ASSIGNMENT - 2
Machine Learning - I
Topic: Bayesian Classification

Total Marks : 100 Deadline: March 5, 2022

Instructions :

1. Do not copy from other students. Any case of plagiarism will result in zero marks.
2. You can refer to codes online (e.g. Github, Kaggle) but do not copy-paste. The resource must be cited in the report if referred.
3. Strictly follow the submission guidelines.
4. Allowed languages: python
5. You can use any inbuilt library if not mentioned to code from scratch.

Submission Guidelines :

1. Submit .py python files or .ipynb for all the questions.
2. Strictly submit a single report (.pdf) for all the questions. No .doc, .docx file will be accepted
3. If you are using colab, then attach your colab link in the report (preferred)
4. Submit a single zip file containing all python files and reports.
5. The name of the zip file should be M21AIEABC.zip, python files should have the name M21AIEABC_qu1.py or M21AIEABC_qu2.py, etc. The report should have the name M21AIEABC.pdf (last three digits of your roll no in place of ABC). If the naming convention is not followed, we will award zero marks.

Question 1:

Download the dataset, where the first four columns are features, and the last column corresponds to categories (3 labels). Perform the following tasks.

1. Split the dataset into train and test sets (80:20)
2. Construct the Naive Bayes classifier from scratch and train it on the train set. Assume Gaussian distribution to compute probabilities.
3. Evaluate the performance using the following metric on the test set
    a. Confusion matrix
    b. Overall and class-wise accuracy
    c. ROC curve, AUC
4. Use any library (e.g. scikit-learn) and repeat 1 to 3
5. Compare and comment on the performance of the results of the classifier in 2 and 4
6. Calculate the Bayes risk.
   Consider,

$\lambda =$

$$\begin{bmatrix} 2 & 1 & 6 \\ 4 & 2 & 4 \\ 6 & 3 & 1 \end{bmatrix}$$

Where λ is a loss function and rows and columns corresponds to classes (c_i) and actions (a_j) respectively, e.g. $\lambda(a_3 / c_2) = 4$.

Question 2: Spam or Ham?

You are requested to use the dataset from here. (Recall the tutorial video shared with you earlier to download the Kaggle datasets). Perform a quick overview of the dataset by using data wrangling techniques. Now perform the following operations to clean the dataset and perform classification tasks:

1. Remove links from the dataset.
2. Remove special characters or symbols from the dataset.
3. Remove numbers or alphanumerical characters from the dataset.
4. Check if your dataset contains sufficient features that are required for the operations.
5. Produce the following visualizations:
    a. Spam vs Ham in the dataset
    b. Word clouds for spam and ham
    c. Find dependencies between different features using correlation matrix, pair plots, and distributions of different features.
6. Perform the following tasks as part of feature extraction:
    a. Tokenization
    b. Lemmatization
    c. Vectorization
    d. TF-IDF weighting
7. Find top N ham and spam words in messages and visualize them either by plot or word cloud.
8. Train-Test Split and perform the classification task using the Naive Bayes classification model.
    a. Find probabilities of the top 30 words.
    b. What problem does smoothing handle? Use the smoothing technique with naive Bayes to predict the sentences in the test split.
    c. Draw confusion matrix, ROC, AUC. (You can use the sklearn library)

Question 3:

Download the dataset containing two columns X and Y. Fix an appropriate threshold and consider Y > threshold as 1 otherwise 0. Perform the following tasks.

1. Create the labels from the given data.
2. Plot the distribution of samples using the histogram.

3. Determine the prior probability for both classes.
4. Determine the class conditional probabilities (likelihood) for the classes.
5. Plot the count of each unique element for each class.
6. Calculate the posterior probability of both classes and plot them.
7. Consider the dataset of question 1 and select any one feature and class label. Repeat 1 to 6.

Question 4:

1. Use Gaussian NB from sklearn on UCI **Dataset**. Plot the decision boundary. Draw analysis on the results by calculating the Roc curve, and classification report.
2. Now consider all features as correlated and construct a classifier using Mahalanobis distance. Draw analysis on the results by calculating the Roc curve, and classification report. Compare the results with part 1. You can refer to this article or book for Mahalanobis distance.

Note: Based on the experiment done, please prepare a report too.

If you have any doubts regarding the assignment, post on Google classroom.