Clustering Part

In this assignment, we consider a set of observations on a number of silhouettes related to different type of vehicles, using a set of features extracted from the silhouette. Each vehicle may be viewed from one of many different angles. The features were extracted from the silhouettes by the HIPS (Hierarchical Image Processing System) extension BINATTS, which extracts a combination of scale independent features utilising both classical moments based measures such as scaled variance, skewness and kurtosis about the major/minor axes and heuristic measures such as hollows, circularity, rectangularity and compactness. Four model vehicles were used for the experiment: a double decker bus, Chevrolet van, Saab and an Opel Manta. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

One dataset (vehicles.xls) is available and has 846 observations/samples. There are 19 variables/features, all numerical and one nominal defining the class of the objects.

Description of attributes:

1. Comp: Compactness
2. Circ: Circularity
3. D.Circ: Distance Circularity
4. Rad.Ra: Radius ratio
5. Pr.Axis.Ra: pr.axis aspect ratio
6. Max.L.Ra: max.length aspect ratio
7. Scat.Ra: scatter ratio
8. Elong: elongatedness
9. Pr.Axis.Rect: pr.axis rectangularity
10. Max.L.Rect: max.length rectangularity
11. Sc.Var.Maxis: scaled variance along major axis
12. Sc.Var.maxis: scaled variance along minor axis
13. Ra.Gyr: scaled radius of gyration
14. Skew.Maxis: skewness about major axis
15. Skew.maxis: skewness about minor axis
16. Kurt.maxis: kurtosis about minor axis
17. Kurt.Maxis: kurtosis about major axis
18. Holl.Ra: hollows ratio
19. Class: type of cars

In this clustering part you need to use the first 18 attributes to your calculations.

1st Objective (partitioning clustering)

You need to conduct the k-means clustering analysis of the vehicle dataset problem. Find the ideal number of clusters (please justify your answer). Choose the best two possible numbers of clusters and perform the k-means algorithm for both candidates. Validate which clustering test is more accurate. For the winning test, get the mean of the each attribute of each group. Before conducting the k-means, please investigate if you need to add in your code any pre-processing task (justify your answer). Write a code in R Studio to address all the above issues (codes/results/discussion are needed to be included in your report). In your report, you need to check the consistency of your produced cluster outcome against the information obtained from 19th column and provide the related results/discussion.

1st Objective (partitioning clustering)

- Find the ideal number of clusters – justify it by showing all necessary steps/methods,
- K-means with the best two clusters,
- Find the mean of each attribute for the winner cluster,
- Check consistency of your results against 19th column,
- Check for any pre-processing tasks (scaling, outliers)

(Marks 50)

Forecasting Part

Time series analysis can be used in a multitude of business applications for forecasting a quantity into the future and explaining its historical patterns. Exchange rate is the currency rate of one country expressed in terms of the currency of another country. In the modern world, exchange rates of the most successful countries are tending to be floating. This system is set by the foreign exchange market over supply and demand for that particular currency in relation to the other currencies. Exchange rate prediction is one of the challenging applications of modern time series forecasting and very important for the success of many businesses and financial institutions. The rates are inherently noisy, non-stationary and deterministically chaotic. One general assumption is made in such cases is that the historical data incorporate all those behaviours. As a result, the historical data is the major input to the prediction process. Forecasting of exchange rate poses many challenges. Exchange rates are influenced by many economic factors. As like economic time series exchange rate has trend cycle and irregularity. Classical time series analysis does not perform well on finance-related time series. Hence, the idea of applying Neural Networks (NN) to forecast exchange rate has been considered as an alternative solution. NN tries to emulate human learning capabilities, creating models that represent the neurons in the human brain.

In this forecasting part you need to use an MLP-NN model to predict the next step-ahead exchange rate of GBP/EUR. Daily data (exchangeGBP.xls) have been collected from January 2010 until December 2011 (500 data). The first 400 of them have to be used as training data, while the remaining ones as testing set. Use only the 2nd column from the .xls file, which corresponds to the exchange rates.

2nd Objective (MLP)

You need to construct an MLP neural network for this problem. You need to consider the appropriate input vector (time-series), as well as the internal network structure (hidden layers, nodes, learning rate). You may consider any de-trending scheme if you feel is necessary. Write a code in R Studio to address all these requirements. You need to show the performance of your network both graphically as well as in terms of usual statistical indices (MSE, RMSE and MAPE). Suggestion: Experiment with various network structures and show a comparison table of their performances. This will be a good justification for your final network choice. Show all your working steps (code & results, including comparison results from models with different input vectors). As everyone will have different forecasting result, emphasis in the marking scheme will be given to the adopted methodology and the explanation/justification of various decisions you have taken in order to provide an acceptable, in terms of performance, solution. The input selection problem is very important. Experiment with various options (i.e. how many past values you need to consider as potential network inputs). Full details of your results are needed in your report.

(Marks 50)

- **2nd Objective (MLP)**
  - Discuss the input selection problem for time series prediction and propose various input configurations (Suggestion: consult literature for system identification configurations)
  - Perform any pre-processing steps (such as normalisation) before training
  - Implement a number of MLPs, using various structures (layers/nodes) / input parameters / network parameters and show in a table their performances comparison (based on testing data) through the provided stat. indices. (10 marks for structures with different input parameters, 10 marks for different internal NN structures and 5 for the comparison table)
  - Provide your best results both graphically (prediction output vs desired output) and via performance indices (5 marks for for the graphical display and 5 marks for showing the requested statistical indices)