Wright State University
Computer Science and Engineering

CS7900-02– Assignment1
Algorithms for Biological Data

Instructor: Dr. Tomojit Ghosh

Due by 02/08/2025 11:59 PM EST

Name: ______________________________

Student Id: ___________________________

This exam contains 2 pages (including this cover page) and 3 questions. Total of points is 100. Good luck and Happy reading work!

Distribution of Marks

Question	Points	Score
1	20	
2	30	
3	50	
Total:	100	

1. (20 points) In this assignment you will work on a subset of MNIST data which has 10 digit classes. The digits and the labels are kept in two separate files as Python's pickle object (*.p). Run PCA on the data. Project the data in two and three dimensional space. Show the projected data (only 100 samples taken randomly per class) using a scatter plot. Your figure should have proper legend for each digit class.

2. (30 points) Consider the same data set. How many eigenvectors are needed to capture the 99% of the total variance of the data? Plot the average reconstruction error for each digit class separately as a function of eigenvectors. Comment on your observation.

3. (50 points) In this assignment you will compare classification results of PCA, LDA and SLCE as a function of embedding dimension. Partition the MNIST data by taking 80% samples from each class randomly into training set and the rest in the test set. Fit each model on the training partition. Then project the training and test samples in $1, 3, 5, \dots 9$-dimensional space and calculate 5NN (five nearest neighbor) accuracy for each embedding dimension. Repeat the experiment ten times and plot the average accuracy curves as a function of embedding dimension and comment on your observation.

   Now, for PCA plot the reconstruction error and the 5NN accuracy as a function of embedding dimension $10, 20, \dots 150$. Comment on you observation.

   Note: Nearest Neighbor is a simple classifier. To predict the class label of a test sample, you need to find out its k-nearest neighbors from the training set, where k is user select (could be 1,3,5,7,...etc). Given the label information of the training set, assign the label of the test samples by majority voting. Example: Assume $k = 5$ and you want to assign the class label of a test sample x^i_{test}. Find out the five nearest neighbors of x^i_{test} from the training set using Euclidean distance. If the majority of the neighbors of x^i_{test} belongs to class C_j, then assign the sample x^i_{test} to the j^{th}-class. For this assignment, use scikit-learn package: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

   Note: You should submit a pdf file with you answers. Make sure to include the code snippet, written in Python/PyTorch for each questions. The figures should have proper legend, labels to X and Y axixes.