KNN Analysis HW #1 – Byron Becker

```
Byrons-MacBook-Pro:knn mmkay12345$ python knn.py
Done loading data
100/10000 for confusion matrix
200/10000 for confusion matrix
.
.
.
10000/10000 for confusion matrix
        0       1       2       3       4       5       6       7       8       9
-----------------------------------------------------------------------------------------
0:      982     0       2       0       1       0       2       0       1       2
1:      0       1060    1       0       1       0       1       1       0       0
2:      3       7       952     2       1       1       1       18      1       0
3:      0       0       4       1001    0       11      2       3       6       3
4:      0       9       0       0       949     0       0       2       0       19
5:      2       0       1       16      2       868     17      3       1       3
6:      1       0       0       1       0       1       964     0       0       0
7:      0       9       0       0       3       0       0       1071    0       6
8:      2       6       1       6       6       16      6       3       947     7
9:      2       2       0       7       11      4       2       6       2       919
Accuracy: 0.971300
```

Analysis

1. What is the relationship between the number of training examples and accuracy?

As the number of training examples goes up, the accuracy goes up as well. This makes sense since with more training data, the general area for each class receives more correctly classified data points, making outliers have less of an weight on estimating the class.

2. What is the relationship between k and accuracy?

Because k compares test data point to the k nearest points around it, the accuracy of k depends on several factors. First, if k is too small, it has a larger chance of having outlier(s) incorrectly classify and estimate the test data example, causing accuracy to be lower. However, if we make k too large compared to the total data set, the knn algorithm will start to include data points from other classes, incorrectly influencing the class of the test data point and lowering the accuracy. If k becomes on the order of m (#of training data samples), then the classifications of the test data points will end up being heavily biased by the majority class, and more inaccurate. Therefore, k is inaccurate at high relative values and very small values, but peaks in accuracy at values just large enough to filter out noise in the data. One method called bootstrapping recommends starting with the sqrt(m) as a good k.

3. What numbers get confused with each other most easily?

From the confusion matrix, we can see that the most common confusions (>10) are the following:

- 2 gets confused for 7,
- 3 gets confused for 5
- 4 gets confused for 9
- 5 gets confused for 3, 6
- 8 gets confused for 5
- 9 gets confused for 4