CARNEGIE MELLON UNIVERSITY
DATA, INFERENCE & APPLIED MACHINE LEARNING (COURSE 18-875)
ASSIGNMENT 7

INSTRUCTIONS

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb]**. In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
- Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_DIAML_AssignmentNo. For example, mcsharry_DIAML_Assignment1, mcsharry_DIAML_Assignment2 and mcsharry_DIAML_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

The student is responsible for checking that their submission is complete. Students will lose 10% as for late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is **on Monday 5th, December 2022 16:59 Eastern Time (ET) / Monday 5th, December 2022 23:59 Rwandan Time (CAT).**

1. PCA (25 points)

1.1 Give a qualitative description of Principal Component Analysis (PCA) and its applications in machine learning. Why might it be useful to consider PCA to transform a set of explanatory variables?
1.2 Write down the mathematical equations for PCA explaining how one transforms the raw input data matrix X into a new set of variables. Give an interpretation of each matrix.
1.3 Use at least one year of daily returns to calculate the correlation matrix for the 30 stocks that are constituents of the Dow Jones Index. MATLAB's "*BlueChipStockMoments*" can be used to calculate the correlation matrix. Use this correlation matrix for PCA and construct bar graphs to show the weight of each stock for the first and second principal components. Is the first or second principal component similar to the market (equal weight on each stock)? Discuss why?
1.4 Calculate the amount of variance explained by each principal component and make a 'Scree' plot. How many principal components are required to explain 95% of the variance?
1.5 Investigate the scatter plot of the first two principal components and calculate the average of all 30 stocks. Based on Euclidean distances away from this average, identify the three most distant stocks. Can you explain why these stocks are unusual?

2. Dendrogram (25 points)

2.1. Describe the components of a dendrogram, how it is constructed and how it is interpreted.
2.2. Given a collection of pairwise dissimilarity values, describe the steps involved in constructing a dendrogram.
2.3. Use the correlation matrix from question (1.3) above to provide pairwise distances between the 30 stocks. Give the formula for this rescaled distance and provide an interpretation of small and large distances.
2.4. Construct a horizontal dendrogram using the average linkage approach, carefully labelling the graphic with the names of the 30 stocks.
2.5. Use the dendrogram to provide a few clusters of stocks and list the stocks that are members of each cluster. Can you provide a description of each cluster and relate it to industrial sectors such as Financials, Energy etc?

3. Ensembles for classification (25 points)

3.1 Name three sources of uncertainty and explain how they impact on the modelling process when using machine learning approaches.
3.2 What is the concept behind model averaging and give some examples of how this technique can be implemented in practice when generating predictions?
3.3 What kind of ensemble methods can be used to reduce the effects of uncertainty and improve on individual models? How do they achieve this goal?
3.4 Construct a random forest (RF) model and apply this to the Titanic dataset. Explain how you selected the optimal number of trees and support your choice using a graph.
3.5 Undertake a ROC analysis and show how the RF performs relative to the previous models (logistic regression, classification tree and KNN). Provide evidence to show as clearly as possible which model is best for classifying survival on the Titanic.

4. Ensembles for regression (25 points)

The wine quality database provides information about the quality of wine. There are two datasets, one for red wine and one for white wine, which contain quality ratings, from one to ten, along with their physical and chemical properties. The challenge is to use these features to predict the rating for a wine and to assess performance. It is advisable to study white and red wine separately:
https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality

4.1 Describe the concept of a random forest (RF) regression model.
4.2 Construct a random forest (RF) model for the red wine dataset and show how the optimal number of leafs was estimated.
4.3 Explain and show how the optimal number of trees was computed.
4.4 Provide a bar graph showing the importance of each feature and compare this with the results from Assignment 6 (using correlation and LASSO).
4.5 What is the performance of the RF model and compare it with the linear regression and KNN models constructed during Assignment 6. Present sufficient information to support your conclusion about the best model for the red wine dataset.

Extra credit: You are encouraged to enter the Kaggle challenge referencing the Titanic data set. At the end of this course, extra-credit will be given to students based on their final score on the challenge, coinciding with the deadline for this final assignment. Go to this link https://www.kaggle.com/c/titanic-gettingStarted and follow the instructions to register and enter the challenge. After this assignment, you should compare all models and decide which is most appropriate for classifying survival on the Titanic. Please submit your score from Kaggle and provide evidence that you have achieved this score (code and printout from Kaggle).