Report
Assess Learners ML4T 2017
yzhu319
Yuanzheng Zhu (yzhu319@gatech.edu, 902974603)

1.
Method:
Using DTLearner, record the RMSE for both in-sample and out-of-sample data for a given leaf_size ranging from 1 to 50.
Conclusion:
Yes. Overfitting occur with respect to leaf_size. As leaf_size decreases from 100 to 1, the RMSE for in-sample data (orange line) decreases, but RMSE for out-of-sample data (blue line) first decreases then increases.
At about leaf_size = 6, overfitting occurs.

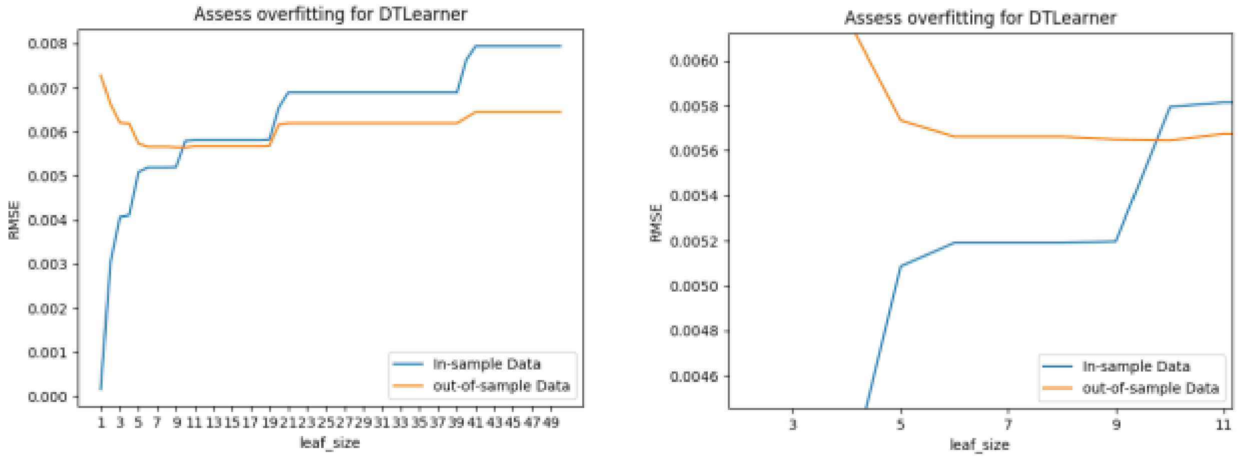

Figure 1. Assess overfitting for DTLearner by testing in-sample and out-of-sample data. Left: leaf size range 1~50. Right: zoom-in at leaf size range 3~10 when overfitting occurs

2.
Method:
Using BagLearner constructed with 200 DTLearner, record the RMSE for both in-sample and out-of-sample data for a given leaf_size ranging from 1 to 100.
Conclusion:
Yes. Bagging can greatly reduce overfitting as leaf_size decreases.
Around leaf_size = 6 (where overfitting occurs in the problem 1), as leaf_size decreases, there is negligible increase in RMSE of out-of-sample data test. This shows bagging can reduce/eliminate overfitting.

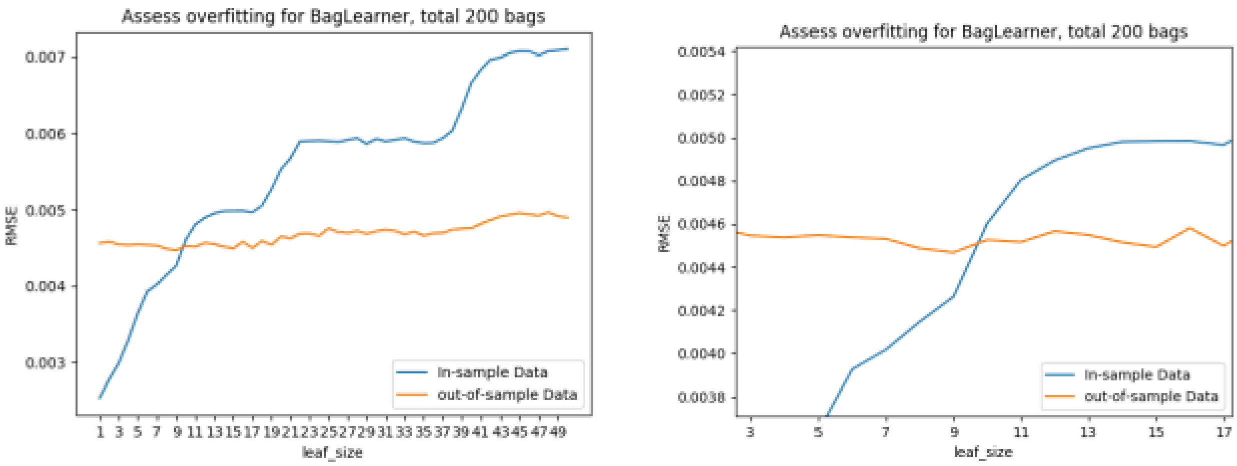

Figure 2. Assessing BagLearner constructed with 200 DTLearner. Left: leaf size range 1~50. Right: zoom-in at leaf size range 3~10 when overfitting occur in the single DTLearner case, in BagLearner the overfitting is eliminated.

3.
Method:
Loop through leaf_size ranging from 1 to 100
Case 1: using 1 DT learner
Case 2: using 1 RT learner
Case 3: using 5 RT and 20 RT learners

Performance:
single random tree perform very poorly compared with a single decision tree with information gain (Figure 3)

Decision tree: select features to split on with information gain, it is an expensive process
Random tree: select features to split on randomly, it is a cheap operation; but if we bag 5 random trees, the time will increase from 0.446 to 5.35s, but noticeable improve on precision (Figure 4)
If we bag 20 random trees, significant impove on precision compared with single DTLearner (Figure 5), but at a cost of computing time increase from 0.446 to 21.33s.

Learner type	time
Single RT learner	0.446 s
Single DT learner	1.56 s
Bag of 5 RT learner	5.35 s
Bag of 20 RT learner	21.33 s

Table 1. Time for a particular learner to test 50 cases from leaf_size=1 to leaf_size=50 (local Linux system)

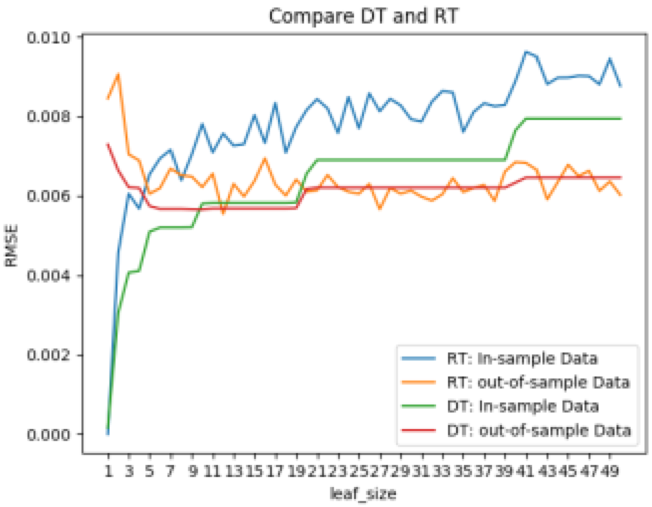

Figure 3. Compare a DTLearner with a RTLearners

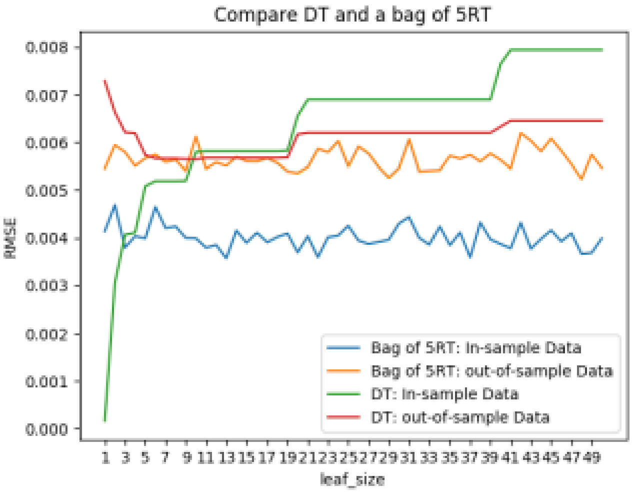

Figure 4. Compare a DTLearner with a bag of 5 RTLearners

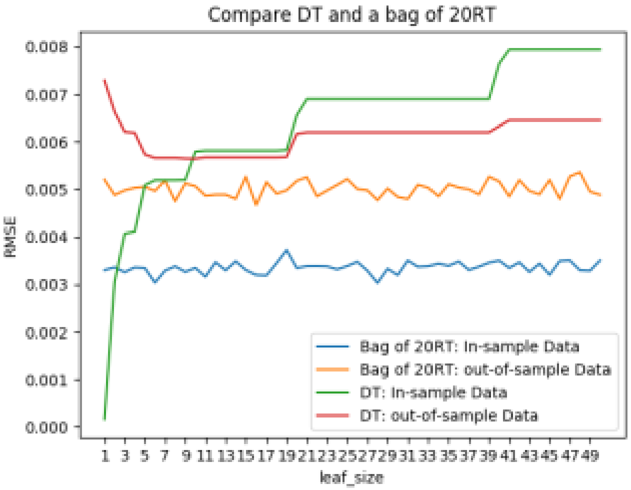

Figure 5. Compare a DTLearner with a bag of 20 RTLearners