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. A description of the data can be found in abalone.info.

Adult

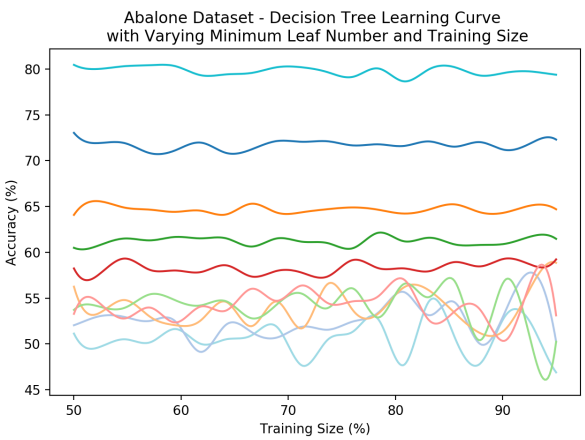
This dataset was obtained to determine whether or not someone's annual salary exceeded $50,000USD based on other aspects of their life, like marital status, country of origin, gender, etc. There are 14 attributes and over 48,000 instances in the dataset. A description of the data can be found in adult.info.

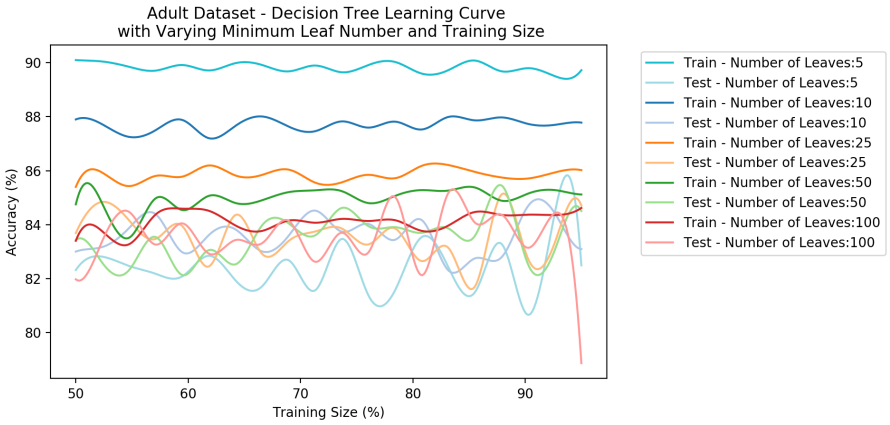
Why Are These Datasets Interesting?

The two datasets are interesting both for their practical uses and for the graphs they output when certain supervised learning algorithms are 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 them! 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 can 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 can be very useful for companies trying to get the most out of their marketing campaigns.

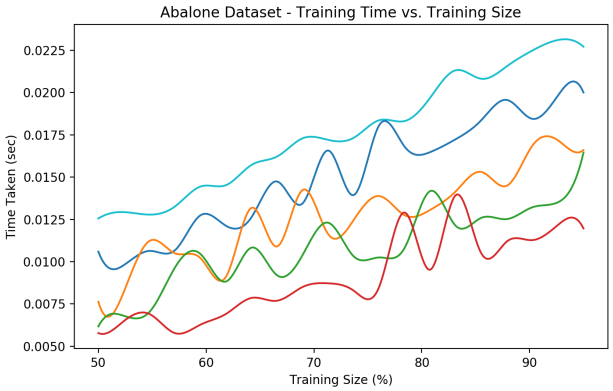
Practically speaking, the two datasets are very interesting. But why are they interesting with respect to machine learning? I find the Adult dataset interesting because it yields good results for four algorithms, namely neural networks, boosting, decision trees, and KNN, but poor results for one, support vector machines with a polynomial kernel. This behavior leads to a great learning opportunity to witness how the accuracies of different algorithms change when an identical dataset is used. The Abalone dataset is interesting because the accuracy for all models never exceeded 60%. After performing the experiment and analysis, I think the relatively low number of samples, around 4,000, does not provide enough insight on the interactions and relationships between the eight features in the dataset. Moreover, the overlapping nature of the data further increases the need for more data-samples (cited in abalone.info). This is a great example of the drawbacks of high dimensional datasets: as the number of features in a dataset increases, the amount of data needed to create accurate models increases exponentially.

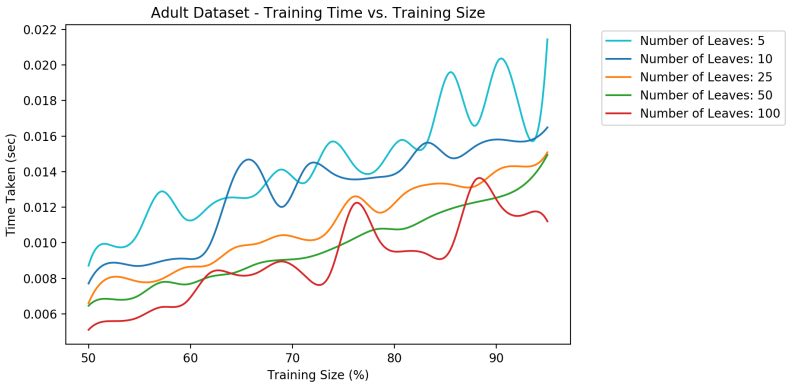
Decision Trees

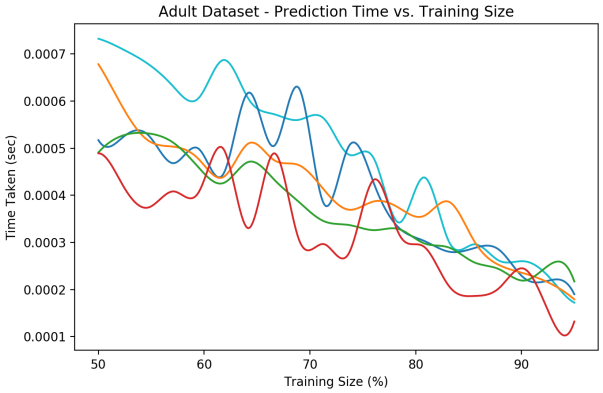


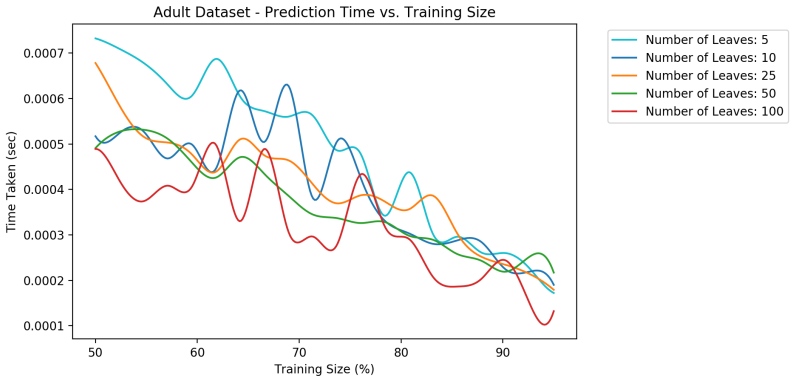


The results shown above reveal the benefits of pruning decision trees. The accuracy of the models for both datasets increases as the pruning becomes more aggressive. It should be noted that pruning does not *always* improve the accuracy of a model; however in this case, I believe pruning the decision tree helps compensate for the overfitting that occurs when a tree has too many leaves. What I find extremely interesting is the 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 even if the data did not overlap a lot.



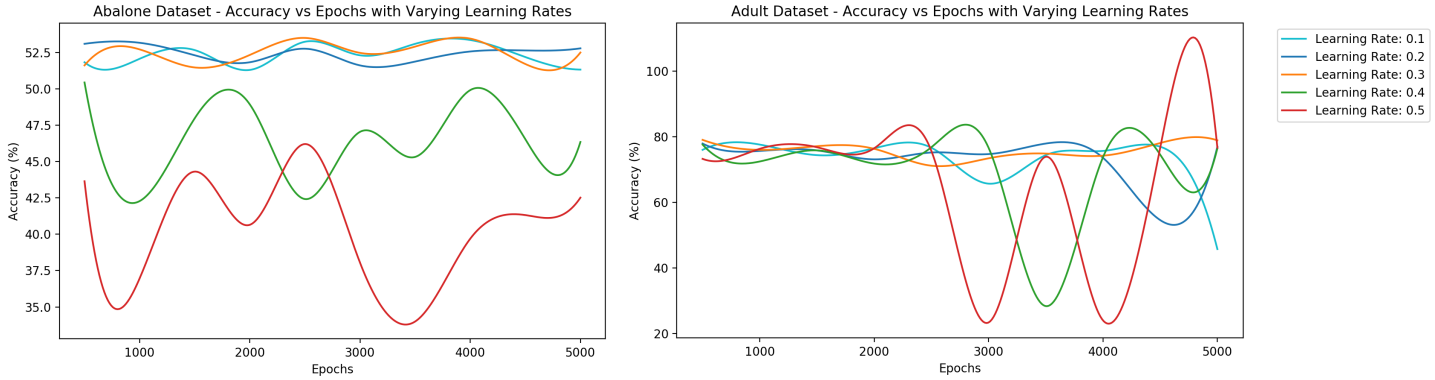






As one would expect, the larger the training sample is the more time it will take to train a model using that data (especially since the decision tree algorithm is an eager learning algorithm). As you can see, the training times took longer than the predicting times as well. One thing that surprised 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.

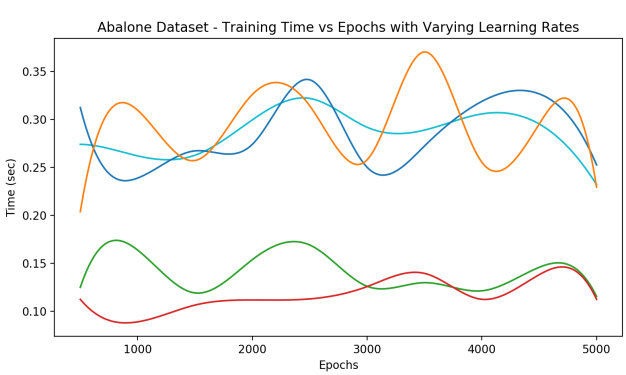
Neural Networks

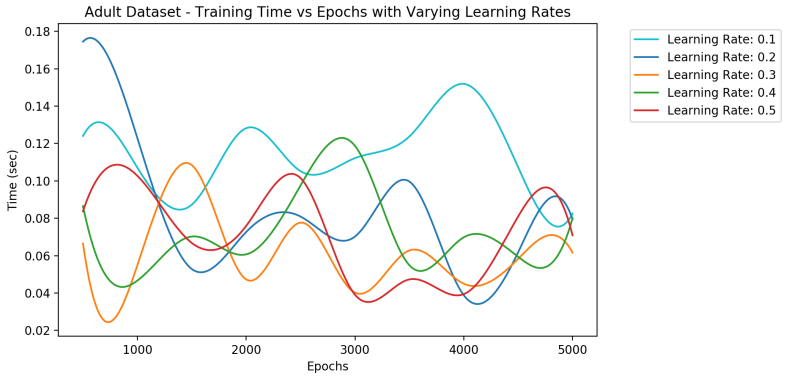


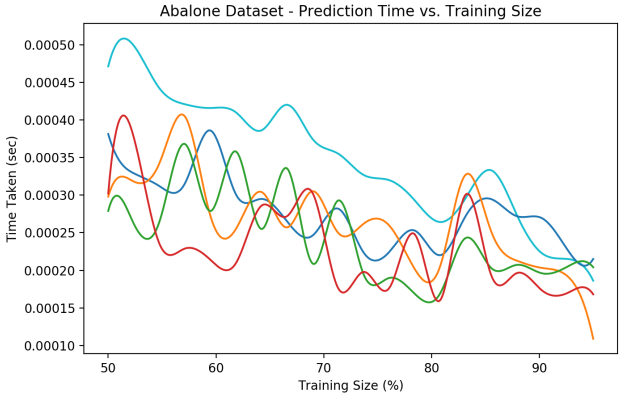
The experiments performed to analyze the performance of neural networks reveal some interesting results, namely, the slight decrease in accuracy for the Abalone dataset when switching from decision trees to neural networks (about a 2-3% drop). I believe this drop in accuracy occurrs because the neural networks are too complex for the task at hand. In other words, the neural networks overfit the abalone training data to such an extent that a decrease in accuracy was realized. This most likely happened because I initialize 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 each network has.

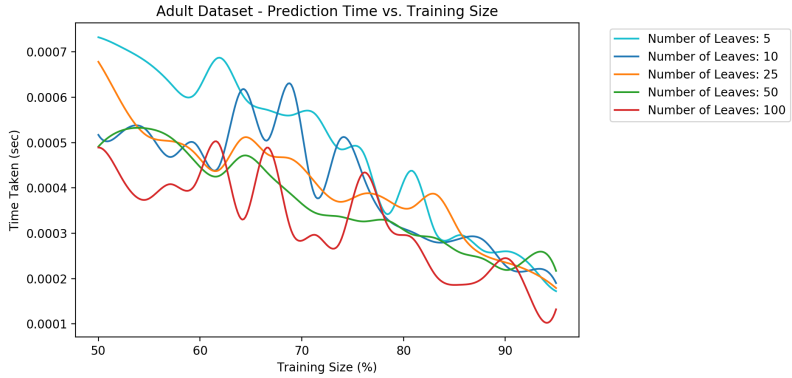
The Adult dataset reveals the dangers of overfitting as a result of running too many epochs. You can see the accuracy of all the learning rates oscillate between 70-80% until 2000 epochs are completed. After that, the accuracies start to oscillate all over the place (more so with the larger learning rates). This is to be expected when choosing extremely large learning rates because the magnitude of change the weights in the network is too large, 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. Also, note the accuracy seems to go above 100% for a learning rate of 0.5. This is simply an approximation error of the line smoothing function used to create the graphs.

It is also important to note K-Fold cross validation with a K value of five was used to average the performance of the neural networks. This cross validation was done in the hopes it would help smooth out the learning curve graphs shown above, and as you can see the graphs are relatively smooth until overfitting takes place.



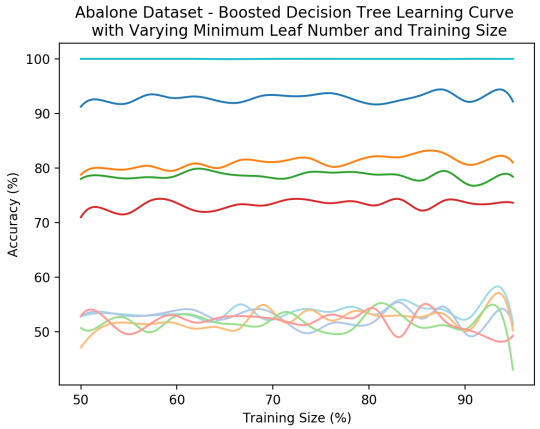


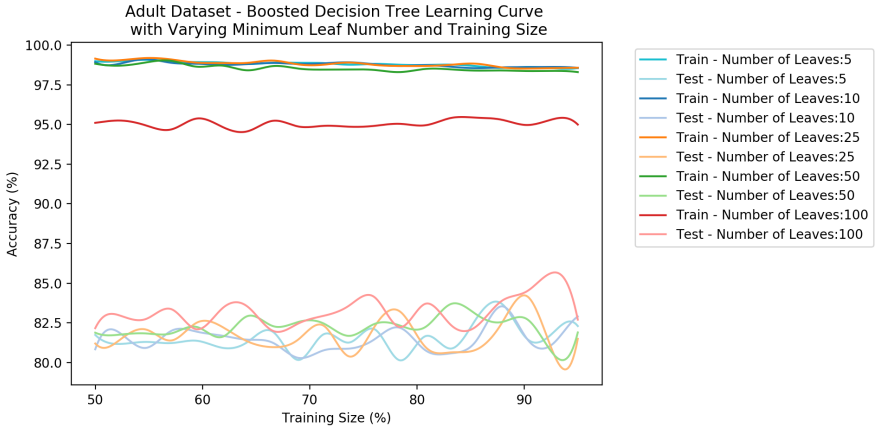




Training the Abalone dataset took *longer* than training the Adult dataset. This is a bit surprising considering the Adult dataset is *larger* than the Abalone dataset. This is most likely the related to the fact that the neural networks created less accurate results for the Abalone dataset, which seems 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.

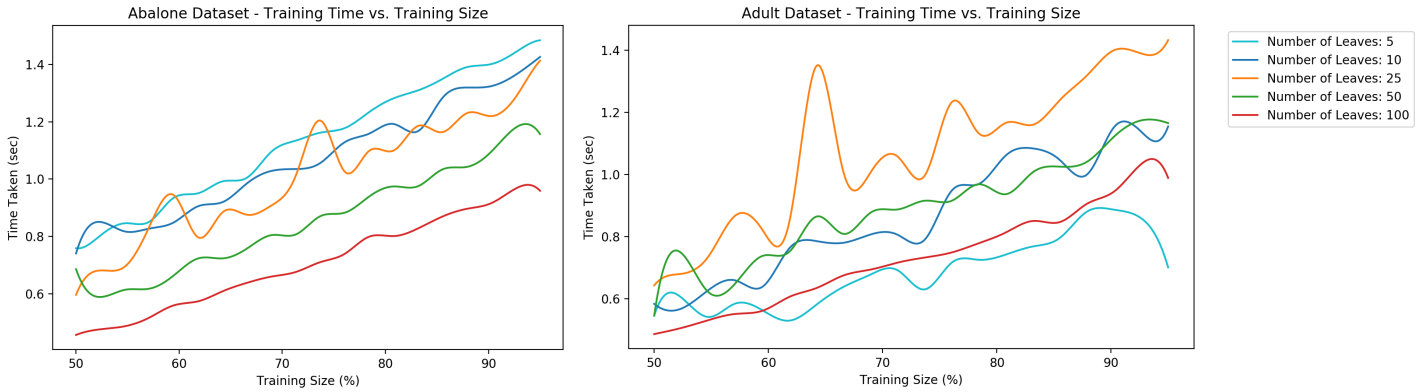
Boosting

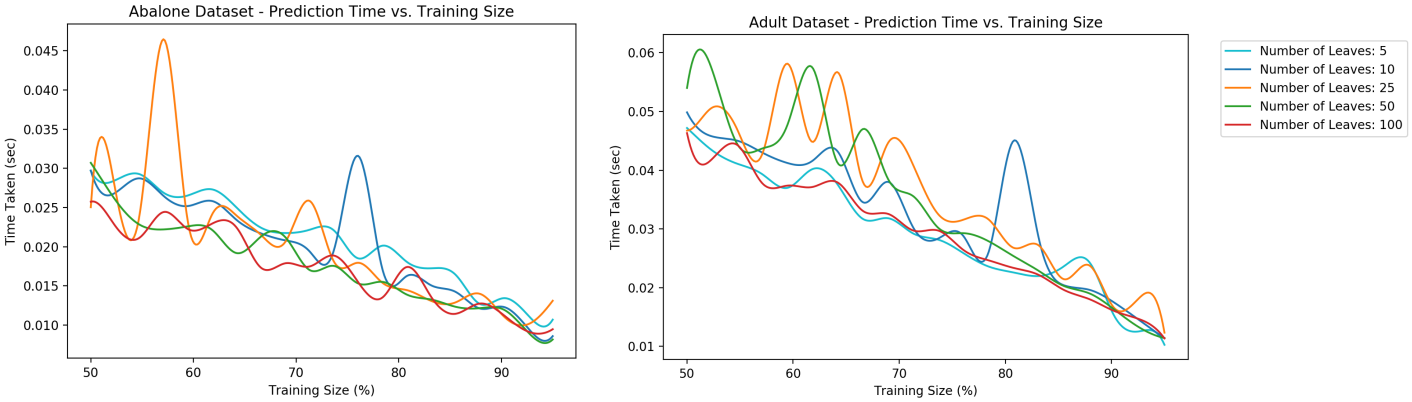




The boosting was done using the same decision tree code used to obtain the results featured in the Decision Trees section of the assignment. As expected, *both* of the models experienced increases in accuracy when using training data for predictions; however, the testing data only experienced a slight 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 maximum depths (from

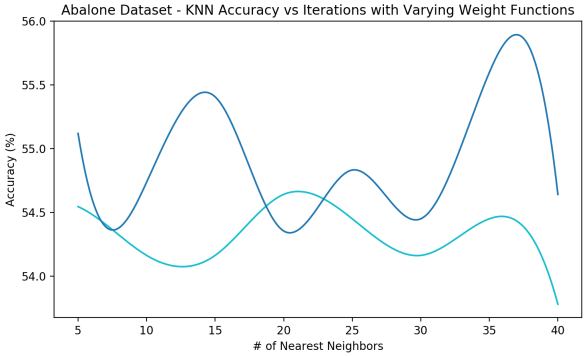
unlimited to four). This highly aggressive pruning and the fact that the Abalone data is “highly over-lapped” (according abalone.info) are probably the reasons why no major increase in accuracy was realized.

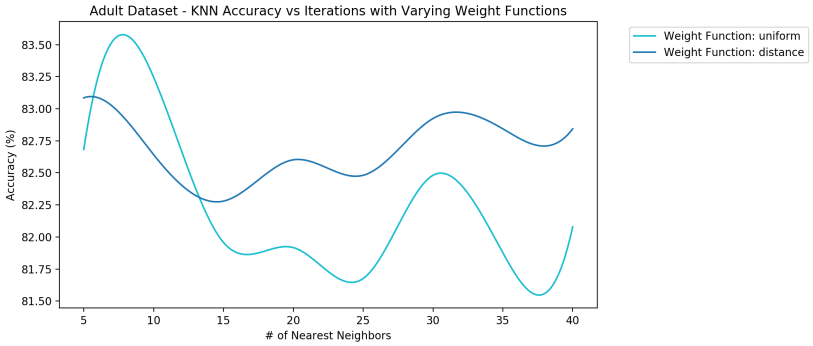




An interesting observation that can be drawn from the timing figures shown above is the Boosting run times look very similar to the timings seen in the Decision Trees section of the assignment. So, it seems the Boosting run times mimic the run times of the algorithm that is actually being Boosted, void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

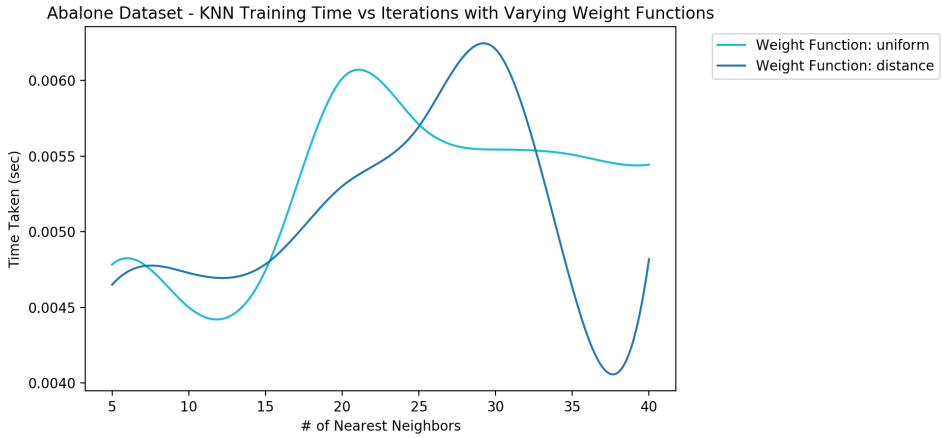
K Nearest Neighbors

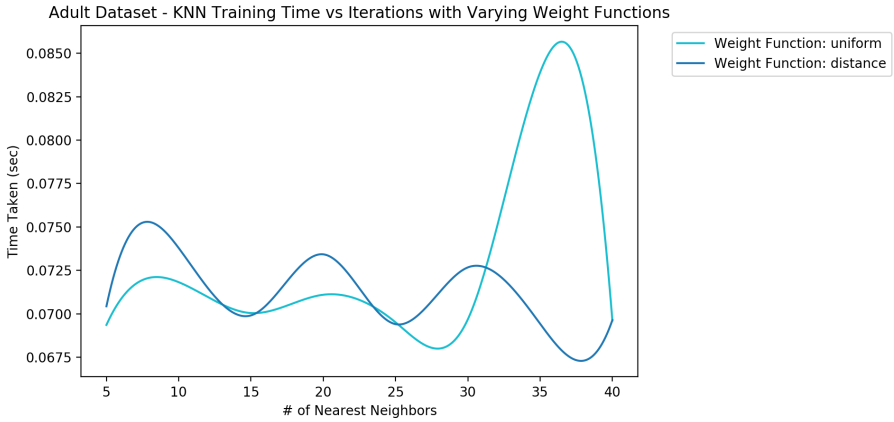


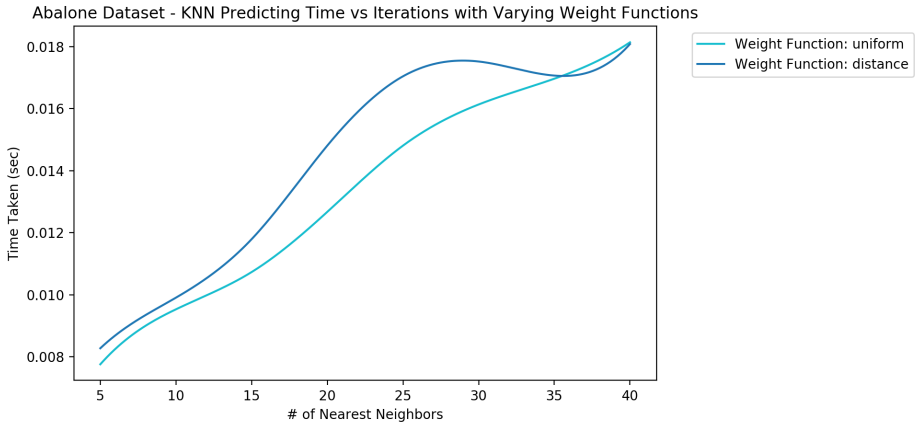


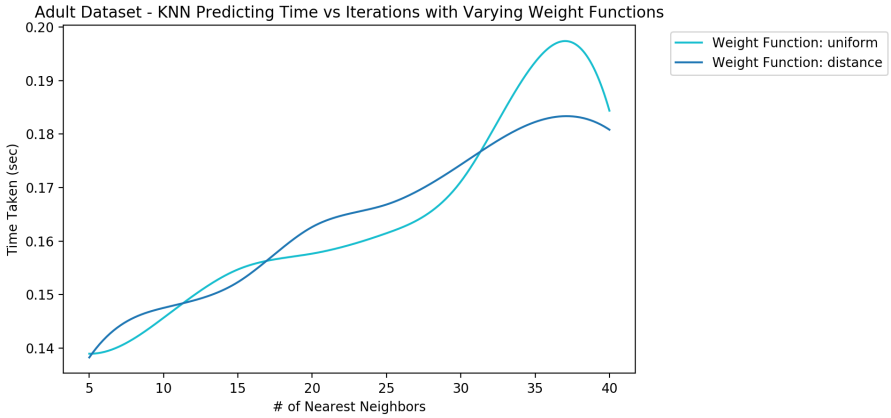
The KNN model with a distance weight function (the weights are scaled to the inverse of the distance) is more accurate for both models. This is to be expected because data points that are closer together are most likely going to be labeled the same way. For the Abalone dataset, the weighted model is more accurate than the uniform model for nearly all values of K (# of nearest neighbors); however, it's interesting to see the uniform model is slightly more accurate than the weighted model for the Adult dataset. I assume this occurs because there is some overlapping of data that is NOT compensated for when K values are less than 15. The higher the K value the less susceptible the model is to overlapping data (similar data with differing target labels).

When comparing KNN's performance to the other algorithms, it seems like the accuracy on the Abalone dataset is very similar. However, the accuracy for the distance-weight function for the Adult dataset is slightly higher than the others.



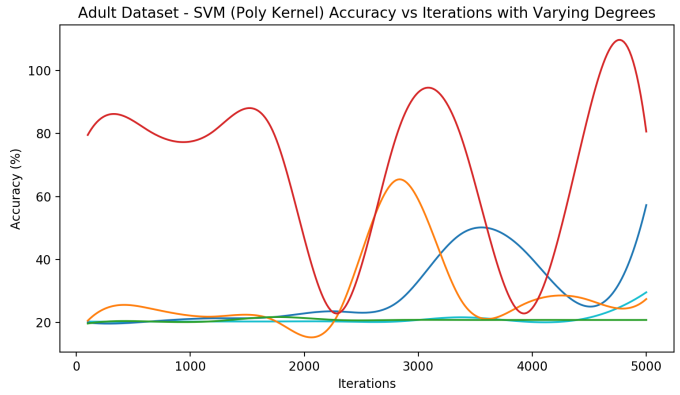


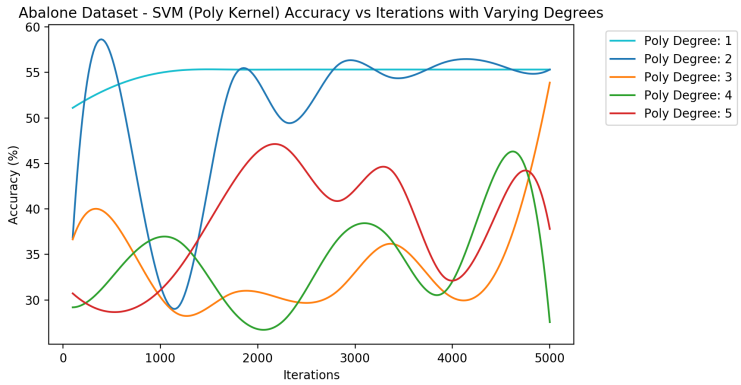


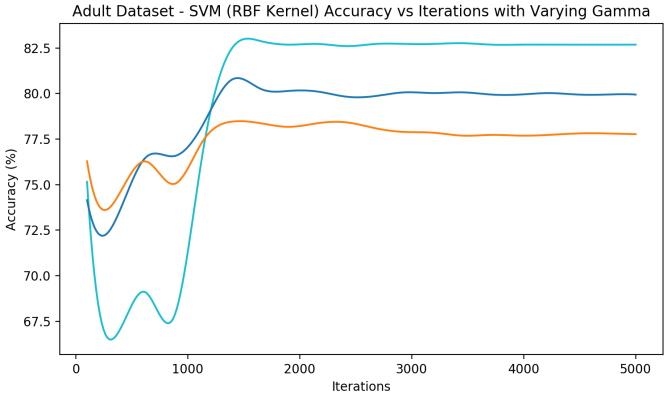


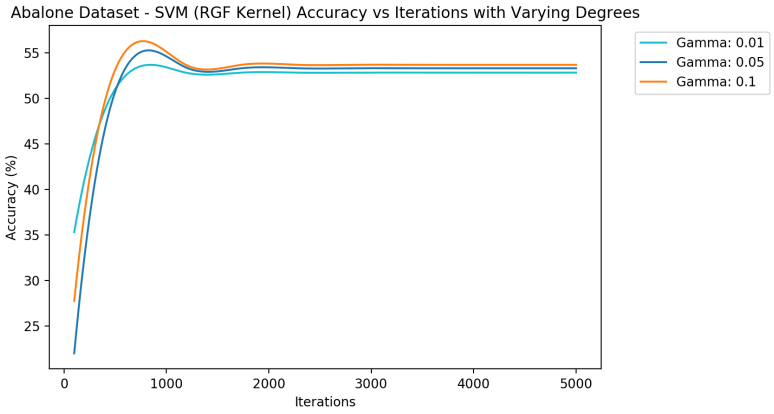
A defining characteristic of the KNN testing/predicting procedure can be seen in the timing graphs above. KNN is a lazy learning algorithm (as opposed to an eager algorithm). That means it is expected to have longer testing times than training times because the generalization beyond the training data is delayed until a query is made to the system.

Support Vector Machines





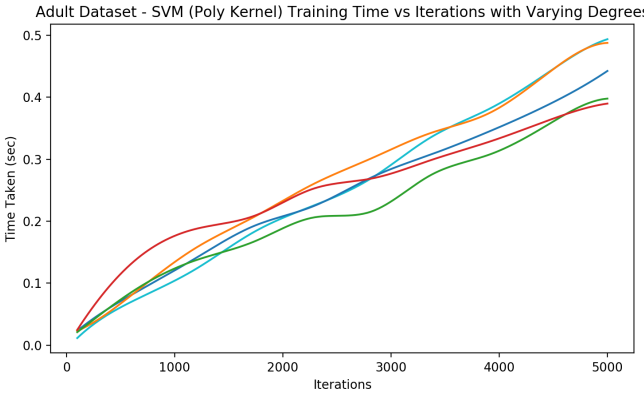


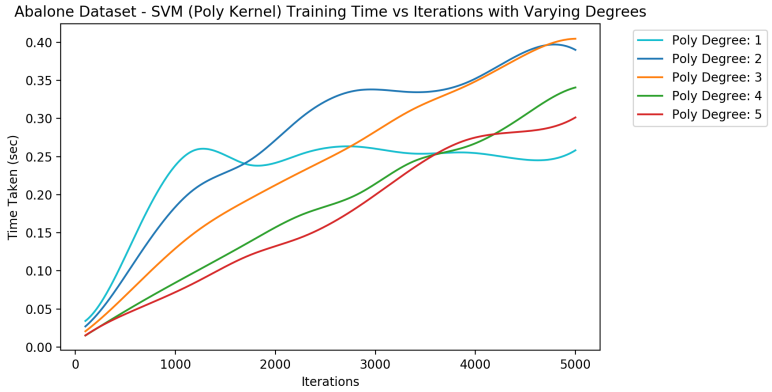


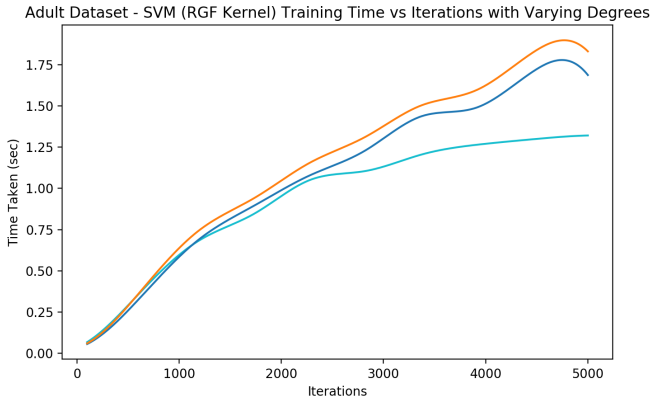
NOTE: The titles for the RGF kernel should say “Varying Gammas” instead of “Varying Degrees”

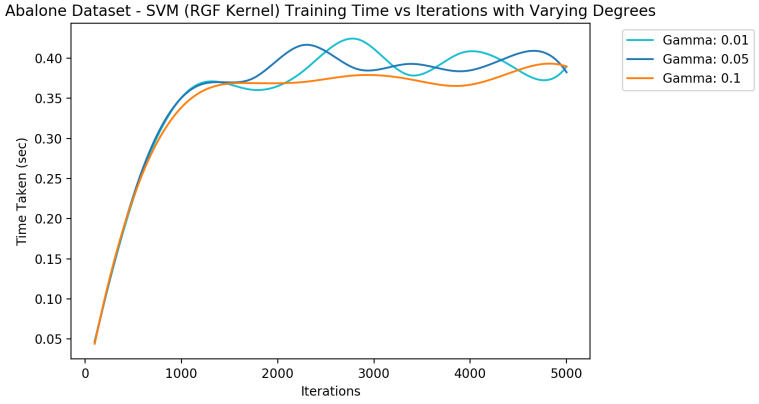
I used two different kernel functions to analyze the performance of SVM algorithms on both datasets: a polynomial kernel with varying exponential degrees and a Radial Based Function kernel with varying gamma values. Increasing the exponential degree of a polynomial kernel function allows the model to become more complex, so one would expect higher degrees of a polynomial kernel to have better performance over kernels with lower degree values. However, this certain is not always the case. The SVM-Poly Kernel for the Adult dataset shows terrible performance for all degrees except for the fifth, which shows a high accuracy until ~1750 iterations are performed where overfitting most likely takes place. It seems like the lower degree kernels do not have a high enough variance to accurately model the data. The SVM-Poly Kernel for the Abalone dataset, however, shows the exact opposite behavior: lower degree models outperforming higher degree models. This is most likely due to the high overlapping of the data, which could make the model overfit even with a polynomial degree of two or higher! You can see the degree one model’s accuracy is the most consistent and highest compared to the other four models.

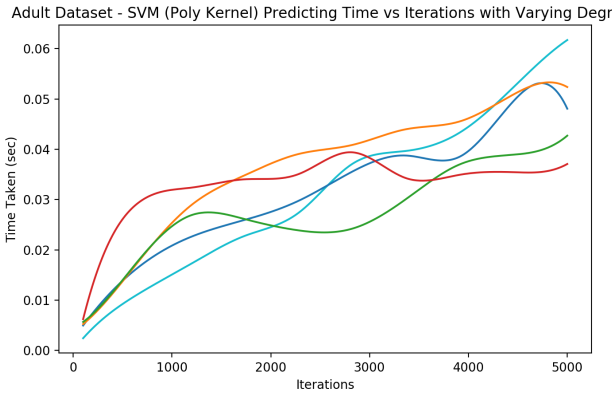
The RBF Kernels showed much better performance compared to the Poly Kernels for the two datasets. RBF Kernels rely on the parameter gamma to define how far the influence of a single training example reaches. With “low” values of gamma meaning “far” and high values meaning “close”. It can be thought of as the inverse of the radius of influence of samples selected by the model as support vectors. For the Abalone dataset, all three values of gamma seem to perform the same, but the Adult dataset tells a different story. After about 1,500 iterations, the lowest value of gamma, 0.01, performs better than the other two values of gamma. In other words, the model became more accurate as the “distance” of influence samples selected by the support vector increased. In terms of overfitting, it makes sense that the model would have a harder time creating a more accurate, generalized model if the support vectors cannot draw influence from a more generalized area of samples.

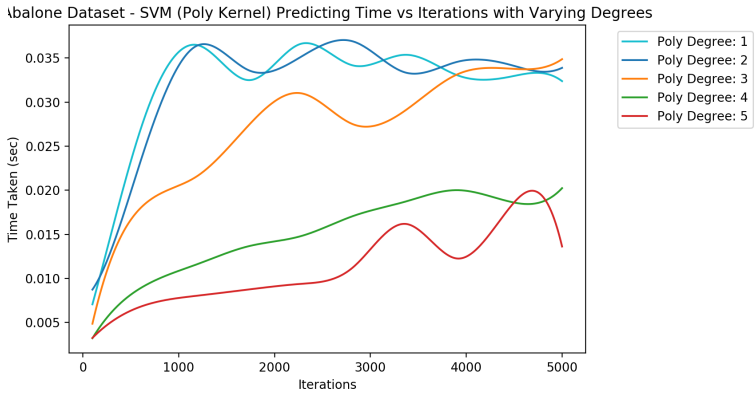


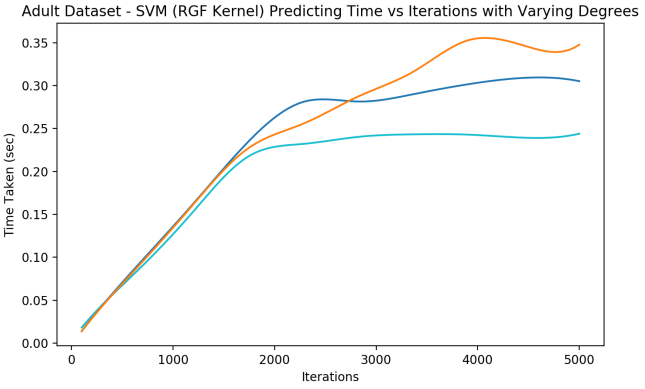


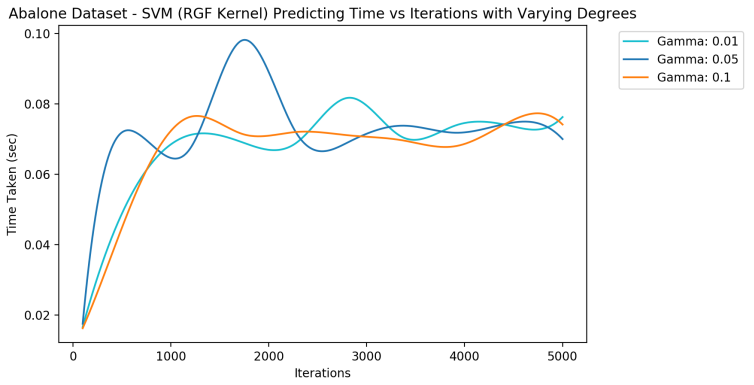












NOTE: The titles for the RGF kernel should say “Varying Gammas” instead of “Varying Degrees”

The runtime graphs for the Adult dataset look how one would expect based on the graphs of the other algorithms used in this experiment; however, the Abalone dataset runtime graphs reveal some interesting behavior. Three of the four graphs seem to flatten out after a certain number of iterations are completed. A possible explanation for this behavior is the converges to a solution in a constant number of iterations, regardless of the number of iterations that can be performed given the initiation parameters of the model. It is also interesting to see the Poly Kernel Predicting Time for the Abalone dataset shows a shorter prediction time for the most complex model (exponential degree of five). One would think a more complex model yields longer prediction times, so I find it difficult to think of a reasonable explanation for this phenomenon. Perhaps the more complex data splits the data so well that the length of the queries actually becomes shorter. In other words, the model has such a high degree of variance the margin between different labels reaches the threshold faster than a model with a lower degree of variance.

Resources

- https://www.sfchronicle.com/bayarea/article/Keeping-endangered-abalone-alive-1-fertilized-6869129.php
- https://archive.ics.uci.edu/ml/datasets/adult
- https://archive.ics.uci.edu/ml/datasets/abalone