University of Westminster

School of Computer Science & Engineering

5DATA002W	Machine Learning & Data Mining – Coursework (2020/21)
Module leader	Dr. V.S. Kontogiannis
Unit	Coursework
Weighting:	40%
Qualifying mark	30%
Description	Show evidence of understanding of various Machine Learning/Data Mining concepts, through the implementation of clustering & regression algorithms using real datasets. Implementation is performed in R environment, while students need to discuss important aspects related to these problems and perform some critical evaluation of their results.
Learning Outcomes Covered in this Assignment:	This assignment contributes towards the following Learning Outcomes (LOs): • Suitably prepare a realistic data set for data mining / machine learning and discuss issues affecting the scalability and usefulness of learning models from that set • Effectively implement, apply and contrast unsupervised/supervised machine learning / data mining algorithms for simple data sets • Evaluate, validate and optimise learned models • Effectively communicate models and output analysis in a variety of forms to specialist and non-specialist audiences
Handed Out:	10/02/2021
Due Date	27/04/2021, 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.
Type of Feedback and Due Date:	Feedback will be provided on BB, after 15 working days

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 40 – 49%, in which case the mark will be capped at the pass mark (40%). 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/discussion). 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. This is a classic multi-dimensional, in terms of features, problem.

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 (desired output)

For this part, you need to use the first 18 attributes to your "clustering"- based calculations.

1st Objective (partitioning clustering)

You need to conduct the k-means clustering analysis of this vehicle dataset problem. As this is a typical multi-dimensional, in terms of features, problem, initially, you need to provide a brief discussion of the methodologies used in reducing the dimensionality for such type of problems and rationale of using them. (**Suggestion:** consult related literature and add some relevant references). In this specific clustering part, however, the analysis will be

performed with all initial features, as the main aim is to assess different clustering results under the initial conditions. Before conducting the k-means, perform the following pre-processing tasks: scaling and outliers removal and briefly justify your answer. (**Suggestion:** the order of scaling and outliers removal is important. The outlier removal issue is not covered in tutorials). Define the number of cluster centres (via manual & automated tools) and perform k-means analysis for each attempt. For each of the above k-means attempts, check your produced cluster outcome against the information obtained from 19th column and provide the related results/discussion (evidence of a "confusion" matrix and calculation of the accuracy/recall/precision indices from it). Choose the best "winner" clustering case (justify your response) and briefly explain the meaning of accuracy/recall/precision indices. Finally, for the "winner" case, provide the coordinates of each centre for each clustering group. Write a code in R Studio to address all the above issues (codes/results/discussion need to be included in your report). 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 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 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 this forecasting part you need to use an MLP-NN to predict the next step-ahead exchange rate of USD/EUR. Daily data (exchangeUSD.xls) have been collected from October 2011 until October 2013 (500 data). The first 400 of them have to be used as training data, while the remaining ones as testing set. Use only the 3rd column from the .xls file, which corresponds to the exchange rates.

2nd Objective (MLP)

You need to construct an MLP neural network for this forecasting problem. The definition of the input vector for NNs is a very important component for time-series analysis. Therefore, initially you need to provide a brief discussion of the various schemes used to define this input vector. (**Suggestion:** consult related literature and add some relevant references). In this specific forecasting part, however, we are going to utilise only the "autoregressive" (AR) approach, i.e. time-delayed exchange rates as input variables. As the order of this AR approach is not known, you need to experiment with various input vectors and for each one of these cases you need to construct an input/output matrix for the MLP (using "time-delayed" rates). Each one of these matrices needs to be normalised, as this is a standard procedure for MLP NN. You need to explain briefly why normalisation procedure is necessary for this specific type of NN. For the training phase, you need to experiment with various MLPs, utilising these input vectors and various internal network structures (such as hidden layers, nodes, learning rate, activation function, etc.). For each case, the testing performance (i.e. evaluation) of the networks will be calculated using the standard statistical indices (RMSE, MAE and MAPE). Create a comparison table of their testing performances (using these specific statistical indices). Briefly explain the meaning of these three stat. indices. From this comparison table, check the "efficiency" of your best one-hidden layer and two-hidden layer networks, by checking the total number of weight parameters per network. Briefly, discuss which approach is more preferable to you. Finally, provide for your best MLP network, the related results both graphically (your prediction output vs. desired output) and via the stat. indices. Write a code in R Studio to address all these requirements. 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. 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 50)

Coursework Marking scheme

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

1st Objective (partitioning clustering)

• Brief discussion of methodologies used for reducing the input dimensionality	5
• Pre-processing tasks (3 marks for scaling and 9 marks for outliers removal)	12
• Define the number of cluster centres by showing all necessary steps/methods (via manual & automated tools).	7
• K-means analysis for each attempt	5
• Evaluation of the produced outputs against 19th column	9
• Define the final "winner" cluster case and provide brief explanation of evaluation indices.	8
• Illustrate the coordinates of each centre for each clustering group	4

2nd Objective (MLP)

• Brief discussion of the various methods used for defining the input vector in time-series problems	5
• Evidence of various adopted input vectors and the related input/output matrices	5
• Evidence of correct normalisation and brief discussion of its necessity	8
• 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. (4 marks for structures with different input vectors, 9 marks for different internal NN structures and 5 for the comparison table).	18
• Discussion of the meaning of these stat. indices	4
• Discuss the issue of "efficiency" with your two best NN structures	4
• 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