Pattern Recognition and Machine Learning

Lab Assignment 7

***Early Bird Submission Deadline**: Tuesday Batch: 20 Mar, 11:59 PM*
Thursday Batch: 22 Mar, 11:59 PM

***Late Submission Deadline**: Tuesday Batch: 21 Mar, 2023, 12:00 Midnight (20% penalty)*
Thursday Batch:Mar 23, 2023, 23:59 (20% penalty)

***Final deadline**: Tuesday Batch: Mar 22, 2023, 23:59 (addl. 20% penalty, total penalty = 40%)*
Thursday Batch: Mar 24, 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. 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 'why' behind whatever you do.
3. Do not paste code snippets in the report.

Question-1: [Principal Component Analysis] :- [60 marks]

In 1990 David, Sterling and Wray Buntine donated an Annealing Dataset in order to study Steel Annealing(a heat treatment that alters the physical and sometimes chemical properties of a material). Classes (1,2,3,4,5,U) hereby act as Label and other parameters as Input Features.

1. From the given link, download "anneal.data", "anneal.names" and "anneal.test", convert them into a readable format (Ex: txt, csv, etc....) and do meaningful Exploratory Data Analysis. **[5 Marks]**

2. Preprocess the data (If any discrepancies/errors, handle them as well) and split the data into [65:35]. **[4 + 1 Marks].** There are two subparts here. You need to write in the report about the difference in the observations and explain it if any.:

- Perform feature standardization and use the standardized data for the rest of the questions
- Do not perform feature standardization and use the original data for the rest of the questions.

3. Train 2-3 Classification Models (studied and implemented so far out of which one has to be *SVM* classifier) with the proper reasoning of choosing them and showing 5-Fold Cross-Validation Plots as well for comparison. **[5 + 5 Marks]**

4. Implement Principal Component Analysis from scratch, with sub-tasks as following:- **[5 + 10 Marks]**

a. Centralize the Data via feature-wise means and standard deviations. Write the code for deriving the covariance matrix from scratch.

b. Compute Eigenvectors, Eigenvalues and Principal Components and comment on what is the role of eigenvectors in the report. You may use sklearn to find the eigenvectors but others are to be found from scratch.

5. Use the above-made PCA to reduce the data upto a chosen dimension/principal-components. Plot a bar graph to show the change in variance as you increase the no. of components. Along with this, plot a scatter plot to show the direction of the eigenvectors along with the data points(you may choose any 2 features among the reduced dataset). **[2+3 Marks]**

6. Train 2-3 chosen classification models alongside 5-Fold Cross-Validation Plots. **[5 Marks]**

6. Show the Test results of Classification Models on both types of datasets (Before and After PCA), via 2-3 Evaluation Metrics of choice (Ex:- Accuracy, Sensitivity, F1-Score, etc.) with the proper reasonings. **[5 + 5 Marks]**

7. Were any changes observed before and after implementing PCA, with respect to the distribution of the dataset? Also, make any suitable graph through which the optimal number of principal components can be decided for optimal results. **[2 + 3 Marks]**

Bonus : Assuming the Naive Bayes assumption, calculate the eigenvector, eigenvalues and principal components. Do part 6 with these new feature vectors and comment on advantages/disadvantages you observed with this assumption.**[5 Marks]**

Question-2: [Linear Discriminant Analysis] :- [40 marks]

LDA is both a classification algorithm and a dimensionality reduction algorithm. In this question, you have three tasks,

- Use LDA as a classifier for a classification task
- Use LDA as a dimensionality reduction technique and use a classifier of your own choice for the classification task.
- Use LDA as a dimensionality reduction technique and compare it with PCA.

Perform the aforementioned tasks on the Wine Classification Dataset as instructed below:(You may use any 2 classification techniques of your choice and perform the classification).

1. Implement Linear Discriminant Analysis from scratch with the following subtasks:-

a. A function for computing within class and between class scatter matrices.

b. A function that will automatically select the number of linear discriminants based upon the percentage of variance that needs to be conserved **[5+5 Marks]**

2. Vary the variance and identify features that have a high impact on the classification tasks using LDA and visualize the feature space for the same using those linear discriminants.

3. Perform PCA on the dataset and compare the results with LDA by using any 2 classification techniques. **[3 Marks]**

4. Create a table to properly note down the accuracies in case of each classifier and the corresponding reduction technique. Show using scatter plot of any two features among the features you chose which contribute to the maximum variance the decision boundary in case of LDA.

5. Using LDA as a classifier, perform 5-fold cross-validation and plot ROC and compute AUC for each fold from scratch **[10 Marks]**