University of Westminster
School of Computer Science & Engineering

5DATA002W	Machine Learning & Data Mining – Coursework (2021/22)
Module leader	Dr. V.S. Kontogiannis
Unit	Coursework The current version of CW can be considered as provisional, as it needs to be moderated by external examiner. Therefore, it may be subjected to slight changes following module leader's agreement for such amendments. If there are any changes, students will be informed.
Weighting:	50%
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 • 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:	4/03/2023
Due Date	3/04/2023 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 online with 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: **https://www.westminster.ac.uk/current-students/guides-and-policies/assessment-guidelines/mitigating-circumstances-claims**

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 white wine varieties involving their chemical properties and ranking by tasters. Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of testing can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled. One dataset (whitewine_v2.xls) is available of which is on white wine and has 4710 varieties. All wines are produced in a particular area of Portugal. Data are collected on 12 different properties of the wines, one of which is Quality (i.e. the last column), based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.

Description of attributes:

1. fixed acidity: most acids involved with wine or fixed or non-volatile (do not evaporate readily)
2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. citric acid: found in small quantities, citric acid can add ‘freshness’ and flavour to wines
4. residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/litre and wines with greater than 45 grams/litre are considered sweet
5. chlorides: the amount of salt in the wine
6. free sulfur dioxide: the free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. total sulfur dioxide: amount of free and bound forms of $S0_2$; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine
8. density: the density of water is close to that of water depending on the percent alcohol and sugar content
9. pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. sulphates: a wine additive which can contribute to sulphur dioxide gas ($S0_2$) levels, which acts as an antimicrobial and antioxidant
11. alcohol: the percent alcohol content of the wine
12. Output variable (based on sensory data): quality (score between 0 and 10)

1st Objective (partitioning clustering)

You need to conduct the k-means clustering analysis of this white wine dataset problem. The dataset of 4710 wine samples is defined by 11 attributes (i.e. input variables) and one output (i.e. quality). There are 4 quality classes. In this assignment, do not attempt any merging of adjacent classes which have few samples. In this specific clustering part, initially the analysis will be performed with all initial features, as the main aim is to assess different clustering results under the initial conditions. In the next phase however, principal component analysis (PCA) will be applied to reduce the input dimensionality and the newly produced dataset will be clustered using the same k as the winner case from the initial phase. 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 topic is not covered in tutorials, so you need to explore it yourself). Define the number of cluster centres (via manual & automated tools). The automated tools should include NBclust, Elbow and one from Gap statistics or silhouette methods. You need to provide the related R-outputs and discussion on these outcomes. Using all input variables, perform a kmeans analysis with k=2, 3 &4. For each of the above k-means attempts, show all related R-based kmeans outputs, including information for the centres as well as the ratio of between_cluster_sums_of_squares (BSS) over total_sum_of_Squares (TSS). In addition, for each of these k-means attempts, check your produced cluster outcome against the information obtained from 12th column and provide the related results/discussion (evidence of a "confusion-like" matrix (CM) 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.
As this is a typical multi-dimensional, in terms of features problem, you need also to apply the PCA method to this wine dataset. You need to show all related to PCA R-outputs. Create a new "transformed" dataset with principal components (PC) as attributes. Choose those PCs that provide a cumulative score > 96%. Apply, kmeans analysis on this "new transformed" dataset using same k as the winner from the previous step. Show the related R-outputs of this kmeans analysis. Discuss the performance of this "PCA-based" kmeans model by calculating the related BSS, ratio BSS/TSS and within_cluster_sums_of_squares (WSS) indices and compare these produced indices against the related ones from the winner model (i.e. all attributes) from the previous stage.
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 40)

Energy Forecasting Part (part of Work Based Learning activity)

Buildings represent a large percentage of a country's energy consumption and associated greenhouse gas emissions. The energy needed in order to maintain internal conditions within buildings, is responsible for a significant portion of the overall energy usage and greenhouse emissions. Thus, improving energy efficiency in buildings is of great importance to our overall sustainability. Over the past few decades, a lot of research has been carried out in order to improve building energy efficiency through various techniques and strategies. The forecasting of energy usage in an existing building is essential for a variety of applications like demand response, fault detection & diagnosis, optimization and energy management. This is a typical time-series based application. Data-driven forecasting models typically include two main approaches; statistical and machine learning based schemes. The statistical approach typically applies a pre-defined mathematical function and has shown good performance for medium to long term energy forecasting. In addition, such models have shown acceptable performance for short-term forecasting of consumption electricity loads. Machine learning approach in contrast, typically applies an algorithmic approach (which may non-linearly transform the data), in order to provide a forecast.

For this forecasting part of the coursework, you will be working on a specific case study, which involves a real-life organisation and a real dataset. More specifically, in collaboration with the Estates Planning & Services Department, at University of Westminster, we have been supplied (via LG Energy Group) with the hourly electricity consumption data (in kWh) for the University Building at 115 New Cavendish Street London for the years 2018 and 2019. Although full data information has been supplied to us, you will use only a small portion of that information in this coursework. The provided (UoW_load.xlsx) file includes daily electricity consumption data for three hours (11:00, 10:00 & 09:00) for the 2018 and partly 2019 periods (in total 500 samples). The objective of this question is to use a multilayer neural network (MLP-NN) to predict the next step-ahead (i.e. next day) electricity consumption for the 11:00 hour case. The first 430 samples will be used as the training data, while the remaining ones will be used as the testing set.

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 energy forecasting analysis. Therefore, initially you need to provide a brief discussion of the various schemes/methods used to define this input vector in electricity load forecasting problems. (**Suggestion:** consult related literature and add some relevant references). In this specific forecasting part, however, you are going to utilise only the "autoregressive" (AR) approach, i.e. time-delayed values of the 11th hour attribute as input variables. As the order of this AR approach is not known, you need to experiment with various (time-delayed) input vectors and for each one of these cases you need to construct an input/output matrix (I/O) for the MLP training/testing (using "time-delayed" electricity loads). Experiment with various input vectors up to (t-4) level. According to literature, the electricity consumption forecast depends also on the (t-7) (i.e. one week before) value of the load. Thus, in your "AR" analysis, you need also to investigate the influence

of this specific time-delayed load to the forecasting performance of your NN models. In addition, to this "classic" AR approach, you need also to consider additional input vectors by including information from the 10th and 9th hour attributes. In that case, your NN models could be considered as NARX (nonlinear autoregressive exogenous) models. Each one of these I/O matrices needs to be normalised, as this is a standard procedure especially for this type of NN. You need to explain briefly the rationale of this normalisation procedure 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 and why. 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 60)

Coursework Marking scheme

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

1st Objective (partitioning clustering)

• Pre-processing tasks (2 marks for scaling and 3 marks for outliers removal)	5
• Define the number of cluster centres by showing all necessary steps/methods via manual & automated tools (2 marks for each one of these automated tools)	6
• K-means analysis for each k attempt (2 marks for each k attempt) (all attributed used)	6
• Evaluation of the produced outputs against 12th column (1 mark for the "CM" table, 6 marks for calculation of these requested indices)	7
• Define the final "winner" cluster case and provide brief explanation of evaluation indices	4
• Apply a PCA for this white wine dataset (2 marks). Create a new dataset with those PCs with a cumulative score > 96%, as attributes (2 marks).	4
• Apply kmeans on this new "PCA-based" dataset	2
• Discuss the performance for this "PCA-based" dataset though the calculation of WSS, BSS, BSS/TSS indices (3 marks) and compare them against the ones produced from the previous "winner" model utilising all attributes (3 marks).	6

2nd Objective (MLP)

• Brief discussion of the various methods used for defining the input vector in electricity load forecasting problems	5
• Evidence of various adopted input vectors and the related input/output matrices for both AR (5 marks) and NARX (4 marks) based approaches	9
• Evidence of correct normalisation (3 marks) and brief discussion of its necessity (3 marks)	6
• Implement a number of MLPs for the AR approach, 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, 8 marks for different internal NN structures and 4 for the comparison table).	16
Repeat the above step for the NARX also approach (2 marks for structures with different input vectors, 4 marks for different internal NN structures and 2 for the comparison table).	8
• Discussion of the meaning of these stat. indices	6
• 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