**University of Westminster**
Department of Computer Science

| 7BUIS008W | Data Mining & Machine Learning – Coursework 1 (2017/18) |
|---|---|
| Module leader | Dr. P.I. Chountas. |
| Unit | Coursework 2 |
| 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):<br>• LO1 critically justify the use of effective and novel data mining and machine learning techniques for Data Science applications;<br>• LO3 critically reflect on the knowledge of how different data mining and machine learning algorithms operate and their underlying design assumptions and biases in order to select and apply an appropriate such algorithms to solve a given problem;<br>• LO5 critically analyse the output of data mining and machine learning algorithms by drawing technically appropriate and justifiable conclusions resulting from the application of data mining and machine learning algorithms to real-world data sets |
| Handed Out: | 28th November 2017 |
| Due Date | 09th January 2018<br><br>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 18th January 2018 (07 working days) |
| BCS CRITERIA MEETING IN THIS ASSIGNMENT | • **7.1.6 Use appropriate processes**<br>• **7.1.7 Investigate and define a problem**<br>• **7.1.8 Apply principles of supporting disciplines**<br>• **8.1.1 Systematic understanding of knowledge of the domain with depth in particular areas**<br>• **8.1.2 Comprehensive understanding of essential principles and practices**<br>• **8.2.2 Tackling a significant technical problem**<br>• **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

## Classifier Evaluation

You have been retained as a data scientist and suppose you have collected a dataset of already-classified instances and you have to build a classifier. Consider the following type of classifiers

1. Decision Tree ( C4.5, Random Forest)
2. Naïve Bayes
3. K-nearest neighbour

**Question**: How will you know how good your classifier is?

### 1st Task: Data Set Selection and Visualisation

You need to select a data set of your own choice (i.e. you may use a dataset already used before in the lab, or from the literature review) for the purposes of building training and validating the above type of classifiers 1-3. With the aid of R package visualise and justify the properties of the selected data set.

**[10 Marks]**

### 2nd Task: Formation of Training and Test Sets

Assuming we have collected one large dataset of already-classified instances, you need to look at three methods of forming training and test sets from this single dataset in R as described below.

**The holdout method**

The simplest method is to take your original dataset and partition it into two, randomly selecting instances for a training set (usually 2/3 of the original dataset) and a test set (1/3 of the dataset). You build the classifier using the training set and then evaluate it on the 'held-out' test set.

This has the advantage of being simple. But it makes poor use of the available data and it raises questions about the representativeness of each dataset (e.g. you may just get lucky with all the 'easy' instances in the test set).

**Cross-validation**

A generalization of the holdout method. N total samples are divided into m groups of equal size. M different classifiers are trained each using m –1 groups, holding out each of the groups. For each of the m classifiers, the group left out is tested.The m test results are averaged. All samples get used for both training and testing. The result is unbiased and with minimum variance. A good method to use for selecting the appropriate classifier type to use and for determining certain classifier "super"-parameters, e.g., k for nearest neighbour.

**Leave-one-out cross-validation (Jack Knife)**

A limiting case of cross-validation. Where m = N. N different classifiers are trained each using N–1 samples. For each of the N classifiers, the one left out sample is tested. The N test results are averaged. Classifiers are very close to optimal. All samples are used for testing. Result is unbiased and with minimum variance. If a fast leave-one-out algorithm is available (to estimate necessary parameters using an update scheme: e.g., mean, covariance and its inverse and determinant). Fast algorithms exist for estimating mean, covariance matrix, as well as inverse and determinant of covariance matrix. So, useful for: Bayes quadratic, k-nearest neighbor (using Euclidean or Mahalanobis distance).

**[15 Marks]**

### 3rd Task: Build Train and Test a Decision Tree type Classifier

You need to construct, train and test Decision Tree type classifier (C4.5, Random Forest) in R. Train and test your decision tree classifier using the training and test sets generated based on the methods tried as part of the **2nd Task**.

**[25 Marks]**

### 4th Task: Build Train and Test a Naïve Bayes type Classifier

You need to construct, train and test Naïve Byes type classifier in R. Train and test your Naïve Byes classifier using the training and test sets generated based on the methods tried as part of the **2nd Task**.

**[15 Marks]**

### 5th Task: Build Train and Test a K-NN type Classifier

You need to construct, train and test K-NN type classifier in R. Train and test your K-NN classifier using the training and test sets generated based on the methods tried as part of the **2nd Task**.

**[15 Marks]**

### 6th Task: Measure Performance

For each type of classifier calculate and display the following performance related metrics in R. Use the library library(ROCR)

1. Confusion matrix
2. Precision vs. Recall
3. Accuracy
4. ROC(receiver operating characteristic curve)
5. RAUC (receiver under the curve area)

**[20 Marks]**

# Coursework Marking scheme

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

## 1st Task: Data Set Selection and Visualisation

- Data Set summary of main properties 5
- Visualisation in R of main data set properties 5

## 2nd Task: Formation of Training and Test Sets

Formation of training and test sets from in R using the methods below.

- The holdout method 5
- Cross-validation 5
- Leave-one-out cross-validation (Jack Knife) 5

## 3rd Task: Build Train and Test a Decision Tree type Classifier

- Building of Decision Tree type classifier (C4.5, Random Forest) in R 8
- Training of Decision Tree type classifier (C4.5, Random Forest) in R 8
- Testing of Decision Tree type classifier (C4.5, Random Forest) 9

## 4th Task: Build Train and Test a Naïve Bayes type Classifier

- Building of Naïve Bayes type classifier in R 4
- Training of Naïve Bayes type classifier in R 5
- Testing of Naïve Bayes type classifier in R 6

## 5th Task: Build Train and Test a K-NN type Classifier

- Building of K-NN type classifier in R 4
- Training of K-NN type classifier in R 5
- Testing of K-NN type classifier in R 6

## 6th Task: Measure Performance

- Confusion matrix estimation 6
- Precision vs. Recall estimation 4
- Accuracy estimation 4
- ROC(receiver operating characteristic curve) plot 3
- RAUC (receiver under the curve area) plot 3