Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT307
Module Title: Applied Machine Learning
Lecturer: Yuhua Li
Assessment Title: Implementation and Evaluation of a Case Study Using Machine Learning Techniques
Assessment Number: 1
Date Set: 3 December 2021
Submission Date and Time: 10 January 2022 at 9:30am
Return Date: 11 February 2022

This assignment is worth **50%** of the total marks available for this module. If coursework is submitted late (and where there are no extenuating circumstances):

1. If the assessment is submitted no later than 24 hours after the deadline, the mark for the assessment will be capped at the minimum pass mark;
2. If the assessment is submitted more than 24 hours after the deadline, a mark of 0 will be given for the assessment.

Your submission must include the official Coursework Submission Cover sheet, which can be found here:
https://docs.cs.cf.ac.uk/downloads/coursework/Coversheet.pdf

Submission Instructions

Your submission must include:

- A filled-in copy of the official Coursework Submission Cover sheet.
- A Jupyter Notebook (.ipynb) file containing all your code and execution outputs/figures.
- A typeset PDF report (see next section for details).

Ensure that your student number is included on the report and as a comment at the top of each Python file that makes up your submission.

You must submit to Learning Central three files (each named using your student number in the format of [student number]_CW1.pdf, e.g., C1234567_CW1.pdf) which contains the following files:

Description		Type	Name
Cover sheet	**Compulsory**	One PDF (.pdf) file	[student number]_CW1.pdf
Assessment 2	**Compulsory**	One Jupyter Notebook (.ipynb) file.	[student number]_CW1_code.ipynb
Report	**Compulsory**	One PDF (.pdf) file containing your calculation for question 1 and your report for question 2.	[student number]_CW1_report.pdf

Before submitting your Jupyter Notebook file (.ipynb), make sure to restart the kernel and execute each cell such that all outputs and figures are visible. Any code submitted will be run in Python 3 and must be submitted as stipulated in the instructions above.

Any deviation from the submission instructions above (including the number and types of files submitted) will result in a mark of zero for the assessment or question part.

You can submit multiple times on Learning Central. ONLY files contained in the last attempt will be marked, so make sure that you upload all files in the last attempt.

Staff reserve the right to invite students to a meeting to discuss coursework submissions

---

Assignment

There are two questions in this coursework, marks for each part are in brackets.

Question 1

Your algorithm gets the following results in a classification experiment, where in the table, 'Id' is the index number, 'Target' is the ground truth that the classifier aims to achieve, 'Prediction' is the predicted results. Please compute the confusion matrix, precision, recall, f1-measure and accuracy **manually** (without the help of your computer/Python, please provide all steps and formulas). Include the process to get to the final results. **[10%]**

Id	Target	Prediction
1	True	True
2	True	True
3	True	False
4	True	True
5	True	True
6	True	False
7	True	True
8	True	True
9	True	True
10	False	False
11	False	False
12	False	False
13	True	True
14	True	False
15	True	True
16	False	False
17	False	False
18	False	True
19	False	True
20	False	False

Question 2

In this question, you will develop machine learning models to predict e-commerce visitors' purchasing intention. The given dataset ***Coursework_1_data.csv***, which can be downloaded from Learning Central, contains shoppers' online activity information including clickstream and session information data, where the last column ***Revenue*** represents visitors' purchasing intention. Your tasks will include data exploration, data pre-processing, machine learning method selection and implementation, and model performance evaluation. In addition to aforementioned tasks, you will write a concise report (around 1000 words, excluding tables and figures) to summarise your work and provide an analysis and discussion of the results.

i) **Data exploration [10%]**:
Conduct exploratory inspection of the dataset to provide a good understanding of data characteristics.

ii) **Data pre-processing [30%]**:
Carry out well thought pre-processing procedures to prepare the data into a form that is likely to lead to better performance.

iii) **Model implementation [30%]**:
Select three representative classification methods with a clear justification of your choice. Implement and optimise the classifiers for your chosen classification methods.

iv) **Performance evaluation [10%]**:
Organise the data in a suitable form to ensure the trained classifiers to provide reliable results. Evaluate models using suitable performance metrics.

v) **Result analysis and discussion [10%]**:
Provide an insightful analysis and comparison on results that you obtained from above steps, draw conclusions based on the results and analysis.

Learning Outcomes Assessed

Completion of this coursework allows students to demonstrate that they can:

1. Implement and evaluate machine learning methods to solve a given task
2. Explain the basic principles underlying common machine learning methods
3. Choose an appropriate machine learning method and data pre-processing strategy to address the needs of a given application setting
4. Reflect on the importance of data representation for the success of machine learning methods

Criteria for assessment

Question 1: Your answer needs to provide all steps and formulas, including the process to get to the final results.

Question 2: Your submitted Jupyter Notebook file (.ipynb) will be marked in conjunction with your Report. Credit will be awarded against the following criteria.

Criteria	Fail (0-49%)	Pass (50-59%)	Merit (60-69%)	Distinction(>=70%)
Data exploration [10%]	No or arbitrary data exploration.	Suitable but limited data exploration.	Good data exploration but miss some insightful analysis.	Thorough and insightful data exploration.
Data pre-processing [30%]	No or very little data pre-processing.	Some necessary pre-processing is conducted.	Adequate pre-processing to prepare the data for model development.	Extensive pre-processing to deal with all aspects of non-ideal characteristics of the data with an aim to achieve a best classification performance.
Model implementation [30%]	Less than 3 classification methods are implemented. Classifiers are not correctly implemented and optimised.	Three different classifiers are implemented but not properly/sufficiently trained.	Three classification methods are selected to give a good representation of various classification techniques. All implemented classifiers are properly trained and optimised.	Three classification methods are carefully selected with a good justification of your selection, which give a good representation of various classification techniques. All classifiers are excellently implemented, properly trained and systematically optimised.
Performance evaluation [10%]	Little performance evaluation. Performance evaluation using arbitrary metrics.	Performance evaluation using metrics without considering data characteristics.	Good performance evaluation with suitable metrics.	Use most suitable performance metrics to assist the selection of the best model on characteristics of this dataset.
Result analysis and discussion [10%]	No or little meaningful result analysis and discussion.	General result analysis and discussion. Vague conclusion.	Good result analysis and discussion but lack of depth.	Insightful result analysis and discussion, conclusion clearly drawn.

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned via Learning Central. There will be opportunity for individual feedback during an agreed time.

Feedback for this assignment will be useful for subsequent skills development, such as data science, natural language processing and deep learning (which will be studied during the second semester).