SYDE: Introduction to Pattern Recognition

Assignment 1

Due: 11:59 PM (EST), Feb 6, 2024, submit on LEARN.

Include your name and student number!

Submit your write-up 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.]

NOTE For all the exercises, you are only allowed to use the basic Python, Numpy, and matplotlib (for plotting) libraries, unless specified otherwise.

Exercise 1: MED and MMD Classifiers (35 pts)

In this exercise, you will be using the MNIST dataset for image classification. This is a common dataset in machine learning and pattern recognition and it is freely available from multiple sources online. 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 two classes in the dataset i.e. the two classes representing numbers 3 and 4.

1. (15 pts) Implement the algorithm for the MED and MMD classifiers on the training set of the MNIST dataset. Determine the decision boundary for the two classifiers. Plot the boundary along with the training data.
2. (2 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 MMD classifiers. Report the classification accuracy of the classifiers using:

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

3. (3 pts) Which of the two classifiers is better? Explain.
4. (15 pts) Now, for the same two classes in the MNIST dataset, use PCA to convert the images to 20×1 vectors. Repeat steps 1-3 for this dataset. For step 1, can you plot the decision boundary for this dataset? Explain.

Exercise 2: Nearest Neighbor Classification and Regression (45 pts)

1. (15 pts) Implement the k-nearest neighbour classifier on the MNIST dataset with 2×1 vectors for the same two classes used in the previous exercise. Use euclidean distance as the distance metric. Compute the kNN solution for each integer k from 1 to 5.
2. (3 pts) Use the test set of the two classes and find the classification accuracy for all kNN classifiers. Plot the accuracy for each value of k.
3. (2 pts) Which k value seems to be producing the best results? Why?
4. (5 pts) How does the kNN classifier compare against the MED and MMD classifiers in the previous exercise?
5. (20 pts) Now let's do kNN regression, which is similar to classification, but instead of aggregating labels by taking the majority, we average the y values of the k nearest neighbours. For this, use the training set of the $d = 20$ mystery dataset F available on the course website, and compute the kNN regression solution with each integer k from 1 to 3. Report the mean squared error using the test set of the mystery dataset F for the three kNN classifiers. Which k value seems to be giving you the best result? Explain.

Exercise 3: ML and MAP Classifiers (20 pts)

In this exercise, you will use the same data as in exercise 1 with 2×1 vectors, for the ML and MAP classifiers.

1. (5 pts) Find the sample mean and covariance for the training set of the two classes in the MNIST dataset and estimate the probability of the two classes as Gaussian distributions. Based on this, develop an ML classifier and report the classification accuracy on the test set of the two classes.
2. (5 pts) Now, let's assume that the prior probabilities for the two classes are $p(C_1) = 0.58$ and $p(C_2) = 0.42$. Using these prior probabilities, and the means and covariances of the two classes, develop an MAP classifier and report the classification accuracy on the test set.
3. (5 pts) Based on the results, do you think that assuming the probability distributions of the two classes as Gaussian was correct? Explain.
4. (5 pts) Compare the ML, MAP, MED, MMD, and kNN classifiers based on the classification accuracy. Which classifier is the best? Could the inferior classifiers be better for different datasets? Explain.