University of Westminster
Department of Computer Science

7BUIS008W	Data Mining & Machine Learning – Coursework 1 (2017/18)
Module leader	Dr. P.I. Chountas. This CW was prepared/written by Dr. V.S. Kodogiannis
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:	9th October 2017
Due Date	22nd November 2017 Submission by 10:00am
Expected deliverables	Submit on Blackboard a zip file containing the required documentation (either in docx or pdf format). All implemented codes should be included in your documentation together with the results/analysis.
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, on 13th December 2017 (15 working days)
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 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**

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 tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.

One dataset (whitewine.xls) is available of which is on white wine and has 4898 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 sulfur dioxide gas ($S0_2$) levels, wich 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)

In this clustering part you need to use the first 11 attributes to your calculations.

1st Objective (partitioning clustering)

You need to conduct the k-means clustering analysis of the white wine sheet. 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. In your report, check the consistency of those produced clusters, with information obtained from column 12.

(Marks 30)

2nd Objective (hierarchical clustering)

You need to conduct the hierarchical clustering (agglomerative) clustering analysis of the white wine sheet. Investigate the hclust() function for single, complete, average methods. Create the visualization of all methods using a dendrogram. Look at the cophenetic correlation between each clustering result using cor.dendlist. Discuss the produced results after using the coorplot function. Write a code in R Studio to address all the above issues.

(Marks 25)

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, recent research has been 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 USD/EUR. Daily data (exchange.xls) have been collected from March 2015 until October 2016 (390 data). The first 320 of them have to be used as training data, while the remaining ones as testing set.

3rd Objective (MLP)

You need to construct an MLP neural network for this problem. You need to consider the appropriate input vector, 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). Hind: 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. 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).

(Marks 30)

4th Objective (SVR)

You need to construct a SVM 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 usual statistical indices (MSE, RMSE and MAPE). Hind: The input selection problem is very important. Experiment with various options. Show all your working steps. 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 15)

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, 8
- K-means with the best two clusters, 8
- Find the mean of each attribute for the winner cluster, 5
- Check consistency of your results against column 12, 4
- Check for any pre-processing tasks 5

2nd Objective (hierarchical clustering)

- Perform hierarchical clustering (for single, complete, etc) 15
- Create a dendrogram 5
- Check the cophenetic correlation and discuss your findings 3
- Coorplot function and discuss your findings 2

3rd Objective (MLP)

- Discuss the input selection problem and propose various input configurations 10
- Design a number of MLPs, using various structures (layers/nodes) / input parameters and show in a table their performances comparison based on provided stat. indices 15
- Provide your best results both graphically (your prediction output vs. desired output) and via performance indices 5

4th Objective (SVR)

- Discuss the input selection problem and propose various input configurations 2
- Design an SVR and use various structures/parameters (linear/nonlinear kernels) 10
- Provide your best results both graphically (your prediction output vs. desired output) and via performance indices 3