Final Project: Status Report

Aidan Garton

April 28, 2022

Members

Just me!

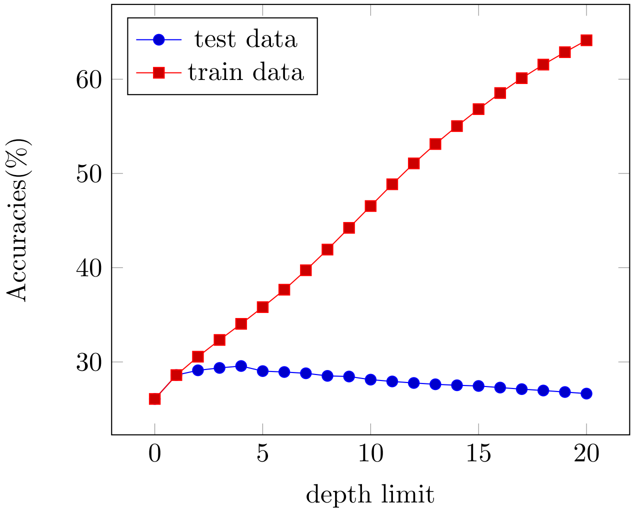
Summary

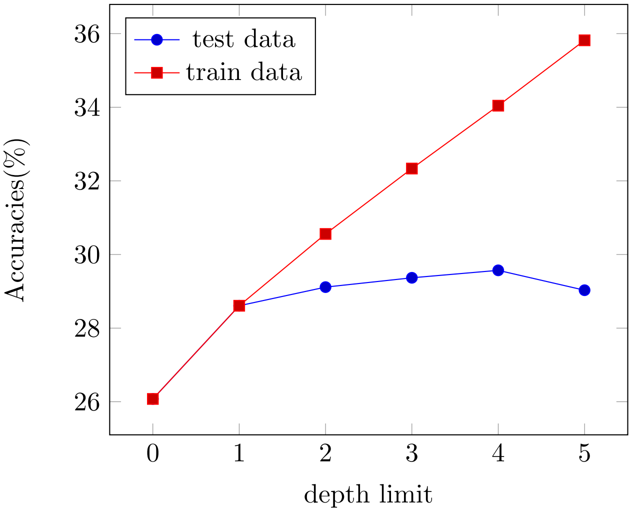
So far I have spent the majority of my time retrieving and cleaning my data set. I decided to try to predict student's final math grades in high school based on a variety of attributes of the students. There are a total of 32 features which were of varying types and values: some were nominal (i.e., a discrete number of values greater than 2), some were numeric, and others were binary. I spent a while converting all of these feature values into nominal ones, determining different cutoffs/ranges for numeric types. For example, the possible values for the "age" feature ranged from 16-21. I converted this to a binary split of > 17 and $<= 17$. I did this for all of the features that needed adjusting. I also removed the columns G1 and G2 which corresponded to the first and second trimester grades of the students because I thought it would be too easy to predict the final grade if these were included in the training process. Lastly, I converted the labels to be from 0-9 instead of 0-19: If a student's final grade was between 0 and 2 they were given the label 0, for grades between 2 and 4 they were assigned, and so on. Thus, the model will be trying to learn what 10-percent cutoff the students' final grades were in.

Next I wanted to get a base line performance for this data set so I imported the starter code for assignment 5 which includes the implementation of a multi-class decision tree. I trained this model on 10 splits of the data for varying depth limits and then measured their prediction accuracy on both the train and test data. The results are given in the section below.

Lastly, I have spent some time researching AdaBoosting to try to understand the concepts behind it and figure out how to implement the algorithm.

Results





The leftmost graph shows performance on the training and testing data for depth limits 0-20 and the rightmost graph shows a zoomed in version of where the accuracies start to diverge, ranging from 0-5. From these results, it seems that the model starts to severely over-fit once the depth limit of the decision tree has exceeded 1-2. The accuracy at this point on the test data was around 29%. Since one of the benefits of AdaBoosting is to minimize over-fitting, I think this will be a good data set/model pair to experiment on for this project.

Problems

One difficulty / point of annoyance is the way the multi-class decision tree decides splits. By only comparing zero and non-zero values to make splits, I'm worried a lot of important information in the data is being lost and is making some features irrelevant. I wonder if I should pick a more simple data set (one's with binary features such as the simple-titanic data set) or if I should try to enhance the current decision tree implementation to allow for multiple nodes at each split. It might also be interesting to pick a Text-based dataset like the wine data set. One other difficulty is that it has been difficult to figure out how to implement the AdaBoost algorithm. The paper's I have read give high level descriptions of the algorithm but I'm not exactly sure how to implement the details. I don't think this will be too big of an issue once I have spent more time doing research and looking through the original papers.

Hours

I have spent about 5-6 hours cleaning the data set, setting up the starting code base, and experimenting on the data with our previously created models. I spent an additional 2-3 hours researching AdaBoosting by reading articles and papers. So in total I have spent around 7-9 hours since the project proposal was submitted.

Code

Link to Github repo: `https://github.com/amgarton47/AdaboostedDTs`