Project 3: Assess Learners

Nathan Riojas  
nathanriojas@gatech.edu

***Abstract—***Tree learners can vary in their implementation and behavior. In this paper, decision trees of varying leaf size are investigated to study overfitting, and bagging is performed to assess its effect on overfitting reduction. Further investigation is also done to compare decision tree and random tree algorithms based on performance and accuracy.

# Introduction

This paper investigates the behavior of decision trees, random trees, and ensemble learners over the course of three experiments with the Istanbul.csv dataset.

## Note on Regression

While tree learners are often used to solve classification problems, the implementations here solve regression problems. As such, the mean of training y values is used when data is aggregated to create leaves. However, were this a classification problem, the approach would call for the use of the mode instead of the mean.

## Learner Summary

Decision trees were implemented using an algorithm developed by JR Quinlan which recursively builds a tree given training data. This type of tree creates nodes by selecting the median of the feature having the highest correlation with the training y values, and continuously does this until the recursive splits arrive at a leaf node. The built tree could then be queried to give the expected result for an instance or several instances of testing data.

Random trees were built in a similar manner as decision trees, except the feature chosen when selecting the median to create a node to further branch the tree was randomly selected, not requiring the correlation calculations of each column with the training y data. This resulted in the tree being built differently than it was for a decision tree. Querying random trees was the same process as it was for decision trees.

Ensemble learners were implemented using boostrap aggregating or bagging. These learners were given an input number of bags or instances to make of a specific learner. In this investigation, the size of the data the bag learner passed to a specified bag (commonly referred to as n’) was equal to the training data size n.

The data passed to the learner within the bag however was created by randomly sampling the input training data with replacement. Thus, when querying the bag learner, it queries the multiple learners or bags inside it, gathers the predicted values of each, and then returns the mean of said values.

## Insane Learner Quick Note

This was a learner that was not studied in further detail but is worth discussing briefly. The insane learner was a bagging learner that implemented 20 bag learners each having 20 bags of linear regression learners.

## Leaf Size and Recursion

Decision Tree and Random Tree learners were further characterized in this study by the size of leaf specified. Leaf size defines the point at which the remaining data is aggregated into a leaf. Thus, the training y values are averaged to produce the value the built tree would return when queried.

It is important to note, in the actual implementation of these tree classes, when recursively creating the tree, leaf size is what is used as the based case to terminate recursion.

Two additional cases were considered. The first being the case in which all y values of a dataset were the same, in which case further recursion was not necessary so a leaf was made and returned. The second was the case in which the median of the feature chosen to split did not adequately split the data. For decision trees, the next best correlated column was used instead, or if no columns could produce valid splits, a leaf was created and returned. For random trees, a leaf was created and returned as well.

## Overfitting Definition

Before detailing the experiments, overfitting must be understood, as this is how decision trees and bag learners were studied. Overfitting is the behavior seen when in sample error decreases as out of sample error increases. This behavior is not desired so determining regions in models where this is present helps in knowing when and how to apply these learners.

## Hypothesis

With regards to decision trees, it is to be expected that overfitting will be inversely proportional to leaf size. This is because the lower the leaf size, the more likely the leaf aggregation values will tend to match the training y values in a one-to-one manner. However, it is expected that bagging will be able to aid in reducing overfitting with respect to decisions trees as leaf size varies.

In comparing random trees and decision trees, the expectation is that decision trees will be more accurate and consistent than random trees. The logic is simply that decision trees carefully select the nodes they create to make the next branch or node in the tree using correlation, so this should produce less error in the tree built. However, random trees exist for a reason, so it is expected that they will run faster than decision trees to some degree.

# Methods

Three experiments were done to study these learners. All investigations were done comparing a specific metric against the learner in question for different leaf sizes.

## Experiment 1

In the first experiment, 100 decision trees were trained and tested. The same data was used to train each of these trees, however, each tree varied in its leaf size (from 1 to 100). The metric studied here was the root mean square error or RMSE, which was calculated using the formula below.



This was done for both in sample and out of sample data by querying the trained tree with training and testing data, and subsequently measuring the RMSE of the output against the expected training and testing results, respectively.

Since RMSE is a metric that is useful for determining the overall goodness of fit of a model, to better understand the accuracy of decision trees, the 100 trees of varying leaf size were plotted against their RSME. This provided a better understanding of how leaf size affected accuracy, in sample and out of sample. From this plot, the region of overfitting was able to be determined.

## Experiment 2

This experiment was virtually identical to experiment 1, with the caveat that instead of a single decision tree at each leaf size, 20 trees or bags were created. This was the application of a bagging learner. The default number of bags selected in this experiment was 20, however, this number was arbitrary and could be increased or decreased as desired.

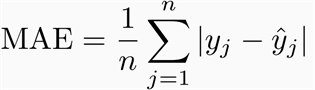
As an example, in the instance of a leaf size of 50, a bag learner was created with the specification of making 20 bags or instances of decision trees. Each tree would then be initialized with the hyperparameter of leaf size set to 50.

The data each decision tree within the bagging learner was trained on varied however as this is how bagging works. The data given to the bag learner to train on was randomly sampled with replacement and then given to each bag instance of the decision tree to train on and develop its tree. The size of data randomly sampled was specified to be the same as the data input. The variation that occurred was a result of sampling with replacement, meaning duplicate values of data could occur. Hence, the 20 bags of decision trees were trained on different versions of the training data.

The same style of plot was generated to study bagging learners and overfitting: RMSE vs leaf size.

## Experiment 3

This experiment diverged from the study of overfitting and instead aimed to compare decision trees and random trees. Two metrics were used to accomplish this. The first being the Mean Absolute Error (MAE) as calculated below.



Compared to RMSE, MAE more directly represents the cumulation of error. While mean square error tends to heavily penalize large prediction error, MAE treats all error the same. With this in mind, the MAE for expected y values versus predicted y values of decision tree and random tree models was calculated in sample and out of sample and plotted against leaf size.

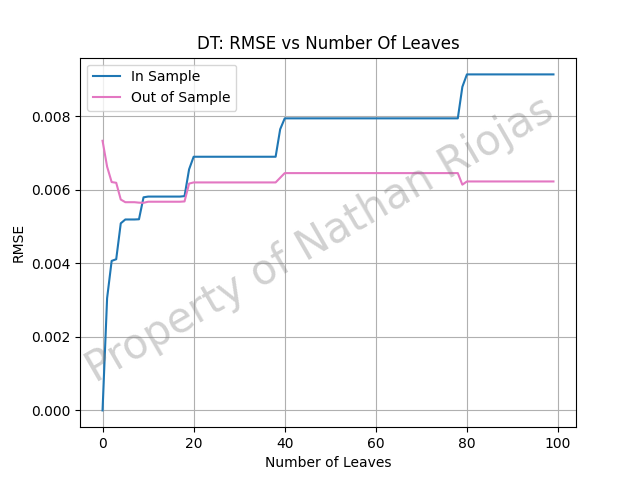
The second metric used to compare decision trees and random trees was time, specifically training time. Perhaps the biggest disadvantage to using tree algorithms for machine learning, is the time it takes to train in comparison to say a linear regression model. Querying time is actually pretty fast, assuming the tree is able to remain relatively balanced. It is therefore of great interest to understand ways to improve the training time, especially if exceptionally large datasets are to be used.

In order obtain this metric, Python’s time module was used to take the time before and after executing the training function (called add evidence in this paper’s implementation) and subtract the two to get the total training time. This was plotted against leaf size for both decision trees and random trees.

# Discussion

## Experiment 1

Upon plotting RMSE vs leaf size for decision tree learners, as seen in Figure 1, it is apparent that the initial hypothesis regarding overfitting was correct. Overfitting is inversely proportional to leaf size. Based on the graph it can be seen that overfitting occurs for leaf sizes less than about 9.



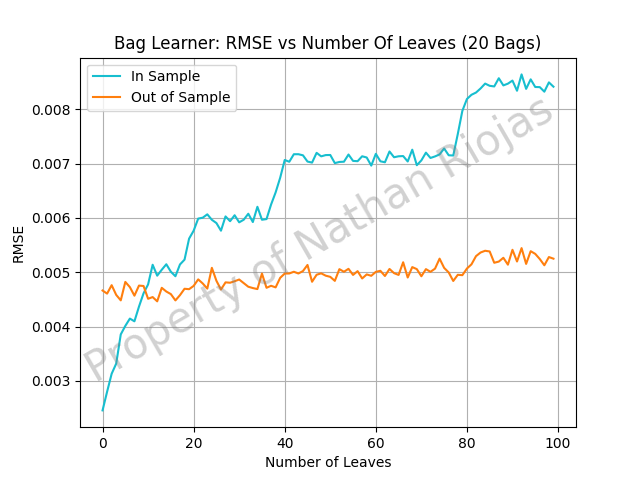
1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).

Figure captions should be placed beneath the corresponding figure, indented 1″ on the left and right sides. The label for the figure, e.g. “Figure 1,” should be set in bold italics followed by an em dash, and the entire caption should be 8.5 points with 14 points of line spacing. The *Figure Caption* paragraph style in Word will number your figures automatically. If need be, you may have one caption corresponding to multiple consecutive figures and use either locational descriptors (e.g. “top left,” “middle”) or labels (e.g. “A”, “B”) to map parts of the caption to parts of the figure. Make sure that caption falls on the same page as the corresponding figure or table; you may need to rearrange text to make this work.

In Microsoft Word, you may need to either change the image’s text wrap settings to “Top and Bottom” or change the line spacing of the image to 1.0.

## Experiment 2

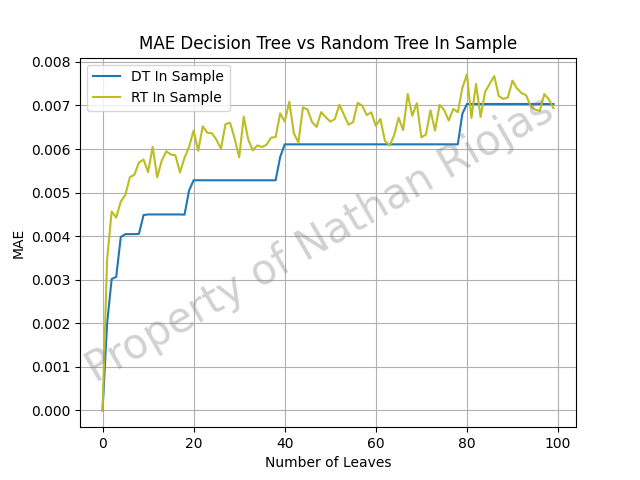
Interestingly, when plotting RMSE vs leaf size for the bag learners, as seen in Figure 2, it appears that the initial hypothesis that overfitting can be reduced using bagging was incorrect, or if correct, correct only to a marginal degree. Overfitting can be observed prior to the point of intersection of the in sample and out of sample lines. This appears to occur up to a leaf size of around 9, the same as was observed in Experiment 1 for decision tree learners.



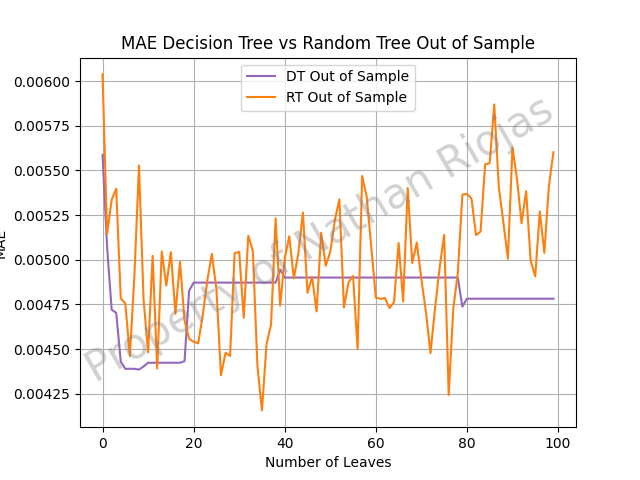
1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).

## Additional elements

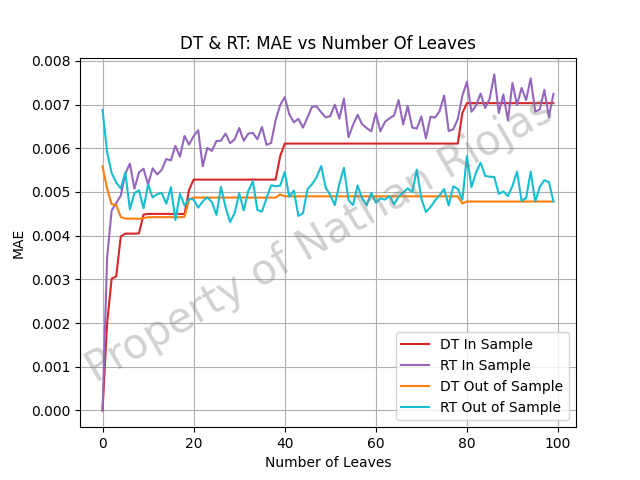
Figure 3, Figure 4, and Figure 5 illustrates the comparison of the MAE of decision trees and random trees in sample and out of sample versus the leaf size. As expected, the random trees produce a highly scattered error trend, though it does seem to follow the same trend as decision trees.



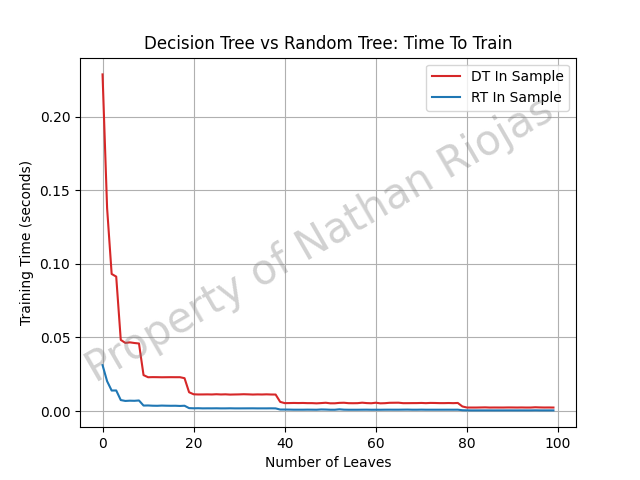
1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).



1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).



1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).



1. Make sure your flowcharts are more useful than this one. Source: [XKCD](https://xkcd.com/1195/).

If you would like to quote an outside source, you may do so in quotation marks followed by a citation. If a quote is fewer than three lines, you may write it in-line. It is acceptable to replace pronouns with their target in brackets for clarity. For example, “Heavy use of peer grading would compromise [the school’s] reputation” (Joyner, 2016). If a quote exceeds three lines, you should set it as its own paragraph with 0.5″ side margins, using the *Blockquote* paragraph style.

“Whether or not the grades generated by peers are reliably similar to grades generated by experts is only one factor worth considering, however. Student perception is also an important factor. […] Reliance on peer grading is one of the top drivers of high MOOC dropout rates. This problem may be addressed by reintroducing some expert grading where possible.” (Joyner, 2016)

### Lists

Bulleted and numbered lists are indented 0.25″ from the left margin, with the bullet or number hanging by 0.25″ (i.e., flush with the left margin).

* Like this
* And this
* And also this

# Procedural elements

## In-line citations

Articles or sources to which you refer should be cited in-line with the authors’ names and the year of publication.[[1]](#footnote-2) The citation should be placed close in the text to the actual claim, not merely at the end of the paragraph. For example: students in the OMSCS program are older and more likely to be employed than students in the on-campus program (Joyner, 2017). In the event of multiple authors, list them. For example: research finds sentiment analysis of the text of OMSCS reviews corresponds to student-assigned ratings of the course (Newman *&* Joyner, 2018). You may also cite multiple studies together. For example: several studies have found students in the online version of an undergraduate CS1 class performed equally with students in a traditional version (Joyner, 2018a; Joyner, 2018b). If you would like to refer to an author in text, you may also do so by including the year (in parentheses) after the author’s name in the text. If a publication has more than 4 authors, you may list the first author followed by ‘et al.’ For example: Joyner et al. (2016) claim that a round of peer review prior to grading may improve graders’ efficiency and the quality of feedback given. This applies to parenthetical citations as well, e.g. (Joyner et al., 2016).

## Reference lists

References should be placed at the end of the paper in a dedicated section. Reference lists should be numbered and organized alphabetically by first author’s last name. If multiple papers have the same author(s) and year, you may append a letter to the end of the year to allow differentiated in-line text (e.g. Joyner, 2018a and Joyner, 2018b in the section above). If multiple papers have the same author(s), list them in chronological order starting with the older paper. Only works that are cited in-line should be included in the reference list. The reference list does not count against the length requirements.

# References

1. Joyner, D. A., Ashby, W., Irish, L., Lam, Y., Langston, J., Lupiani, I., Lustig, M., Pettoruto, P., Sheahen, D., Smiley, A., Bruckman, A., & Goel, A. (2016). Graders as Meta-Reviewers: Simultaneously Scaling and Improving Expert Evaluation for Large Online Classrooms. In *Proceedings of the Third Annual ACM Conference on Learning at Scale*. Edinburgh, Scotland.
2. Joyner, D. A. (2017). Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth Annual ACM Conference on Learning at Scale*. Cambridge, Massachusetts.
3. Joyner, D. A. (2018a). Intelligent Evaluation and Feedback in Support of a Credit-Bearing MOOC. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.
4. Joyner, D. A. (2018b). Toward CS1 at Scale: Building and Testing a MOOC-for-Credit Candidate. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*. London, United Kingdom. ACM Press.
5. Newman, H. & Joyner, D. A. (2018). Sentiment Analysis of Student Evaluations of Teaching. In *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom. Springer.

# Appendices

You may optionally move certain information to appendices at the end of your paper, after the reference list. If you have multiple appendices, you should create a section with a *Heading 1* of “Appendices.” Each appendix should begin with a descriptive *Heading 2;* appendices can thus be referenced in the body text using their heading number and description, e.g. “Appendix 5.1: Survey responses.” If you have only one appendix, you can label it with the word “Appendix” followed by a descriptive title, e.g., “Appendix: Survey responses.”

These appendices do not count against the page limit, but they should not contain any information *required* to answer the question in full. The body text should be sufficient to answer the question, and the appendices should be included only for you to reference or to give additional context. If you decide to move content to an appendix, be sure to summarize the content and note it in relevant place in the body text, e.g., “The raw data can be viewed in *Appendix 5.1: Survey responses.*”

1. In-line citations are preferred over footnotes, and we favor APA citation format for both in-line citations and reference lists. Refer to the [Purdue Online Writing Lab](https://owl.purdue.edu/owl/research_and_citation/apa_style/apa_formatting_and_style_guide/in_text_citations_the_basics.html), or follow the above examples. You should use the *Footnote* paragraph style, with 8.5 point text and 14 point line spacing. [↑](#footnote-ref-2)