Statistical Machine Learning (SML)

Winter 2021

Assignment 2

Maximum Marks - 100

Due Date: 24th Feb, 23:59 hrs

Instructions:

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 (eg. 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. **[30 Marks] In this problem, you will explore classification based on discriminant analysis.**

1. **[4 Marks]** Generate 200 multivariate (dimension = 2) normally distributed samples (**Note**: you can use inbuilt functions to generate these samples), 100 of those samples (**Class 1**) should be generated from $\mathcal{N}(\mu1=\begin{pmatrix}0\\0\end{pmatrix}, \Sigma1=\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix})$ and the other 100 (**Class 2**) should be generated from $\mathcal{N}(\mu2=\begin{pmatrix}1.5\\1.5\end{pmatrix}, \Sigma2=\begin{pmatrix}0.8 & 0\\0 & 0.8\end{pmatrix})$.

2. **[4 Marks]** Select 50 samples out of 100 samples from class 1, call them **training set** and use those samples to compute the class conditional parameters ⇒ Mean(μ) and Covariance (Σ) for class 1. The remaining 50 samples will be used later for classification, call them **test set**.

3. **[4 Marks]** Repeat step (b) for class 2 samples.

4. **[4 Marks]** Plot training samples using scatter plot (**Note**: choose a different color for different classes).

5. **[8 Marks]** Use appropriate **Discriminant Function** for normal density discussed in class to classify the samples from the test set.

6. **[2 Marks]** Compute the number of samples correctly classified in each class.

7. **[4 Marks]** Plot the decision boundary.

B. [30 Marks] In this problem you will explore Maximum Likelihood Estimation (MLE) and Maximum a posteriori (MAP) estimation for Bernoulli distribution.

1. **[4 Marks]** Generate 1000 samples from Bernoulli distribution with p(x=1) = 0.2. (**Note:** you can use inbuilt functions for this part)

2. **[4 Marks]** Compute MLE using the samples generated in part (a).

3. **[8 Marks]** Now compute MLE using only n samples at a time where n = 1,2,3,.....,1000. Analyze MLE vs n plot and report your observations.

4. **[14 Marks]** Now assume that prior follows beta distribution with alpha=2 and beta=5. Compute and plot MAP estimate using n (=1,2,3,4,....,1000) points at a time. Compare the results with MLE and report the number of samples required for a good estimate in MLE and MAP.

C. [40 Marks] In this problem you will implement and analyze Principal Component analysis (PCA).

1. Generate 100 multivariate Gaussian points (dimension=2) with $\mu=\begin{pmatrix}-1\\-1\end{pmatrix}$ and $\Sigma=\begin{pmatrix}2 & 0.5\\0.5 & 1\end{pmatrix}$ Call it X1.
   Generate 100 more points with $\mu=\begin{pmatrix}1\\1\end{pmatrix}$ and $\Sigma=\begin{pmatrix}2 & 0.5\\0.5 & 1\end{pmatrix}$ Call it X2.
   Form Z = [X1 X2] (horizontal concatenation), check that Z is a d X n matrix where d=2 and n=200. **[4 Marks]**

2. **[4 Marks]** Plot X1 and X2 using scatter plot (**Note:** use different colors for X1 and X2).

3. **[4 Marks]** Compute the mean of Z as μz and create a new matrix X = Z- μz (Centralized data). Now compute Covariance of X and call it S.

4. **[6 Marks]** Find eigenvalues and eigenvectors of S (**Note**: You are allowed to use inbuilt functions) and arrange eigenvectors in non-increasing order of their eigenvalues, create a matrix U with these ordered eigenvectors as its columns.

5. **[4 Marks]** Project X using only the first column of U, let the projection be Y and it can be computed as follows, Y= U [first column]**'** X. Note that your projection is 1 dimensional whereas your data was 2 dimensional.

6. **[8 Marks]** Similar to part (5), find projections for X1 and X2 and call them Y1 and Y2 respectively.

   Now plot Y1 and Y2 using scatter plots and with different colors. Write down your observations from the plot, can you discriminate between these two classes in the plot?

7. **[10 Marks]** Perform reconstruction of data using U and Y, it can be done as follows.
   X_reconstructed =U [First Column] Y
   Z_reconstructed= X_reconstructed + μz (Note: μz was subtracted from Z to get X)
   (Note:
   Compute Mean Squared Error (MSE) between Z and Z_reconstructed as follows.
   MSE = mean (mean ((Z_reconstructed - Z)^2)).
   The MSE should be very low since the reconstruction is lossless. (**Note:** you are not allowed to use any inbuilt functions for MSE) Write all your observations in the report.