IITM-CS5691 : Pattern Recognition and Machine Learning
Assignment 1

Release Date: August 30, 2023
Due Date : September 14, 2023, 23:59

Roll No: CS23E001
Name: Shuvrajeet Das
Collaborators (if any):
References/sources (if any):

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting pdf files (one per question) at Crowdmark by the due date. (Note: **No late submissions** will be allowed, other than one-day late submission with 10% penalty or four-day late submission with 30% penalty! Instructions to join Crowdmark and submit your solution to each question within Crowdmark **TBA** later).
- For the programming question, please submit your code (rollno.ipynb file and rollno.py file in rollno.zip) directly in moodle, but provide your results/answers (including Jupyter notebook **with output**) in the pdf file you upload to Crowdmark.
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material or LLMs (Large Language Models like ChatGPT) for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words (this also means that you cannot copy-paste the solution from LLMs!). Please be advised that *the lesser your reliance on online materials or LLMs* for answering the questions, *the more your understanding* of the concepts will be and *the more prepared you will be for the course exams*.
- Points will be awarded based on how clear, concise and rigorous your solutions are, and how correct your answer is. The weightage of this assignment is 12% towards the overall course grade.

1. (16 points) [LET'S ROLL UP YOUR CODING SLEEVES...] (Note: You should follow instructions in the preamble on how to submit notebook with output/results, as well as the code source files, to get full credit for this programming question.)
You are supposed to build Bayesian classifiers that model each class using multivariate Gaussian density functions for the datasets assigned to you (under assumptions below and employing MLE approach to estimate class prior/conditional densities). This assignment is focused on handling and analyzing data using interpretable classification models, rather than aiming solely for the best classification accuracy.

   Build Bayesian models for the given case numbers (you may refer to the Chapter 2 of the book "Pattern Classification" by David G. Stork, Peter E. Hart, and Richard O. Duda):

   Case 1: Bayes classifier with the same Covariance matrix for all classes.

   Case 2: Bayes classifier with different Covariance matrix across classes.

   Case 3: Naive Bayes classifier with the Covariance matrix $S = \sigma^2\mathbf{I}$ same for all classes.

   Case 4: Naive Bayes classifier with S of the above form, but being different across classes.

   Refer to the provided dataset for each group, which can be found here. Each dataset includes 2D feature vectors and their corresponding class labels. There are two different datasets available:

   1. Linearly separable data.
   2. Non-linearly separable data.

   There are 41 folders in each dataset, but you need to look at only one folder – **the folder number assigned to you** being $\texttt{RollNo}\%41 + 1$.

   Plots/answers Required: For your assignment, you need to provide the following plots/answers (refer to the "Sample Plots" folder: link):

   (a) (4 points) The plot of Gaussian pdf for all classes is estimated using the train data (train.txt). (4 Cases $\times$ 2 Datasets = 8 plots in one page)

   (b) (4 points) The classifiers, specifically their decision boundary/surface as a 2D plot along with training points marked in the plot (again 8 plots in one page).

   (c) (1 point) Report the error rates for the above classifiers (four classifiers on the two datasets as a 4×2 table, with appropriately named rows and columns).

   (d) (1 point) Answer briefly on whether we can use the most general "Case 2" for all datasets? If not, answer when a simpler model like "Case 1" is preferable over "Case 2"?

   (e) (6 points) Ensure that the properly running code files that generate the above plots, etc., are submitted according to the detailed instructions in the preamble.

   (Not)Allowed Libraries: You are not allowed to use any inbuilt functions for building the model or classification using the model. However, you can use inbuilt functions/libraries for plotting and other purposes.

Solution: The solution of question (a)

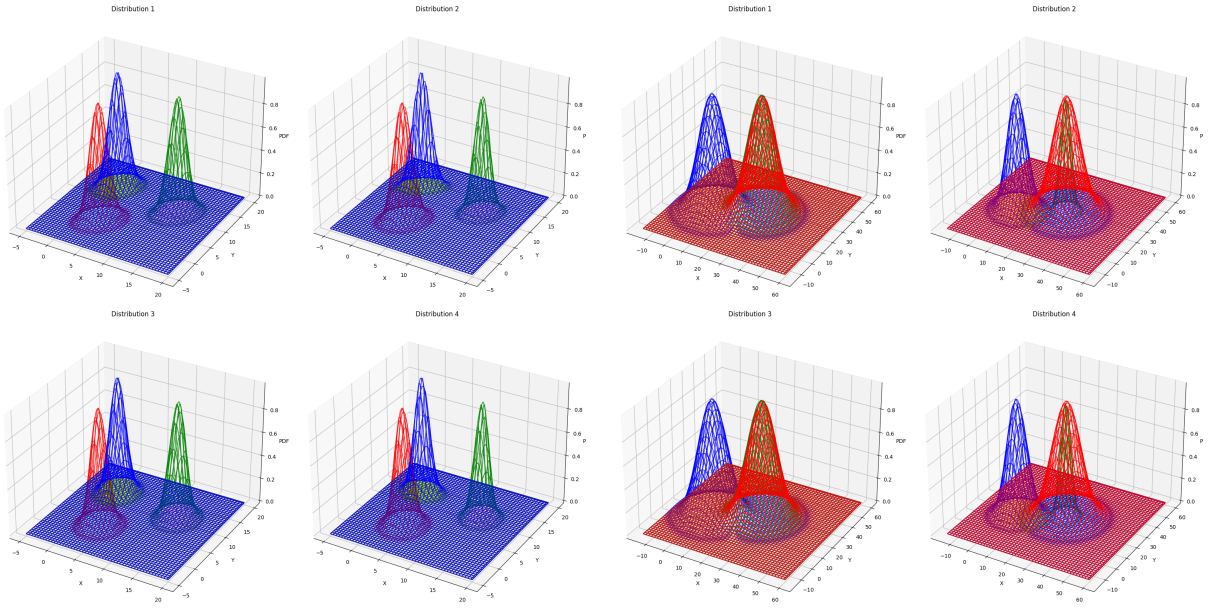

Figure 1: Figures generated Left 2 for linear data and Right 2 for non linear data

Description The first 2x2 plot is for the linearly separable data and the next 2x2 is for Non-Linearly separable data. The 8 graphs are produced after running the 4 classifier algorithm over the dataset(train.txt) for getting the plot of Gaussian pdf for all classes.

The solution of question (b)

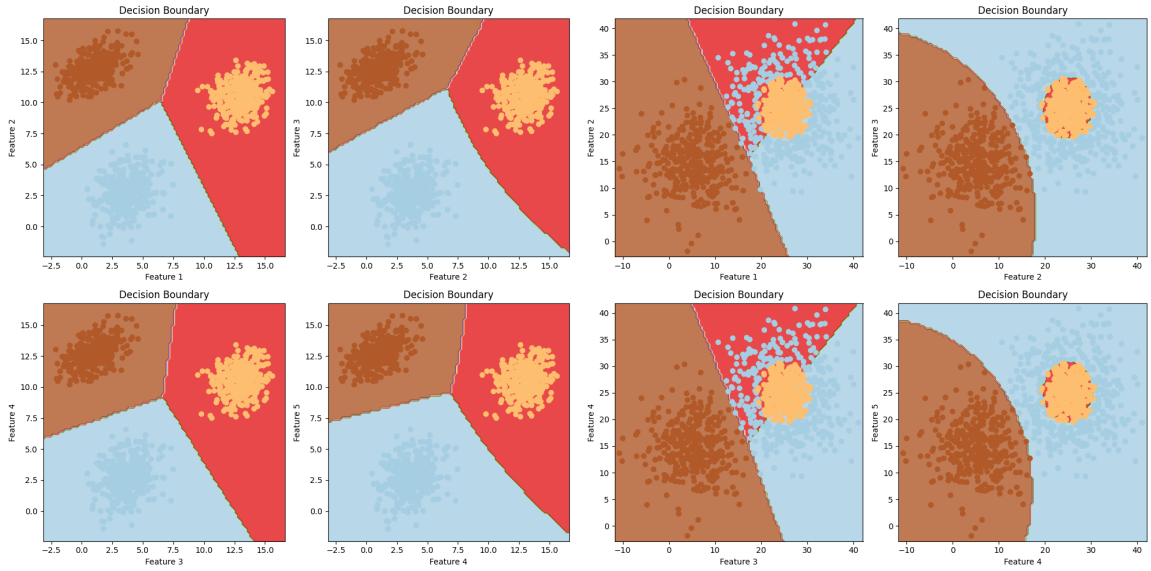

Figure 2: Figures generated Left 2 for linear data and Right 2 for non linear data

Description The first upper 2x2 plot is for the linearly separable data and the next lower 2x2 is for Non-Linearly separable data. The 8 graphs are produced after running the 4 classifier algorithm over the dataset(train.txt) for getting the classifier's decision boundary.

The solution of question (c)

Model	**Correct**	**Wrong**
Classifier 1	1050	0
Classifier 2	1050	0
Classifier 3	1050	0
Classifier 4	1050	0
Classifier 1	747	303
Classifier 2	1033	17
Classifier 3	748	302
Classifier 4	1030	20

The solution of question (d)

Yes we can use the most general case "Case 2" for all datasets.

The solution of question (e)

Reffer codes.