Qingyang Li
Report: Project 03 – Assess learners
CS7646: ML4T - Spring 2019
Feb 10, 2019

- 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).

  Answer: Figure 1 shows the in-sample and out-of-sample errors when using decision tree (DT) learner to learn from the Istanbul dataset.

  60% of the Istanbul data were used for training and rest of the data were used for testing. The same training and testing data were used to train and test 80 DT learners. The learners are exactly the same except that their leaf sizes. The leaf sizes are from 1 to 80. The in-sample and out-of-sample root mean squared errors were calculated and plotted for each learner.

  As seen in Figure 1, the out-of-sample errors are the smallest when leaf size is between 6 and 19. As the leaf size decreases (the degree of freedom increases), the in-sample error is always decreasing, and it decreases sharply when leaf size decreases from 10 to 1. The out-of-sample error however, started to increase when leaf size goes from 6 to 1.

  Overfitting happens when the in-sample error is decreasing and the out-of-sample error is increasing. So, Yes, overfitting does occur with respect to leaf size. In this case, overfitting starts to occur when leaf size is 6 or less where the in-sample error decreases sharply as leaf size decreases.

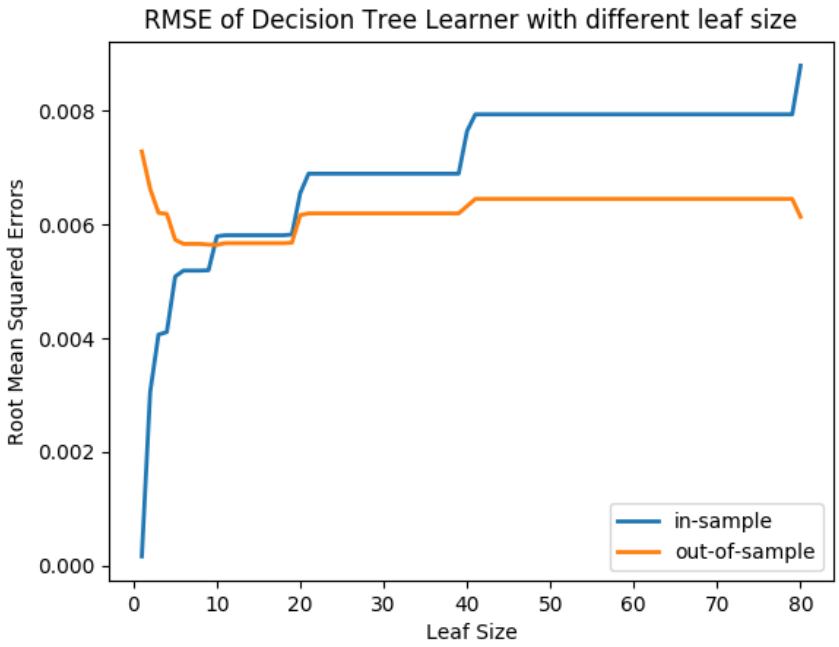


Figure 1: Errors of Decision Tree learners with different leaf sizes

- 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 metric.
  Answer: The same data set used in answering the first question is used here again. This time we feed the same training and testing data to train and test a bagging learner with 20 bags which uses the DT learner as the underlying learner for the bag learner. Leaf size were varied from 1 to 80 for the DT learner used in the bag learner. Figure 2 shows the in-sample and out-of-sample errors for all the learners with Istanbul data.

  As we can see in Figure 2, the in-sample errors (as measured as the root mean squared errors) were not affected by bagging. It still decreases as the leaf size decreases. The decreasing rate of the in-sample errors is similar between the DT learner and the DT learner with bagging (see the blue lines in Figure 1 and figure 2).

  The out-of-sample errors of the bag learners, however, are not changing much with respect of the leaf size. That is, no matter what leaf size is, the out-of-sample error or the testing error is not changing much. Bagging essentially eliminated overfitting, at least for the leaf sizes tested.

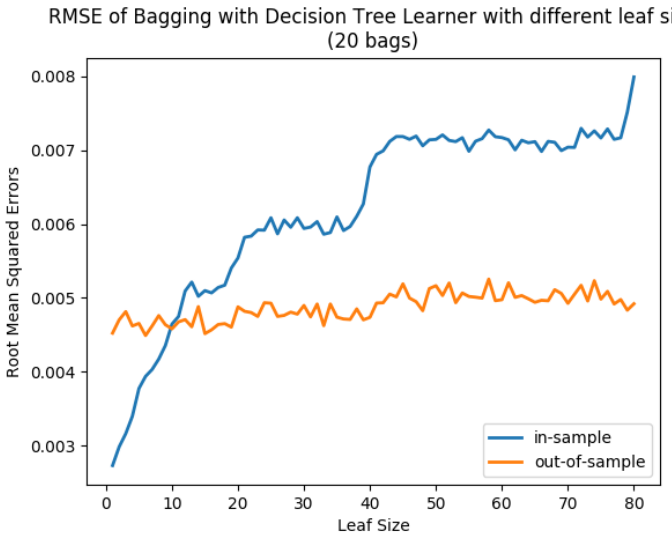
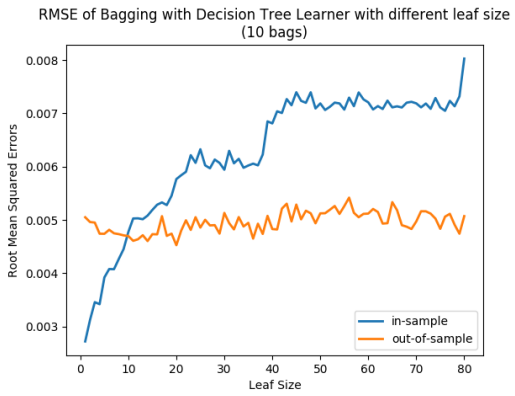


Figure 2: Errors of Bag learners with decision tree learners of different leaf sizes

Now, a natural question is that does the number of bags affect the conclusion. To answer this question, the bag learners were tested with two different number of bags (bags = 10 vs. bags = 100). The result is showed in Figure 3.

With bag learners with 10 bags (Figure 3a) and 100 bags (Figure 3b), the similar trend we see in bag learners with 20 bags were observed. Number of bags did not affect the conclusion that bagging can eliminate overfitting in this case because it stabilized testing error.

a.



b.

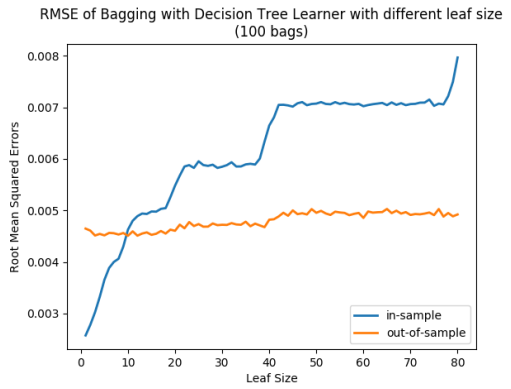


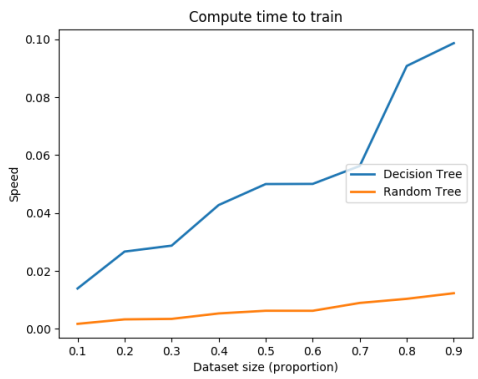
Figure 3: Errors of Bag learners with decision tree learners of different leaf sizes. a) bags = 10; b) bags = 100

- 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.

Answer: I compared the DTLearner and RTLearner for their efficiency for training and testing with respect to the dataset size. I used the Istanbul data, and use 10%, 20%, …, 90% of the dataset to train the DT and RT learner (both have their leaf size set to 6), and then use the rest of the data (90%, 80%, …, 10%, respectively) as testing set to run query the learner. The observed the compute time for training and querying was plotted in Figure 4.

In Figure 4a, the DTLearner runs faster than the RTLearner for training when the same data was given to the two learners. Further, as the data size increases, the DTlearner take more time to train while the training time of the RTLearner stays relatively stable.

a.



b.

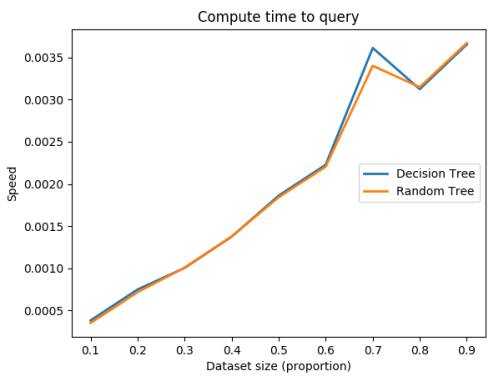


Figure 4: Compute time of DT and RT learners. a) Training time; b) Querying time

As seen in figure 4b, the querying time increases as the input dataset size increases for both the learners. The querying times do not differ between the two learners.