CS 391L: Machine Learning **Summer 2024**

Homework 5 - Programming

Lecture: Prof. Qiang Liu

k-means Clustering on Images

In this problem, you'll do some basic exploration of the clustering techniques on the MNIST[1] dataset. This dataset contains a set of handwritten digits (see Figure 1), within 10 different classes (0, 1, ..., 9). Each image has a size of $784 = 28 \times 28$ (width $\times$ height). For this problem, we will use a subset of $n = 1000$ images that includes 4 types of digits $(0, 1, 2, 3)$.

Figure 1: MNIST samples

In the original dataset, each image is represented by a 784×1 vector. In order to make the task easier, we will project the images into a two-dimensional space with principle component analysis (PCA). We have implemented this in the notebook. Run the corresponding block in the notebook. You will find a two-dimensional scatterplot of the images, in which the coordinates of each point are the top two principle components of an images, and the color of each point represents its label in the dataset. We will work on this simpler two-dimensional dataset in our problem. Please follow the instruction in the notebook to complete the following tasks.

(a) **[5 points]** Implement the standard k-means algorithm. You are NOT allowed to directly copy any existing code of k-means for this problem.

(b) **[15 points]** Run your k-means function on the 2D dataset (of the top two PCA components). Set the number of clusters to be $k = 4$. Visualize the result by coloring the 2D points in (a) according to their clustering labels, returned by your k-means algorithm.

Because k-means is sensitive to initialization, repeat your k-means code for at least 5 times with different random initializations and show the plot of each initialization.

[1] http://yann.lecun.com/exdb/mnist/

To quantitatively evaluate the clustering performance, we evaluate the *unsupervised clustering accuracy*, which can be written as follows,

$$\text{accuracy} = \max_{\mathcal{M}} \frac{\sum_{i=1}^{n} \mathbb{I}(y_i = \mathcal{M}(z_i))}{n}, \quad n = 1000,$$

where y_i is the ground-truth label, z_i is the cluster assignment produced by the algorithm, and $\mathcal{M}$ ranges over all possible one-to-one mapping between clusters and labels and $\mathbb{I}(x)$ is a indicator function ($\mathbb{I}(x) = 1$ if $x = 1$; otherwise 0). Calculate the best clustering accuracy you get out of 10 random initializations.

(c) [**10 points**] We have been testing k-means on the top two principal components for visualization purpose. Please run k-means on the (784 dimensional) original image dataset (use again $k = 4$ clusters). Try at least 10 different random initializations and show the best accuracy as above.