**University of Westminster**

Department of Computer Science

|  |  |
| --- | --- |
| **7BUIS008W Data Mining & Machine Learning – Coursework 2 (2019/20)** | |
| Module leader | Dr. Amitha Caldera |
| Unit | Coursework 2 |
| Weighting: | 50% |
| Qualifying mark | 35% |
| Description | Students are expected to critically justify the use of effective and novel data mining and machine learning techniques for a specific problem domain and definitely reflect on the knowledge of how different data mining and machine learning algorithms operate in terms of their underlying design assumptions and biases for a given problem domain. Students expected to methodically analyse the output of data mining and machine learning algorithms by drawing technically appropriate and sound conclusions resulting from the application of data mining and machine learning algorithms to the given problem |
| Learning Outcomes Covered in this Assignment: | This assignment contributes towards the following Learning Outcomes (LOs):   * LO1 critically justify the use of effective and novel data mining and machine learning techniques for Data Science applications; * LO3 critically reflect on the knowledge on 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; * 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: | 24th of April 2020 |
| Due Date | 22nd of May 2020  Submission by 13:00 hours |
| 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. |
| 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 50 – 59%, in which case the mark will be capped at the pass mark (50%). 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**

**Ensemble Classifier Evaluation**

You have been retained as a data scientist and suppose you have collected a dataset from [UCI](https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table), excluding IRIS, of already-classified instances and you have to build an ensemble type classifier.

Ensembles can give you a boost in accuracy on your dataset. You can create ensembles of machine learning algorithms in R. There are three main techniques (Boosting, Bagging and Stacking) that you can create an ensemble of machine learning algorithms in R.

The three most popular methods for combining the predictions from different models are:

* **Bagging**: Building multiple models (typically of the same type) from different subsamples of the training dataset.
* **Boosting**: Building multiple models (typically of the same type) each of which learns to fix the prediction errors of a prior model in the chain.
* **Stacking**. Building multiple models (typically of differing types) and supervisor model that learns how to best combine the predictions of the primary models.

You can combine the predictions of multiple *caret* models using the ***caretEnsemble***package. Given a list of caret models, the ***[caretStack()](https://www.machinelearningplus.com/machine-learning/caret-package/)***function can be used to specify a higher-order model to learn how to best combine the predictions of sub-models together

This assignment is focused on **Bagging** and **Stacking** and on how you can continue to ratchet up the accuracy of the models on your own datasets.

Bagging Algorithms

The base type bagging machine learning algorithms that will be examined in this assignment are:

* Bagged CART,
* Random Forest

Stacking Algorithms

The base type stacking machine learning algorithms that will be examined in this assignment are

* Classification and Regression Trees (CART),
* K-Nearest Neighbors (KNN),
* Naïve Bayes (NB)

*Question: How will you know how good your ensemble classifier is? Under which conditions ensemble learning is useful?*

**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 (Bagging, Stacking). With the aid of R package visualise and justify the properties of the selected data set.

**[15 Marks]**

2nd Task: Formation of Training and Test Sets

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

## **Repeated k-fold Cross Validation**

The process of splitting the data into k-folds can be repeated a number of times; this is called Repeated k-fold Cross Validation (repeatedcv). The final model accuracy is taken as the mean from the number of repeats.

**[10 Marks]**

**3rd Task: Build Train and Test a Bagging type Classifier**

You need to construct, train and test a Bagging type classifier in R, based on Bagged CART and Random Forest base classifiers. Train and test the Bagging classifier using the training and test sets generated based on the method tried as part of the 2nd Task.

**[20 Marks]**

**4th Task: Build Train and Test a Stacking type Classifier**

You need to construct, train and test a Stacking type classifier in R, based on (CART, KNN, NB). Train and test your Stacking classifier using the training and test sets generated based on the method tried as part of the 2nd Task.

**[25 Marks]**

**5th Task: Measure Performance**

For each type of ensemble type classifier calculate and display the following performance related metrics in R. Critically comment on the importance of each metric for each type of ensemble type classifier. 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)
6. Training time
7. Testing time

**[30 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
* Feature Selection 5

2nd Task: Formation of Training and Test Sets

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

* Repeated CV for Bagging type classifier 5
* Repeated CV for Stacking type classifier 5

**3rd Task: Build Train and Test a Bagging type Classifier type Classifier**

* Building of Random Forest type classifier in R 5
* Building of Bagged CART type classifier in R 5
* Testing of Bagging type classifier 10

**4th Task: Build Train and Test a Stacking type Classifier**

* Building of Stacking CART classifier in R 5
* Building of Naïve Bayes type classifier in R 5
* Building of K-NN type classifier in R 5
* Testing of Stacking type classifier in R 10

**5th Task: Measure Performance**

* Confusion matrix estimation 6
* Precision vs. Recall estimation 4
* Accuracy estimation 4
* ROC(receiver operating characteristic curve) plot 4
* RAUC (receiver under the curve area) plot 4
* Training time 4
* Testing time 4