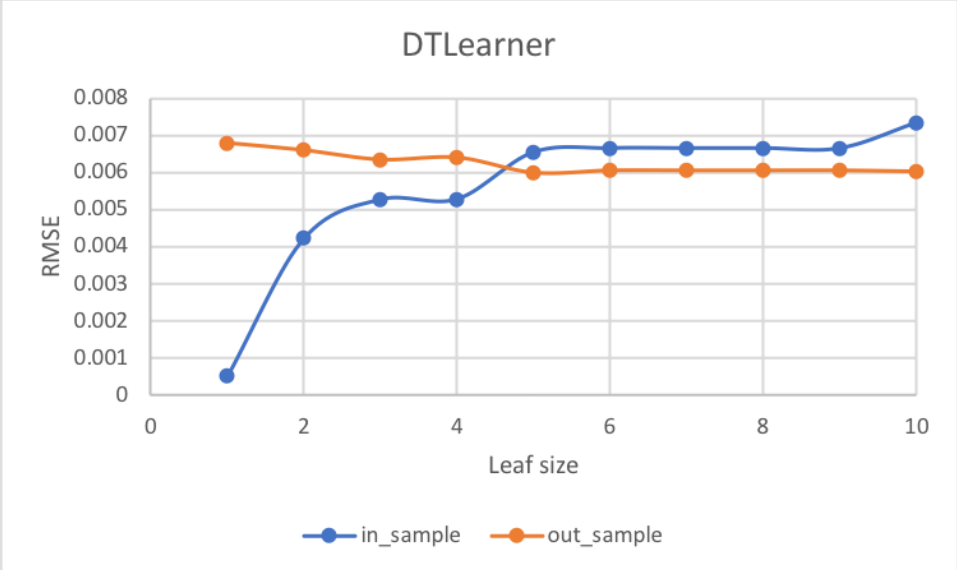
Assess Learners

Student name: Tao Peng
Gatech ID: tpeng38

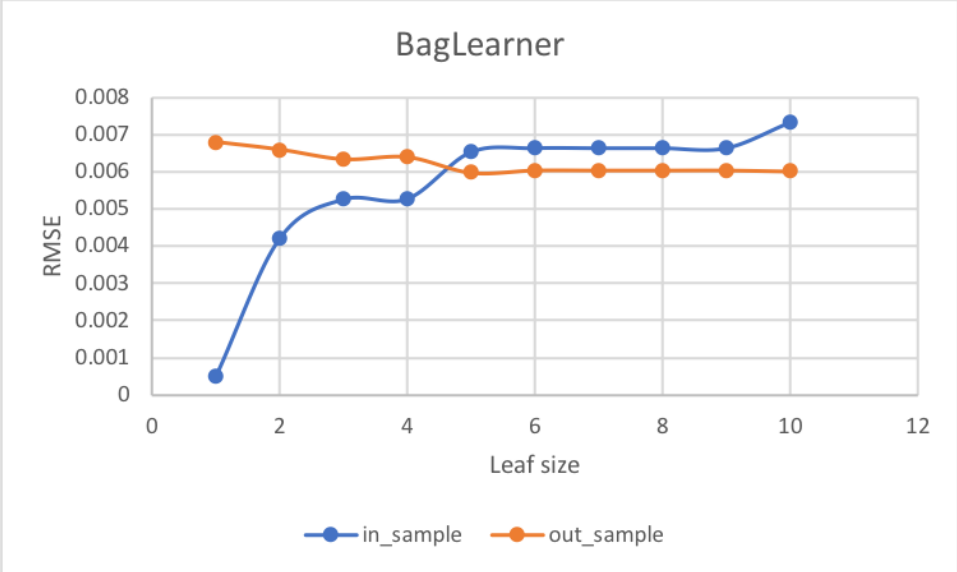
Question 1: Does overfitting occur with respect to leaf_size? Consider the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).

Anwser: Yes, according the following figure, overfitting does occur with respect to leaf_size. I applied DTLearner to Instanbul.csv, and calculated RMSE (both in and out of sample) for different values of leaf size. From the figure, we see that the in sample RMSE increases as the leaf size (thus the degree of freedom) increases. This is expected. The out of sample RMSE first decreases, but then gradually increases a bit as leaf size gets bigger, which is a sign of overfitting. The overfitting occurs when the out sample RMSE begins to increase, which roughly leaf size = 5.



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.

Answer: According to the following figure, bagging does not eliminate overfitting. I applied DTLearner with bagging (bag size = 30) to Istanbul.csv, and calculated RMSE (both in and out of sample) for different values of leaf size. We saw the out of sample RMSE first decreases as leaf size increases, but again increases a little as leaf size continue to increase, which is a sign of overfitting. And the increase begins at a little lower than leaf size = 5, so it reduces the overfitting a little.



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.

Answer: I used two quantitative measures to compare DTLearner with RTLearner: correlation and training time, as shown in the following two figures. We see that for most of the leaf size, the correlation of RTLearner is smaller than that of DTLearner, so this means somewhat higher accuracy for RTLearner. We also see that RTLearner generally needs a shorter training time, in this regard, RTLearner is also better (faster) than DTLearner. So considering these two measuers, the RTLearner is better than DTLearner.