6/8/2019

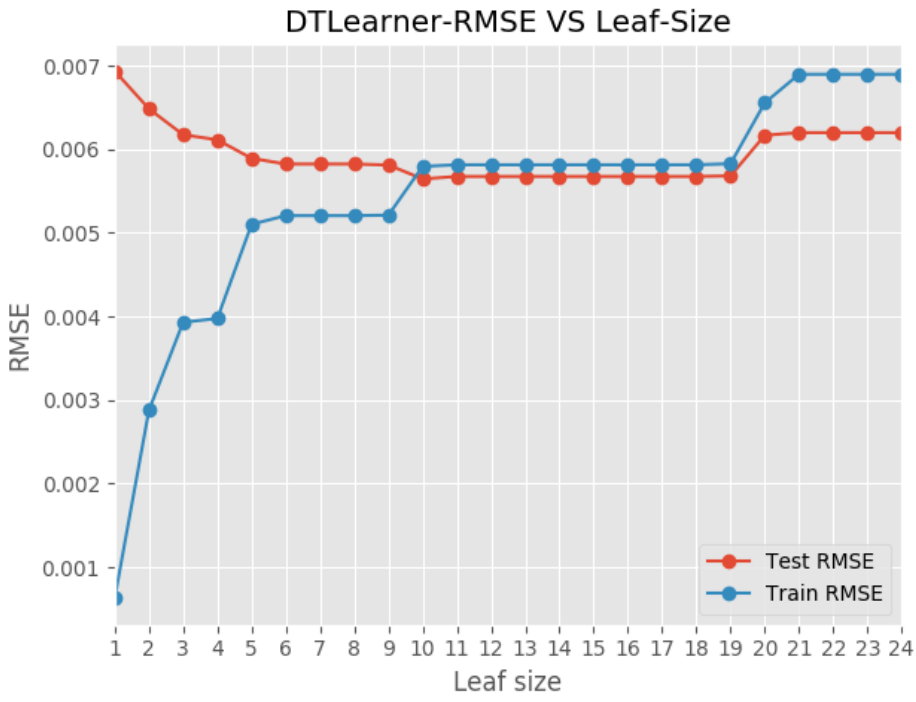
Assess Learners

Machine Learning for Trading

Josh Adams

CS 7646: MACHINE LEARNING FOR TRADING

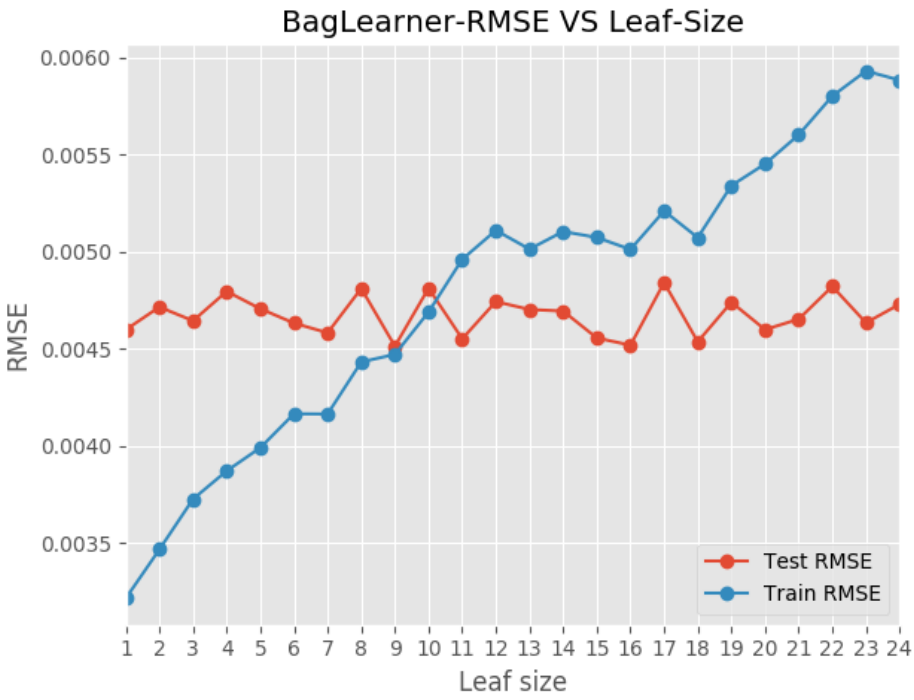
Overfitting according to Wikipedia is "the production of an analysis that corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably". We first train our decision tree using the data in '*istanbul.csv*', then compare the root mean squared error (rmse) with regards to the size of the leaf. We can see a clear correlation between the leaf size and the rmse. I compared leaf sizes from 1 to 24, when the leaf size is 1 that is when there is the largest difference between the rmse in the test and training set. The testing rmse is the highest at 0.007 and the training rmse is lowest at 0.001. As we increase the leaf size the rmse of the two sets begin to converge. The rmse for the testing and training sets cross once we reach a leaf size of 10. The rmse of the two sets stabilize and as such, diminishing returns begin. There is not a notable advantage to increase the leaf size from this a leaf size of 19. In CS7646 overfitting is defined as the point at which the in-sample error is decreasing, and the out-sample error is increasing (03-03 Assessing a Learning Algorithm – 12.Overfitting 1:50-2:10). Overfitting, according to our definition, starts around leaf 10 and steadily gets worse as we approach a leaf size of 1, where the model is almost perfectly fit to the training set.



JOSH ADAMS
6/8/19

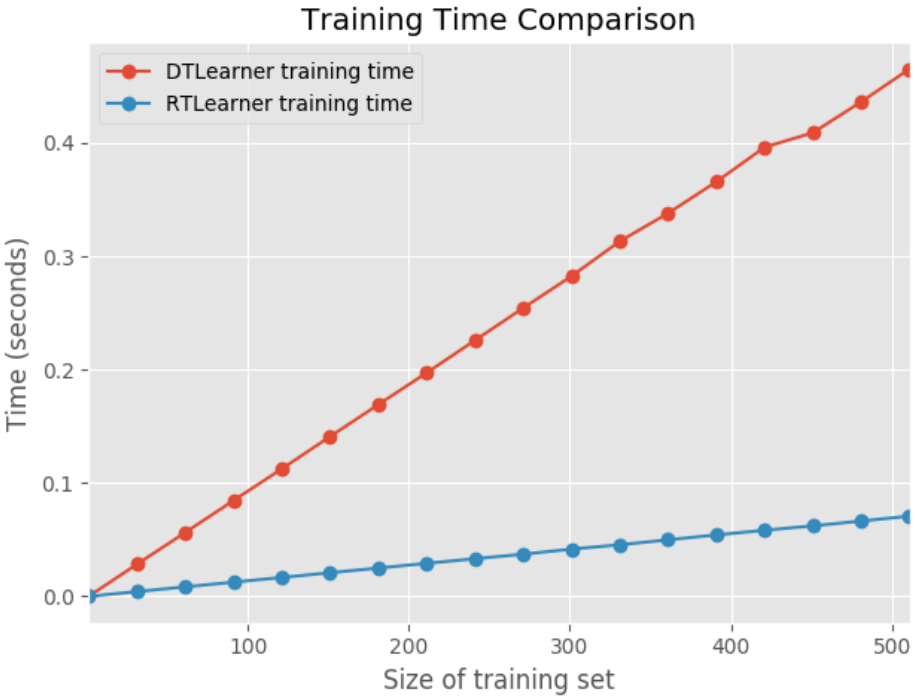
ASSESS LEARNERS

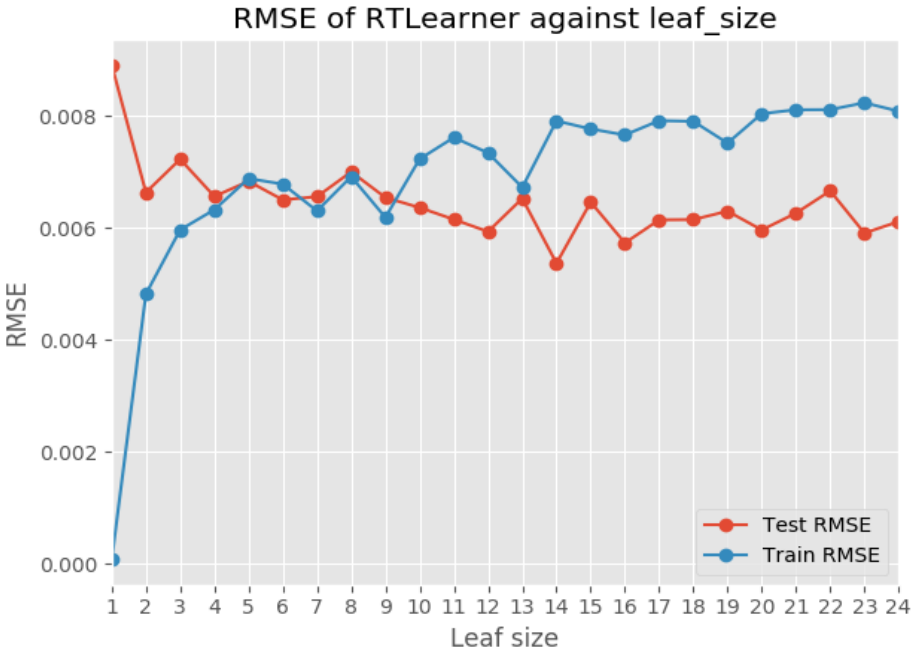
Bagging's reduction of overfitting with respect to leaf size. In bagging we typically use the same learner but train each learner on a different set of data. I am not convinced that bagging will eliminate overfitting, but it is a method to reduce overfitting. When we compare the bag learner to the classic decision tree from the first question, we can see clear differences in performance. In the classic decision tree graph, the testing error has a much larger variance than the bag learner. The classic decision tree suffers from overfitting, but we do not see the same thing occur in the bag learner. The testing error averages around 0.0046 in the bag learner, while in the classic decision tree we can see that the lowest error achieved is near 0.0058. When the leaf size is one in the classic decision tree, we reach a point of being nearly perfectly overfit to the training dataset. When using bootstrap aggregating (bagging), in the same situation where leaf size is reduced to one, the error is still roughly the same as the previous tests. To be able to say that bagging eliminates overfitting would require more testing because I would expect there to be situation where overfitting can occur as m approaches infinity. Mapping the test errors for different combinations of bag and leaf sizes, generates an interesting chart. There is a reduction in error as we increase the bag size. Reducing the leaf size, the changes are not as pronounced and it would appear that the error is reducing as we shift towards the bottom left.



Quantitatively

comparing the random decision tree to the classic decision tree to determine which aspects of the two types of trees are better. I compared the time required to train the random decision tree vs the classic decision tree. The training time for the random decision tree is much lower than the classic decision tree due to not having to calculate correlation. If training time is very important, then using the random decision tree would be the better solution between the two.





When we plot the root mean squared error (rmse) for the random decision tree, compared to leaf size, the random decision tree still is affected by overfitting. Bootstrap aggregating would be a solution for the overfitting of the random decision tree but that is not the focus of this question. When we compare the rmse of the two types of trees we do see that the classic decision tree has lower error but is affected by overfitting sooner than the random decision tree. When considering accuracy, the classic decision tree is better. Both random and classic decision trees have their own set of benefits and should be evaluated based on the needs of the experiment.