SYDE: Introduction to Pattern Recognition

Assignment 1

Due: 11:59 PM (EST), Sep 30, 2022, submit on LEARN.

Include your name and student number!

Submit your writeup in pdf and all source code in a zip file (with proper documentation). Write a script for each programming exercise so that the TAs can easily run and verify your results. Make sure your code runs!

[Text in square brackets are hints that can be ignored.]

Exercise 1: Nearest Neighbor Classifier (35 pts)

In this exercise you will be using the MNIST dataset for image classification. To work with this dataset, you will first need to flatten your images from 28×28 to 784×1 vectors. Next, use the PCA in scikit learn to convert the 784×1 vectors to 2×1 vectors. Note that the dataset consists of the training and the test sets. Use the training set for implementing the classifiers in the exercise. Also, use only the first two classes in the dataset i.e. the two classes representing numbers 0 and 1.

1. (15 pts) Implement the k-nearest neighbour classifier on the MNIST dataset. Use euclidean distance as the distance metric. Compute the kNN solution for each integer k from 1 to 5. Plot the classification boundaries between the two classes for the kNN classifier for each value of k between 1 and 5.
2. (15 pts) Use the test set of the two classes in the MNIST dataset, and make label predictions for all the test vectors for the two classes using your kNN classifier. Find the prediction error for your classifier as:

$$error = \frac{Number\ of\ incorrect\ predictions}{total\ number\ of\ data\ points\ in\ the\ test\ set.} \tag{1}$$

   Plot the test set error for each value of k.
3. (5 pts) Which k value seems to be producing the best results? Why?

[NOTE You are not allowed to use any special libraries in Python unless mentioned otherwise. You are allowed to use the basic Python and Numpy libraries.]

Exercise 2: MED and GED Classifiers (65 pts)

In this exercise you will be using the MNIST dataset as in the previous exerciese. However, unlike the previous exercise, use the PCA in scikit learn to convert the 784×1 vectors to 20×1 vectors. Use the training set for implementing the classifiers in the exercise. Also, use only the first two classes in the dataset.

1. (20 pts) Implement the MED and GED classifiers on the MNIST dataset. Determine the decision boundaries for the two classifiers. Can you plot this decision boundary? Explain.
2. (10 pts) Use the test set of the two classes in the MNIST dataset, and make label predictions for all the test vectors for the two classes using the MED and GED classifiers. Plot the test set error for both your classifiers in two separate plots.
3. (5 pts) Which classifier is better? Why?
4. (15) Convert the training set images for the two classes in MNIST to 2×1 vectors, and plot the decision boundaries for the MED and GED classifiers. Do you think MED and GED classifiers are better than the kNN classifier? Explain.
5. (15 pts) Find the confusion matrices for the MED, GED, and kNN classifiers (k=1 and k=5). Find the confusion matrices using the test set of the two MNIST classes. Use the 2×1 vectors for the data.

[NOTE You are not allowed to use any special libraries in Python (such as Scikit Learn) unless mentioned otherwise. You are allowed to use the basic Python and Numpy libraries.]