Sujit Kanti Biswas
GT User ID: sbiswas67
GT ID: 903549376

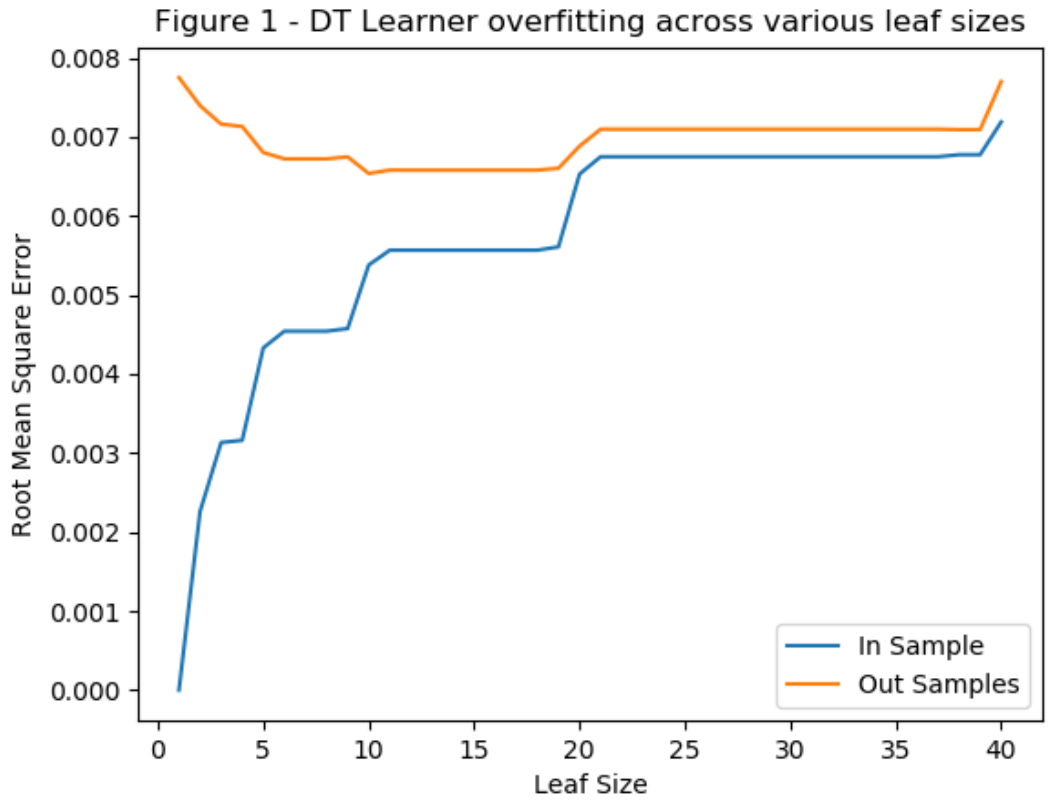
Project 3: Assess Learner Report

1. We are trying to test whether overfitting occurs with respect to the leaf_size of the Decision Tree Learner.

The DT Learner module has been executed with a varying leaf_size from 1 to 40 and the corresponding RMSE (Root Mean Square Error) has been recorded for each leaf_size. This process is carried out for both In-Sample results and out-sample results using the Istanbul.csv file as input data. The results are plotted with varying RMSE on y-axis and no. of leaves on x-axis.

From the graphical plot we can see that the in-sample RMSE (blue line) is increasing as the leaf_size increases. The out-sample RMSE (orange line) is decreasing with the increase in leaf_size and at leaf_size=10, the out-sample RMSE is the least and then it starts increasing as the leaf_size increases.

This behaviour of the out-sample RMSE corresponds to overfitting of ML algorithms. So, we can say that overfitting does occur with respect to the leaf_size and in this experiment the overfitting point is leaf_size = 10.

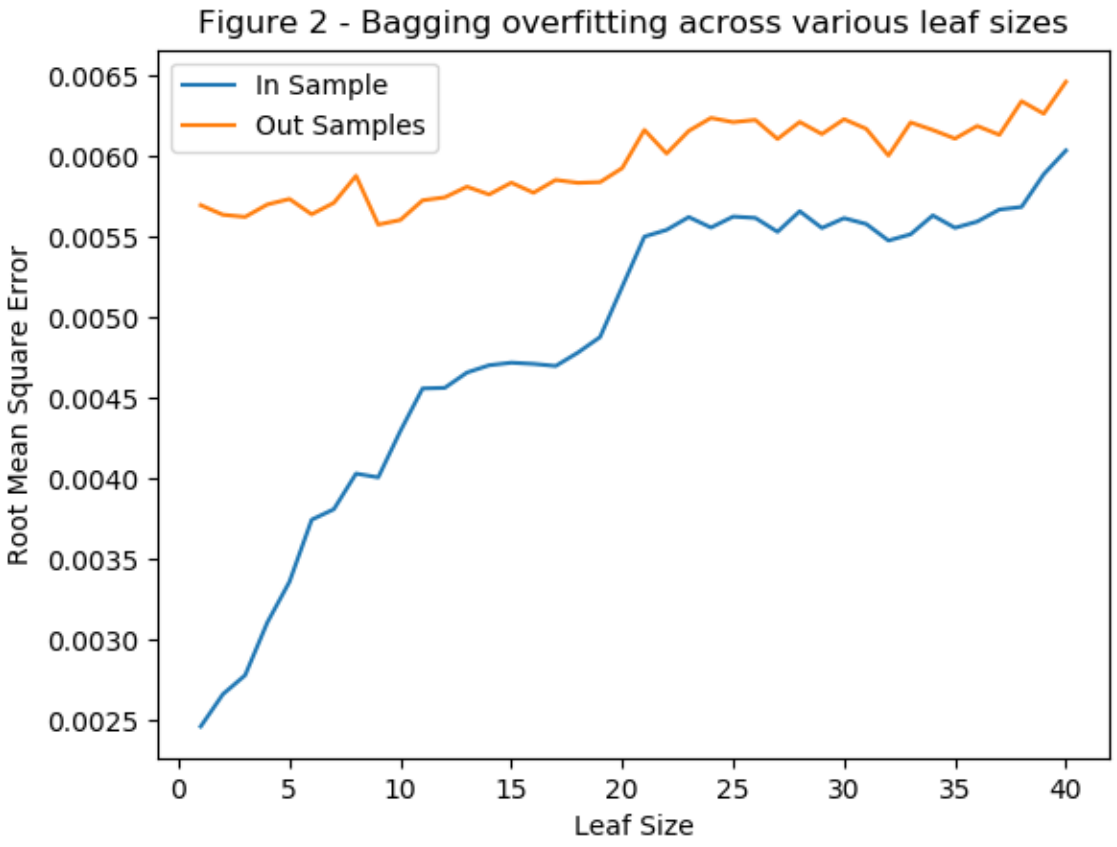


2. We are trying to test whether overfitting occurs with respect to the leaf_size when we are using Bagging of twenty DTLearners and observe if bagging can reduce or eliminate overfitting.

The BagLearner module has been executed with a varying leaf_size from 1 to 40 consisting of 20 bags and the corresponding RMSE (Root Mean Square Error) has been recorded for each leaf_size. This process is carried out for both In-Sample results and out-sample results using the Istanbul.csv file as input data. The results are plotted with varying RMSE on y-axis and no. of leaves on x-axis.

From the graphical plot we can observe that the same trend (as in experiment 1) where the in-sample error increases with the increase of leaf sizes.
With respect to overfitting, we see that the out-sample (orange line) error rates have gone down at leaf_size = 8 and then it starts to increase as the increase in leaf_size. So, we can say the overfitting point in this case is leaf_size = 8.

But at the same time, we can observe that the overall increase in RMSE for out-sample data is considerably lower than what we have seen in experiment-1 using only one DTLearner which tells us that though overfitting can occur, the overall error rate is considerably lowered by the use of bagging.

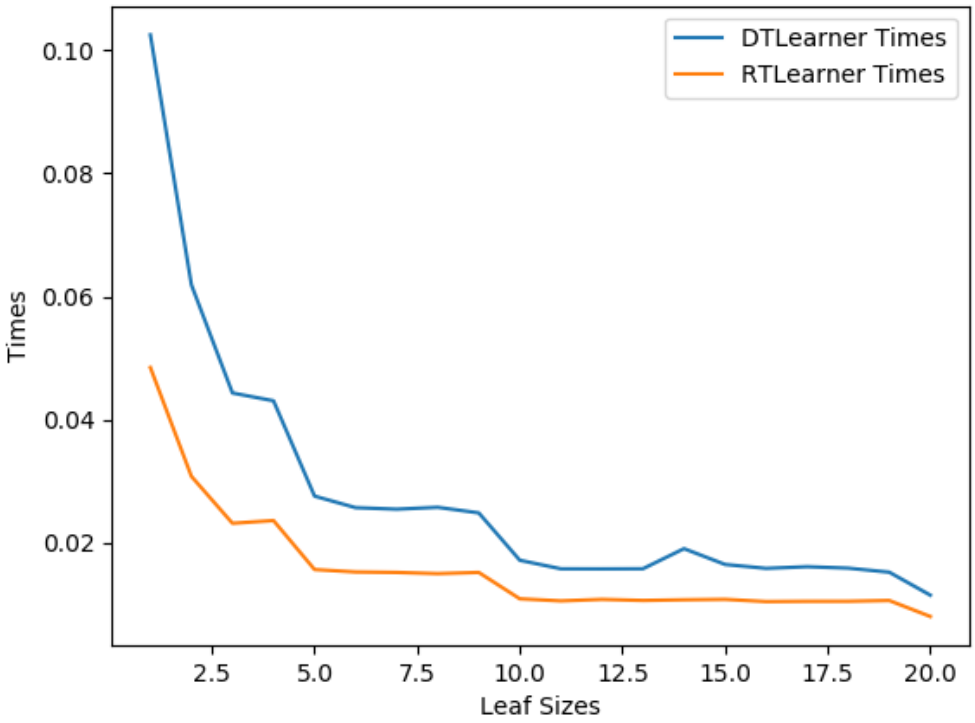


3. We shall be using 2 different metrics to compare DTLearner with RTLearner as follows.

Metric 1 – Total Run Time for Training + In-sample prediction + Out-sample prediction:
Here we are trying to compare run time. The running time of each algorithm is captured with increasing the size of the leaf nodes from 1 to 20 using the Istanbul.csv file as input data.

Here we can observe that the random tree is pretty much faster than the classic decision tree irrespective of the leaf_size. This completely makes sense as it takes time for the classic decision tree to evaluate the correlation function to find the median for splitting. The random tree on the other hand just picks a random factor to split. The random tree is much faster when the leaf_size was smaller, especially at leaf size = 1, but the difference became very narrow as the size of the leaves grew bigger since, when the leaf sizes are smaller the tree has to make more splits and so the classic tree loses more time finding median value a greater number of times.

Figure 3 - DT Learner v/s. RT Learner w.r.t Run Time



Metric 2 – Mean Absolute Error (MAE) on Out-sample prediction:
The MAE of each algorithm is captured on out-sample prediction with increasing the size of the leaf nodes from 1 to 20 using the Istanbul.csv file as input data.

We are sure that the classic decision tree would perform better in terms of accuracy as it actually finds the best factor to split the tree on using some metric like correlation.
This is evident from the graphical plot (Figure 4) as shown below. When the leaf_size=1, the MAE for RTLearner (orange line) is pretty high compared to the MAE of DTLearner (blue line). This shows that in terms of prediction accuracy DTLearner performs much better compared to RTLearner.

We can also see that the MAE for DTLearner has much less fluctuations and shows more consistency in prediction even when the leaf_size continues to increase. Whereas RTLearner shows very inconsistent MAE with lot of fluctuations.

Figure 4 - DT Learner v/s. RT Learner w.r.t Mean Absolute Error (MAE)