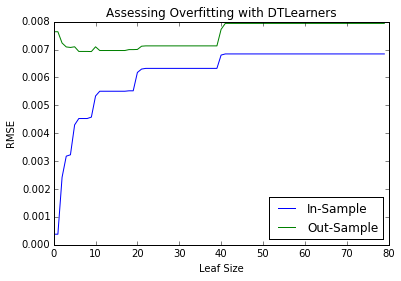
Seungkwan Bryan Baek

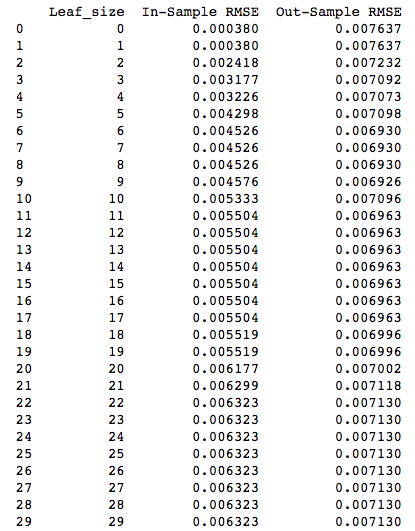
ML4T

Assesss\_Learners Report

**Question 1: Decision Tree Learner**

In the dataset Istanbul.csv with DTLearner, overfitting occurs when the leaf\_size decreases after leaf\_size = 6, where the trend of overfitting for out-of-sample data (slope) is opposite that of in-sample data. Overfitting would occur when the In-Sample RMSE is decreasing and the Out-of-sample RMSE is increasing. Overfitting happens as the leaf\_size decreases. Thus, when looking at the graph shown, it helps to look at the trend from the right to left.

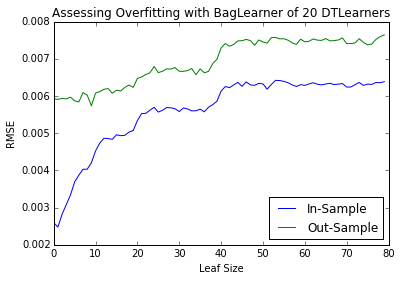
The dataset was randomly shuffled and split into Training and Testing sets. After training the learner on the training sets, I ran the learner on the training set (in-sample) and on the test set (out-of-sample).

To determine where the learner overfits, I used Root Mean Square Error (RMSE) as a metric. Looking at chart and the table, it's clear that In-Sample RMSE is strictly decreasing as leaf\_size decreases, and it makes sense. A leaf size of 1 for In-Sample RMSE is very close to 0 (0.000380) because the data is almost perfectly fit. As the leaf\_size increases, the model becomes more generic and not as precise when fitting the data.

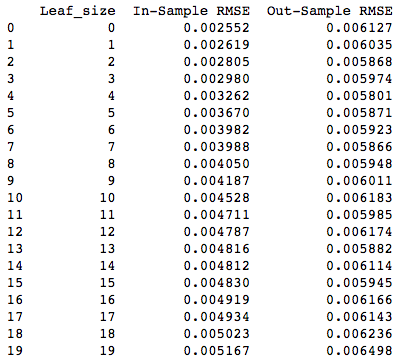
For the out-of-sample trained learner, RMSE trend is similar as that of the in-sample trained learner. Because this is based on out-of-sample data, the RMSE is a bit higher. As the leaf\_size decreases, there is a noticeable diverging pattern. Unlike the in-sample data, the out-of-sample performance is worse. The RMSE actually increase when leaf\_size is smaller than 6. Therefore, the leaf\_size of 1 does not yield a good fit, because the model was built on the train set.

After a certain point (when leaf\_size=40), RMSE for both in-sample and out-of-sample learners increase because as leaf\_size gets bigger, the models become generic.

**Question 2: Bag Learner vs. Decision Tree Learner**

I tested 20 bags of Decision Tree Learners against the Istanbul.csv dataset, while varying the leaf\_size. Bagging reduces overfitting with respect to leaf\_size. Same technique was applied to the ensemble learner of decision tree learners as the Question1.

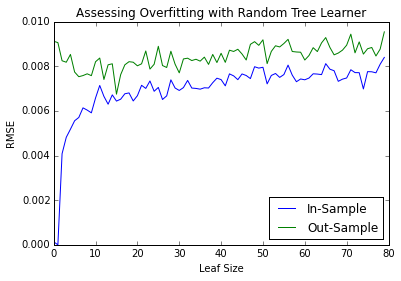
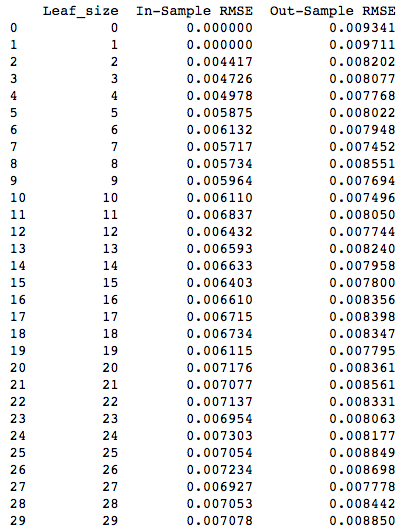
Training an ensemble learner resulted in lower minimum RMSE than just training a single Decision Tree Learner. In this BagLearner, 0.005801 was the minimum RMSE which occurred when leaf\_size was 4. In comparison, 0.006930 was the minimum RMSE when trained on a single Decision Tree.

The Ensemble of learners typically does not overfit as much as any individual learner by itself. Each kind of learner has an own intrinsic bias. But when put together, they tend to reduce the overall individual biases.

Therefore, there is not as much diverging between in-sample and out-of-sample compared to the decision tree. Also, there is no definitive region in which the out-of-sample learner is overfitting. They both have a general decreasing trend of RMSE as the leaf\_size decreases.

**Question 3:** **Random Tree Learner vs. Decision Tree Learner**

Random Tree Learner is based off of Decision Tree Learner. For this experiment, each splitting in the decision tree learner was based on which factor at the time of splitting had the highest absolute value of correlation against the output data. With the random tree learner, however, a factor is randomly chosen at the time of splitting with no regard for the absolute value of correlation.

When the Istanbul.csv was trained with a Random Tree Learner, the performance was weaker than that of a Decision Tree Learner. The lowest RMSE for Random Tree Learner was 0.007496 when the leaf\_size was 9. In comparison, 0.006930 was the lowest RMSE when trained on a single Decision Tree.

Another thing to notice is the general trend of RMSE for the out-of-sample dataset as the degree of freedom (leaf\_size) decreases. Up until the point of diverging, the overall trend of RMSE for the out-of-sample trained learner was constant with the exception of random shocks. When using Decision Tree Learner, there was a clear local min, after which the model was overfitting as the leaf\_size decreased. This indicates that when time/resources/memory is emphasized and the experiment has to be ran with bigger leaf\_size, you are not as concerned about underfitting. In the plot for decision tree learners, the RMSE at the point of splitting is a local min. In contrast, random tree’s RMSE at the point of splitting is not clearly a local min.

The criteria for which the nodes are split is random. Thus, the “randomness” addresses the inherent bias of decision tree learner, and it may hide the bias that gets more evident as overfitting happens. On the other hand, decision tree learner does have a local min of RMSE that is lower than minimum RMSE for a random tree learner. Thus, if the designer is careful about overfitting and there is not as much constraint on time/resources/memory to allow for smaller leaf\_size, using decision tree learner may be better as it may produce a better performing model.