Machine Learning Lab 7

K-Means Clustering

In previous labs, you have practiced supervised learning techniques such as KNN or Naïve Bayesian Classifier. In this lab, we will exercise one classical unsupervised learning algorithm: K-means clustering. As for the dataset, we still use the subset of MNIST which we used in our previous labs. However, this lab does not use any of the labels in the dataset. This is because unsupervised learning mines the data without labels. You should finish the following task:

1. Implement K-Means clustering as a function by your own instead of using built-in function from Matlab/Python. Note that your function should allow the user to easily choose methods for initialization (random centroid or any specified centroid) and the stopping criteria (those mentioned in the lecture notes).

2. Perform K-Means clustering on the MNIST data using your function with different Ks, initialization methods, and stopping criteria. You can calculate the distribution of the ten digits for each cluster to see how well your clustering performs. You may use `hist` in Matlab for plotting a histogram for distribution. Note that although there are ten digits in the dataset, you can try different Ks other than 10 to see whether you can cluster different digits with similar shape together.