Machine Learning DRP project
mentor: Zhaoqi Li
mentee: Zhijun Peng

During the quarter, we talked about two method of categorizing in machine learning: nearest neighbor and K-means.

Nearest neighbor is the process of determining the label of the data by looking at its K closest data's label. And K-means is calculating centroid and repeatedly reassign data to each group and recalculating centroid to find the final clustering which each data is really close to its within cluster centroid.

We are using neatest neighbor and K-means clustering to categorize the data in MINIST dataset, which is a lot of images of handwritten digits from 0-9. We fit a multinomial logistic regression with L1 penalty on a subset of the MINIST digits classification task.

first using K-Means,
we got an average of squares of the distances of points from their respective cluster centriods is 4174.00157
and the square root of that is 64.6065134

then use nearest neighbor methods, we choose k equals 1 to 9, and calculate their accuracy scores.
then we got result

Test score with L1 penalty: 0.8027
Accuracy score for K-nearest neighbors : 0.8995
Accuracy score for K-nearest neighbors with k = 1 : 0.8982
Accuracy score for K-nearest neighbors with k = 2 : 0.8788
Accuracy score for K-nearest neighbors with k = 3 : 0.8995
Accuracy score for K-nearest neighbors with k = 4 : 0.8974
Accuracy score for K-nearest neighbors with k = 5 : 0.8999
Accuracy score for K-nearest neighbors with k = 6 : 0.8954
Accuracy score for K-nearest neighbors with k = 7 : 0.8957
Accuracy score for K-nearest neighbors with k = 8 : 0.894
Accuracy score for K-nearest neighbors with k = 9 : 0.8934

we can see the accuracy is highest when K= 1,
and the accuracy is decreasing as we increasing k.