Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT307
Module Title: Applied Machine Learning
Lecturer: Jose Camacho-Collados, Yuhua Li
Assessment Title: Coursework 1
Assessment Number: 1
Date Set: Monday, October 28th
Submission Date and Time: Tuesday, January 14th at 9:30am
Return Date: Friday, February 7th

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

This coursework consists of a portfolio divided into two parts with equal weight:

- Part (1) consists of **selected homework** handed in throughout the course. The final deliverable consists of a single PDF file, which may include reflective answers to theoretical exercises, snippets of Python code and solved exercises.

- Part (2) consists of a **machine learning project** where students implement a basic machine learning algorithm for solving a given task. The deliverable is a zip file with the code, and a written summary (up to 1200 words) describing solutions, design choices and a reflection on the main challenges faced during development.

Description		Type	Name
Cover sheet	**Compulsory**	One PDF (.pdf) file	[student number].pdf
Part 1	**Compulsory**	One PDF (.pdf) file	part1_[student number].pdf
Part 2	**Compulsory**	One ZIP (.zip) file containing the Python code	part2code_[student number].pdf
Part 2	**Compulsory**	One PDF (.pdf) file for the reflective report	part2report_[student number].pdf

Any code submitted will be run in Python 3 (Linux) 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.

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

Assignment

In this coursework, students demonstrate their familiarity with the topics covered in the module via two separate parts with equal weight (50% each).

Part 1

In Part 1, students are expected to answer two types of questions: theoretical and practical. Please answer the questions with your own words and provide short answers (fewer than 100 words) for the theoretical questions.

1. Theory (15%)

1. What is the difference between a rule-based system and a machine learning system? **(5%)**
2. What is the difference between unsupervised and supervised learning? **(5%)**
3. What do we mean when we say that a machine learning system is overfitting? **(5%)**

2. Practice (85%)

1. Your algorithm gets the following results in a classification experiment. Please compute the precision, recall, f-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 result. **(20%)**

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

2. You are given a dataset (named Wine dataset) with different measured properties of different wines (dataset available in Learning Central). Your goal is to develop a machine learning model to predict the quality of an unseen wine given these properties. Train two machine learning regression models and check their performance. Write, for each of the models, the main Python instructions to train and predict the labels (one line each, no need to include any data preprocessing) and the performance in the test set in terms of Root Mean Squared Error (RMSE) **(30%)**

3. Train an SVM binary classifier using the Hateval dataset (available in Learning Central). The task consists of predicting whether a tweet represents hate speech or not. You can preprocess and choose the features freely. Evaluate the performance of your classifier in terms of accuracy using 10-fold cross-validation. Write a table with the results of the classifier (accuracy, precision, recall and F-measure) in each of the folds and write a small summary (up to 500 words) of how you preprocessed the data, chose the feature/s, and trained and evaluated your model **(35%)**

Part 2

In Part 2, students are provided with a sentiment analysis dataset (IMDb). The dataset contains positive and negative movie reviews. Training, development and test splits are provided. Based on this dataset, students will be asked to preprocess the data, select features and train a machine learning model of their choice to solve this problem. Students should include at least three different features to train their model, one of them should be based on some sort of word frequency. Students can decide the type of frequency (absolute or relative, normalized or not) and text preprocessing for this mandatory word frequency feature. The remaining two (or more) features can be chosen freely. Then, students are asked to perform feature selection to reduce the dimensionality of all features.

Deliverables for this part are the Python code including all steps and an essay of up to 1200 words. The Python code should include the Python scripts and a README file with instructions on how to run the code in Linux. Jupyter notebooks with clear execution paths are also accepted. The code should take the training set as input, and output the results in the test set. The code will consist of **25%** of the marks for this part and the essay the remaining **75%**. The code should contain all necessary steps described above: *to get the full marks for the code, it should work properly and clearly perform all required steps*. The essay should include:

1) Description of all steps taken in the process (preprocessing, choice of features, feature selection and training and testing of the model). **(25% - The quality of the preprocessing, features and algorithm will not be considered here)**
2) Justification of all steps. Some justifications may be numerical, in that case a development set is included to perform additional experiments. **(25% - A reasonable reasoned justification is enough to get half of the marks here. The usage of the development set is required to get full marks)**
3) Overall performance (precision, recall, f-measure and accuracy) of the trained model in the test set. **(10% - Indicating the results, even if very low, is enough to get half of the marks here. A minimum of 65% accuracy is required to get full marks)**

4) Critical reflection of how the deliverable could be improved in the future and on possible biases that the deployed machine learning may have. **(15% - The depth and correctness of insights related to your deliverable will be assessed)**

The essay may include tables and/or figures.

Extra credit (15% extra marks in the second part): For this second part students can get extra credits by writing an essay on one specific task related to Part 2 (except for option d, see instructions below). The essay will need to contain a maximum of 500 words (figures/tables are allowed and encouraged) and will deal with one of the following four specific topics:

a. **Error analysis:** Check the types of errors that the system submitted for Part 2 makes and reflect on possible solutions. Qualitative analysis with specific examples is encouraged.

b. **Literature review:** Write an essay about the state of the art of the field (i.e. automatic hate speech detection). Retrieve relevant articles and digest them, connecting them with your proposed solution to the problem in Part 2.

c. **Model comparison:** Propose and evaluate machine learning systems of different nature from the ones taught during the course. Write a table with all results and analyze the strengths and limitations of the approaches.

d. **Code release:** Create a GitHub or Bitbucket repository with the data and Python code used for Part 2, with very clear instructions on how to run the code from the terminal and about its different functionalities/parameters. Include all necessary data, provide full documentation and comment on the code. Students only need to include the link to the repository in the pdf.

Learning Outcomes Assessed

This coursework covers the 7 LOs listed in the module description. Specifically:
Part 1: LO1, LO2
Part 2: LO1, LO3, LO4, LO5, LO6

Criteria for assessment

Credit will be awarded against the following criteria.Credit will be awarded against the following criteria.

➢ **Part 1.** The main criteria for assessment in based on the correctness of the answer, unless a written reflection is required, in which case correctness/performance and written justification weigh 50% each.

- **Part 2.** This part is divided into Python code (25%) and an essay (75%). The code will be evaluated based on whether it works or not, and whether it minimally contains the necessary steps required for the completion of Part 2. Four items will be evaluated in the essay, whose weights and descriptions are indicated in the assessment instructions. The main criteria to evaluate those items will be the adequacy of the answer with respect to what was asked, and the justification provided.

The grade range is divided in:
Distinction (70-100%)
Merit (60-69%)
Pass (50-59%)
Fail (0-50)

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned between February 3rd and February 7th 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).