University of Westminster

School of Computer Science & Engineering

5DATA002W	Machine Learning & Data Mining – Coursework 2023-2024
Module leader	Dr. V.S. Kontogiannis IIT Module Leader: Nipuna Senanayake
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:	60%
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:	19/02/2024
Due Date	02/05/2024, Submission by 13:00 IST
Expected deliverables	The preferable submission delivery via Blackboard is only one pdf file containing the required details (results/analysis/discussion). All implemented R codes should be included in your documentation (in an Appendix). Alternatively, you may not include the implemented R codes in your pdf document. But in this case, you need to create and submit a zip/rar file which will contain the pdf document and the scripts of R codes. Results/ana-lysis/discussion should be on the pdf document.
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

Partitioning 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.

Objectives/Deliverables (partitioning clustering)

One dataset which includes samples of white wine (whitewine_v6.xls) is available and has 2700 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. For this clustering part, you need to use only the first 11 attributes to your calculations. Clustering is an unsupervised scheme, thus, the information included in the “quality” attribute can’t be used.

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)

The work in this part is divided into two subtasks:

- In the 1st subtask, the analysis will be performed with all initial attributes (i.e. the first 11 features), as the aim is to assess clustering results using all input variables.
- In the 2nd subtask, however, principal component analysis (PCA) will be applied to reduce the input dimensionality and the newly produced dataset will be again clustered. The aim in this 2nd subtask is to help students understand the principles and effects of reducing dimensionality in multi-dimensional problems.

1st Subtask Objectives:

a. Before conducting the k-means, perform the following pre-processing tasks: scaling and outliers detection/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). Obviously, you can implement this clustering task without exploring this "outlier" component, however, you will not be awarded the allocated marks for this component!

b. You need then to determine the number of cluster centres via four "automated tools". The "automated tools" should include NBclust, Elbow, Gap statistics and silhouette methods. You need to provide, in your report, the related R-outputs and your discussion on these outcomes.

c. The next step is the kmeans clustering investigation. Using, again, all input variables, perform a kmeans analysis using the most favoured "k" from those "automated" methods. For this k-means attempt, show the related R-based kmeans output, including information for the centres, clustered results, as well as the ratio of between_cluster_sums_of_squares (BSS) over total_sum_of_Squares (TSS). It is also important to calculate/illustrate the BSS and the w**ithin_cluster_sums_of_squares** (WSS) indices (internal evaluation metrics).

d. Following the kmeans attempt, provide the silhouette plot (another internal evaluation metric) which displays how close each point in one cluster is to points in the neighbouring clusters. Provide the average silhouette width score and your discussion on this plot, which should include your comments on the level of "quality" of the obtained clusters.

2nd Subtask Objectives:

e. 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 R-outputs related to PCA analysis, including eigenvalues/eigenvectors, cumulative score per principal components (PC). Create a new "transformed" dataset with principal components as attributes. Choose those PCs that provide at least cumulative score > 85%. Provide a brief discussion for your choice to choose specific number of PCs.

f. In reality, as we have practically a new dataset, we need to find an appropriate k for our "new" kmeans clustering attempt. Like previously, apply the same four "automated" tools to this new pca-based dataset. You need to provide, in your report, the related R-outputs and your discussion on these "new" outcomes.

g. Using this new pca-dataset, perform a kmeans analysis using the most favoured k from those "automated" methods. For this k-means attempt, show the related R-based kmeans output, including information for the centres, clustered results, as well as the ratio of between_cluster_sums_of_squares (BSS) over total_sum_of_Squares (TSS). It is also important to calculate/illustrate the BSS and the w**ithin_cluster_sums_of_squares** (WSS) indices (internal evaluation metrics).

h. Following this "new" kmeans attempt, provide the silhouette plot which displays how close each point in one cluster is to points in the neighbouring clusters. Provide the average silhouette width score and your discussion on this plot, which should include your comments on the level of "quality" of the obtained clusters.

i. Following the kmeans analysis for this new "pca" dataset, implement and illustrate the Calinski-Harabasz Index. This is another well-known internal evaluation metric and it has not been covered in tutorial sessions. Provide, a brief discussion on the outcome of this index.

Write a code in R Studio to address all the above issues (related 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 for all these tasks. The usage of kmeans R function is compulsory.

(Marks 50)

Financial 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.

Objectives/Deliverables (Multi-layer Perceptron Neural Network)

For this forecasting part of the coursework, you will be working on a specific case study, which is related to exchange rates forecasting. The task of this question is to use a multilayer neural network (MLP-NN) to predict the next step-ahead (i.e. next day) 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 (i.e. USD/EUR) column from the .xls file, which corresponds to the exchange rates. In this part, the task of the one-step-ahead forecasting of exchange rates will utilise only the "autoregressive" (AR) approach (i.e. time-delayed values of the 3rd column attribute as input variables).

a) Before, you attempt any analysis on this dataset, you need to provide a brief discussion on the type of input variables used in MLP models for exchanges rates forecasting. The definition of the input vector for NNs is a very important component for time-series analysis. Therefore, briefly discuss the various schemes/methods used to define this input vector in this financial domain. The AR approach used in this CW is obviously one of such schemes. (**Suggestion:** consult related literature in exchange rates forecasting and add some relevant references to this domain).

Task Objectives:

In this specific task, utilise only the "autoregressive" (AR) approach, i.e. time-delayed values of the exchange rates (i.e. 3rd column) attribute as input variables. Experiment with various input vectors up to (t-4) level.

b) As the order of this AR approach is not known, you need to experiment with various (time-delayed) input vectors and for each case chosen, you need to construct an input/output matrix (I/O) for the MLP training/testing (using "time-delayed" exchange rates)
c) Each one of these I/O matrices needs to be normalised, as this is a standard procedure especially for this type of NN. Explain briefly the rationale of this normalisation procedure for this specific type of NN (i.e. why do we need to normalise data before using them in an MLP structure?)
d) For the training phase, you need to experiment with various MLP models, utilising these different input vectors and various internal network structures (such as hidden layers, nodes, linear/nonlinear output, 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, MAPE and sMAPE – symmetric MAPE).
e) Briefly explain the meaning of these four stat. indices.
f) Create a comparison table of their testing performances (using these specific statistical indices). Add a column in this matrix, where you will provide a brief description of the specific NN structure. As, the number of potential NN structures (with various input vectors and internal structures) that can be created can be huge, in this exercise, restrict your total number of developed NNs to 12-15 models. Obviously, these models will have differences in terms of input vector and internal structure. The main aim of this task, by providing such different models, is to understand how such differences may have effect in the forecasting accuracy.
g) 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/structure is more preferable to you and why.
h) Finally, provide for your best MLP network, the related results both graphically (your prediction output vs. desired output) and via the stat. indices. In terms of graphics, you can either use a scatter plot or a simple line chart.

Write a code in R Studio to address all the above issues/tasks. 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 for these tasks. The usage of neuralnet R function for MLP modelling is compulsory. 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.

(Marks 50)

Coursework Marking scheme

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

1st Objective (partitioning clustering)

1. Pre-processing tasks (2 marks for scaling and 5 marks for outliers detection/removal) 7
2. Determine the number of cluster centres by showing all necessary steps/methods via "automated" tools (1.5 mark for each one of these "automated" tools) 6
3. K-means analysis for the chosen k (all attributes used) and show all requested outputs 5
4. Show the silhouette plot (3 marks) and provide related discussion on this output, following this Kmeans attempt (2 marks) 5
5. Apply a PCA for this wine dataset and show all related R-outputs (4 marks). Create a new dataset with those PCs with a cumulative score at least > 85%, as attributes and provide a discussion for your choice (4 marks). 8
6. Determine the number of cluster centres by showing all necessary steps/methods via "automated" tools (1.5 mark for each one of these "automated" tools) 6
7. K-means analysis for this "pca"-based dataset for the chosen k and show all requested outputs 5
8. Show the silhouette plot (3 marks) and provide related discussion on this output, following this "pca-based" Kmeans attempt (2 marks) 5
9. Implement and show the Calinski-Harabasz index. Provide, a brief discussion on the outcome of this index. 3

2nd Objective (MLP NN)

1. Brief discussion of the various methods used for defining the input vector in exchange rates forecasting problems 5
2. Evidence of various adopted input vectors and the related input/output matrices for this "AR" based approach 4
3. Evidence of correct normalisation (3 marks) /de-normalisation (3 marks) and brief discussion of its necessity for MLP networks (3 marks) 9
4. Implement a number of MLPs for the "AR" approach, using various internal structures (layers/nodes)/input variables/network parameters and show in the comparison table, their performances (based on testing data) through the provided stat. indices. (4 marks for structures with different input vectors, 8 marks for different internal NN structures). 12
5. Discussion of the meaning of these four stat. indices (2 marks for each index) 8
6. Creation of the comparison matrix for this exchange rates case 4
7. Discuss the issue of "efficiency" with your two best NN structures 4
8. Provide your best results both graphically (your prediction output vs. desired output) and via performance indices (2 marks for the graphical display and 2 marks for showing the requested statistical indices) 4