CS554 Homework 2: *k*-Means Clustering for Unsupervised Learning
Spring 2023/2024

In this homework, your task is to implement *k*-means clustering on two different datasets, a synthetic 2D dataset and the MNIST dataset. You are provided with a compressed file **datasets.zip** that contains both data sets.

Part 1. The **data.csv** file contains two-dimensional instances in its rows.
Implement *k*-means clustering using the **data.csv** file for $k = [1, 2, 3, 4, 5, 6]$:
- For each *k* value, repeat the training process ten times starting from different initial centers.
- Plot the mean reconstruction loss over those ten runs vs. *k* in a single plot.
- For each *k*, for the run with the smallest reconstruction loss (among ten), plot the clustering assignments at the end.

Part 2. The **mnist.csv** file contains the MNIST dataset composed of 28x28 images from 10 different classes. Each class represents a different type of digit item but in this homework, you are not going to use the class information. Each row of the file corresponds to one instance and there are 10,000 instances. The first column of each row contains the label of the image (which you will ignore), and the remaining 784 columns that represent the grayscale value of each pixel constitute the input image that you will use to calculate the cluster centroids.

Implement *k*-means clustering using the **mnist.csv** file for $k = [10, 20, 30]$:
- Plot the reconstruction loss for all three *k* in a single plot.
- Visualize the centroids for each *k* as grey-scale images.

You can use the pandas and numpy libraries for your implementation.

This homework is due **March 26th (Tuesday), 23:00**.
Your submission should include a short report of your findings, the plots, and your source code.
Upload your report **as a pdf file** to LMS alongside your .py/.m code file. Do not compress your submission.