Michael Knapp

Data mining CSC 535

Homework # 2

This homework assignment was the implementation of the K nearest neighbor algorithm. This was a straight forward concept of the processing of the dataset. I did however have some trouble getting the data in a format that was able to process and conceptualize how I needed to apply the Euclidian distance formula in the Knn function that I have in my program. After many headaches and some twists and turns I completed finding K at 90% accuracy with k= 7 or k = 9.

Marcello Palillo is the person who invented the concept of classifying the Nearest neighbor rule. This is completely explained to in “Patterns and Recognition letters”. Pelillo refers to the famous Cover and Heart Paper (1967), that shows what happens if a very large training set is used. The result is the classification error is smaller than twice the Bayes error. Before the Cover and Heart rule was mentioned by Nilsson(1965) who called it “minimum distance classifier” and by Sebestyen(1962), who called it “proximity algorithm”. The fundamental principle known as Ockham’s razor: “Select the hypothesis with the fewest assumptions” or the most obvious the answer is most likely the most correct one. It is however, not formulated in terms of observations. Ockham worked in the 14th century and emphasized observations above ideas.

The fist problem I had to tackle was how to get the Training data in a format that I needed to implement the Knn algorithm. I tried several ways before settling on using the CSV reader library. I tried to build a control flow graph as a outline but this wasn’t much help because I had to my logical path several times before finding a good solution.

After the data was parsed correctly the Knn algorithm was straight forward in comparing each data sample with the test data sample. I just hard coded k as 5 because the run time was so long that I didn’t want to wait for all data samples to be tested.

Next, I implemented the weighted voting function. This was easy I created a dictionary and just kept count of each classification with the same distance. Instead of using a list I thought a dictionary would be easier to use.

Last I updated the main to make sure all the output was in the correct format requested in the homework doc. I also took the time to loop though k and see which k would find the best accuracy rate. I found that when K=7 or K=9 the reach 90% accuracy rate. This is the beat rate I found in my program.

In conclusion I found that that this assignment was easier than the last especially since I was able to use methods like the CSV reader to help manipulate the data to it can be processed in in the KNN algorithm. I have a better understanding of how long it actually takes to process small data sets against one test sample. The amount of computing time it would take on large data sets is mindboggling.