Statistical Machine Learning (SML)

Winter 2023

Assignment 1

Maximum Marks - 50

Instructions For Theory Questions:

1. We will place a submission box, you will need to submit the handwritten copy of your solutions.
2. You also have to submit a scanned copy of your handwritten solution on google classroom.

Instructions For Coding Question:

1. You are free to use either python or MATLAB for this assignment.
2. You can use inbuilt libraries for Math, plotting, and handling the data (for example, NumPy, Pandas, Matplotlib).
3. Usage instructions for other libraries can be found in the question.
4. Only (*.py) and (*.m) files should be submitted for code.
5. Create a (*.pdf) report explaining your assumptions, approach, results and any further detail asked in the question.
6. You should be able to replicate your results if required.

A. Coding.

[30 Marks] In this problem, you will explore classification based on discriminant analysis.
Dataset : MNIST

(a) Download the dataset, and visualize 5 samples from each class in the form of images.

(b) Implement Linear Discriminant analysis(LDA) from scratch.Follow these instructions:

  (i) Compute a covariance matrix using weighted average of covariance of each class and let's call it Σg.

  (ii) Assume Σg to be the covariance for all classes while implementing the Linear Discriminant function. (Hint: recall the discriminant function with assumption of same covariance matrix for each class)

(c) Use LDA parameters calculated from training data to predict the class of samples from testing data.

(d) Implement the Quadratic discriminant analysis(QDA) from scratch. Follow these instructions:

(i) You will need to calculate the covariance matrix of each class separately and they will be different.
(ii) Implement Quadratic discriminant function.(Hint: recall the discriminant function with different covariance matrices for each class)

(e) Use QDA parameters calculated from training data to predict the class of samples from testing data.
(f) Implement an accuracy function (Not allowed to use inbuilt library) and report accuracy on testing data for both LDA and QDA.
(Accuracy = No of correctly classified samples/total samples)
(g) Use LDA and QDA functions of the sklearn library (or matlab function) and compare the results with your implementation.

B. Theory

1. **[3 Marks]**Suppose we have three equi-probable categories in two dimensions with the following underlying distributions:
P(x|w1) ~ N (mean =[0,0] , covariance = I)
P(x|w2) ~ N(mean =[1,1], covariance = I)
P(x|w3) ~ 0.5N(mean=[0.5,0.5],covariance= I) + 0.5N(mean=[-0.5,0.5],covariance= I)

   Where I is an identity matrix and N is normal distribution. By explicit calculations classify the point x=[0.3,0.3] for minimum probability of error.

2. **[5 Marks]**Suppose two equally probable one-dimensional densities (Laplacian pdf) are of the form

$$p(x|\omega_i) = \frac{1}{2b} e^{-\frac{|x-a_i|}{b}}$$

   for i=1,2 and a1 = 0, a2 = 1, b = 1.

   a. Calculate likelihood ratio and plot against x.
   b. Compute optimal/minimum error rate decision boundary for zero-one loss.
   Note: This can be obtained by equating both posterior densities. Using a sketch indicate how posterior densities will appear and also mark the decision boundary.
   c. Compute average probability of error.

3. **[5 Marks]** Prove that the below given estimate of variance is unbiased:

$$\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

(Note: x bar is mean MLE)

4. **[2 Marks]** Let's say you have a one dimensional random variable X with mean μ and standard deviation σ and scalars a,b. Find
    a. E [a*X+b] (Note: E denotes expectation)
    b. Variance(a*X+b)

5. **[5 Marks]** Let's define "probability of correct given x" as P(c|x) = 1 – P(error|x). Plugin this formulation into probability of error and explicitly derive "probability of correct P(c)" for Cauchy distribution example covered in class.