MNIST Digit Classification
Approximating KNN with K-Means

Summary

K-Nearest Neighbors (KNN) is an accurate classifier, but it is just too slow. The MNIST handwritten digits have 60,000 training images, and 10,000 test images. A KNN classifier would need to compare each of 10,000 images against each of 60,000 images to find the K nearest neigbhbors. That's 600,000,000 image comparisons! Which is just too slow.

The purpose of project 3 is to speed up KNN, by approximating the MNIST training data using K-means clustering. You will split the MNIST training dataset (60,000 images) into groups based on it's label digit "0-9". Then you will run K-means with K=9 separately on each group. The result is an "Approximate Training Set" with only 90 images.

Once you have your "Approximate Training Set", you will classify the "Test Data" (10,000) using the 1-Nearest Neighbor model from each of the 90 cluster centers.

This algorithm is a fast approximation of KNN, because you compare each of 10,000 test images with only a total of 90 "approximate" train images to find the closest one for classification.

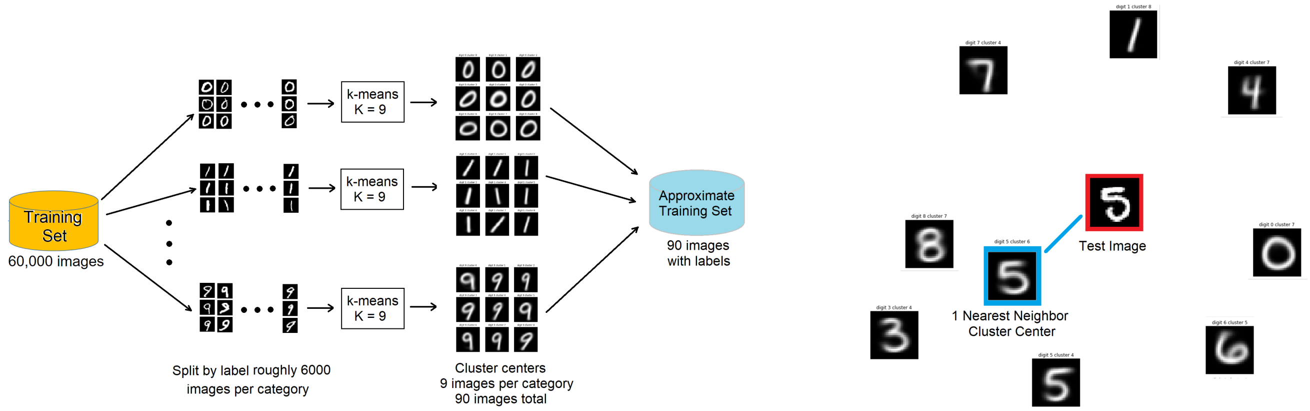

Figure 1. Illustration of the Classification pipeline (Left) constructing the *Approximate Training Set* (Right) 1NN classification using *Approximate Training Set*

Left: MNIST digit "training data" is first split into 10 groups, each with a different digit
then a separate k-means cluster is run to create 9 clusters per digit
the result is an "Approximate Training Set" with only 90 images.

Right: Classification with the 1-Nearest Neighbor (KNN with K=1) model.
Each test image is "classified" using L2 distance from the closest "cluster center"
from the Approximate Training Set.