Report of Assess_Learners

Course: CS 7646 ML4T
Name: Yichao Zhang
GT User ID: yzhang3414

1. **Does overfitting occur with respect to leaf_size? Use 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).**

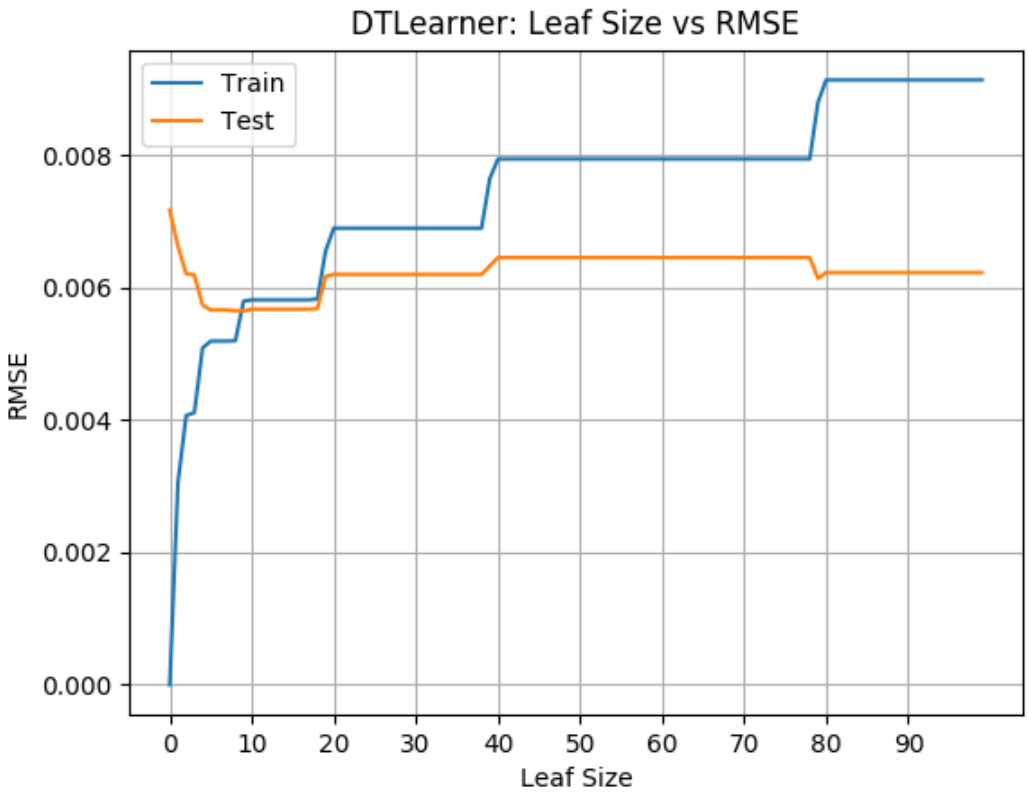

Figure 1

We tune the hyperparameter leaf_size from 0 to 50 and calculate the RMSE of the model on Training data and Testing data. Then make a graph as Figure 1.

Overfitting occurs when leaf_size is small, where leaf_size < 8, we can find from the data that RMSE of Train is smaller than EMSE of Test. That indicates the overfitting.

However, as the leaf_size increase, the RMSE of Test decrease fast and EMSE of Train increase fast. When leaf_size in [9, 18], both RMSE are small and similar. That's the optimal point of leaf_size.

After that, RMSE of Train increases and become larger than RMSE of TEST, however, the RMSE of test doesn't decrease anymore, so too big leaf_size is not needed.

2. **Can bagging reduce or eliminate overfitting with respect to leaf_size? Again use 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 metri**

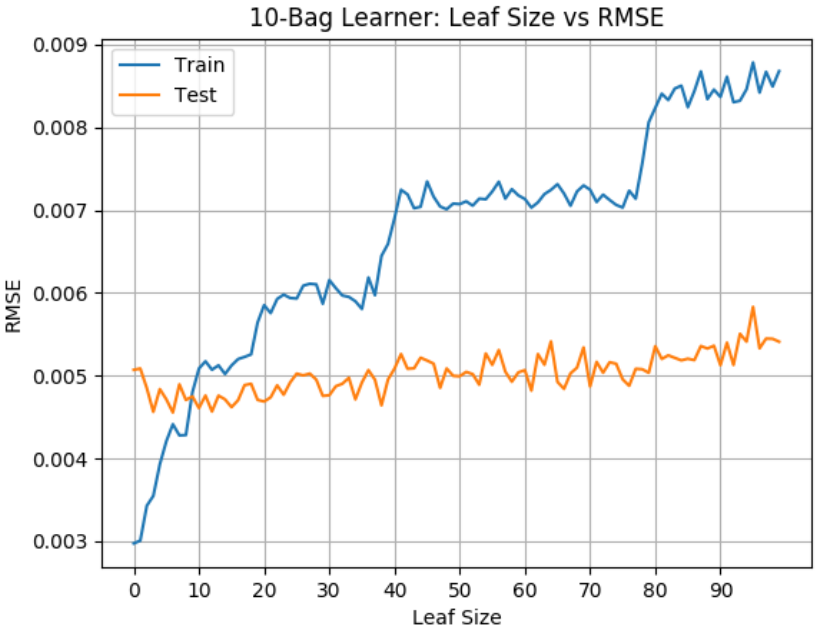

Figure 2-1

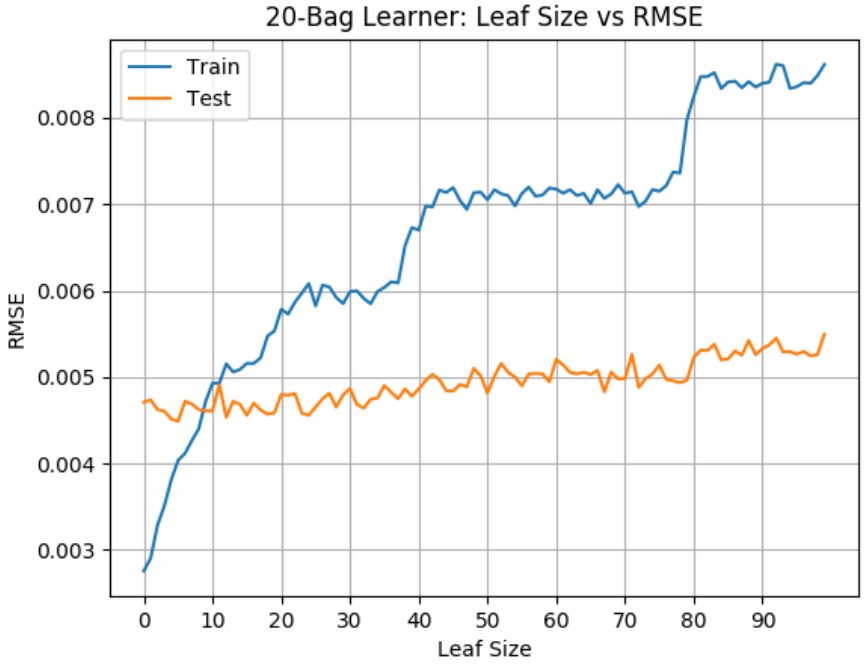

Figure 2-2

After adding the bagging method, we still tune the hyperparameter leaf_size from 0 to 50 and calculate the RMSE of the model on Training data and Testing data. Then make a graph as Figure 2-1 with 10 bags, and Figure 2-2 with 20 bags.

Similarly, overfitting occurs when leaf_size is small, where leaf_size < 8, we can find from the data that RMSE of Train is smaller than EMSE of Test. That indicates the overfitting.

And similarly, the optimal point is on leaf_size in [9, 17]. Too large leaf_size is not needed since the RMSE of Test will never decrease.

Compare with non-bagging learner, the bagging learner has a lower optimal RMSE of Test, which is < 0.005, while the optimal RMSE of Test in DTLearner is about 0.006. That means bagging learner avoids overfitting better than single learner.

Compare Figure 2-1 and 2-2, we can find that 10 bags and 20 bags have similar RMSE. So too many bags are not needed.

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. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, do not use two measures for accuracy.) Note for this part of the report you must conduct new experiments, don't use the results of the experiments above for this.**

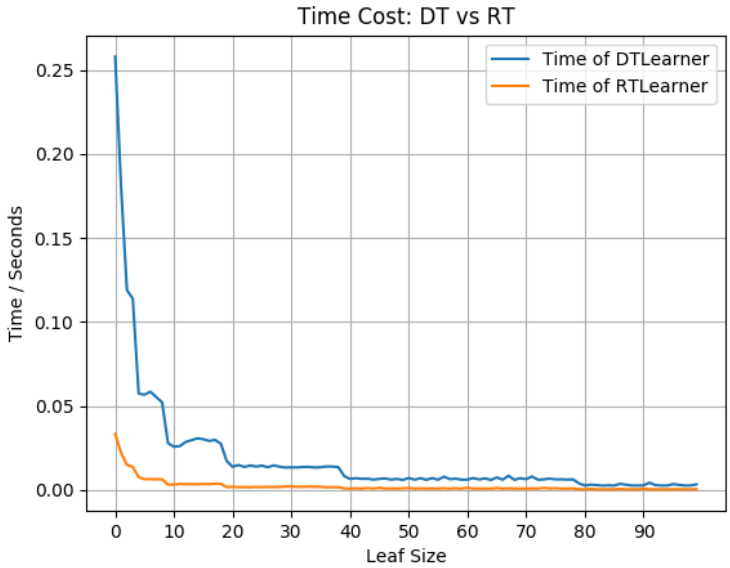

Figure 3-1

Here we compare the running time of DT and RT, we tune the hyperparameter leaf_size from 0 to 50 and record the running time.

From Figure 3-1 we can find that the RTlearner is much faster than DTlearner. So RTlearner is more efficient

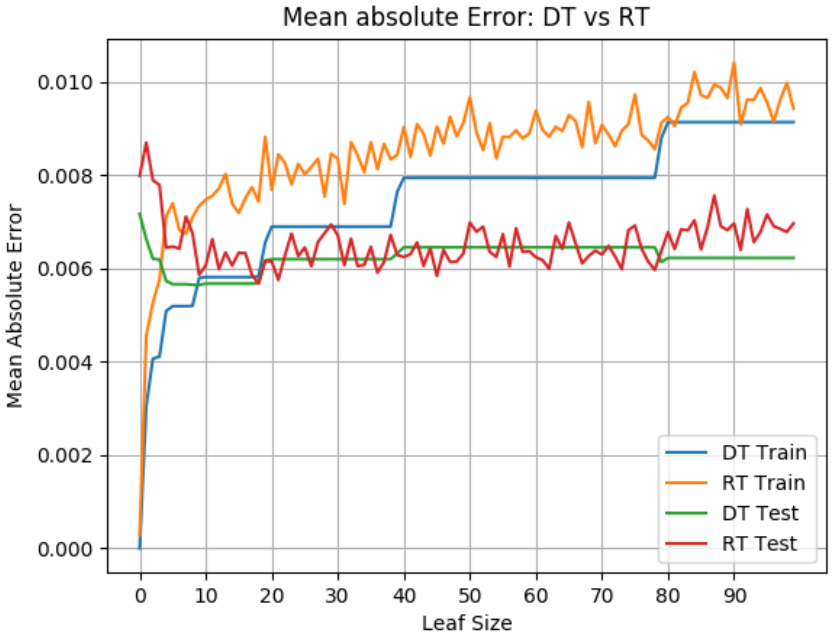

Figure 3-2

Now we compare the Mean Absolute Error time of DT and RT, for both Train and Test, we tune the hyperparameter leaf_size from 0 to 50 and record the running time.

From Figure 3-2 we can find that the MAE of DTlearner is more stable than RTlearner. Both the values are similar in Test.