Pattern Recognition and Machine Learning
(Winter 2022)
Assignment 4: Bayes Classification

Submission Deadline: Feb 13, 2022, 23:59

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. Submit a zip with the colab file [.ipynb] and report [.pdf] on the classroom
5. Submit the .py file on the floated form for the laboratory
6. Plagiarism will not be tolerated

Guidelines for Report:

1. The visualization of the dataset required in problem 1 should be computed as subplots in the colab file and relevant features should be added in the pdf.

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

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

4. Do not paste code snippets in report.

Question 01: **[60 marks]**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. 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 - [7 marks].

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 from scratch. [you may not use any 3rd-party library's classifier function, such as scikit-learn. However, Built-in functions, such as train/test split, can be used for supplementary tasks.] - [20 marks]

4. Perform 5-fold cross-validation using the entire training feature set - [5 marks]

5. Visualize 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]

6. Compare your scratch implementation with scikit-learn in terms of the performance - [5 marks]

7. Implement any other model of your choice [not necessarily from scratch] and perform 5-fold cross-validation and summarize the results. Compare it with the Naive Bayes Classifier you have implemented and justify your results - [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. Analyse the plot. [10 marks]
