Introduction to Machine Learning, Spring 2023
Homework 4 Coding

(Due Friday, May. 5 at 11:59pm (CST))

April 22, 2023

1. [50 points] Given a dataset $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n \in \mathbb{R}^d$ and an integer $1 \leq k \leq n$, recall the following k-means objective function

$$\min_{\pi_1,\cdots,\pi_k} \sum_{i=1}^{k} \sum_{j\in\pi_j} \|x_j - \mu_i\|_2^2, \mu_i = \frac{1}{\pi_i} \sum_{j\in\pi_j} x_j$$

Above, $\{\pi_i\}_{i=1}^k$ is a partition of $\{1, 2, \cdots, n\}$. The objective is NP-hard to find a global minimizer. Nevertheless the commonly used heuristic is Lloyd's algorithm.

Implement Lloyd's algorithm for solving the k-means objective function. Do not use any off the shelf implementations, such as those found in scikit-learn.

(a) Run your algorithm on MNIST with $k = 5, 10, 20$ cluster centers. Use the image data at MNIST data/images.mat, which contains 60000 unlabeled images, and each image contains 28×28 pixels. Each pixel represents one coordinate in the k-means algorithm. [25 points]

(b) Visualize the centers you get, viewing each coordinate as a pixel (i.e., each center is represented by an image of 28×28 pixels). What are the differences between results with different numbers of cluster centers? [25 points]