University of Westminster
School of Computer Science & Engineering

7BUIS008W	Data Mining & Machine Learning – Coursework 1 (2020/21)
Module leader	Dr. P.I. Chountas. This CW was prepared/written by Dr. V.S. Kontogiannis
Unit	Coursework 1
Weighting:	50%
Qualifying mark	35%
Description	Show evidence of understanding of the clustering and modelling concepts, through the implementation of requested algorithms using real datasets. Implementation is performed in R environment, while students need to perform some critical evaluation of their results.
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): • LO2 fully implement data mining/machine learning projects, focused on problem analysis, data pre-processing, data post-processing by choosing and implementing appropriate algorithms; • LO4 fully implement encode and test data mining and machine learning algorithms using the programming language (such as Python) and standard packages and toolkits (such as R). • LO6 perform critical evaluation of performance metrics for data mining and machine learning algorithms for a given domain/application
Handed Out:	18/02/2021
Due Date	30/03/2021, Submission by 13:00
Expected deliverables	Submit on Blackboard only one pdf file containing the required details. All implemented codes should be included in your documentation together with the results/analysis/discussion.
Method of Submission:	Electronic submission on BB via a provided link close to the submission time.
BCS CRITERIA MEETING IN THIS ASSIGNMENT	• **7.1.6 Use appropriate processes** • **7.1.7 Investigate and define a problem** • **7.1.8 Apply principles of supporting disciplines** • **8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas** • **8.1.2 Comprehensive understanding of essential principles and practices** • **8.2.2 Tackling a significant technical problem** • **10.1.2 Comprehensive understanding of the scientific techniques**

Assessment regulations

Refer to section 4 of the "How you study" guide for undergraduate students for a clarification of how you are assessed, penalties and late submissions, what constitutes plagiarism etc.

Penalty for Late Submission

If you submit your coursework late but within 24 hours or one working day of the specified deadline, 10 marks will be deducted from the final mark, as a penalty for late submission, except for work which obtains a mark in the range 50 – 59%, in which case the mark will be capped at the pass mark (50%). If you submit your coursework more than 24 hours or more than one working day after the specified deadline you will be given a mark of zero for the work in question unless a claim of Mitigating Circumstances has been submitted and accepted as valid.

It is recognised that on occasion, illness or a personal crisis can mean that you fail to submit a piece of work on time. In such cases you must inform the Campus Office in writing on a mitigating circumstances form, giving the reason for your late or non-submission. You must provide relevant documentary evidence with the form. This information will be reported to the relevant Assessment Board that will decide whether the mark of zero shall stand. For more detailed information regarding University Assessment Regulations, please refer to the following website:**http://www.westminster.ac.uk/study/current-students/resources/academic-regulations.**

Instructions for this coursework

During marking period, all coursework assessments will be compared in order to detect possible cases of plagiarism/collusion. For each question, show all the steps of your work (codes/results). In addition, students need to be informed, that although clarifications for CW questions can be provided during tutorials, coursework work has to be performed outside tutorial sessions.

Coursework Description

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 (i.e. centres) of each group. Before conducting the k-means, please investigate if you need to add in your code any pre-processing task (scaling and/or outliers detection and justify your answer). Write a code in R Studio to address all the above issues (codes/results need 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 (evidence of a "confusion" matrix and extracted information from it). At the end of your report, provide also as an Appendix, the full code developed by you. The usage of kmeans R function is compulsory.

(Marks 40)

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 behavior. 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 addition, research has been also directed to Support Vector Machine (SVM) which has emerged as a new and powerful technique for learning from data and in particular for solving classification and regression problems with better performance. The main advantage of SVM is its ability to minimize structural risk as opposed to empirical risk minimization as employed by the NN system.

In this forecasting part you need to use an MLP-NN and a SVM-based regression (SVR) 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 (such as 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 the following statistical indices (RMSE, MAE and MAPE). Suggestion: Experiment with various network structures as well as various input vectors and show a comparison table of their performances (using these specific statistical indices). 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 and internal structure). 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/codes/discussion are needed in your report. At the end of your report, provide also as an Appendix, the full code developed by you. The usage of neuralnet R function for MLP modelling is compulsory.

(Marks 35)

3rd Objective (SVR)

You need to construct a SVR model to address this forecasting problem. You need to consider the appropriate input vector. Write a code in R Studio to implement this SVR scheme. You need to show the performance of your model both graphically as well as in terms of the following statistical indices (MSE, RMSE and MAPE). Suggestion: The input selection problem is very important. Experiment with various SVR parameters. 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. Full details of your results/codes/discussion are needed in your report. At the end of your report, provide also as an Appendix, the full code developed by you.

(Marks 25)

Coursework Marking scheme

The Coursework will be marked based on the following marking criteria:

1st Objective (partitioning clustering)

- Find the ideal number of clusters – justify it by showing all necessary steps/methods (via manual & automated tools) 8
- K-means with the best two clusters, 8
- Find the mean of each attribute for the winner cluster, 6
- Check consistency of your results against 19th column and provide relevant discussion, 8
- Check for any pre-processing tasks (scaling, outliers) 10

2nd Objective (MLP)

- Discuss the input selection problem for time series prediction and propose various input configurations (Suggestion: consult literature for system identification configurations) 8
- Perform any pre-processing steps (such as normalisation) before training 5
- 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. (6 marks for structures with different input parameters, 6 marks for different internal NN structures and 4 for the comparison table) 16
- Provide your best results both graphically (your prediction output vs. desired output) and via performance indices (3 marks for the graphical display and 3 marks for showing the requested statistical indices) 6

3rd Objective (SVR)

- Discuss the input selection problem and propose various input configurations 3
- Design an SVR and use various structures/parameters (incl. linear/nonlinear kernels)/input parameters and show in a table their performances comparison (based on testing data) through the provided stat. indices. (6 marks for structures with different input parameters, 6 marks for different internal SVM structures/parameters and 4 for the comparison table) 16
- Provide your best results both graphically (your prediction output vs. desired output) and via performance indices (3 marks for the graphical display and 3 marks for showing the requested statistical indices) 6