Final Project

Group 40: Xinzhe Xie Zhihao Meng Zihan Xiao

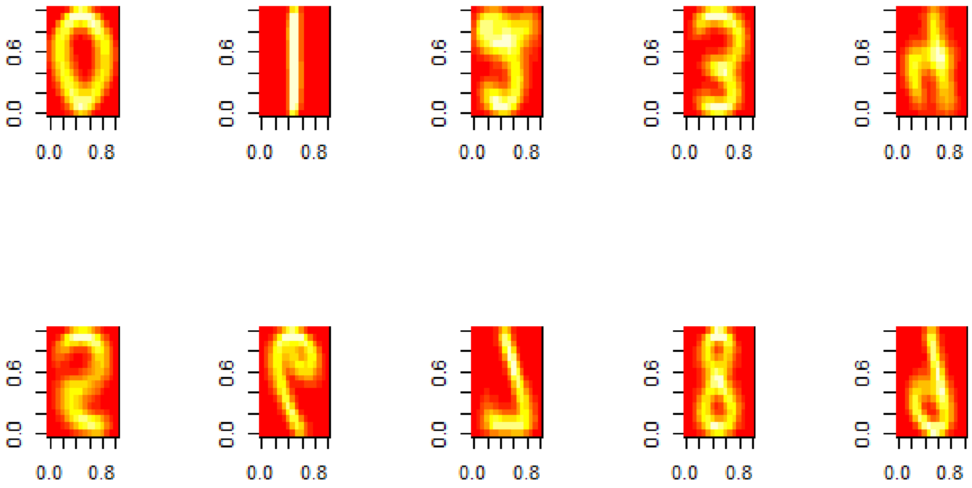
Questions:

1. Please see the Appendix R code.

2. Please see the Appendix R code.

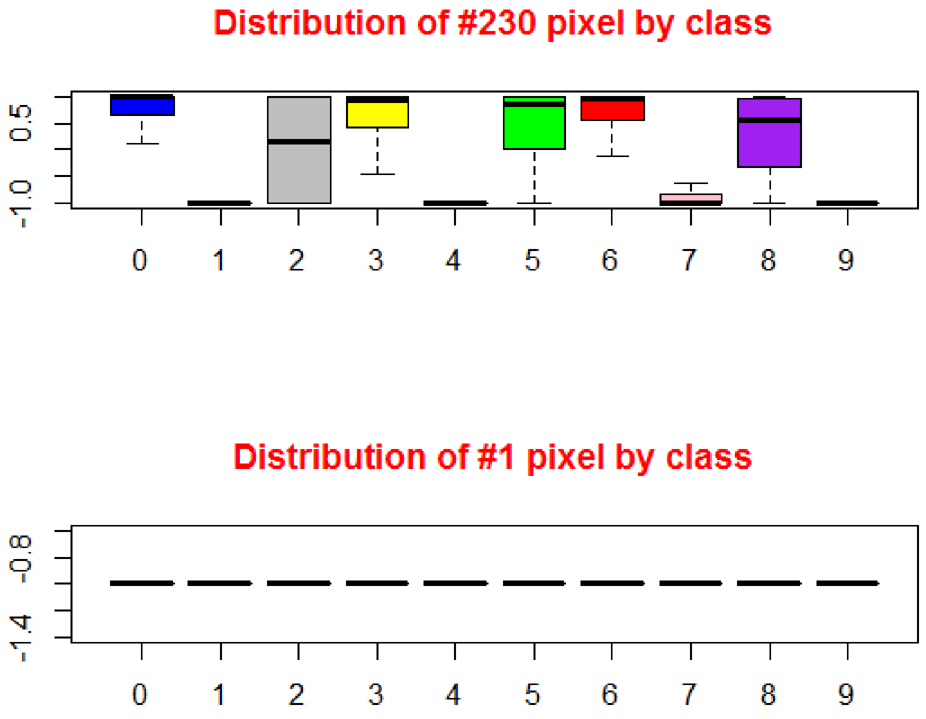
3.

(a) Graph 1 is what each digit (0 through 9) looks like on average.



Graph 1 The image of each digit (0 through 9) on average

(b) The #230 pixel seems the most likely to be useful. The #1 pixel seems the least likely to be useful. The most useful features for classification are the ones that take different values depending on the class of the observation. So we calculate the variance of the mean of every pixel for each class. The one with most variance is the one fluctuates most by class and observations can be classified by this pixel. By contrast, the one with least variance is the one fluctuates least by class and observations are hard to be classified by this pixel. The Graph 2 are the distributions of #230 and #1 pixel by class.



Distribution of #230 pixel by class

Distribution of #1 pixel by class

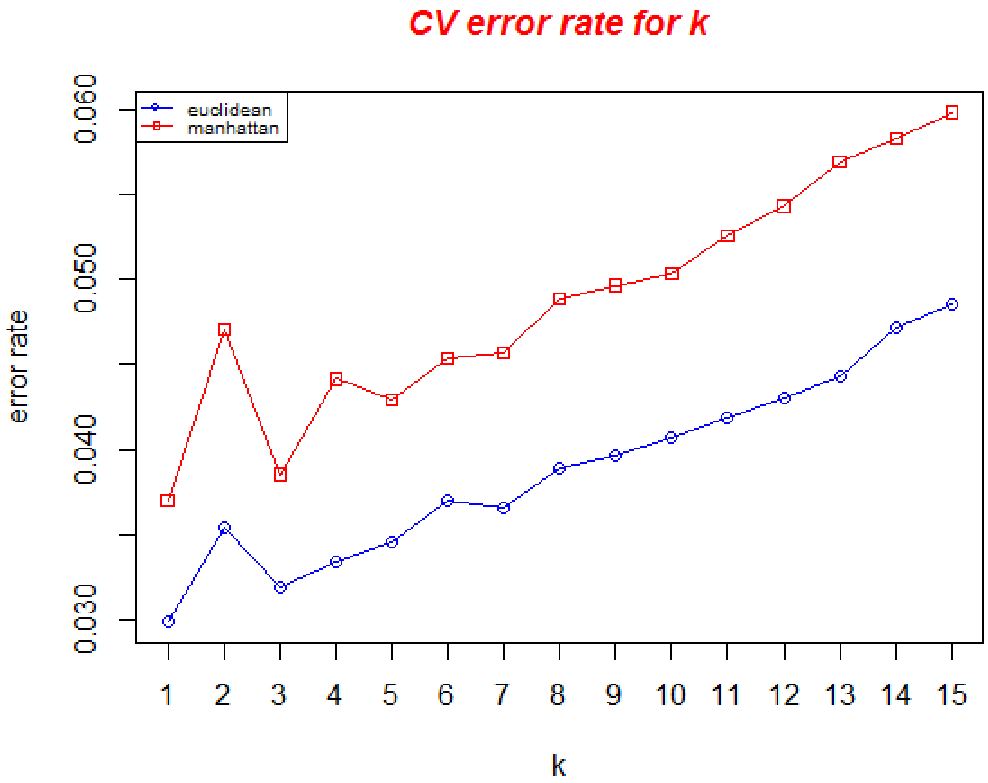
Graph 2 Distributions of #230 and #1 pixel by class.

From Graph 2, we can see that the distributions of #230 pixel are very different for each class and the distributions of #1 pixel are nearly the same for each class.

4. Please see the Appendix R code.

5. The estimated error rate for k-nearest neighbors is 0.04073398.
   The strategies we used to make the function runs more efficiently are:
   (a) The dist function wastes too much time but we only used it one time.
   We calculate the distance between each observations in the whole data firstly and save them in one matrix. We only need to extract some of the distances every time.
   (b) We divided the whole data into 10 groups and we labeled the observations in each group. When we predicted the class of one observation, we can match its group label and its actual class and compared the predicted class and actual class.

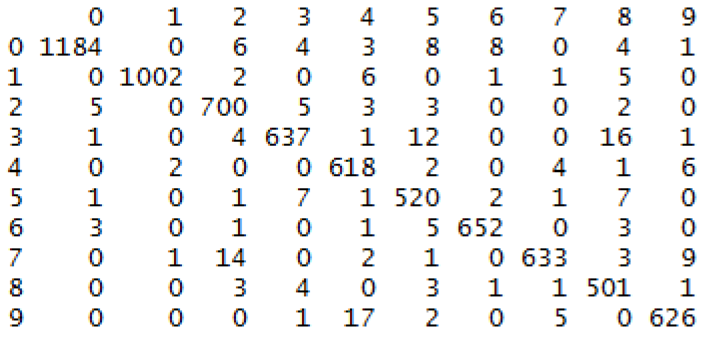
6. The Graph 3 are the results for 10-fold CV error rates for k=1,2………..15 with distance metrics Euclidean and Manhattan.



Graph 3 CV error rate for K with Euclidean and Manhattan

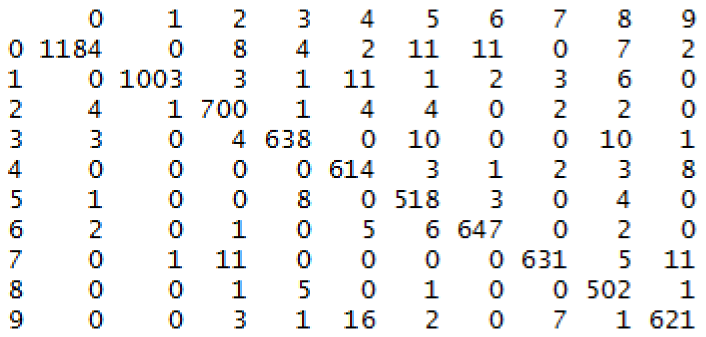
From the Graph 3, we can see that the combination K=1 with Euclidean is the best. The error rates with Euclidean are always lower than that with Manhattan and the error rates increase as K increases except for K=2. So it is not useful to consider additional values for K.

7. This is the confusion matrix for K=1 with Euclidean.



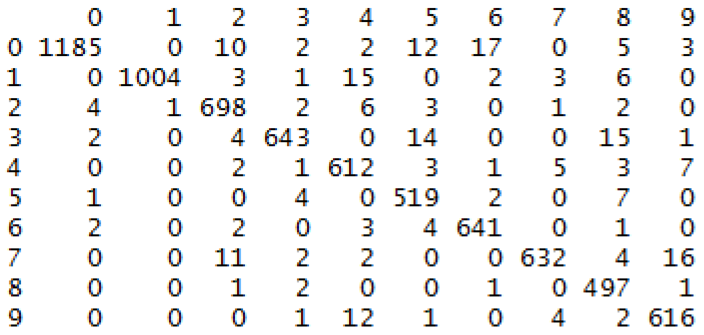
```
     0    1    2    3    4    5    6    7    8    9
0 1184    0    6    4    3    8    8    0    4    1
1    0 1002    2    0    6    0    1    1    5    0
2    5    0  700    5    3    3    0    0    2    0
3    1    0    4  637    1   12    0    0   16    1
4    0    2    0    0  618    2    0    4    1    6
5    1    0    1    7    1  520    2    1    7    0
6    3    0    1    0    1    5  652    0    3    0
7    0    1   14    0    2    1    0  633    3    9
8    0    0    3    4    0    3    1    1  501    1
9    0    0    0    1   17    2    0    5    0  626
```

This is the confusion matrix for K=3 with Euclidean.



```
      0    1    2    3    4    5    6    7    8    9
0  1184    0    8    4    2   11   11    0    7    2
1     0 1003    3    1   11    1    2    3    6    0
2     4    1  700    1    4    4    0    2    2    0
3     3    0    4  638    0   10    0    0   10    1
4     0    0    0    0  614    3    1    2    3    8
5     1    0    0    8    0  518    3    0    4    0
6     2    0    1    0    5    6  647    0    2    0
7     0    1   11    0    0    0    0  631    5   11
8     0    0    1    5    0    1    0    0  502    1
9     0    0    3    1   16    2    0    7    1  621
```

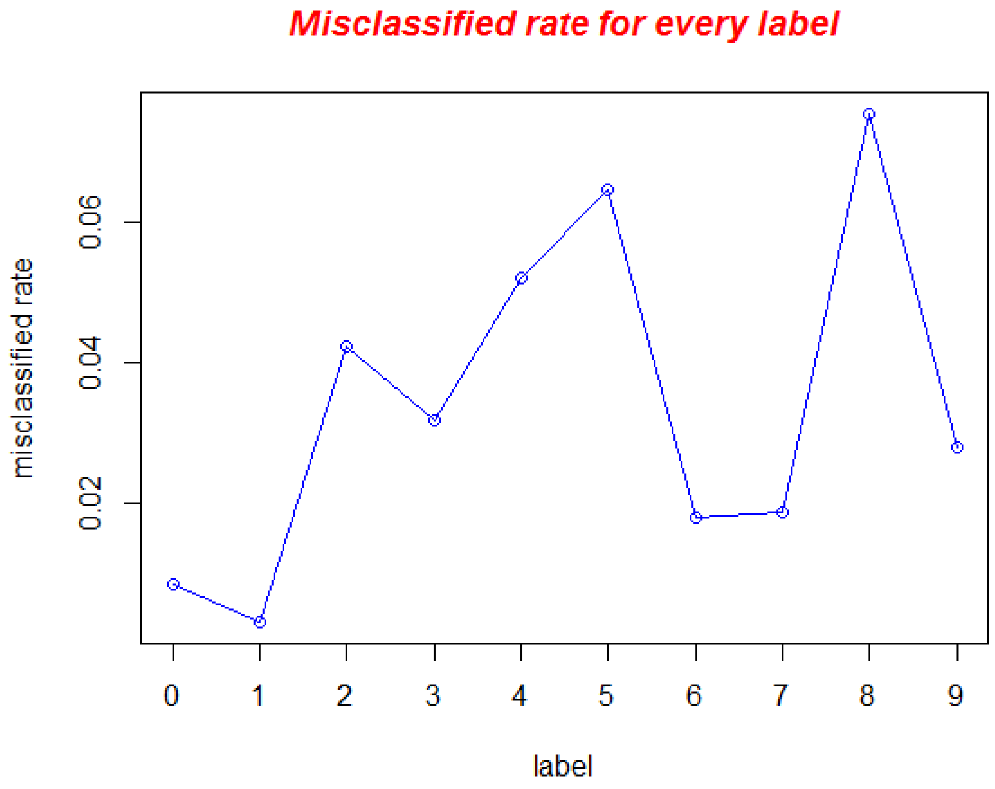
This is the confusion matrix for K=4 with Euclidean.



```
      0    1    2    3    4    5    6    7    8    9
0  1185    0   10    2    2   12   17    0    5    3
1     0 1004    3    1   15    0    2    3    6    0
2     4    1  698    2    6    3    0    1    2    0
3     2    0    4  643    0   14    0    0   15    1
4     0    0    2    1  612    3    1    5    3    7
5     1    0    0    4    0  519    2    0    7    0
6     2    0    2    0    3    4  641    0    1    0
7     0    0   11    2    2    0    0  632    4   16
8     0    0    1    2    0    0    1    0  497    1
9     0    0    0    1   12    1    0    4    2  616
```

From the three confusion matrix, we can see that the error rates for each K are very low and we don’t change the result of best combinations in question 6.

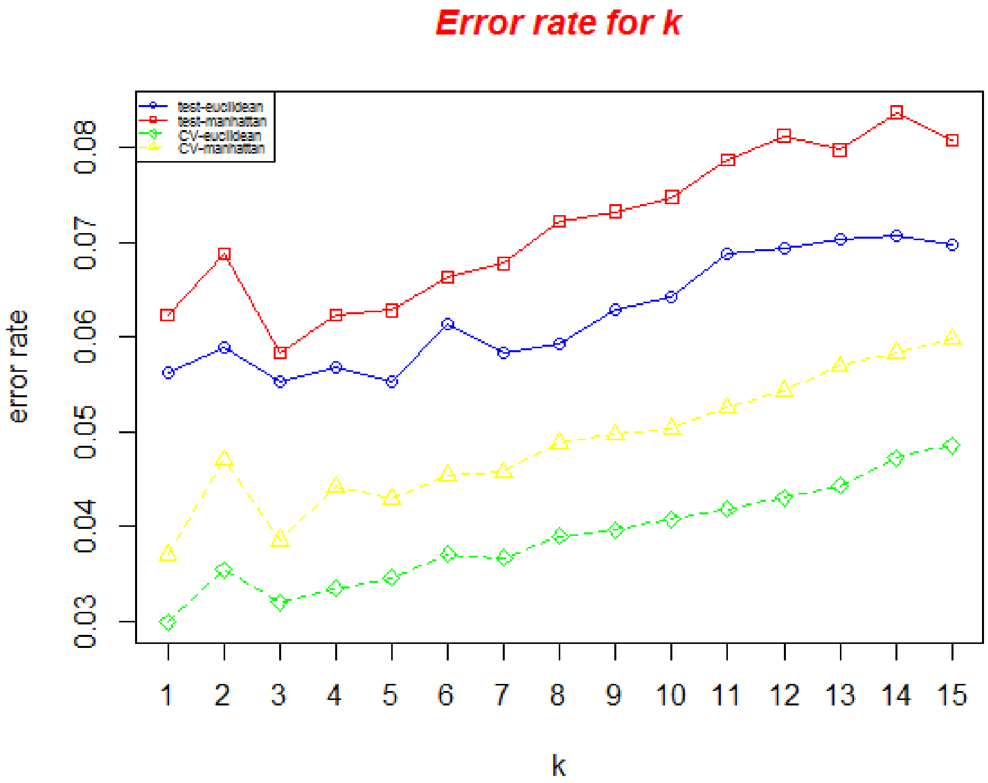
8. The Graph 4 are the results of misclassification during cross-validation.



Graph 4 Misclassified rate for each label

From Graph 4, we can see that the label 8 has the highest misclassified rate and the label 1 has the lowest misclassified rate.

9. The Graph 5 shows the test set data error rates for K=1,2………15 with Euclidean and Manhattan and the comparation with CV error rates.



Graph 5 Error rate for K

From Graph 5, we can see that the test set data error rates are always higher than CV error rates. And the features of test set data error rates are the same as CV error rates which are the error rates with Euclidean are always lower than that with Manhattan and the error rates increase as K increases except for K=2.

10. We all contribute very much to the group. We write the R codes and report together and there are no concrete assignments for everyone.