Assess Learners

Jenny Goldsher
jgoldsher3@gatech.edu

***Abstract*—**The following investigation aims to assess the effectiveness of various tree model classifiers. The investigation will address comparisons between Decision tree and Random tree models as well as cover the topics of overfitting and bagging. Root-mean squared error (RSME), leaf size, bag size, training time, and coefficient of determination (R-squared) were used to compare the effectiveness of the various tree models.

Assess Learners

INTRODUCTION

When creating and refining machine learning models, it is important to make comparisons across models to determine their relative effectiveness and best use-cases. Depending on the goals and limitations of the data investigation, different models may be more or less appropriate. In the case of the present investigation, we will look at the effectiveness of Decision tree models, Random tree models, and Bagging models using Decision trees.

Overfitting is a major consideration when training a machine learning model. While machine learning algorithms aim to bet represent data, if they account for variance in the training set rather than actual trends, they can mis-represent the dataset. With this in consideration, it is important to split data into training and testing sets and compare error figures to examine if and to what magnitude overfitting is occurring in a given model. Comparisons across models can also be made when examining overfitting by altering parameters to see when overfitting occurs in various models.

When computation time is relevant in picking a model, it is important to consider the methodologies needed to construct a given model. Relevantly, Decision tree models test the correlation amongst a group of columns at each node whereas Random tree models select columns to split at random. Since selecting a column at random requires minimal computation, I hypothesize that Random tree models will train faster than Decision tree models at the expense of prediction accuracy.

Using Bagging as a modelling technique entails sampling the training dataset with replacement to create a collection of models, then using the mean of their resulting outputs as a final model. The combination of sampling with replacement and using the mean is theoretically effective in minimizing the effects of overfitting by reducing the inclusion of modelling sample variance. Bagging models can use different tree models with various leaf sizes in the creation of the final model. Compared to a single tree model, I hypothesize the effects of overfitting are to be lesser but will not completely eliminate overfitting. Despite this, one limitation is that bagging models are more conceptually and computationally dense, making them harder to explain and having higher computation times. The present study aims to investigate the characteristics of various tree models using python implementations and the Istanbul dataset.

METHODS

Experiment 1: Investigating overfitting

To investigate the relationship between leaf size and overfitting in decision trees, I computed RSME in-sample and out-of-sample across various leaf sizes. The cleaned Istanbul dataset (dates and column headers removed) was used to train (60%) and test (40%) across leaf sizes. To assess the occurrence of overfitting, in-sample and out-of-sample RSME was tracked to see exactly where overfitting occurred. The leaf sizes tracked were 1, 5, 10, 20, 30, 40, and 50. At each leaf size, a new instance of a Decision Tree learner was created, trained, and applied to the testing dataset and RSME was calculated.

Experiment 2: Bagging and overfitting

In an attempt to look at the relationship between bagging and overfitting, the BagLearner class was instantiated across various leaf and bag sizes while tracking in-sample and out-of-sample RSME. Each Bag Learner class used Decision tree models at leaf sizes of 1, 5, 10, 20, 30, 40, and 50. The Bag sizes tested were 10, 20, 30, 40, 50, and 100. RSME was tracked across bag and leaf size for comparison to determine if and where overfitting may have occured.

Experiment 3: Decision trees versus random trees

To examine the differences between Decision trees and Random trees, training time and coefficient of determination were calculated for each model across leaf sizes. At leaf sizes 1, 5, 10, 20, 30, 40, and 50 instances of decision trees and random trees were created, trained, and tested. At each increment of leaf size, the time it took to train each model was tracked and the resulting coefficient of determination was calculated.

DISCUSSION

Experiment 1

The point at which overfitting occurs can be determined as the out-of-sample error (RSME) begins to increase as the in-sample error continues to decrease across degrees of freedom. In the present study, degrees of freedom were relative to leaf size. As seen in Figure 1, out-of-sample error begins to increase as leaf size becomes greater than 40 while in-sample error continues to decrease. Based on this data, we can conclude that overfitting does occur with respect to leaf size.

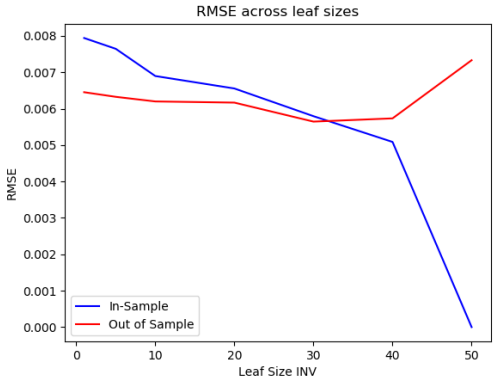
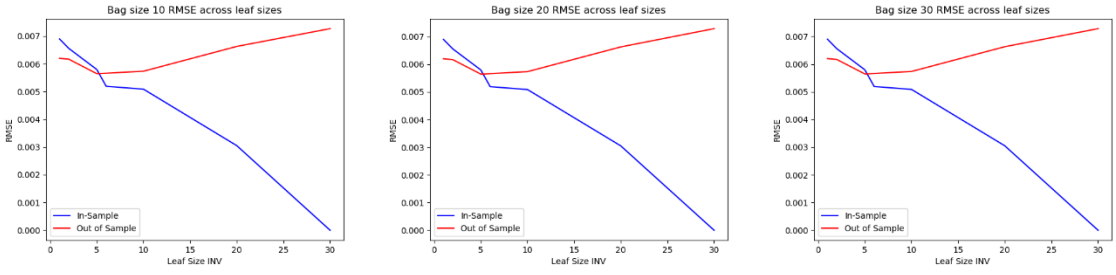


Figure 1—The figure above shows the RSME of Decision Trees across degrees of freedom (INV of leaf size).

Experiment 2

Across bag sizes, overfitting begins to occur after at leaf sizes greater than 10. Compared to a single decision trees, where overfitting occurs with leaf sizes at 40, this allows for more precise distinctions to be made throughout the tree building process. Further, where decisions grouping at a minimum 40 values to a grouping in a single decision tree could be made, decisions grouping a minimum of 10 values can be made without resulting in overfitting. Although this minimizes the effects of overfitting, overfitting is not completely eliminated when using bagging. As seen in the figures below, while bagging has a general effect on overfitting, bag size did not have a meaningful effect on overfitting.



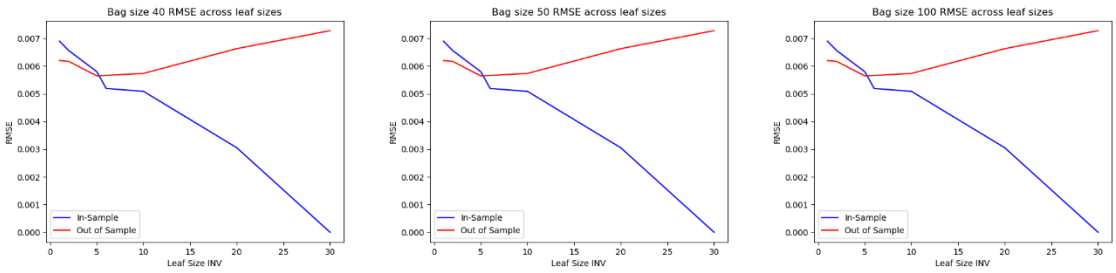


Figure 2—The collection of figures above show the outputs of RSME across leaf sizes separated by bag size.

Experiment 3

Two additional quantitative measures to compare decision trees and random trees were computed: training time and the coefficient of determination. Across all leaf sizes, the Random tree model has a smaller training time than the Decision tree model. This difference is at its greatest at smaller leaf sizes. The difference in training time is likely due to the computational time needed for decision trees to calculate the correlation between each column and the response variable, whereas the random tree simply selects a column. This is congruent with the initial hypothesis that Random Tree training would be quicker than Decision Trees.

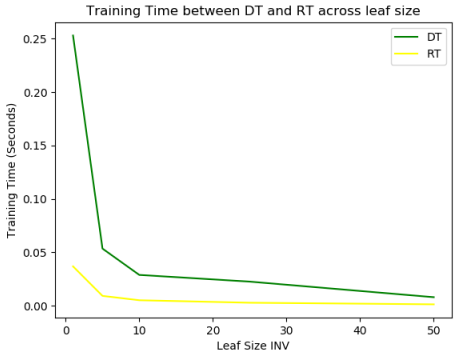


Figure 3—The figure above shows the training time between Decision tree and Random tree models at various leaf sizes.

The coefficient of determination encapsulates what proportion of the variation in the predictions is accounted for by each tree. Having a greater coefficient of determinations indicates a better modelling process, in our case indicating that whichever tree has a higher coefficient of determination more accurately modelled the dataset. As seen by figure 4, at leaf sizes of 5 or greater the decision tree clearly accounts for more variability in predictions. Since decision trees choose splitting columns based on correlation rather than randomness, this is congruent with our hypothesis that they would better model the data.

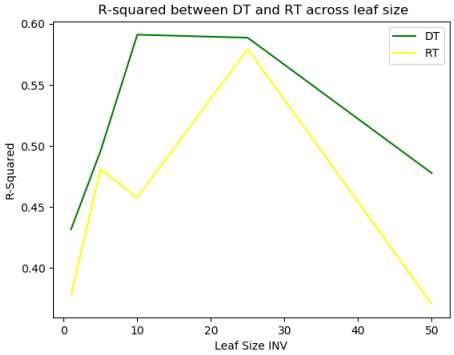


Figure 4—The figure above shows the R-squared (coefficient of determination) values between Decision tree and Random tree models at various leaf sizes.