CS 422: Project 1 Write-up

Sophie Liu

September 25, 2022

1 Decision Tree Implementation

I implemented my decision tree functions recursively and chose a nested list as my data structure. I'm quite happy with my choice of data structure. It was quite simple to add elements to the list and access the tree. It would likely be far more difficult if the decision tree were not binary, but as it was binary, it was rather easy to implement.

2 Random Forest Implementation

As for the random forest, the reason that the individual trees have such variance in their accuracy is probably that they're all built on such different parts of the training data set (each being just a random 10% of the training data), they all look very different, meaning that they'll also make rather different predictions, as well. That means that their accuracy on the entire training data set will vary quite a bit.

I could reduce this variance by increasing sampling size that each decision tree is built on. After all, a larger sampling size would mean two decision trees will have more samples in the training data set in common, which means that the decision trees will be more similar and thus their predictions will be more similar. With the predictions more similar, the accuracies will have less variance. Doing so might increase accuracy, as well, as a larger data set may improve the accuracy.

As an example, when I increased the sampling size to 50% of the training data set, the accuracy of the individual decision trees was as follows: [0.7352941176470589, 0.738562091503268, 0.7320261437908496, 0.7352941176470589, 0.738562091503268, 0.738562091503268, 0.7352941176470589, 0.7320261437908496, 0.7320261437908496, 0.7352941176470589, 0.738562091503268].

Meanwhile, the variance when the sampling size was 10% was as follows: [0.7352941176470589, 0.5751633986928104, 0.7352941176470589, 0.7320261437908496, 0.6928104575163399, 0.6209150326797386, 0.7352941176470589, 0.7352941176470589, 0.7352941176470589, 0.7287581699346405, 0.696078431372549].

Clearly, there is a vast difference in the variance.

Finally, the reason that it is beneficial for the random forest to use an odd number of individual trees is so that, when making predictions, there can't be a tie. With an even number of trees, there is the possibility that half the trees make one prediction and the other half make another. With an odd number of trees, that is impossible.

3 Overall Thoughts

I started the assignment thinking it would be a daunting task and not knowing where to begin. It seemed as if a lot was involved, particularly since I haven't touched Python in years. Now, I am much more comfortable with the langauge. I'm sure I'll still need some practice for actually working with classes (and probably file i/o), but overall, I'm feeling pretty confident.