MC3 - P1: CS7646 Machine Learning for Trading

Saad Khan (skhan315@gatech.edu)

October 2, 2016

Introduction

The intention of this project report is to experimentally highlight, with the help of graphs, if there is overfitting when the random tree learner and bagging are implemented as part of Mini-Course 3 - Project 1. For this purpose, three different scenarios were tested using white and red wine variants of the wine quality data sets[1] and Root Mean Square (RMS) Error and correlation coefficient values were noted and plotted against varying leaf size and no. of bags. Results are presented in the later sections.

Overview of Overfitting : Predictive models are trained using data that is usually split into 2 sets, a train and a test set. Train set is used to train the model and test is used to verify it to a certain degree of desired accuracy. Overfitting tends to occur when the predictive model fits the training too well, i.e. the model created is over complicated and encapsulates all what is in the training data but fails to do well on unseen examples (test set).

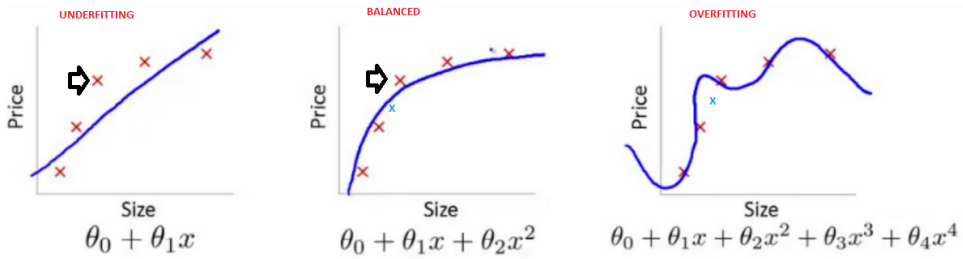

Figure 1: Underfitted model, Balanced model, Overfitted model

An example to illustrate overfitting is shown above in Figure 1. Regression Model created on the left is linear and does not encompass the variance in the data that well. It might fit the training data points that are close to the line but not to the ones away from it (like the one shown with an arrow). This is underfitting (opposite of overfitting). On the other hand, example of overfitting is the model created on the right side, a 4th degree polynomial. Although, it is not biased towards some training points like the previous model on the left, however, it is over complicated and tries to capture all the variance in the training data points. As a result, if a new data point (x:blue) is introduced to be tested, the model would predict poorly, hence it is overfitting in this case. A balanced predictive model, should lie somewhere in between, doing well against both training and testing data, as shown by the 2nd degree polynomial curve in the middle.

Experiments

Experiment 1 : RT Learner (White Wine Dataset) - varying Leaf Size

In this experiment, Random Tree learner (RT Learner) was inputted with data from the white wine dataset with varying leaf sizes outputting RMSE and correlation coefficient which were plotted as shown in Figure 2. Figure 2 [LEFT] shows RMSE from training and test set with variation in leaf size while Figure 2 [RIGHT] shows training and test set correlation coefficient w.r.t leaf size.

Observation : By looking at the plots for both RMSE and corr. coef. it can be observed that overfitting tends to occur for lower values of leaf sizes, highlighted by yellow regions. To be more precise, the range of values of leaf size for which overfitting seems to occur is from 1 to approximately 15. Figure 2 [LEFT] shows that as the leaf size increases, i.e. for leaf size greater than 20, test set fits the tree as well as the training set, apparent by the overlapping training and test set curves. This is not completely but some what indicative in the correlation plot in Figure 2 [RIGHT] and when both plots are analyzed together, the region of overfitting is the roughly the one highlighted in yellow.
NOTE : For smoother curves and reduced randomization, experiments were run 10 times for each leaf size and output was saved in MS Excel for plotting.

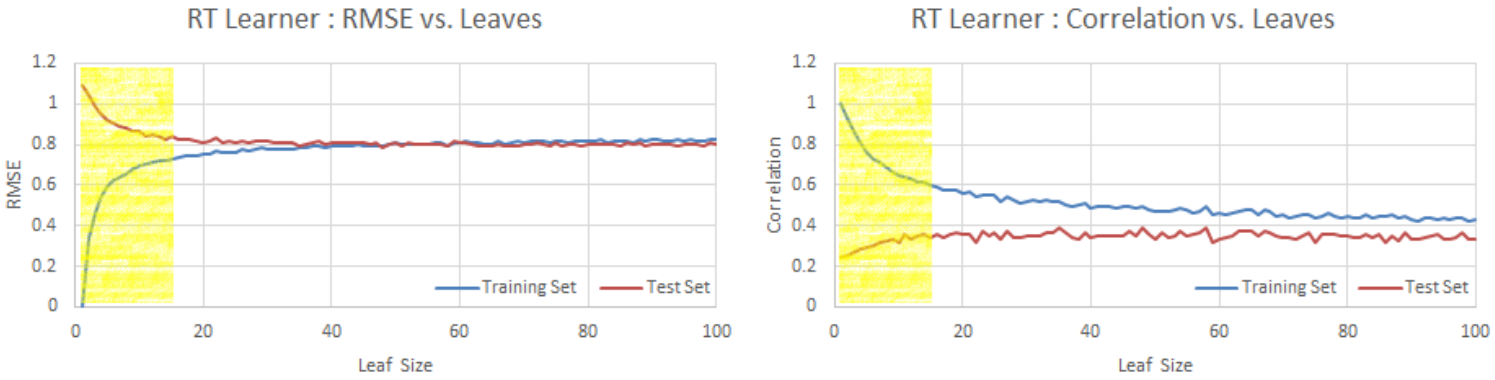

Figure 2: [RT Learner] (LEFT : RMSE vs. Leaves), (RIGHT : Correlation vs. Leaves)

Experiment 2 : Bag Learner (Red Wine Dataset) - varying No. of Bags - fixed Leaf Size

Second set of experiments were performed on the red wine dataset where in addition to performing random tree search, bagging was also implemented. Performance of the random tree was noted with increasing no. of bags and results were plotted for variation in RMSE and correlation as shown in Figure 3. Here the leaf size for the random tree was fixed at 5.

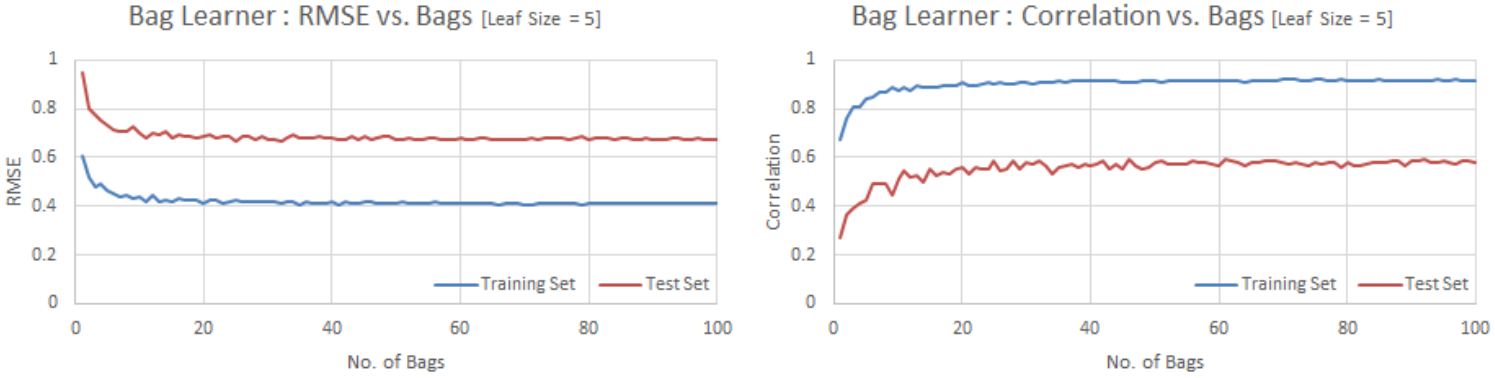

Figure 3: [Bag Learner] (LEFT : RMSE vs. No. of Bags), (RIGHT : RMSE vs. No. of Bags) - Leaf Size = 5

Observation : It can be seen for both RMSE and the corr. coef. curves that level of overfitting seems to be more or less constant throughout with increasing no. of bags. There might be a very slight reduction in the level of overfitting from bag = 1 up until bag = 5 or 6 but overall effect or overfitting is constant. A balanced model in this case would have been where training set RMSE would be very close to the test set RMSE plot (just like in experiment 1), which is not the case here. To investigate this further, a fairly larger value of fixed leaf size = 50 was chosen and the experiments were re-run to see if there was any change in the observation. Results are shown in Figure 4 below and plots are scaled as per the y-axis in Figure 3 for apples to apples comparison.

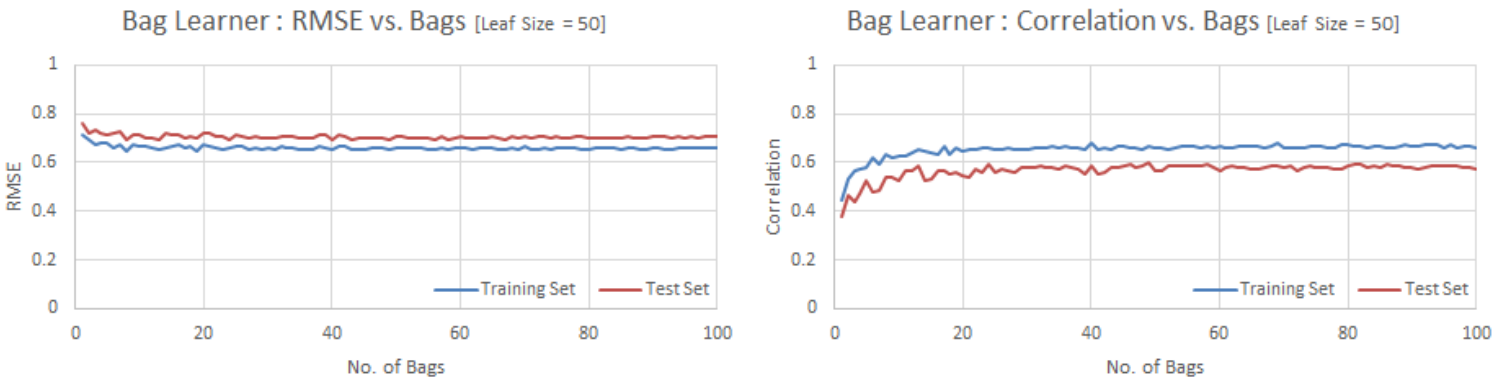

Figure 4: [Bag Learner] (LEFT : RMSE vs. No. of Bags), (RIGHT : RMSE vs. No. of Bags) - Leaf Size = 50

Changing leaf size to 50, drastically reduces the extent of overfitting but the overall gap between the training set RMSE/correla and the test set RMSE/correlation, although very small, is still constant throughout with increasing no. of bags. Another interesting point to note is that making the leaf size bigger has reduced overfitting but increased test set RMSE. Never

the less, based on observations in both Figure 3 and 4, it can be said that bagging is not drastically effecting the level of overfit, rather it is the leaf size which is effecting overfitting.

Experiment 3 : Bag Learner (White Wine Dataset) - varying Leaf Size - fixed No. of Bags

Final set of experiments were performed using the white wine dataset, keeping the no. of bags constant while varying leaf size. At first, no. of bags were kept constant at 5 and results for RMSE and corr. coef. were plotted as shown in Figure 5. Here it can be seen that even with fewer no. of fixed bags at 5, overfitting starts to reduce as the leaf size increases, again highlighted by the yellow region. This is some what consistent with the results of RT learner in experiment 1.

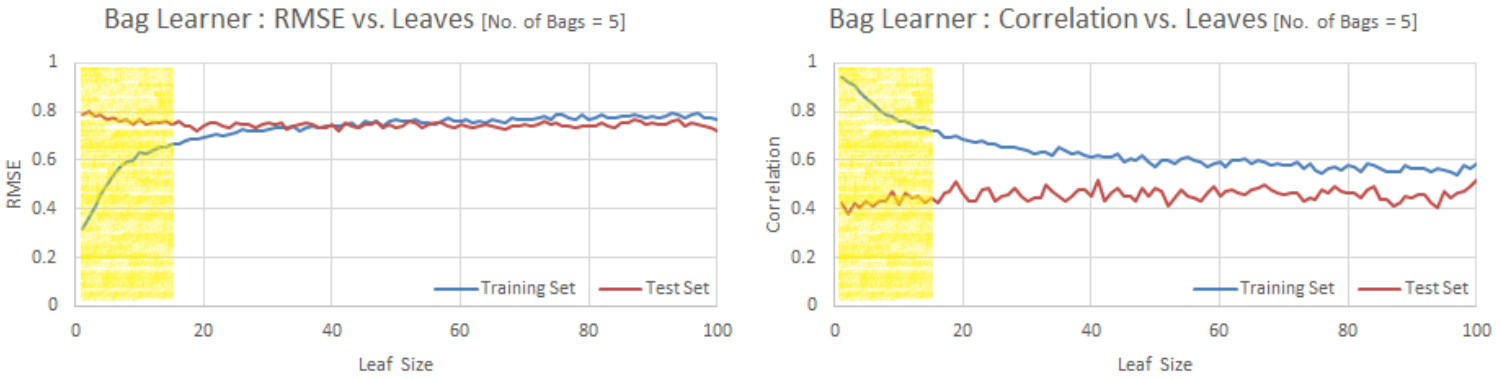

Figure 5: [Bag Learner] (LEFT : RMSE vs. Leaves), (RIGHT : RMSE vs. Leaves) - Bags = 5

Further investigation was also carried out with a larger fixed bag size of 50 and results are shown in Figure 6. It can be seen that even with a larger fixed bag size, leaf size is still commanding the extent of overfitting as the general trend of the curves for RMSE and correlation are almost identical in Figure 5 and 6. The region of maximum overfitting in Figure 6 (highlighted in yellow) is also similar to that in Figure 5 and it can be said that the overfitting region is approximately between leaf size 1 to 15. The most significant effect of increasing the fixed bag size has been the straightening of the test set RMSE/correlation curves, specially when compared to RT learner results in experiment 1 where for lower values of leaf size the curves were slightly asymptotic towards the y-axis and here in experiment 3 they are not. This leads to another interesting observation that overall gap between $Train_{RMSE}$ and $Test_{RMSE}$ has reduced with introduction of bagging, i.e. at leaf size = 1, without bagging $Train_{RMSE}$ - $Test_{RMSE}$ = 1.09, whereas with bagging its 0.47.

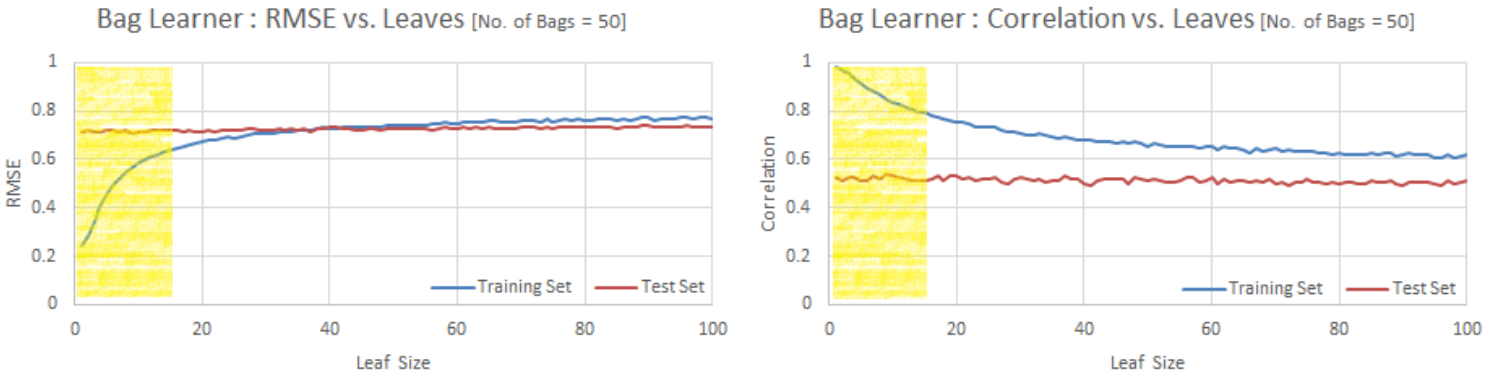

Figure 6: [Bag Learner] (LEFT : RMSE vs. Leaves), (RIGHT : RMSE vs. Leaves) - Bags = 50

Conclusion

Overall this was a very challenging project, if it had not been for some of the fellow course mates on piazza and Professor's suggestion for implementing querying before building the tree, I would not have made it so far. I learned a lot about implementing ML models from scratch as part of this project.

References

[1] **Wine Quality Data Set**

https://archive.ics.uci.edu/ml/datasets/Wine+Quality