Mikey Ling
CS 7641
Assignment 1
GTID: 903278121

Supervised Learning

Datasets

Abalone

This dataset was taken from a field survey of abalone, a member of the mollusk family. We will run multiple ML algorithms to attempt to determine the gender of an abalone based on the eight features obtained in the 4177-sample study (description of the data can be found in abalone.txt)

Adult

This dataset was obtained to determine whether or not someone's annual salary exceeded $50,000USD based on other aspects of their life, like marital status, country of origin, gender, etc. There are 14 attributes and over 9,000 instances in the dataset. The goal is to determine whether or not someone makes more than $50,000 on an annual basis.

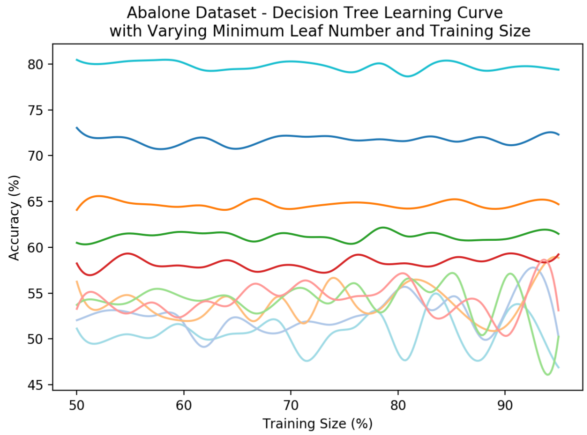
Why Are These Datasets Interesting?

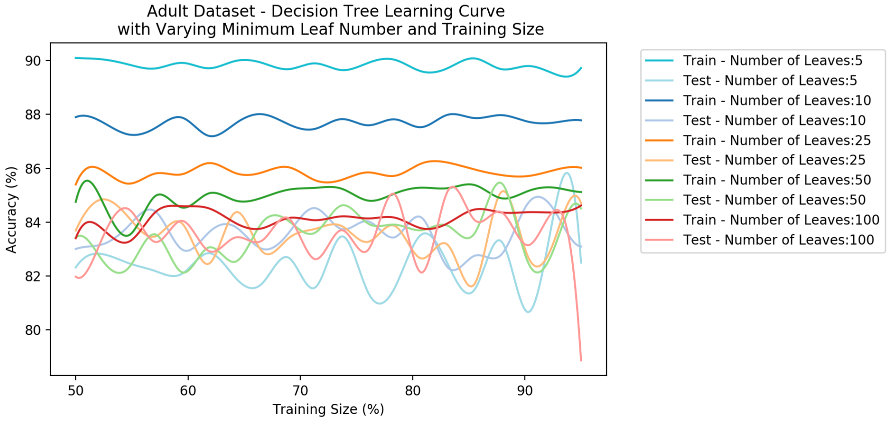
The two datasets are interesting both for their practical uses and for the graphs they output when certain supervised learning algorithms were used to analyze them. The net population of abalone all around the world has been on a steady decline due environmental changes and human interaction. Unfortunately, the very act of obtaining information about the population of abalone has the potential to hurt the abalone! For example, it is often the case that an abalone dies or suffers some sort of physical harm when humans are sexing it. If we could employ ML techniques to determine the gender of abalone without causing them harm, then perhaps the abalone population can make a comeback. I find the Adult dataset interesting because it can reveal monetary data of people within a certain demographic. This type of information could be very useful for companies trying to get the most out of their marketing campaigns.

Practically speaking, the two datasets are very interesting. But why are they interesting with respect to machine learning?

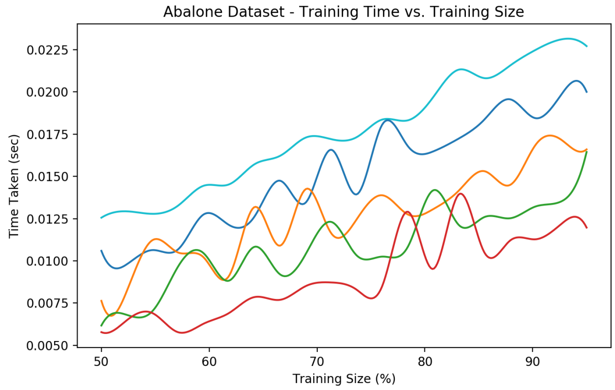
FINISH THIS PART!!!

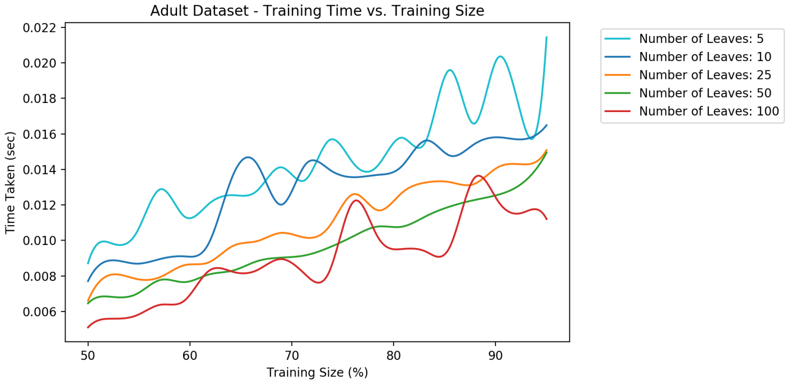
Decision Trees

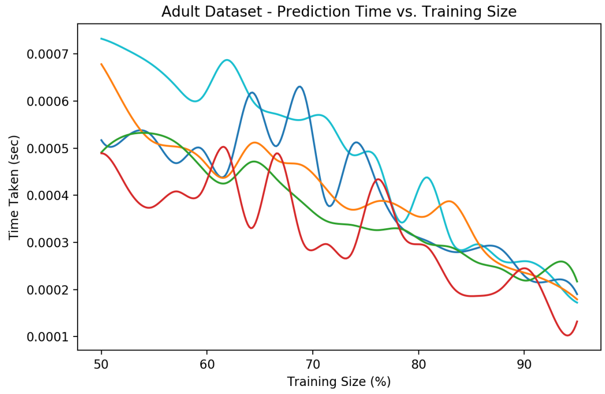


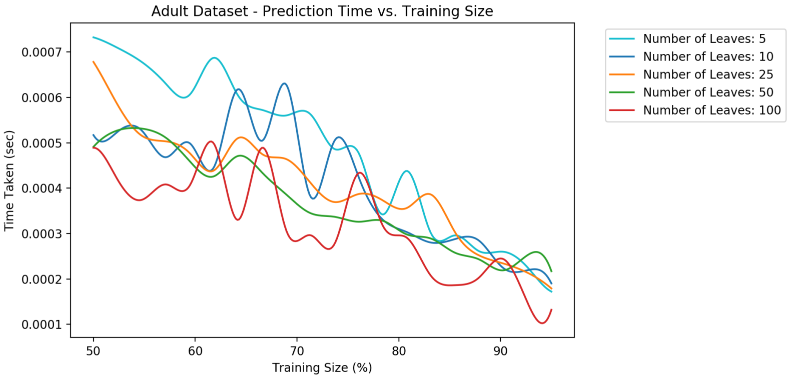


The results shown above reveal the benefits of pruning decision trees. The accuracy of the models for both datasets increased as the pruning became more aggressive. It should be noted pruning does not always improve the accuracy of a model; however in this case, I believe pruning the decision tree helps compensate for the overfitting that occurs when a tree has too many leaves. What I find extremely interesting difference in accuracies between the two datasets. The adult dataset model is nearly 30% more accurate than the abalone model; however, I initially expected the abalone dataset to have more accurate results because the physical traits of an abalone would be highly-correlated with its gender. And the amount of money one makes is an extremely complicated measurement that *seems* like it would be much harder to predict. Upon further research, though, there has been some research conducted that may suggest the abalone can change genders to increase the probability of reproductive success. This would make predicting the gender of an abalone very difficult.



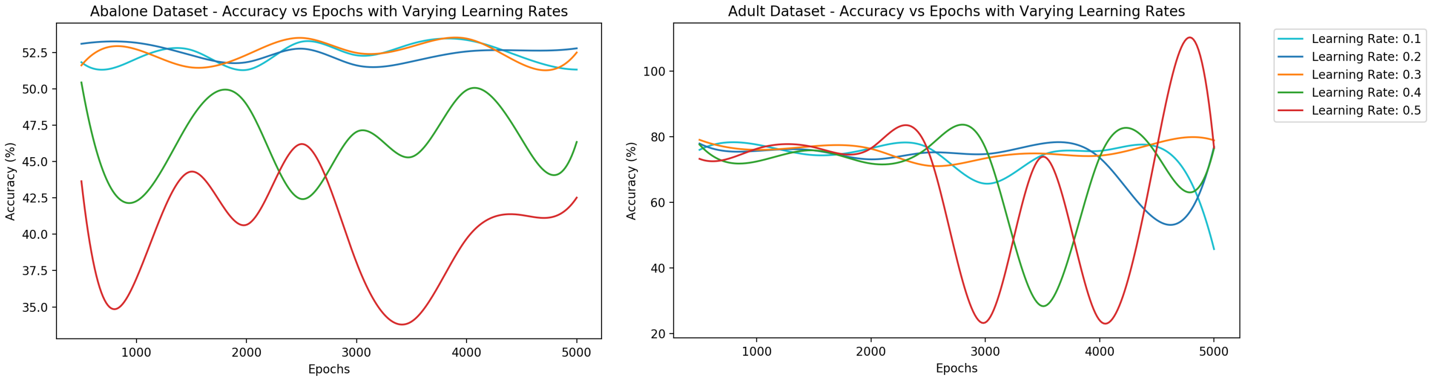






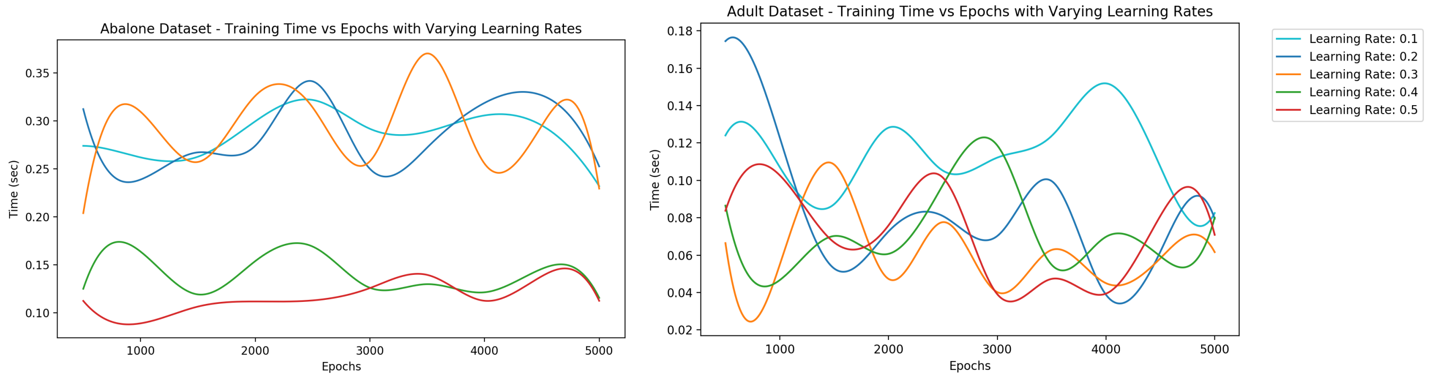
As one would expect, the larger the training sample is the more time it will take to train a model using that data. As you can see, the training times took longer than the predicting times as well. One thing that surprised me during this analysis was how *fast* training the model was done. A possible explanation for the extremely fast training speeds is the target feature is highly correlated to a small number of training features. This would allow models to converge rather quickly upon training.

Neural Networks



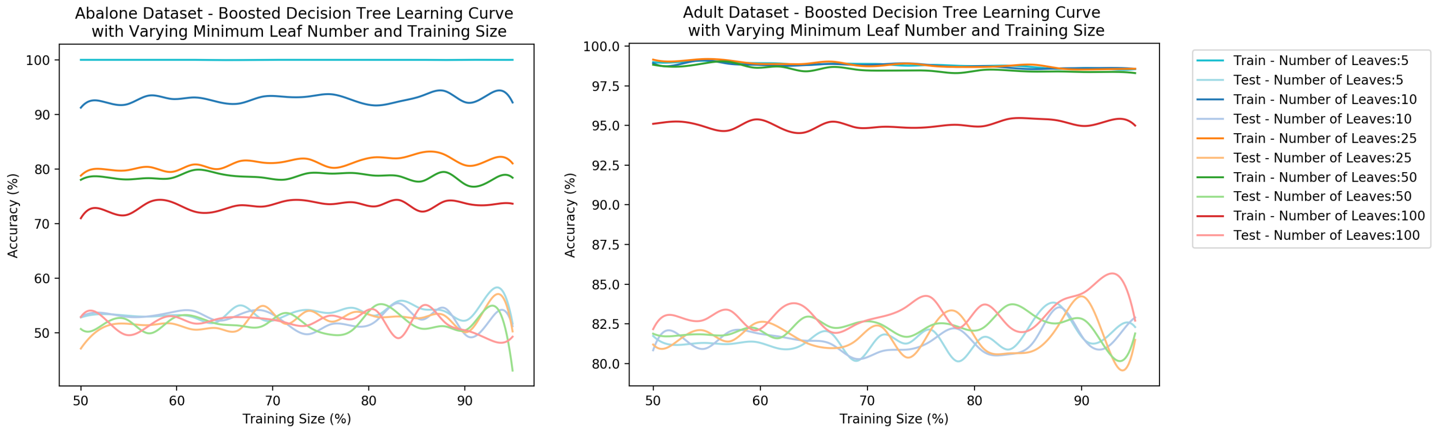
The experiments performed to analyze the performance of neural networks revealed some interesting results, namely, the decrease in accuracy for the Abalone dataset when switching from decision trees to neural networks (nearly a 30% drop!). I believe this major drop in accuracy occurred because the neural networks were too complex for the task at hand. In other words, the neural networks overfit the abalone training data to such an extent that a large decrease in accuracy was realized. This most likely happened because I initialized the neural networks used in both models to have eight hidden layers. After seeing these results, the next thing I could do to increase the accuracy of the abalone dataset would be decrease the number of hidden layers a network has.

The Adult dataset reveals the dangers of overfitting as a result of running too many epochs. You can see the accuracy of all the learning rates oscillate between 70-80% until 2000 epochs are completed. After that, the accuracies start moving all over the place (more so with the larger learning rates). This is to be expected when choosing extremely large learning rates because the weights of the perceptrons within the network change values too much, which results in inconsistent results in between epochs. It would be an interesting experiment to see the weights of the perceptrons change with respect to the number of epochs that were performed; however, it was beyond the scope of my knowledge.

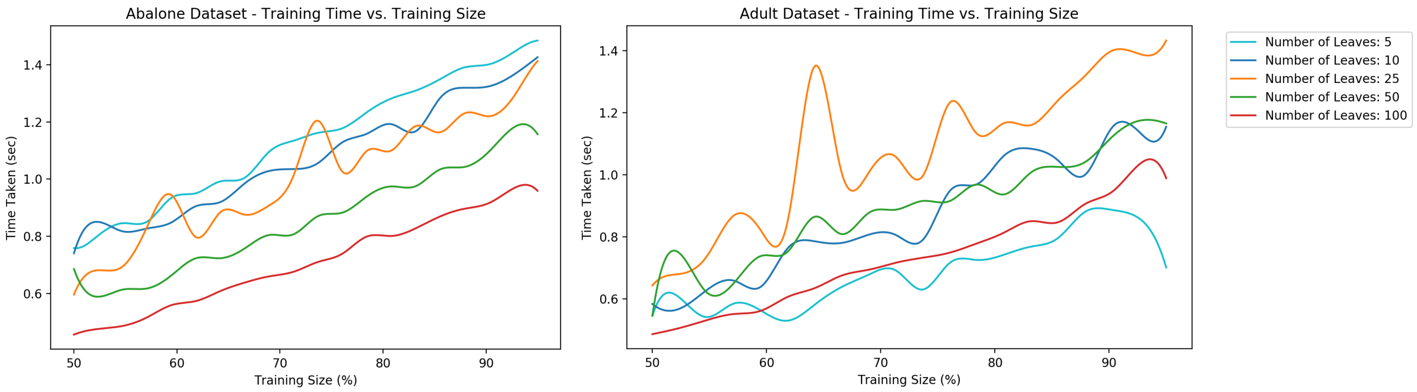


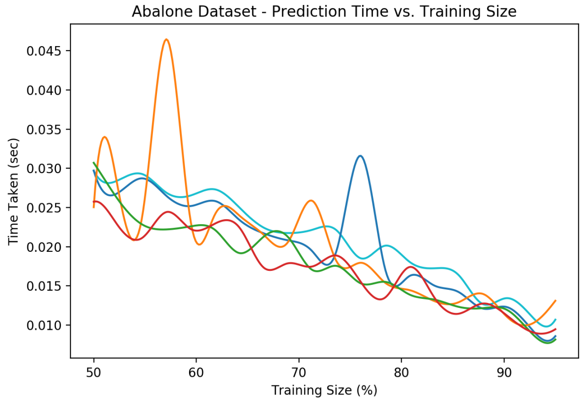
Training the Abalone dataset took *longer* than training the Adult dataset. This is a bit surprising considering the Adult dataset is *larger* than the Abalone dataset. This is most likely the related to the fact that the neural networks created less accurate results for the Abalone dataset, which seems to indicate the model had a harder time converging to a reliable set of weights during training. So, despite the larger number of samples in the Adult dataset, the information in training features can be used to predict the target feature more effectively and efficiently than the Abalone dataset. The predicting time graphs looked very similar to the prediction time graphs features in the Decision Trees section of the assignment, so they were omitted.

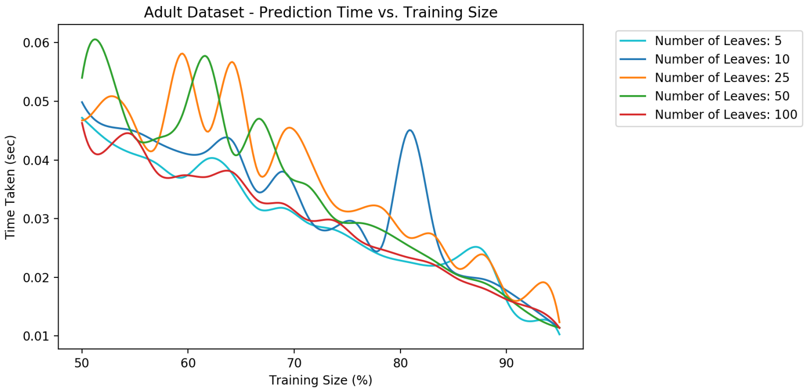
Boosting



The boosting was done using the same decision tree code used to obtain the results featured in the Decision Trees section of the assignment. As expected, *both* of the models experienced significant increases in accuracy when using training data for predictions; however, the testing data only experienced a slight increase in accuracy (a percent of two). The pruning performed in the Decision Tree section of the assignment was relatively aggressive already; however, I pruned the trees even more by limiting the their maximum depths (from unlimited to four). This highly aggressive pruning and the fact that the Abalone data is “highly over-lapped” (according abalone.info) are probably the reasons why no major increase in accuracy was realized.

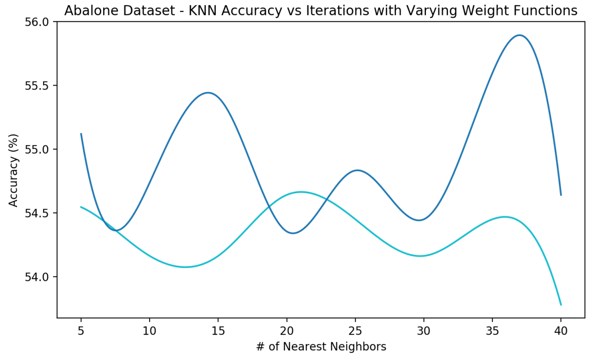


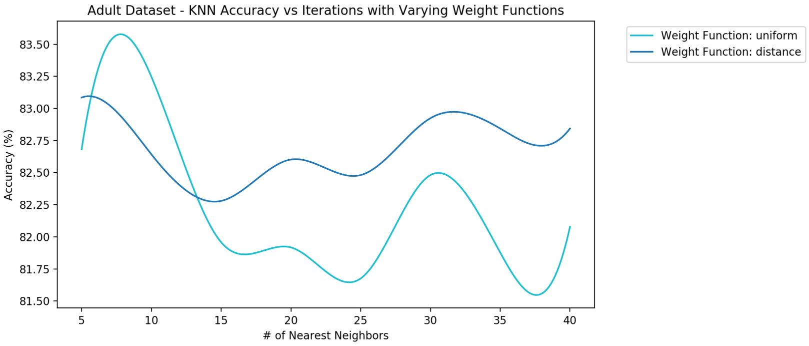




An interesting observation that can be drawn from the timing figures shown above is the Boosting run times look very similar to the timings seen in the Decision Trees section of the assignment. So, it seems the Boosting run times mimic the run times of the algorithm that is actually being Boosted, void of the more pronounced vertical displacement between curves in the Decision Tree run time graphs.

K Nearest Neighbors





Resources

https://www.sfchronicle.com/bayarea/article/Keeping-endangered-abalone-alive-1-fertilized-6869129.php