Spencer Phillips

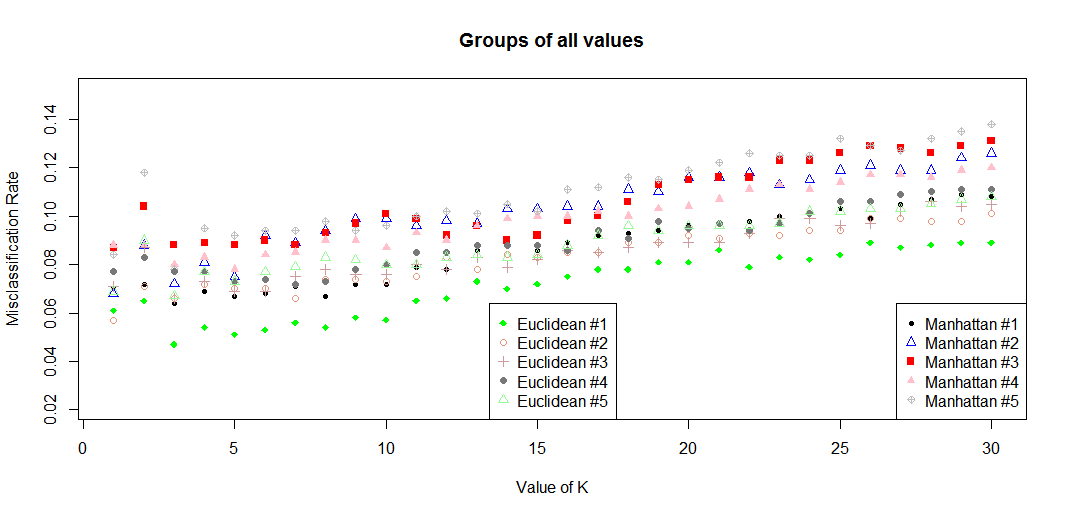
998887568

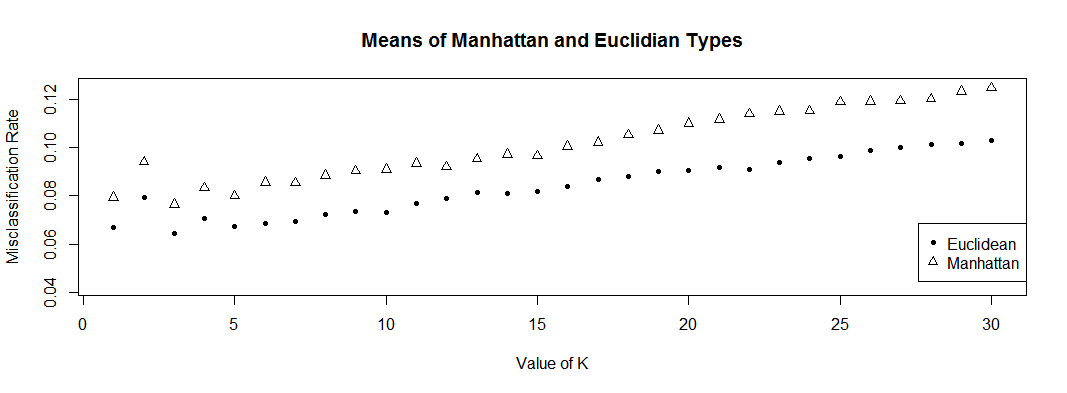
11/3/15

STA 141 Assignment 3

Part 1:

From the following graphs, it is clear that the most effective method to set up the model in KNN is to use k = 3 and a Euclidean method of finding the distance.



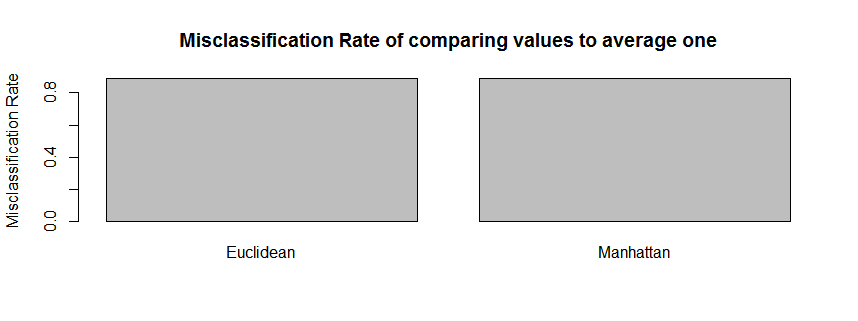


The confusion matrix is as follows (tested as Euclidian with k = 3)This showed us that the digits that were classified best were the 0s, 2s, and 8s while the ones that were classified the worst were the 1s, 3s, 7s, and 9s. The digits that were generally confused with one another were the 4s being called 9s (and visa versa), the 7s being called 1s, 5s being called 3s, and 8s being called 3s.

Some of the digits that were misclassified and were difficult to intemperate are as follows.

The top left value is supposed to be an 8 and was thought to be a 0, an obvious mistake by looking at the strange handwriting. By eye, you could almost mistake it as a 6 even! The top right value is a 5 but was mistaken as a 3 by the algorithm, perhaps a victim of a tie, the bottom part does look similar to a three, possibly causing some errors. The bottom left value is a 3 mistaken as a 9, with the top half of the 5 being so close together, it could be interpreted as a 9 with the top half forming the top loop of the 9. Finally, the bottom right value is supposed to be a 7 which was mistaken as a 9. I could also see it being mistaken as a 0 due to it almost being a full closed circle, but the shape of the value does look like a 9 with the top part of the 9 not completed all the way, an example of poor handwriting.

Part 2:

For this part, I had to create averages for the values and then run it against the test set. I did this and it led to there being a very poor success compared to the success of the knn model. This is shown in the following graph with there being tests for Euclidean and Manhattan after being split 5 times with these values being the means of them. barplot(c(Part2meanE,Part2meanM), names.arg = c("Euclidean", "Manhattan"), ylab = "Misclassification Rate", main = "Misclassification Rate of comparing values to average one")

With misclassification rates above 80% and with both basically being the same (Manhattan is slightly lower) we can mark this type of classification as inferior to the previously mentioned knn function. Meaning that the knn function should always be used over the average one just due to the massive amounts of variance in the way people write numbers