Statistical Machine Learning **Assignment 3 (Max Marks: 100)**

Instructions: You are allowed to discuss but the final answer should be your own. Any instance of cheating will be considered as academic dishonesty and penalty will be applied.

1. Restrict to using only Python for coding assignments.
2. You are free to use math libraries like *Numpy, Pandas*; and use *Matplotlib, Seaborn* library for plotting.
3. Add all the analysis on the question in the written format, anything not in the report is not marked.
4. Use of inbuilt function for any evaluation metric is not allowed unless stated otherwise. Each of the metrics needs to be implemented from scratch.
5. Implement code that is Modular in nature and generalized to be executed for any input.
6. Code should be submitted in Python file format only(.py)

DIMENSIONALITY REDUCTION

1. **(60 points)** Download the databases from the links supplied. Database 1 contains 713 face images of 11 subjects (people). Each folder contains images of one particular subject, and hence considered as a class. Database 2 contains images 60000 colour images in 10 classes, with 6000 images per class.
   Database 1 (Face Dataset)
   Database 2 (CIFAR 10)

   Protocol:
   - Convert images to grayscale, in the dataset 1 and dataset 2.
   - Any classifier from sklearn library is allowed for Question 1.
   - For dataset 1, randomly split it in 70% training set and 30% testing set.
   - For dataset 2, use training set and testing set as defined in the dataset only.
   - Perform 5 fold cross validation on each dataset. In each fold, put 80% of the training data in training, and the rest for validation.

   Perform the following tasks:
   a. Implement PCA and LDA from scratch.
   b. Classify the dataset as it is and use the ***same classification algorithm*** for the parts below.
   c. Use Linear discriminant analysis (LDA) to find out the best projection directions.
   d. Project your data on the new projection matrix given by LDA.
   e. Classify the projected data using 5 fold cross-validation, report the mean and standard deviation of classification accuracy. Use the best model, to classify the test set and plot the ROC curve and confusion matrix. How does the results on the 2 databases differ and why?
   f. Perform PCA on the same data by preserving 95% eigen energy and report the mean accuracy and standard deviation over the 5 folds. Use the best model, to classify the test set and plot the ROC curve and confusion matrix. How does the results on the 2 databases differ and why?
   g. Compare and analyze the results obtained by PCA and LDA.
   h. Separately for PCA do the following:
      i. Visualise and analyse the eigenvectors obtained using PCA (only for eigenvectors obtained in part f).

ii. Perform classification by preserving the below stated eigen energy and compare and analyse the classification performance obtained for each part and best model for part f:
    1. 70% eigen energy
    2. 90% eigen energy
    3. 99% eigen energy

i. If you perform LDA on the PCA projected data, find out the classification performance for both the databases. Analyze the result. If we perform the process the other way round, what performance do you get and why?

ENSEMBLE LEARNING

Dataset: [*ftp://ftp.ics.uci.edu/pub/machine-learning-databases/letter-recognition*](ftp://ftp.ics.uci.edu/pub/machine-learning-databases/letter-recognition)

2. **(40 points)** Perform the following using Decision Tree classifier with upto 2 levels of tree and 5 nodes as a weak classifier. *(Decision tree classifier can be used from sklearn).*

   Boosting:

   a. Implement a function for AdaBoost Algorithm which should take N, Train, Test as it's parameter. Here N determines the number of rounds of Boosting, Train for Training set and Test for Testing Set.

   b. The function should return final predictions for Training and Test Set from the combined classifier with Accuracy and Error Rate on these set.

   c. Use 5 fold Cross validation for Training set and report Mean and Standard deviation of accuracy and error rate. Also report the classification accuracy on test set.

   Bagging:

   d. Repeat the above parts by implementing Bagging procedure using Decision Trees as Weak Learner. The input parameters and returned values of this function should be just like those for your boosting function.

   e. With the trained classifiers in part (d), perform score normalisation by *min-max*, *z-score* and *tanh* followed by fusion of the normalised scores by sum of scores method. Perform classification on the basis of the final score value. Also compare and reason out the performance of the approach under the various score normalisation schemes.

   For Part b of the above question, randomly split your data set into 70% training and 30% testing. Report your results and write your analysis.

Submission format: Please submit a report for all your analysis and observations only and only in the PDF format. Other formats will not be evaluated. All the graphs should have labels (on the axis), legends, title. You should also try to combine the graphs and plots for comparison and better representation. All the python code files need to be submitted with the following naming format: "**<Roll_Number>_file.py**" in a zipped folder with source folder named as "**A3_<Roll_Number>.zip**"