Recitation

Friday 25th November 2022

Question 1

1.1 Give a qualitative description of Principal Component Analysis (PCA) and at least its two applications in machine learning. Describe why it might be useful to consider PCA to transform a set of explanatory variables. Use lecture slides.

1.2 Write down the mathematical equations for PCA, explaining how they transform the raw input data matrix X into a new set of variables. Give an interpretation of each matrix.

1.3 Load the Dow Jones Index (***BlueChipStockMoments*** in MATLAB and ***BeautifulSoup*** in Python) dataset directly into your programming environment. **Hint:** use ***load*** in MATLAB and ***import*** in Python. The data is available on Yahoo Finance too.

Question 1 – cont'd

- Calculate the correlation matrix for the 30 stocks that are constituents of the Dow Jones Index. You can get the correlation matrix from the covariance matrix.

- Use the correlation matrix for PCA and construct bar graphs to show the weight of each stock for the first and second principal components. You can produce 2 bar graphs: 1 per principal component.

- Say if there is a similarity between the first or second principal component and the market (equal weight on each stock). **Hint:** Compare values from PCA to the actual values of the stocks. Provide comments and explanations on any similarity observed.

Question 1 – cont'd

1.4 Make a scree plot of the amount of variance explained by each principal component. Determine how many principal components are required to explain 95% of the variance. **Hint:** the amount of variance explained can be directly obtained from the output of the principal component analysis.

1.5 Produce a scatter plot of the first two principal components and then calculate the mean of all 30 stocks for each principal component. Next, calculate Euclidean distances from the mean then identify the three most distant stocks for each principal component.

Question 2

2.1 Describe the components of a dendrogram - how it can be constructed - how it can be interpreted. Use lecture slides and do more research.

2.2 Explain how a dendrogram can be built from a collection of dissimilar pairwise values. Do this by listing all the steps involved in constructing a dendrogram.

2.3 Compute pairwise distance using correlation matrix and explain what the distances mean (see lecture slides for formula).

Question 2 – cont'd

2.4 Make a dendrogram using the linkage approach (check linkage method). Use the correlation matrix from Question 1. You may use ***dendrogram()*** in MATLAB and ***dendrogram()*** from Scipy in Python. Label the dendrogram with the names of all the 30 stocks

2.5 Create clusters (3 probably) using as input the output from the linkage method used in creating the dendrogram. Derive the clusters from the first and second principal component coefficients and label them accordingly. Give them separate colors too. Provide a qualitative description of each cluster and relate them to industrial sectors such as Financial, Energy, Pharmaceutical, etc.

Question 3

3.1 There are many forms of uncertainties. Name three sources of uncertainty and describe their impact on the modelling process when using machine learning approaches.

3.2 Describe the underlying concept of model averaging and give some examples of how this technique can be implemented in practice when generating predictions. Use lecture slides and find additional online resources.

3.3 Describe some kinds of ensemble methods and how they work to reduce the effects of uncertainty and improve on individual models. Use lecture slides and online resources.

Question 3 – cont'd

3.4 Process dataset and construct a random forest model. In MATLAB you may use ***TreeBagger()***. In Python you may use ***RandomForestClassifier()*** from Scikit-Learn. Provide a graph to justify your choice of the optimal number of trees.

3.5 Perform ROC analysis for Logistic Regression, Decision Tree, Random Forest and KNN, in MATLAB you may use ***perfcurve()***. In Python, you may use ***roc_curve()*** from Scikit-Learn. The AUC and/or other metrics could help conclude which is the best model.

Question 4

4.1 Describe the concept of Random Forest (RF) regression model

4.2 Construct different random forest trees with different number of leaves and plot their errors to determine optimal number of leaves. In MATLAB you can use ***Treebagger*** and ***oobError***. Construct the RF model with the optimal number of leaves found.

4.3 Give the optimal number of trees and explain how it was computed. (cf. Question 3.4)

4.4 Provide a bar graph of the features showing the importance of each. You can use ***OOBPermutedVarDeltaError*** in MATLAB. Compare them with those of the last assignment (Assignment 6).

4.5 Calculate the performance of the RF model, you can use the mean squared error then compare it with linear regression and KNN regression (cf. Assignment 6). Choose the model you would use for the red wine dataset and clearly explain why.

Kaggle challenge (Optional)

Link : https://canvas.cmu.edu/courses/31496/files/folder/Assignments/Kaggle%20Competition?preview=8869023

The challenge is optional and extra- credit will be given to students based on their final score.

Submission deadline : 9th December 2022

Submission Files (Python)

- **Single** Python code file(.ipynb) - **andrewID_DIAML_AssignmentNo.ipynb**
- Assignment report(.pdf) - **andrewID_DIAML_AssignmentNo.pdf**
    - 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:

- Put all data files and the source code in a **single folder** named with your **andrewID**
- Zip this folder and submit the zipped **(.zip)** with your report (**.pdf**)to CANVAS

Submission Files (MATLAB)

- **Single** MATLAB code file (.m) - **andrewID_DIAML_AssignmentNo.m**
- Assignment report(.pdf) - **andrewID_DIAML_AssignmentNo.pdf**
- Data files (as given)

Submission process:

- Put all data files and the source code in a **single folder** named with your **andrewID**
- Zip this folder and submit the zipped **(.zip)** with your report (**.pdf**)to CANVAS