SYDE: Introduction to Pattern Recognition

Assignment 3

Due: 11:59 PM (EST), Nov 3, 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: Non-Parametric Estimation (60 pts)

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.
In this exercise you will be using the MNIST dataset. Use PCA in Scikit learn to convert the 784×1 vectors to 1×1 vectors (1-d data). Use the training set for implementing the classifiers in the exercise. Also, use only the first two classes in the dataset.

1. (25 pts) Use histogram based estimation with region sizes of {1,10,100} to estimate the probability distribution of your dataset. Note that you will need to estimate the distribution of each class separately. Plot your probability distributions for all the region sizes. Which region size seems to be the best for estimating the probabilities? Explain.

2. (5 pts) Use the estimated distributions with an ML classifier to predict labels of your test data. Report the test error for using the distributions estimated with all the region sizes. Which region size seems to be the best in terms of the test error? Explain.

3. (20 pts) Repeat the above two steps using the kernel-based density estimation. You can use a Gaussian kernel with $\sigma = 20$

4. (5 pts) Which of the two approaches (histogram or kernel) for probability estimations best represent your data and produce the lowest error? Explain.

5. (5 pts) Compare your results in this exercise with the results for parametric estimations of your data in exercise 2 of assignment 2. Do you think non-parametric approaches are better than the parametric approaches for density estimation? Explain.

Exercise 2: K-means clustering (40 pts)

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. You are also allowed to compare your results with the Scikit implement of k-means, but you cannot use the Scikit implementation as your own solution.
In this exercise you will be using the MNIST dataset as in the previous exercise. You will also be using 784×1 directly without any PCA. You can use the training data for this exercise. Use all the classes in the dataset.

1. (10 pts) Implement the k-means clustering algorithm with euclidean distance as the distance metric.

2. (10 pts) Apply your k-means implementation to the MNIST dataset. Use k={5,10,20,40}. Also, you will not be using class labels during the clustering process as it is an unsupervised learning problem.

3. (7 pts) As we do not use class labels for clustering we cannot use test error to evaluate the k-means algorithm. Instead, we can test our clustering implementation using cluster consistency: all data points in a cluster should belong to the same class. For example, if your cluster number 1 has all the data points belonging to class "5" then this cluster has perfect consistency. You can find the cluster consistency Q_i for each cluster i as follows:

$$Q_i = \frac{m_i}{N_i} \tag{1}$$

where, N_i is the total number of data points in cluster i, and m_i represents the total number of data points belonging to the most common class in cluster i. You can find m_i by first counting the total number of data points belonging to each of the 10 classes in cluster i, and then taking the max of this

list. To find the overall clustering consistency:

$$Q = \frac{1}{k} \sum_{i=1}^{k} Q_i \tag{2}$$

Report the cluster consistency of your k-means clustering on the MNIST dataset for all four values of k.

4. (8 pts) Which k value produces the best results? Explain. Can the results from cluster consistency be misleading? Explain. [HINT Intuitively, what k value should produce the best results on the MNIST dataset?]

NOTE: If timing is an issue, you can use 100 data points for each class.