Everardo Villasenor

GT User: evillasenor3

GT ID: 903389317

CS 4646/7646

Assess Learners Report

1. Does overfitting occur with respect to leaf_size?

My initial hypothesis is that the DTLearner will have greatest overfit with a leaf size of 1 because each sample is aggregated at a leaf and the model is more specific because of that. To test, I increase the leaf size as my testing variable and analyze the in sample and out of sample root mean square error. If the difference between the two errors decreases as leaf size increases, then I would suspect that there is some correlation with leaf size and overfitting.

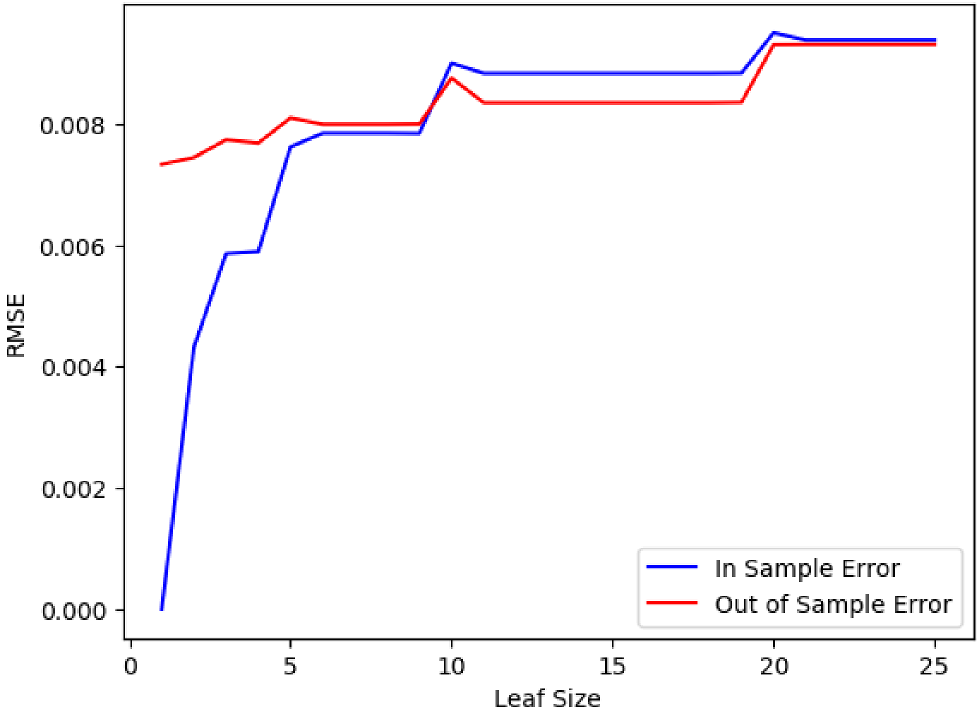

Figure 1: RMSE v Leaf Size for DTLearner class

Leaf Size	Difference in RMSE
1	0.007332
2	0.003131
3	0.001876
4	0.001790
5	0.000479
6	0.000145
7	0.000145

Table 1: Difference in RMSE for Overfitting Region

Given that the largest difference between errors occurs at the smallest leaf size. This affirms my initial hypothesis because the difference between the two errors decreases as leaf size increased. It seems that the difference stops around the leaf size of 6, anything below that would be the region I would consider to be overfitted.

2. Can bagging reduce or eliminate overfitting with respect to leaf size?

My initial hypothesis is that bagging can reduce overfitting with respect to leaf size because we are using an ensemble of models, but I still think for small leaf sizes there will be some overfitting. To test this, I will use 20 bags in my BagLearner and vary the leaf size like I did for the first experiment.

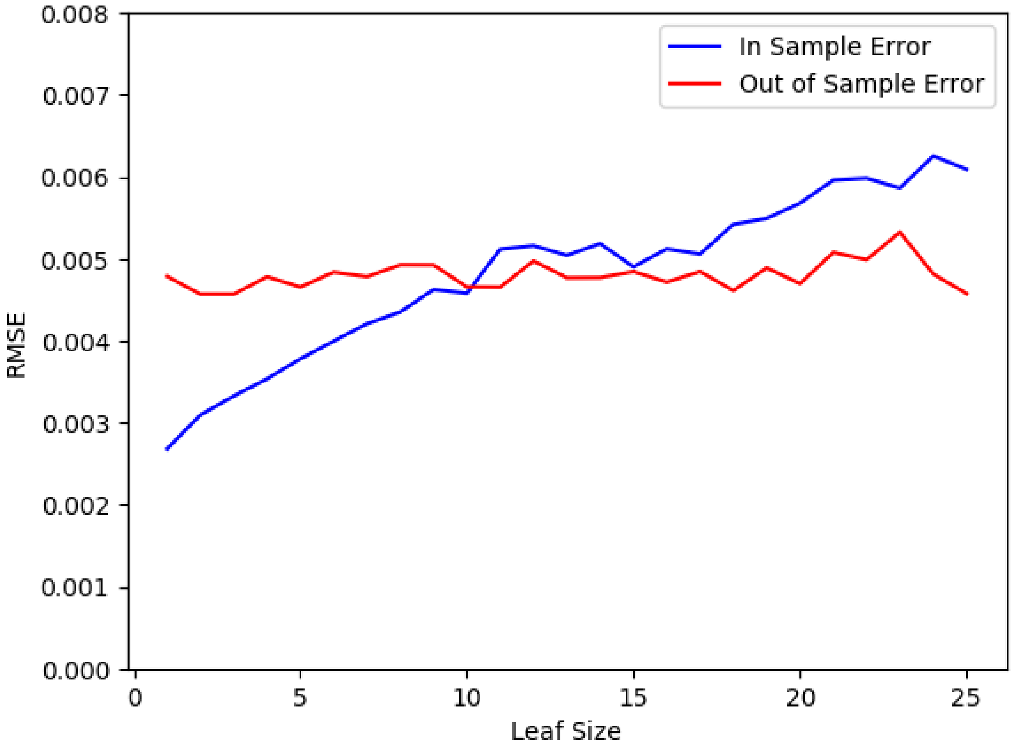

Figure 2: RMSE v Leaf Size for BagLearner using 20 bags on a DTLearner

Mean Difference of RMSE: No Bagging	Mean Difference of RMSE: Bagging
0.000815	0.000726

The chart shows that my initial hypothesis was correct because the overfitting region from the previous experiment is reduced. The interesting point is that leaf size of around 18, overfitting starts to increase again as the in sample and out sample errors start to diverge. This is unlike the previous example where increasing leaf size let to a converging of the errors. The mean difference of RMSEs are also shown and we can see that bagging reduced this difference. Given the chart and differences in error, I do not think bagging eliminated overfitting, but it did reduce it.

3. Comparison of decision trees versus random trees.

To compare the two tree learners I chose to look at the error versus the leaf size, the amount of time it took to add evidence versus leaf size, and error versus different sized training sets. I hypothesize that since the selection of the feature to split on is random it will be worse than the normal decision tree and more variability in changes to leaf size because it is random. I can also foresee that it takes longer to learn and less accurate with different sizes of training sets.

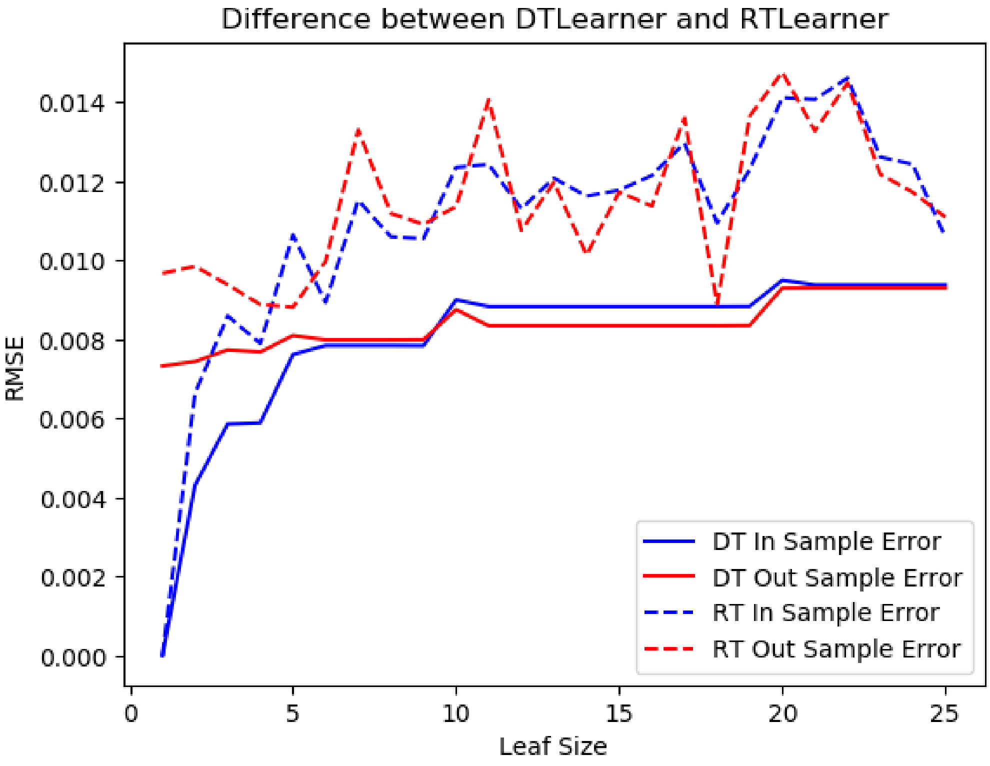

Figure 3: RMSE vs Leaf Size for a Decision Tree Learner and Random Tree Learner

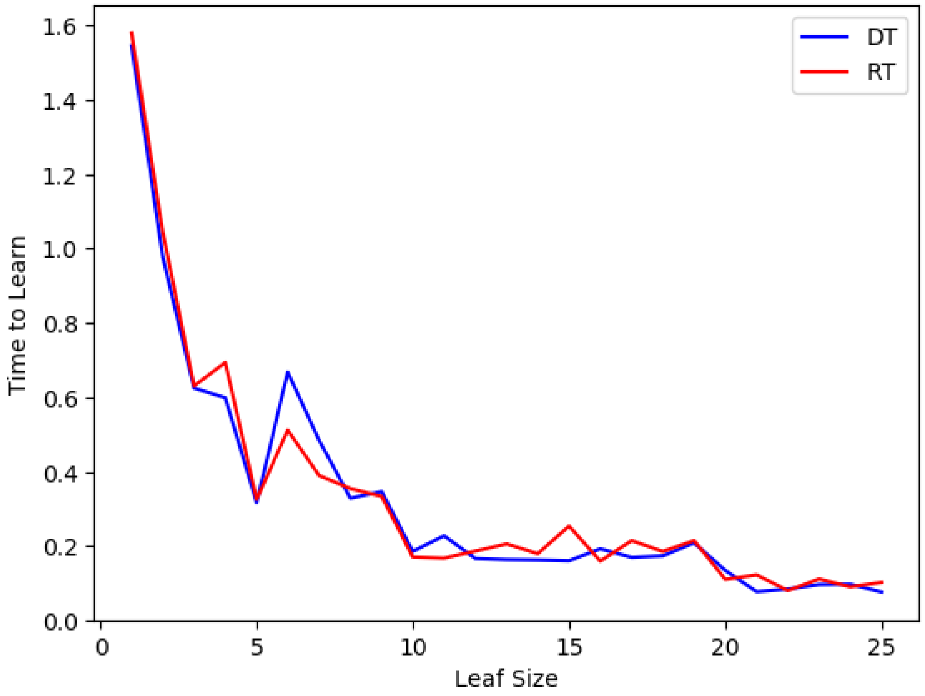

Figure 4: Time to Learn vs Leaf Size for a Decision Tree Learner and Random Tree Learner

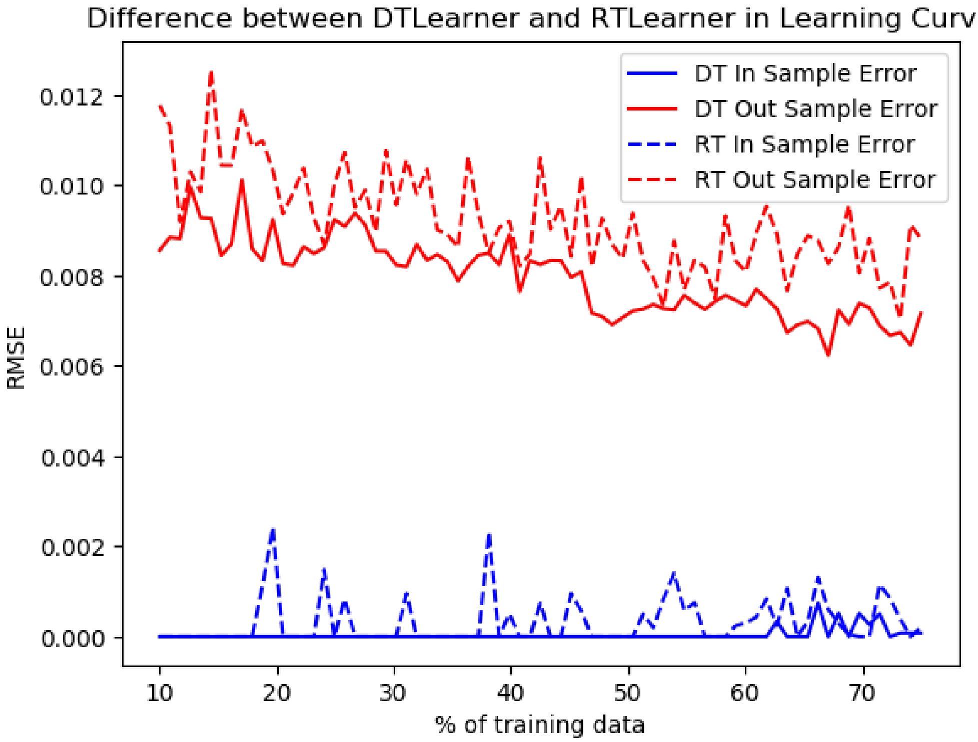

Figure 5: Error vs. % of training data for the Decision Tree Learner and Random Tree Learner

Figure 3 shows that a random tree learner overall has a higher error value and more variability than the best feature decision tree. The time to learn is also not significantly varied as I thought it would be as shown by Figure 4. Figure 5 again helps visualize the fact that the random tree learner has a more variable performance and overall it has a worse performance on it’s out of sample error than the decision tree learner. It is also interesting to see how the more training data is used the better the learner is which makes common sense.