Shu Liu

Gt User: sliu654

Gt id: 903459561

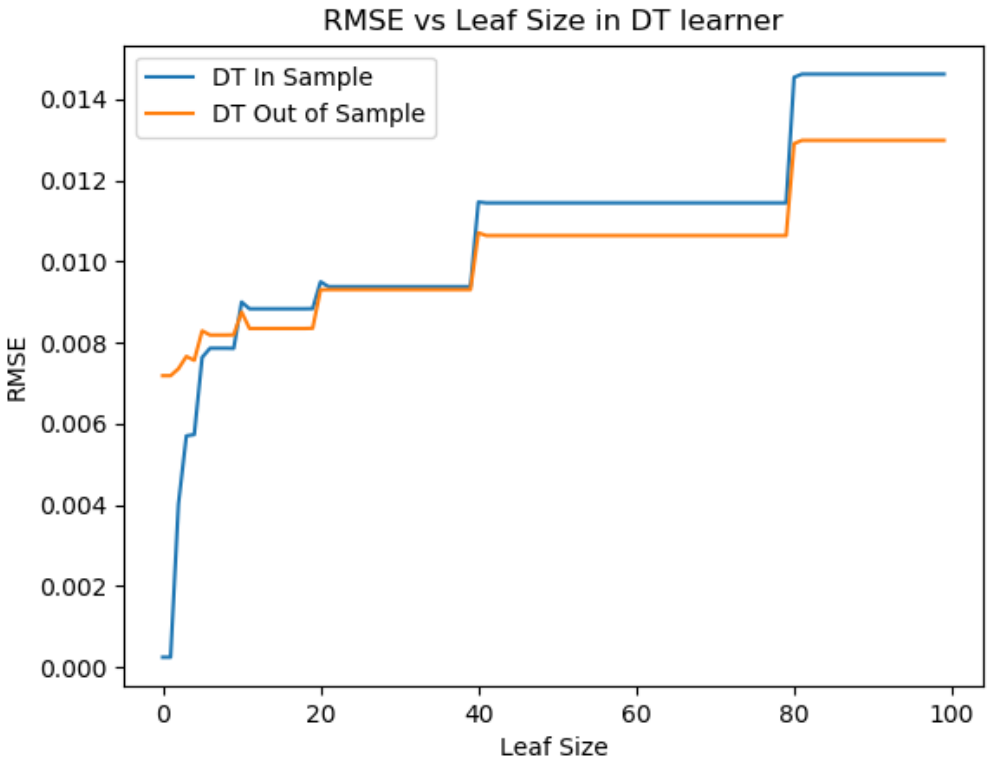
Assess Learners

Q1

Does overfitting occur with respect to leaf_size?

By definition, overfitting happens when the root mean square error (RMSE) is very small in sample but very large out of sample. From Figure 1, we can see that when the leaf_size is smaller than 10, RMSE of in sample decreases significantly but out of sample doesn't change quite much, and that causes the sample to be overfitting. When the leaf_size is larger than 20, the RMSE in sample and out of sample form the same trend, which makes the overfitting disappear.

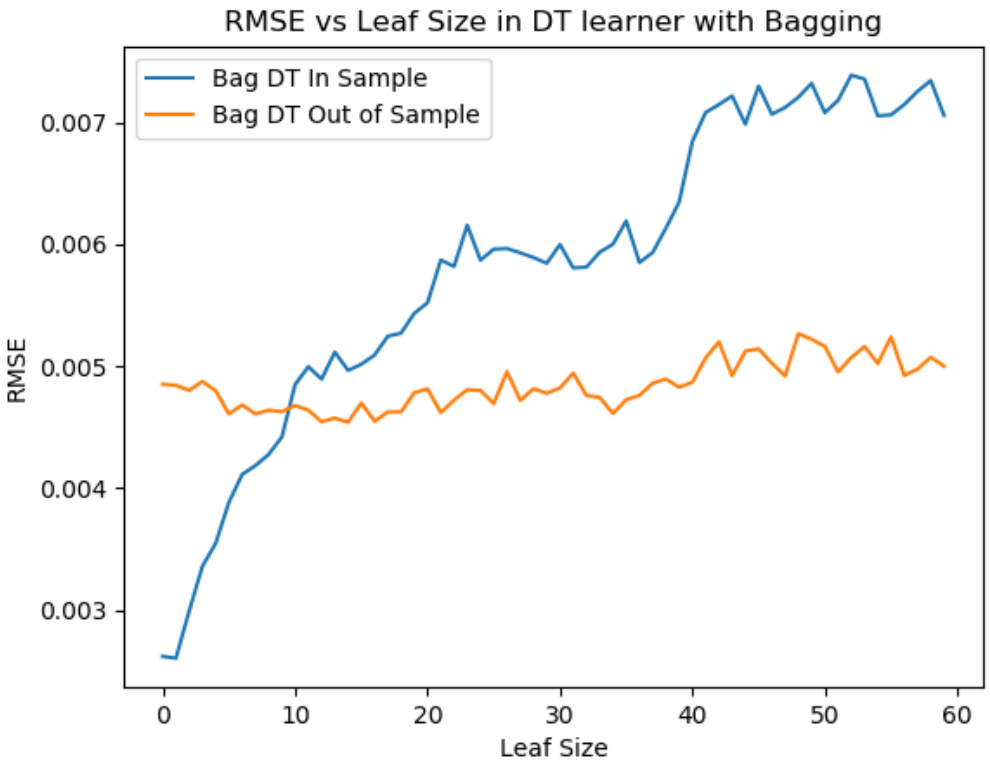
So, we can make a conclusion that using Decision Tree method, the overfitting does occur with respect to leaf_size. And From the graph, it is interesting to point out that after the leaf_size approaching to 10, the RMSE in sample is high than out of sample. I think that's because the data we use to train the sample is not good enough, namely it is not well correlated between each other.



Q2

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

To address the problem, I implement the bagging method on Decision Tree learner with 60 bags. Comparing with Figure 1, bagging makes the two curves smoother as the leaf size increases, but the overfitting still happens when the leaf size is below 10. From the graph, we can see that bagging can reduce the overfitting with respect to leaf size, but it cannot eliminate overfitting especially in small leaf sizes



Q3

Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other?

The mainly difference between the DTLearner and RTLearner is how they select the feature to spit the data. DTLearner calculates the correlation value between the input and output and select one with the largest value to continue build the tree. While the RTLearner just select the feature randomly. From Figure 3, we can see that random learner has higher variance than DTLearner. Namely, the RTLearner has more random curve than DTLearner as the leaf size increases. And for the same leaf size, the RMSE in sample and out of sample using RTLearner is larger than that of DTLearner. Both learner method have strong overfitting when the leaf size is smaller than 5. But overall it seems that RTLearner has less overfitting than DTLeaner.

From the Figure 4, we can see the time difference between DTLearner and RTLearner. When the leaf size is smaller than 20, it takes more running time for RTLearner comparing with DTLearner. And DTLearner shows smoother curve for running time than RTLearner.