Pattern Recognition and Machine Learning (Winter 2022)
Assignment 6: Dimensionality Reduction

Submission Deadline: **Mar 10, 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 and unlabelled preprocessing plots.
3. Try to modularize the code for readability wherever possible
4. Submit 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 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-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]**

3. Train 2-3 Classification Models (studied and implemented so far) with the proper reasoning of choosing them and show 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. Make a function Singular_Value_Decomp from scratch in order to compute Eigenvectors, Eigenvalues and Principal Components.

5. Use the above-made PCA to reduce the data upto a chosen dimension/principal-components and train 2-3 chosen classification models alongside 5-Fold Cross-Validation Plots. **[5 + 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]**

<u>Question-2:</u> ***[Linear Discriminant Analysis]*** :- [40 marks]

Perform feature extraction using LDA on the aforementioned Dataset. 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. Perform PCA and compare the results with LDA **[3 Marks]**
3. Identify features having a high impact on classification tasks using both PCA and LDA and visualize the sample space using the first two principal components and first two linear discriminants and comment your observations **[2+5 Marks]**
4. Using any 2 classification techniques make a 2 * 2 table with column headers as classification techniques used and row headers as feature extraction methods used. The values inside the table should be the accuracy achieved in each case. Compare the results of the table. **[10 Marks]**
5. Using LDA as a classifier, perform 5-fold cross-validation and plot ROC and compute AUC for each fold from scratch **[10 Marks]**