ASSESSING LEARNERS

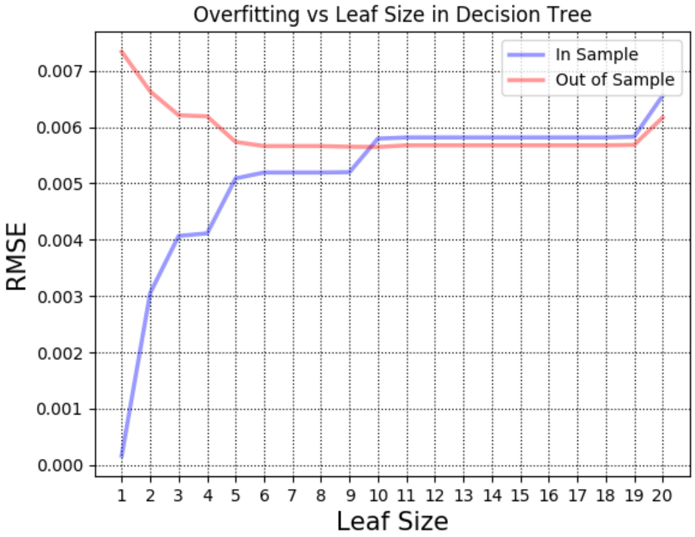
Allan Reyes, reyallan@gatech.edu

INTRODUCTION

Decision Trees are a powerful type of Machine Learning algorithms that allow solving both classification and regression problems by creating models that are more intuitive to the human designer than more mathematically-oriented ones like Neural Networks. There are different variations of these trees: classic decision trees, random trees, bagged trees, among others. Each one of them has its advantages and disadvantages, and sometimes are more appropriate for certain classes of problems. To assess them quantitatively, different experiments were performed to compare characteristics such as overfitting, generalization and correlation. In this report a description of these experiments, their results, and the conclusions derived from each one of them, are presented.

OVERFITTING IN DECISION TREES

For the first experiment the effects of the size of the leaves, in a classic decision tree, on overfitting were studied. 20 learners were trained on 60% of the Istanbul data each with a different leaf size ranging from 1 to 20. After each training session, the in-sample and out-of-sample errors were computed using RMSE[1] as the metric. The following figure shows the results from comparing both errors for each of the learners.



As it can be seen from the graph, there is a clear manifestation of overfitting. Overfitting can be identified as the region where the in-sample error decreases, but the out-of-sample increases. In this particular case,

[1] Root Mean Squared Error

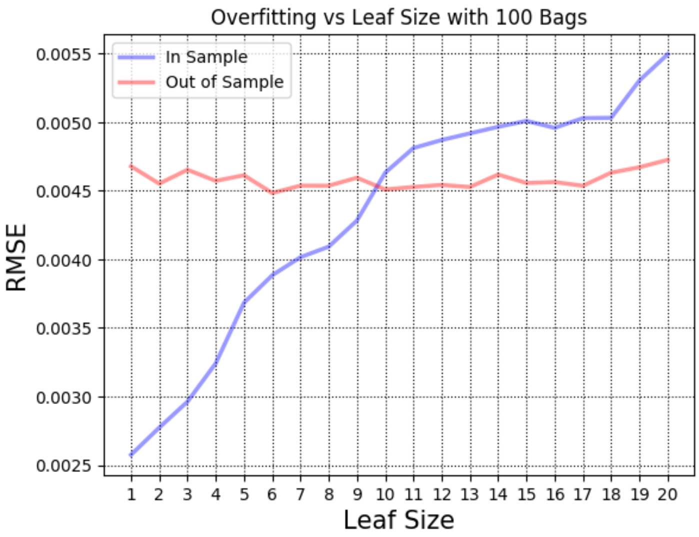
it can be observed that as the leaf size gets smaller, overfitting starts to occur. Concretely, the algorithm starts failing to generalize with a leaf of size 3.

This behavior is expected because as the size of the leaf decreases, the algorithm is forced to keep partitioning the samples into smaller groups until either the desired size is matched or a unique label is encountered. For example, in a decision tree with a leaf of size 1, an extreme case would be partitioning the samples such that each one of them is grouped in its own node. The training data would be fit perfectly, but the model would completely fail to generalize as new samples are not going to have the exact same characteristics. This can be directly observed in the graph with the lowest in-sample error and highest out-of-sample error occurring with a leaf of size 1.

Finally, it is important to note that underfitting is also observable in the graph. Once a tree gets constructed with a leaf of size 19 or more, the generated model is not able to fit the data correctly anymore; shown as an increase in both in-sample and out-of-sample errors. This is due to the fact that more samples are grouped together, but they not might be completely related or similar.

OVERFITTING IN BAGGED TREES

The second experiment involved studying the effects of the size of the leaves on overfitting too, but using bagged trees instead of classic decision trees. As in the previous experiment 60% of the Istanbul data were used for training 20 bagged learners, each with a fixed number of 100 bags[2] and varying leaf size from 1 to 20. Since bagging involves some randomness when sampling the data, a seed of 0 was used to allow for reproducibility. The results of comparing all learners using in-sample and out-of-sample RMS errors are shown in the following figure.



Analyzing this graph, it can be noticed that, in contrast with the first experiment, no overfitting is present in the learned model: when the in-sample error starts decreasing, the out-of-sample error does not increase. The immediate conclusion is that bagged trees definitely helped remove the overfitting problem that classic decision trees have as the size of the leaves start to decrease. The reason behind this behavior

[2] The number of bags was chosen arbitrarily

is that, though internally each bag contains a tree that completely fits the data with nodes of potentially a single sample, the aggregation of these many trees helps smooth the predicted value; allowing the model to generalize better.

There is another aspect of this experiment that is worth noting. Observe how the in-sample error quickly grows as the size of the leaf increases. Internally each tree is underfitting the data, a behavior that was identified from the previous experiment, and the bagging increases this effect by smoothing the result predicted from each one of them. In other words, the model is not fitting the training data enough to predict more accurately.

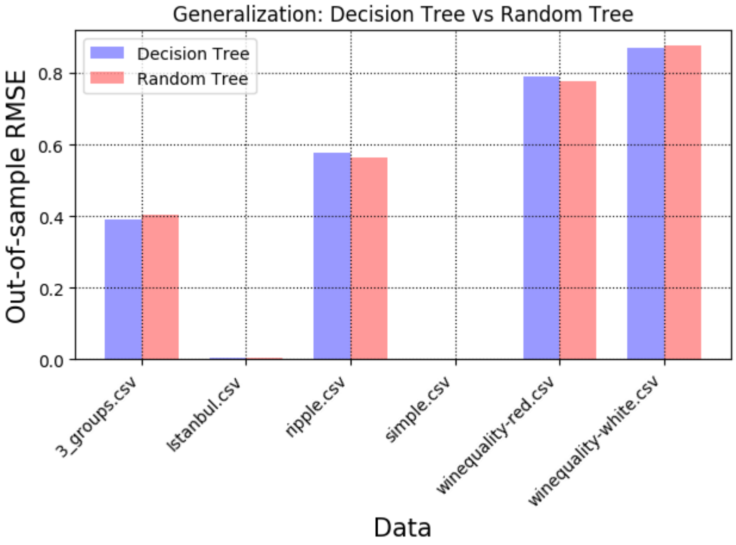
However, it can be seen that the out-of-sample is not actually increasing, in fact, it stays practically the same independently of the size of the leaf. This indicates that the model is not underfitting completely. The effect of aggregating the predictions from all the trees, even if the result is not fitting well the underlying data, allows the model to generalize appropriately and predict with high accuracy the values of unseen samples. This leads to the possible conclusion that the in-sample error increase might be due to the data not being evenly distributed among all the bags which is an expected behavior when bootstrapping, i.e. sampling with replacement.

COMPARING DECISION TREES AND RANDOM TREES

The third experiment involved a quantitative comparison between classic decision trees and random trees. Concretely, three aspects were used to evaluate these two learning algorithms: generalization power, correlation and overfitting effects. In all of the sub-experiments performed, two learners, one classic tree and one random tree, were trained with each of the datasets available for this project. As with the other experiments, 60% of the data was used for training and to ensure reproducibility a value of 0 was used for the random seed. The next sections describe the specific details of each of the sub-experiments.

GENERALIZATION

To evaluate the generalization power of each algorithm, the out-of-sample error was computed after each training session for both of the trees. As with all previous experiments, RMSE was used as the metric. The following bar chart compares the results after training with all the datasets.

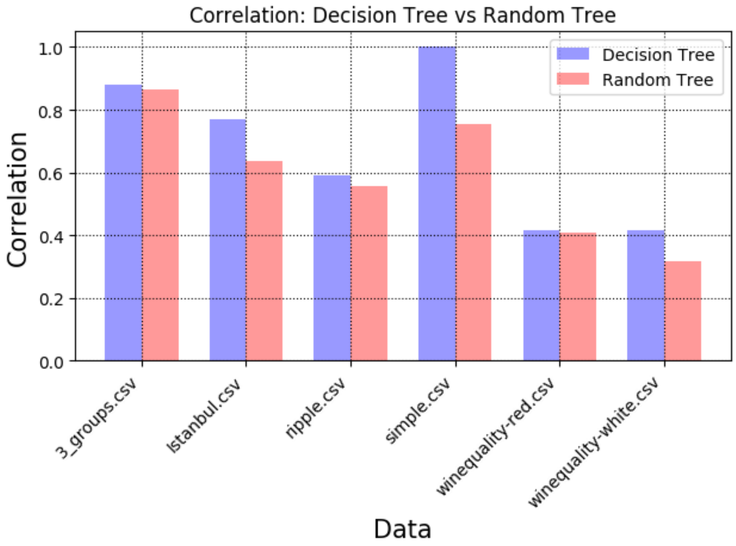


The first thing to notice is that the out-of-sample error of these two learners varies widely between datasets. This is an expected behavior of any Machine Learning problem since the characteristics of the data influence greatly the success of an algorithm. In particular, notice how both kinds of trees do a bad job of predicting the quality of both white and red wines.

The second fact that can be derived from this graph is that neither of the two types of trees is particularly better than the other one. For all datasets, both models have a similar out-of-sample error. The similarity between these two algorithms (the only real difference is their choice of feature: best vs random) might explain why their generalization power is also very similar.

CORRELATION

Correlation was the second aspect evaluated for these two types of trees. A similar approach was followed as with the previous sub-experiment; however, instead of using RMSE as the metric, the correlation between the predicted values and the actual values was computed. This correlation was also done out-of-sample, i.e. using the testing data. The results are shown in the following figure.

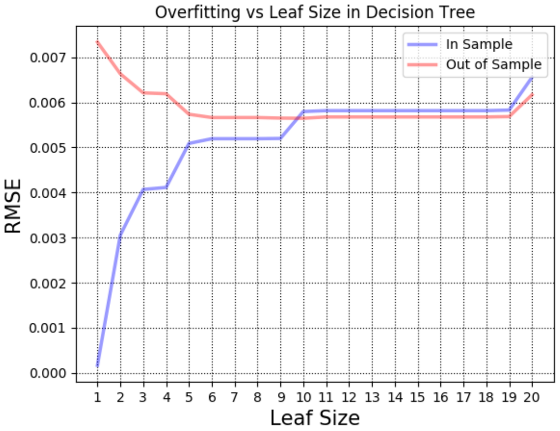


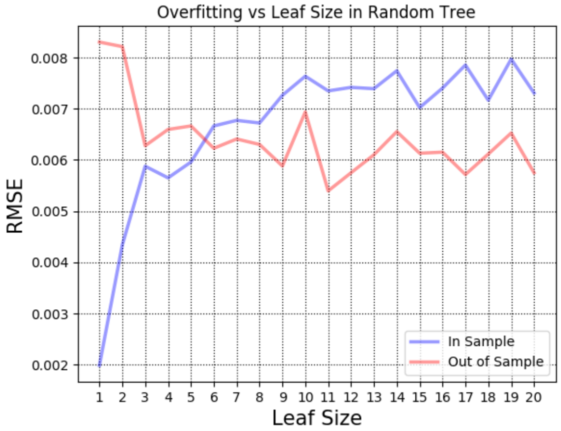
From this bar chart, it can be observed that decision trees provide a clear advantage over random trees when evaluating correlation. The model learned by the decision tree was able to outperform the random tree model in all of the datasets. Notice how for the Simple dataset, the decision tree is able to achieve a perfect correlation in contrast to less than 80% by the random tree.

These results could be explained as exposing the key internal difference between the two implementations of the algorithms. For the classic decision tree, correlation was in fact used as the metric for determining the best feature to split on which of course leads to a model that will predict values that will be intrinsically correlated with the target values. If other metric had been used, e.g. entropy or Gini index, the results might have been different. Experimenting with other metrics for the classic decision tree was out of scope for this report.

OVERFITTING

The last aspect evaluated for this experiment was the effect of the size of the leaf on overfitting for both types of trees. The results for classic decision trees had already been obtained from the first experiment described at the beginning of this report. So, to compare the two algorithms, the same test was run using a random tree: the learner was trained with 60% of the Istanbul data and the in-sample and out-of-sample errors were collected. The following two figures show the results of the effects on overfitting for various leaf sizes for both trees side by side.





Analyzing the two charts it is clear that both types of trees suffer from overfitting as the leaf size decreases. Interestingly, both models start overfitting the data as soon as the size of the leaf is 3; as it can be seen in both graphs by the region where the in-sample error decreases but the out-of-sample error increases. This leads to the conclusion that though random trees try to reduce some bias by introducing that randomness, the construction of the tree is inherently the same as with classic decision trees and thus suffers from the same overfitting problems.

It is important to note, however, that the randomness of this type of trees might help maintain some degree of generalization power because there is no clear indication of underfitting, at least until leaves of size 20, since there is no region where both error metrics increased; in contrast with the behavior noticed in classic decision trees with leaves of size 19 or more.

CONCLUSION

In this report, different types of decision trees were evaluated. The results obtained from these tests showed that the size of the leaf has a clear effect on overfitting as soon as it decreases towards 1 affecting both classic and random trees. Furthermore, both types of trees were shown to be equally able to generalize; however, classic trees were found to generate models more correlated with the data though it might be an expression of the correlation metric used internally. Finally, the results demonstrated the capabilities of bagged trees to eliminate overfitting and maintain a consistent generalization power independently of the size of the leaf. This shows, at a very high level, why bagged algorithms like Random Forests are being used with great success in different industries.