Final Project

Adam Atamian

Student ID last 4#: 7719

Lucy Chen

Student ID last 4#: 6330

Lijing Xu

Student ID last 4#: 7513

STA 141A

Prof. Gupta

December 5, 2017

Question 1.

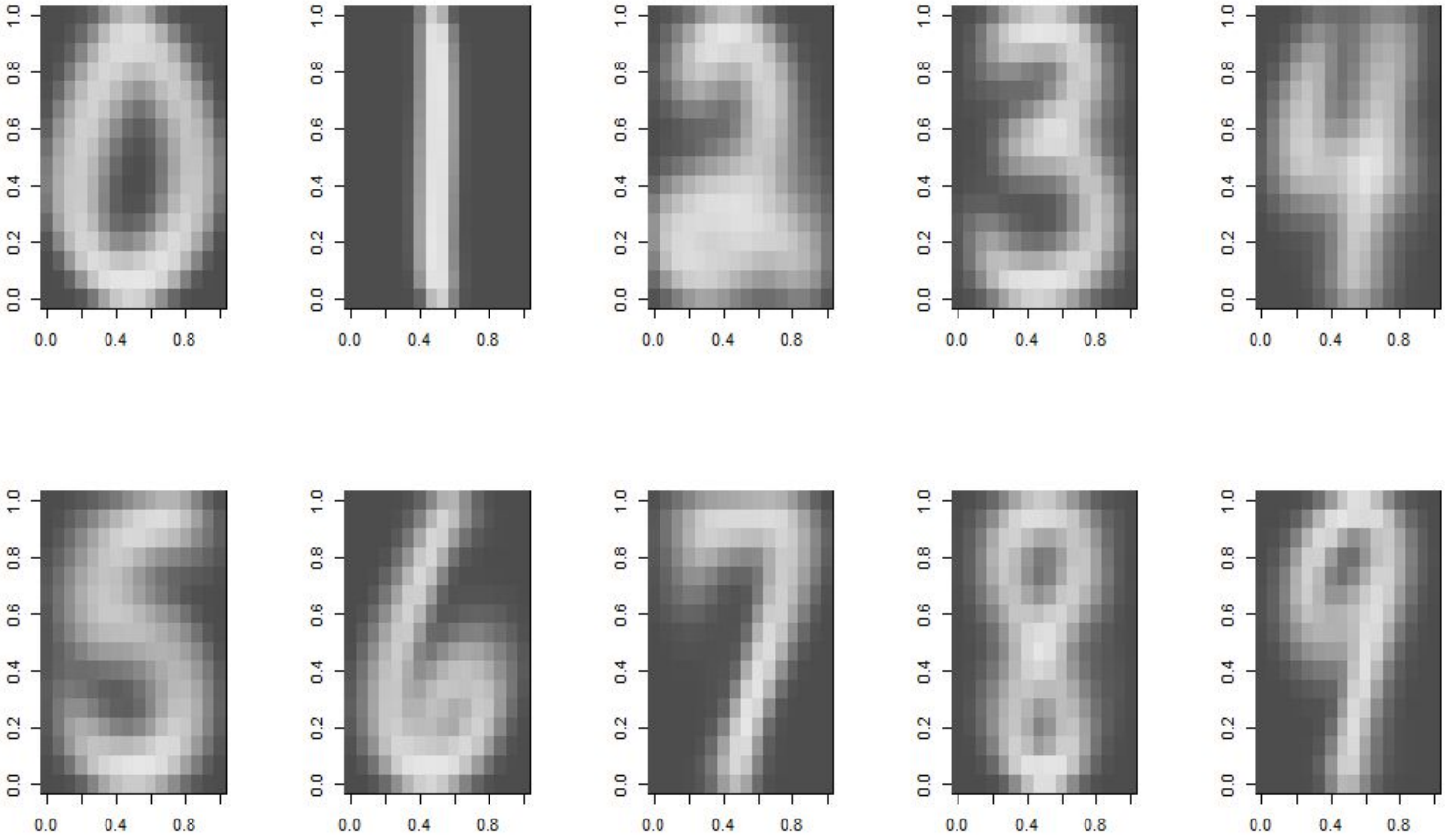
(See Appendix)

Question 2.

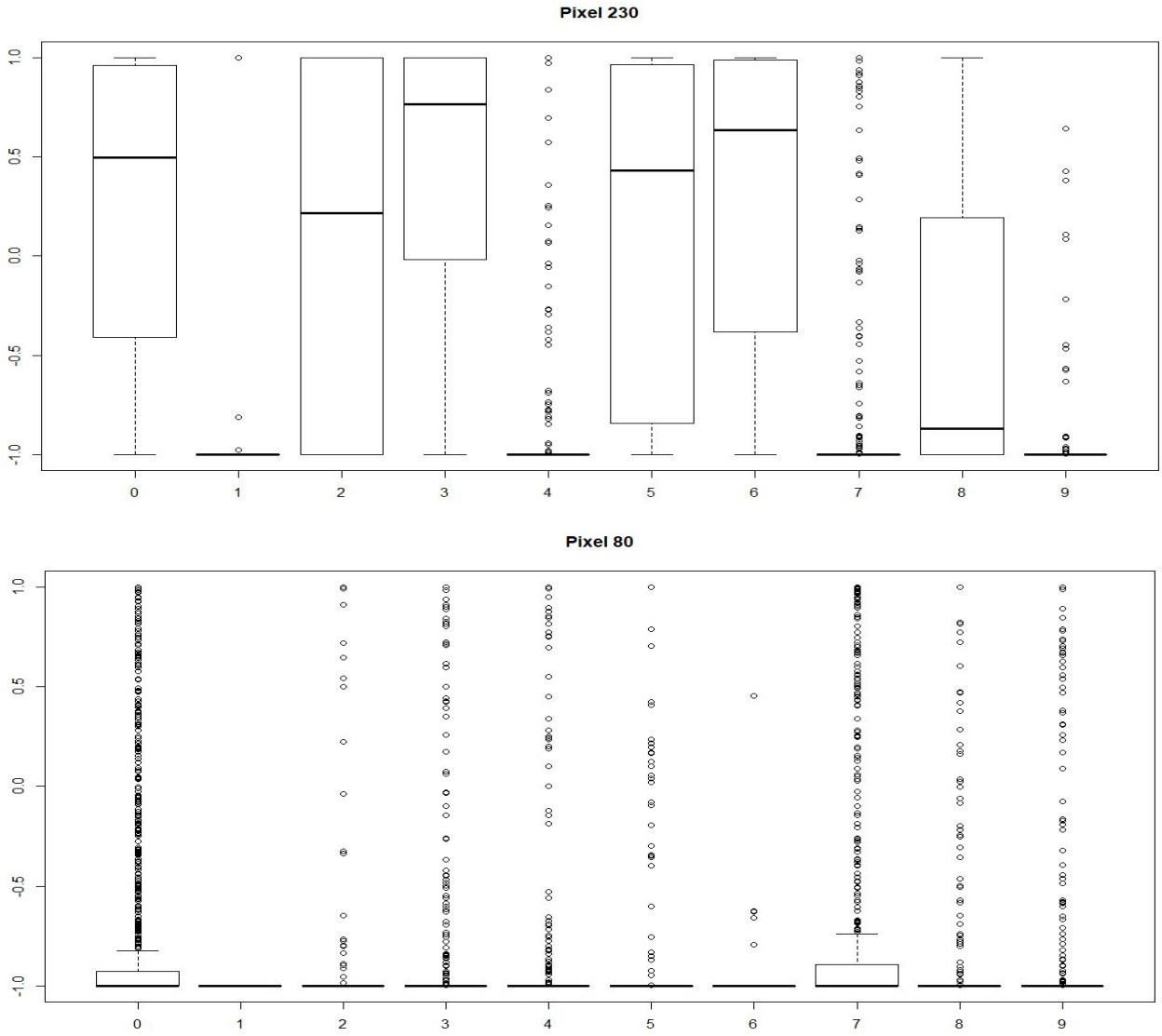
(See Appendix)

Question 3.

(a)



(b)



Pixel 230 and 213 are the most useful pixels for classification. Numbers 3, 6, 0, 5, and 2 uses both pixel 230 and 213 the most, since the pixels located near the bottom middle of the image. Pixels 1, 16, 65, and 80 are the least useful pixels for classification because these pixels are on the top right corner of the image, and most of the written numbers would not be written off to the side.

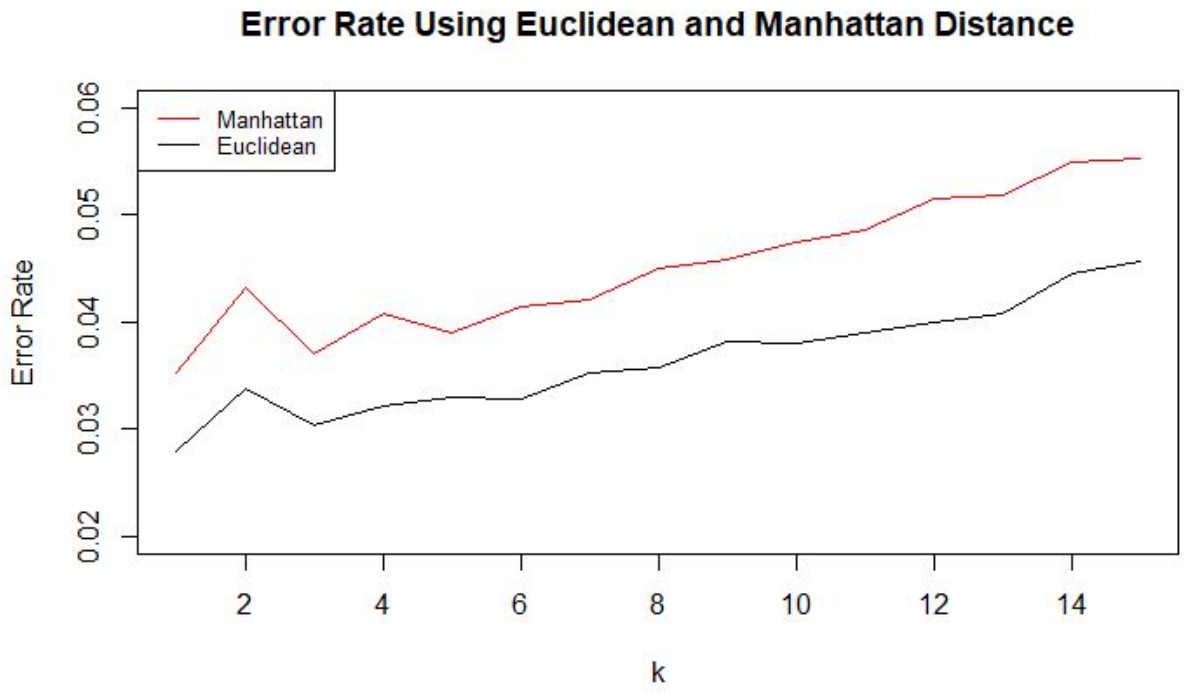
Question 4.

(See Appendix.)

Question 5.

We calculated the distance matrix of each training point beforehand, and then subsetted from the matrix when calculating the error rate, so the distance calculations only need to be made once per distance metric. We used the sapply function instead of a for loop to run each of the ten-folds for cross validation. We also used the sapply function instead of a for loop to calculate each of the predictions using a similar function from question 4, and we did not put any other calculations into the loops, so the function is fairly efficient in terms of runtime.

Question 6.



The error rates have a similar trend between the Euclidean (black line) and Manhattan (red line) distances, but Euclidean has a consistently lower error rate for each k. The best combination of k and distance metric is the Euclidean distance and k = 3. For the Manhattan distance k = 3 is also the a good combination, but the error rate for Euclidean distance is the smallest. It appears that it would not be useful to consider additional values for k because as k increases, the error rate also increases. More values for k would only produce higher error rate values.

Question 7.

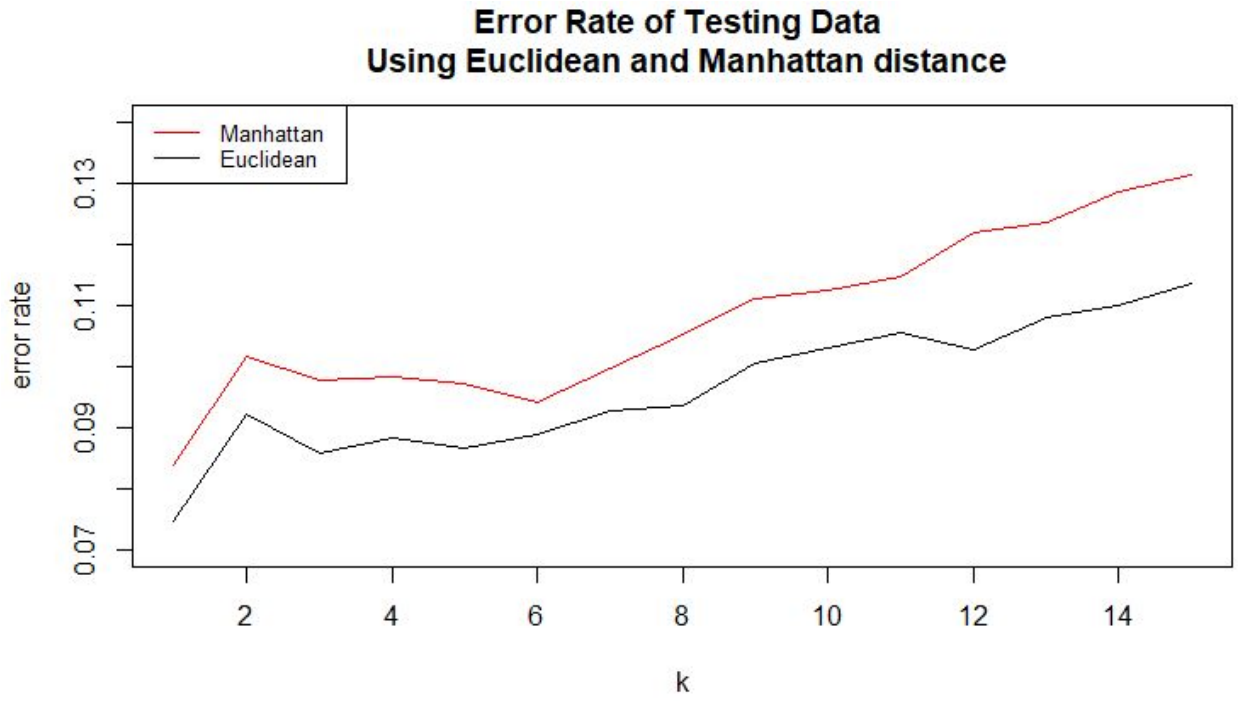
From the results in question 6, the three best k values using the Euclidean distance is k = 3, 4, and 5, and the three best k values using the Manhattan distance is also k = 3, 4, and 5, as these three k values give the smallest error rate, excluding k=1. We chose not to include k=1 because the value of this error rate does not match with the overall trend of the graph, which starts slightly high, then decreases down to about 0.03 and 0.035, and then increases again. The confusion matrices (see appendix for code) shows that the common misclassifications of digits are mistaking 7 as a 2, 1 and 9 as a 4, and 3 as an 8. These numbers have similar structures, so it is understandable that these particular values are easily misclassified. Out of all six combinations, we still believe that three-nearest-neighbors using Euclidean's distance is the best combinations because it yields the most amount of accurate predictions.

Question 8.

Pred/Acc	0	1	2	3	4	5	6	7	8	9
0	1185	0	4	3	2	8	10	0	5	1
1	0	1003	1	0	11	0	1	2	6	0
2	2	1	701	2	3	4	0	1	1	0
3	3	0	6	638	0	9	0	0	11	1
4	0	0	1	0	611	3	0	3	1	5
5	1	0	1	8	0	522	1	0	4	0
6	3	0	1	0	5	7	649	0	2	0
7	0	1	14	0	2	0	0	632	5	11
8	0	0	1	5	0	1	3	1	505	1
9	0	0	1	2	18	2	0	6	2	625

The confusion matrix above shows the accuracy of our predictions using 10-fold cross validation with k=3 and the Euclidean distance. This combination generally has the highest accuracy rate out of all the other combinations from question 7. Although most of the digits have accurate predictions, there are a few digits that have a slightly higher misclassification rate, such as predicting 9 as 5, 7 as 2 and 9, or 1 as 4. As mentioned earlier, these digits may have fairly similar structure when drawn, so it is reasonable that these digits have higher misinterpretations. From this matrix, we can see that although k=3 is a good classifier, there are still certain values that have a higher chance of error.

Question 9.



The error rates for the test data shows a similar trend to the train data set, although Manhattan distance with k=6 has a lower error rate than the other k values. The overall error rates are also higher than the error rates from the train data set because we only calculated the error rates once with the whole data set instead of averaging over a 10-fold cross validation that has more data points, so there is a higher variance using the test data.

Question 10.

For this final project all three of the group members contributed work towards its completion. For each problem we worked together simultaneously in order to understand the problem, create code, debug code, create visuals, and interpret results. Lijing specialized in drafting code and optimization. Lucy specialized in debugging code and writing interpretations. Adam specialized in formatting code and presenting results. The report was a combination of our efforts in order to assure that information was correct and visually consistent with our code and interpretations.