CS 410/510: Deep Learning: Computational Structures and Programming
Winter 2021
Assignment 4
Due Thursday March 4th by 11:59pm

1. *k-means clustering:* From the MNIST data set, pick 100 samples from each of the 10 classes. Take all these 1,000 images and run them through a k-means clustering algorithm (k=10). You can use the scikit-learn library. Here is a link: https://scikit-learn.org/stable/modules/clustering.html#k-means

   The examples provided should be sufficient to enable you to write the code in python and run it.

   a. *What is the accuracy of the clustering? To measure accuracy, simply count the number of images from class a that are clustered into class b (for all a and b in 0..9). Produce a table of size 10x10. Each column and row is labeled 0-9. An entry (i,j) is a count of how many members from class i are clustered into cluster for class j. How would you cmpare the accuracy of this method with what you achieved previously in the first assignment?*

2. *k-means clustering using feature vectors:* This time we will take 1,000 images from each of the 10 MNIST classes giving us 10,000 images. Build an autoencoder and train it. You can use the code in the second part of this page: https://www.cs.toronto.edu/~lczhang/360/lec/w05/autoencoder.html

   After training, you obtain the feature vectors by using the encoder part alone (model.encoder(x)). Then use k-means clustering of these feature vectors.

   a. *As above, produce a table and then explain any differences between these results and the ones obtained in part 1 above.*

   b. *(Bonus 5 points) Run PCA on the feature vectors, reduce the dimensionality to just a few most important dimensions and then do the clustering.*