Report for Project assess_learners

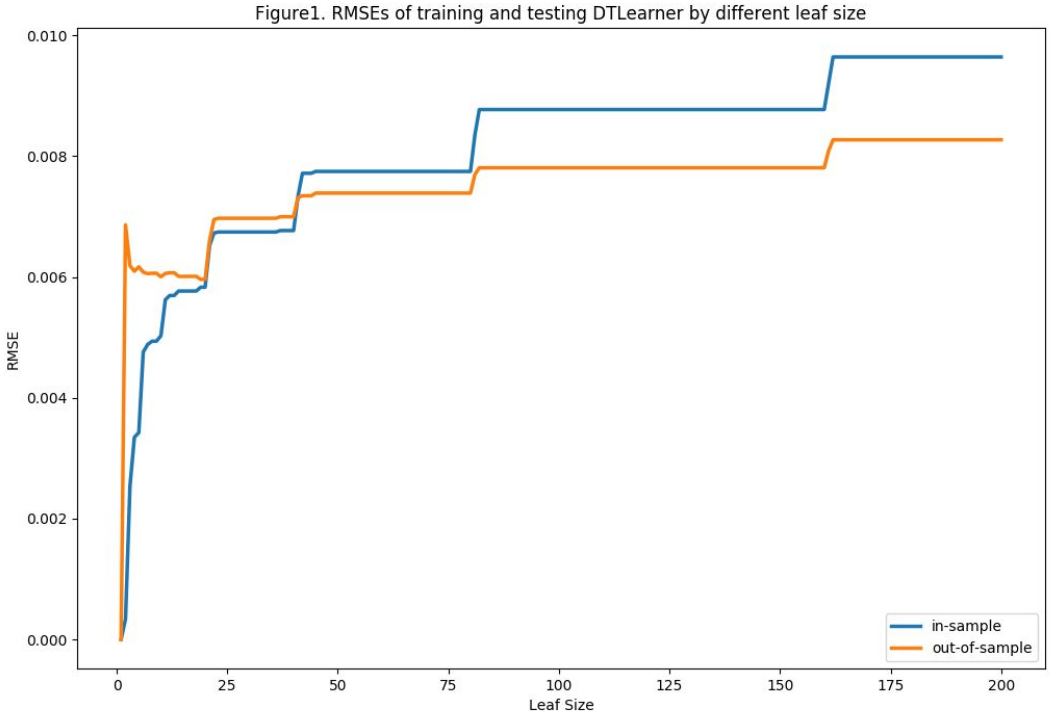
ZHENG FU
ZFU66@GATECH.EDU

Question 1: Does overfitting occur with respect to leaf_size? Consider the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).

In this experiment the data was randomly shuffled and 60% was used to train the DTLearner (in-sample) and 40% was used to test the model (out-of-sample). The range of leaf_size is from 1 to 200, and for each leaf_size, 10 iterations were carried out to obtain the average RMSE value and smooth the curve. Here I define overfitting as the RMSE of in-sample is less than the RMSE of out-of-sample. Figure 1 shows the trends of RMSE value changing while leaf_size increasing for both in-sample and out-sample. It is clear that these two lines are crossing between leaf_size 25 ~ 50. Then I use the following python code to detect the crossing point:

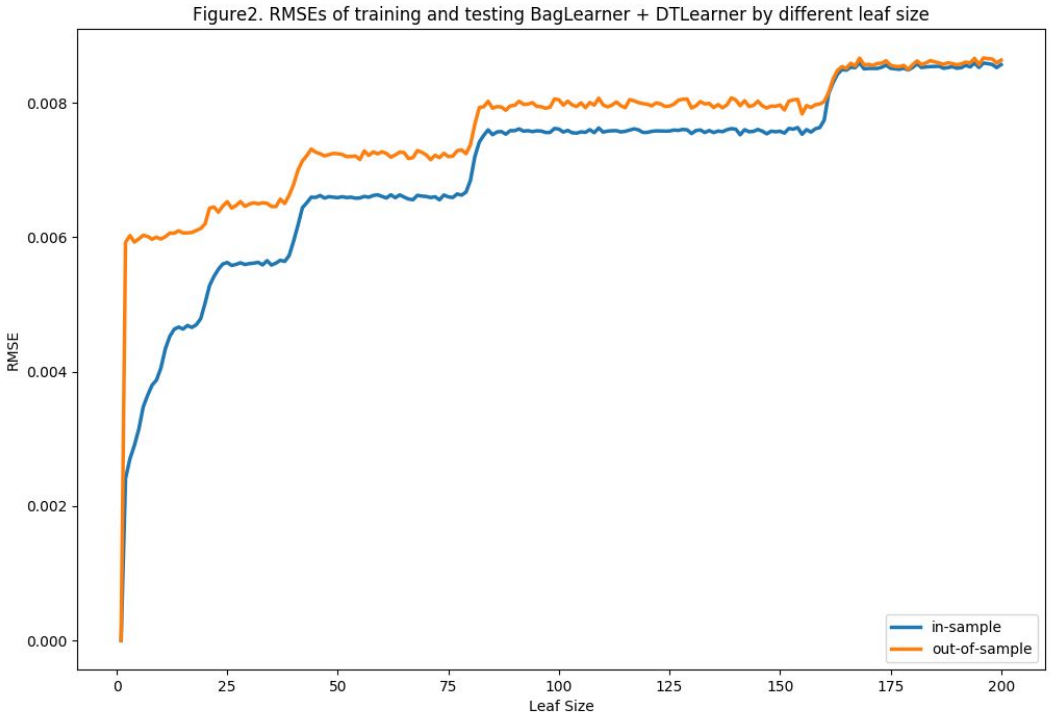
```
RMSE_diff = np.subtract(RMSE_in_mean, RMSE_out_mean)
RMSE_diff_max = np.argmax(RMSE_diff > 0)
```

And the results is RMSE_diff_max = 40. That means when leaf_size is no more than 40, the overfitting occurs.

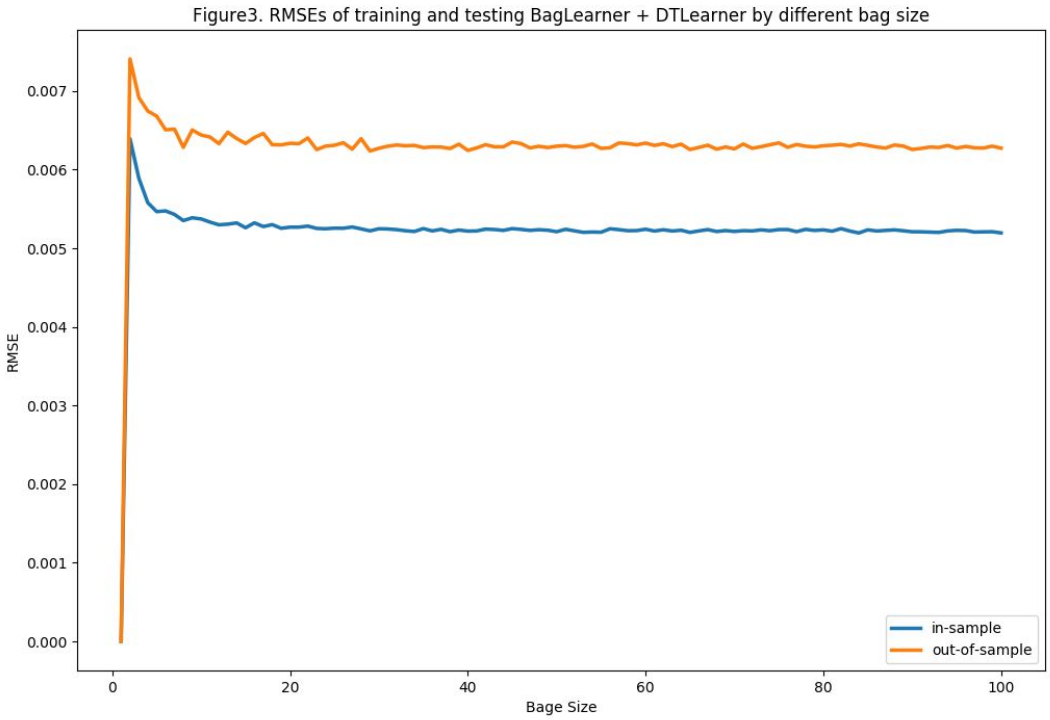


Question 2: Can bagging reduce or eliminate overfitting with respect to leaf_size? Again consider the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.

In the first experiment, the data was randomly shuffled and 60% was used to train the BagLearner with DTLearner (in-sample) and 40% was used to test the model (out-of-sample). The number of bags was fixed to 20. The range of leaf_size is from 1 to 200, and for each leaf_size, 10 iterations were carried out to obtain the average RMSE value and smooth the curve. Figure 2 shows the trends of RMSE value changing while leaf_size increasing for both in-sample and out-of-sample. It is clear that the RMSE of in-sample is less than the RMSE of out-of-sample for all leaf_size no less than 1, which indicates the occurrence of the overfitting.



In the second experiment, the data was randomly shuffled and 60% was used to train the BagLearner with DTLearner (in-sample) and 40% was used to test the model (out-of-sample). The leaf_size was fixed to 20. The range of number of bags is from 1 to 100, and for each bag_size 10 iterations were carried out to obtain the average RMSE value and smooth the curve. Figure 3 shows the trends of RMSE value changing while bag_size increasing for both in-sample and out-of-sample. It is clear that the RMSE of in-sample is greater than the RMSE of out-of-sample for all bag_size no less than 1, which indicates the occurrence of the overfitting.



If we compare Figure 2 with Figure 1, we could see the blue line and yellow line are much closer to each other when leaf_size > 175. Meanwhile Figure 3 reveals the blue line and yellow line are almost overlapped when bag_size no more than 5. All these indicate that bagging could somewhat reduce the overfitting and smaller bag size may have better results.

Question 3: Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Note that for this part of the report you must conduct new experiments, don't use the results of the experiments above for this.

In the first experiment the data was randomly shuffled and 60% was used to train the DTLearner (in-sample) and 40% was used to test the model (out-of-sample). The range of leaf_size is from 1 to 200, and for each leaf_size, 10 iterations were carried out to obtain the average RMSE value as well as the average Pearson Correlation Coefficients (PCC) value to smooth the curves. Figure 4 shows the trends of RMSE value changing while leaf_size increasing for both in-sample and out-of-sample, and Figure 5 shows the trends of PCC value changing while leaf_size increasing for both in-sample and out-of-sample.

In the second experiment the data was randomly shuffled and 60% was used to train the RTLearner (in-sample) and 40% was used to test the model (out-of-sample). The range of leaf_size is from 1 to 200, and for each leaf_size, 10 iterations were carried out to obtain the

Figure4. RMSEs of training and testing DTLearner by different leaf size

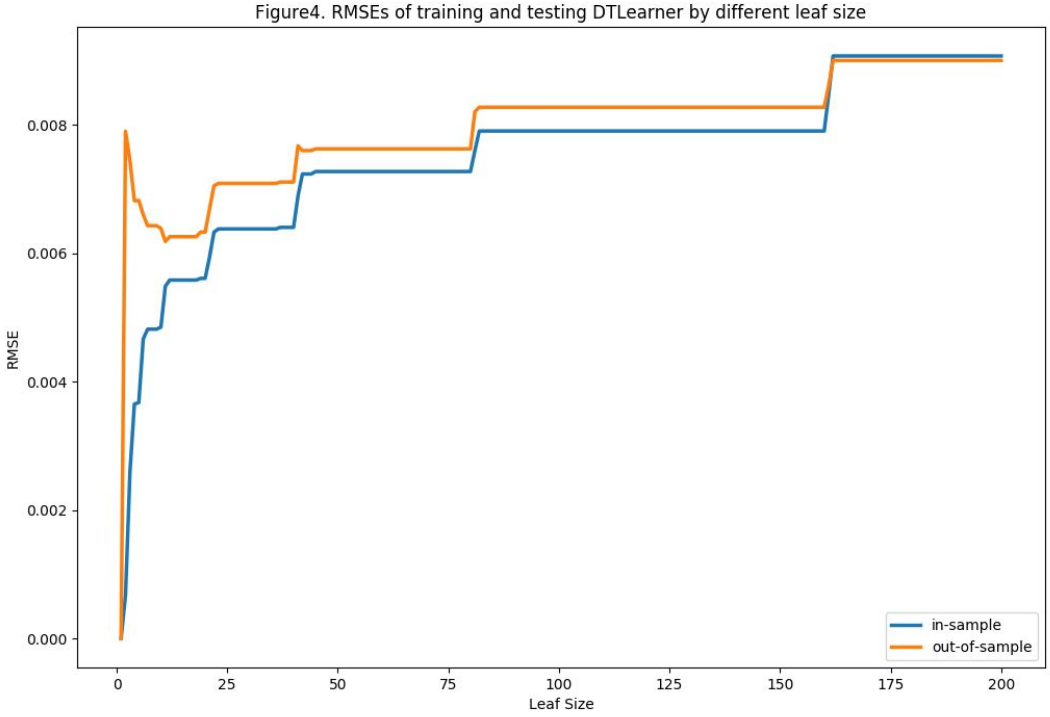


Figure5. PCC of training and testing DTLearner by different leaf size

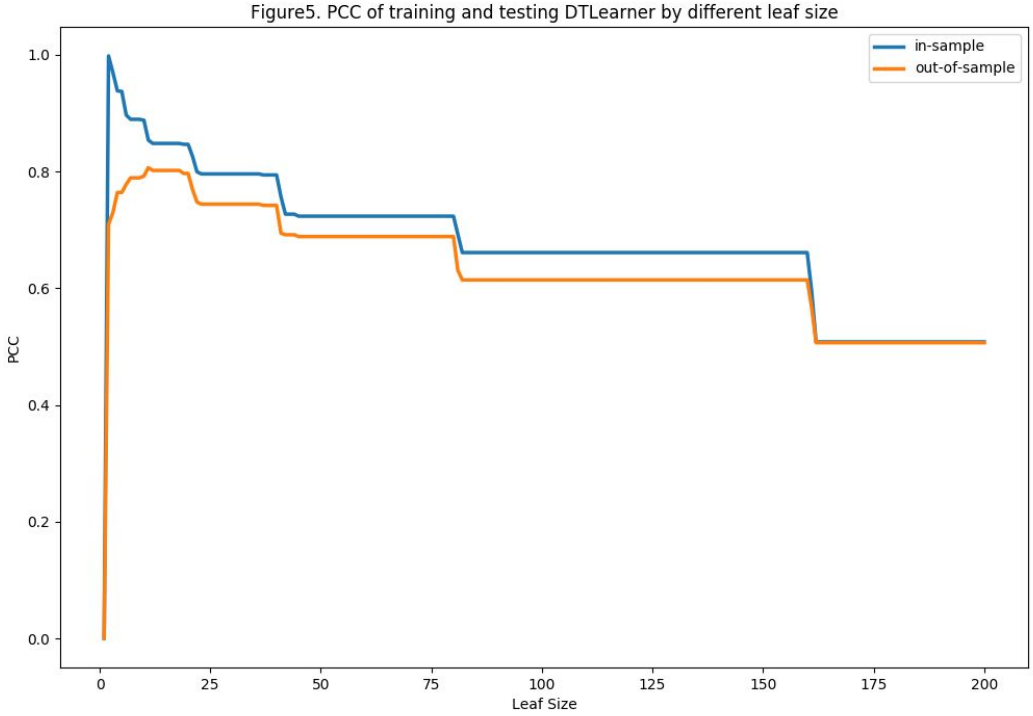


Figure6. RMSEs of training and testing RTLearner by different leaf size

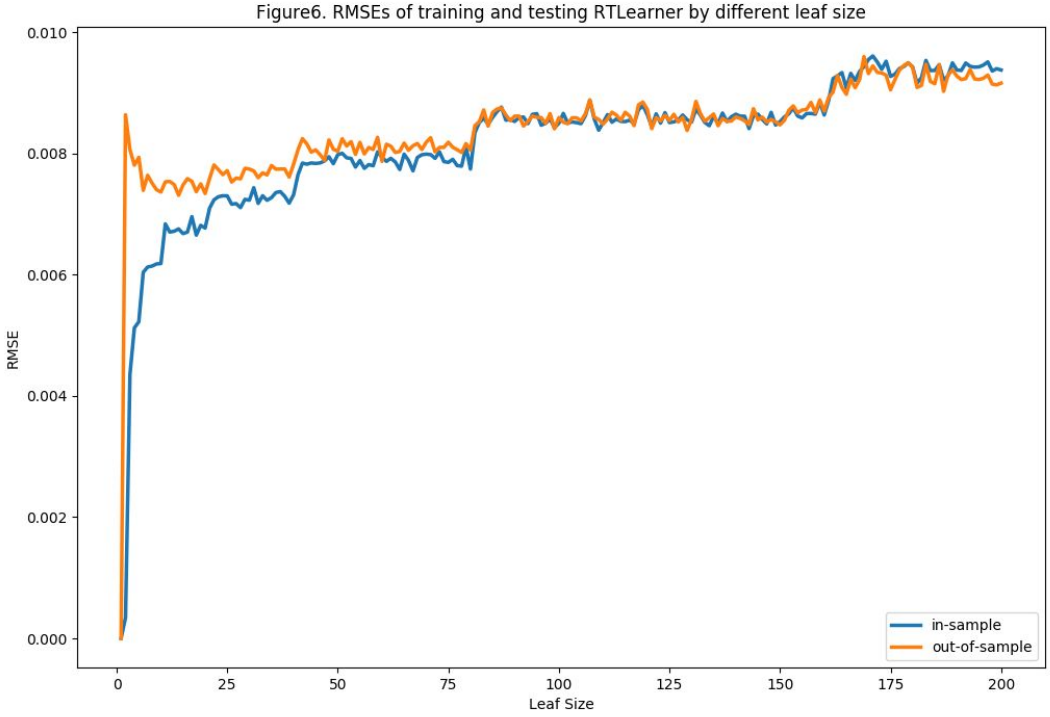


Figure7. PCC of training and testing RTLearner by different leaf size

average RMSE value as well as the average Pearson Correlation Coefficients (PCC) value to smooth the curves. Figure 6 shows the trends of RMSE value changing while leaf_size increasing for both in-sample and out-of-sample, and Figure 7 shows the trends of PCC value changing while leaf_size increasing for both in-sample and out-of-sample.

If we compare Figure 4 with Figure 6, we could see in Figure 4 the blue line and the yellow line are almost overlapped for leaf_size no less than 170, whereas in Figure 6 the blue line and the yellow line are almost overlapped for leaf_size no less than 75. Similarly, if we compare Figure 5 with Figure 7, we could see in Figure 5 the blue line and the yellow line are almost overlapped for leaf_size no less than 170, whereas in Figure 7 the blue line and the yellow line are almost overlapped for leaf_size no less than 75. That indicates when leaf_size is between 75 and 170, RTLearner has better performance than DTLearner in terms of reducing overfitting.