Statistical Machine Learning (SML)

Winter 2023

Assignment 1

Maximum Marks - 50

---

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 PCA and FDA.

Dataset : MNIST (Note: for every part just consider class 0 and 1 of MNIST, both for training and testing)

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

(b) Implement PCA from scratch.
(Note: PCA matrix will be calculated from the training data only)

(c) Take n= 2,3,5,8,10,15 to project your input data (Note: here n is the number of eigenvectors).

(d) Reconstruct the original data from the projected data you got in part c. Plot reconstruction error vs n.

(e) Apply pca on Input data and use LDA to classify the testing samples. Perform this part for all n given in part (c) and report accuracy on the test set for all cases.
(Note: you can use sklearn for LDA in all parts, parameters of LDA needs to be calculated using training data only)

(f) Implement FDA from scratch. (FDA vector will be calculated using training data only)

(g) Perform FDA on input data and use LDA to classify testing samples, report the accuracy.

(h) Apply pca with best n value from part (e) and then apply FDA, now use LDA to classify testing samples and report the accuracy.

B.Theory (Submit the hardcopy in the box outside B-611 R&D block)

1. [5 Marks]Compute pca weight matrix for the matrix given below.(Show all the steps involved)

$$\begin{vmatrix} 1 & 3 \\ 4 & 7 \end{vmatrix}$$

2.[5 Marks] Suppose there is a system that uses FDA before binary classification. Assume that you have complete access to the system but your role is that of an attacker. As an attacker, you would want that the classification performance is as poor as possible. Since you know that the system applies FDA, you decide to find a vector w which is obtained in the following way. Given the data matrices for two classes **X1** and **X2**, first project them using w. Assume the number of samples in each class to be N/2. Then, assume that the projected samples (including both the classes that is all N samples) follow a Bernoulli distribution with known parameter θ as well as these samples are iids. In order to attack the system, a constraint is also introduced which is $\mathbf{w^T\mu 1 = w^T\mu 2}$, where μ1 and μ2 are the respective means of the two classes. Find an expression for w.Express w in terms of μ1, μ2, θ, N.

3. [10 Marks] Let

$p(x|\omega 2) \sim N(\mu, I)$
$p(x|\omega 3) \sim N(-\mu, I)$

where I is 2 × 2 identity matrix and x is a 2-dimensional vector. Let us project x using u. Let the projections be called y. We want that the probability of projections belonging to respective classes be maximum. This mean we have to perform

$$\max_{\mathbf{u}} \ln p(y|\omega_2; \mathbf{x}) p(y|\omega_3; \mathbf{x})$$

$$p(y|\omega_2; x) = \frac{1}{Z} exp\{-.5(y - \mathbf{u}^\top \boldsymbol{\mu})^\top (y - \mathbf{u}^\top \boldsymbol{\mu})\}$$
$$p(y|\omega_3; x) = \frac{1}{Z} exp\{-.5(y + \mathbf{u}^\top \boldsymbol{\mu})^\top (y + \mathbf{u}^\top \boldsymbol{\mu})\}$$

Z is a normalization constant. $p(y|\omega i)$ i = 2, 3, denotes the probability of

projected encodings of x conditioned on respective classes. However, this does not capture FDA assumptions which can help in better discrimination. Using Equation 1 and knowledge of FDA, find an expression for u. Note you need to create a function of u which has both Equation 1 and objective of FDA. Assume the total scatter matrix is known to be S. u should be expressed in terms of S, μ, number of samples in each class N.