Pattern Recognition and Machine Learning

Lab - 3 Assignment
Bayes Classification

Early Bird Submission Deadline: Tuesday Batch: 30 Jan, 11:59 PM
Thursday Batch: 1 Feb, 11:59 PM
Late Submission Deadline: Tuesday Batch: 31 Jan, 2023, 12:00 Midnight (20% penalty)
Thursday Batch:Feb 2, 2023, 23:59 (20% penalty)
Final deadline: Tuesday Batch: Feb 1, 2023, 23:59 (addl. 20% penalty, total penalty = 40%)
Thursday Batch: Feb 3, 2023, 23:59 (addl. 20% penalty, total penalty = 40%)

Guidelines for submission

1. Perform all tasks in a single colab file.
2. Create a report regarding the steps followed while performing the given tasks. The report should not include excessive unscaled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Link for In-Lab Submission: Link
5. Submit the colab[.ipynb], python[.py] and report[.pdf] files here : Link
6. Plagiarism will not be tolerated

Guidelines for Report:

1. The report should be to the point. Justify the space you use!

2. Explanations for each task should be included in the report. You should know the 'why' behind whatever you do.

3. Do not paste code snippets in the report.

Question 01: [60 marks]

This question is meant to help you comprehend how a naive bayes classifier works. Download the Titanic Dataset and implementation includes the following tasks:-

1. Perform pre-processing and visualization of the dataset. Split the data into train and test sets. Also identify the useful columns and drop the unnecessary ones - **[10 marks]**. **[In-Lab Submission]**

2. Identify the best possible variant of naive bayes classifier for the given dataset. Justify your reason for the same - **[3 marks].**

3. Implement the identified variant of Naive Bayes Classifier using scikit learn,report its performance based on appropriate metrics.(ROC AUC etc) - **[20 marks]**

4.Perform 5 fold cross validation and summarize the results across the cross-validation sets. Compute the probability of the top class for each row in the testing dataset. - **[5 + 5 marks]**

5. Make contour plots with the data points to visualize the class-conditional densities. What can you say about the assumption Naive Bayes model is based on from these plots? Explain in your report. **[7 marks]**

6. Compare your model with the Decision Tree classifier on the same dataset by performing 5-fold cross-validation and summarizing the results. Justify why one of them works better on this numeric dataset.**[10 marks]**

Note: *Implementing the wrong variant of naive bayes classifier will attract a loss of credit in the above question.*

Question 02: [40 Marks]

Only Numpy, Pandas, Seaborn and Matplotlib are allowed.

Dataset - Link

There are 210 rows with 7 input variables and 1 output variable. The variable names are as given:

1. Area.
2. Perimeter.
3. Compactness
4. Length of kernel.
5. Width of kernel.
6. Asymmetry coefficient.
7. Length of kernel groove.
8. Class (1, 2, 3).

a. Use histogram to plot the distribution of samples. **[3 marks]**
b. Determine the prior probability for all the classes. **[3 marks]**
c. Discretize the features into bins from scratch. Use of pandas, scikit learn and scipy is not allowed for this subpart. **[12 marks]**
d. Determine the likelihood/class conditional probabilities for all the classes. **[9 marks]**
e. Plot the count of each unique element for each class. Compare the plot with the plot of distribution. **[3 marks]**
f. Calculate the posterior probabilities and plot them in a single graph. Analyze the plot. **[10 marks]**