SYDE: Introduction to Pattern Recognition

Assignment 2

Due: 11:59 PM (EST), Oct 20, 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: Normal Parametric Estimation (40 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 (2-d data). 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. (14 pts) Assume that probability distribution of your dataset is a multi-variate Gaussian Distribution. Using maximum likelihood estimation (MLE), derive the formulae for estimating the two parameters of the Gaussian distribution i.e. mean and the covariance matrix. Make sure that this is a general expression that can be applied to any d dimensional data.

2. (2 pts) Using the derived expressions for the two parameters of the Gaussian distribution, find the probability distribution of the MNIST dataset. Note that you will have to estimate the distributions of the two classes separately. So, you will find two mean vectors and two covariance matrices.

3. (5 pts) Use the maximum likelihood (ML) classifier on the estimated class distributions $p(X|C)$, to make predictions on your test set. Find the prediction error for your ML classifier.

4. (3 pts) Plot the decision boundary for your classifier. Note, you don't have to derive the boundary mathematically.

5. (1 pt) Now, let's find the prior probabilities for the two classes ($p(C_1)$ and $p(C_2)$) using the class statistics: $p(C_k) = N_k/N$, where N_k is the number of training data points for class C_k, and N is the total number of training data points.

6. (10 pts) Using the prior probabilities, and the estimated class distributions $p(X|C)$, consider the MAP classifier. Make predictions using the MAP classifier on the test set. How does this compare to the ML classifier? Explain.

7. (5 pts) Compare the results for your MAP and ML classifiers with MED and GED classifiers from the previous assignment. Explain.

[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: Parametric Estimation with Multiple Distributions (40 pts)

In this exercise you will be using the MNIST dataset as in the previous exercise. However, unlike the previous exercise, use the 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. (10 pts) Let's assume that the probability distribution of your dataset is an exponential distribution (Wiki). In this case, you only need to estimate a single parameter λ. Use MLE to derive an expression for finding λ.

2. (10 pts) Now, let's assume the probability distribution of your dataset is uniform (Zhang Ch. 19.8). In this case, you will estimate two parameters a and b. Use MLE to derive an expression for finding the two parameters.

3. (15 pts) Use the ML classifier on the estimated class distributions from exponential estimation, uniform distribution, and Gaussian distribution (you will need to recompute the mean and covariance matrices for 1-d data), to make predictions on the test set. Find the prediction error for the three different ML classifiers (for the three estimated probability distributions).

4. (5 pts) Which of the estimated probability distributions best represent your data? Explain.

[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 3: Parametric Estimation with Two Gaussian Distributions (20 pts)

For this exercise, consider that you have 1-d data.

1. Let's assume the probability distribution of your dataset is not just a single Gaussian distribution but a product of two Gaussian distributions:

$$p(X|\theta) = \frac{1}{2}[\mathcal{N}(X|\theta_1, \theta_2) \times \mathcal{N}(X|\theta_3, \theta_4)] \tag{1}$$

   The only difference between this expression and the one in your notes for parametric estimation, is that instead of estimating the mean and variance for one Gaussian, you will be estimating the mean and variances of two Gaussian distributions. Using MLE, derive an expression for finding the four parameters of your probability distribution defined by eq. (1).