of obvious random nature, while the Classical Decision Tree is more stable.

Figure 4 shows the plot of the construction time of both Classical Decision Tree and Random Decision Tree as the leaf size changes. The Classical Decision Tree takes more time to construct because of the extra computation required to select the best feature from the data. Random Decision Tree just picks a feature randomly reducing the computational time. As the leaf size increases the Construction time for both the Trees converge. From this aspect the Random Decision Tree is better than the Classical Decision Tree.

Overall using the two different quantitative measures Random Decision Tree is better than Classical Decision Tree when working with huge amounts of data. The error difference between the two is very small considering the speed difference between them. Hence, the Random Decision Tree would be better than Classical Decision Tree.