Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT316
Module Title: Applications of Machine Learning: Natural Language Processing and Computer Vision
Lecturer: Jose Camacho-Collados
Assessment Title: Coursework 1
Assessment Number: 1
Date Set: Monday, March 1st
Submission Date and Time: Tuesday, April 20th at 9:30am
Return Date: Friday, May 14th

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** similar to the one handed in throughout the course. The final deliverable consists of a single PDF file, which may include methodology, 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].zip
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 (first part: 40%; second part: 60%).

Part 1 (40%)

In Part 1, students are expected to answer two practical questions.

---

1. Practice

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. **(15 points)**

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	True
16	False	False
17	False	False
18	True	False
19	True	False
20	False	False

2. You are given a dataset (named "real_state") with different house properties (dataset available in Learning Central). Your goal is to train machine learning models in the training set to predict the house price of a unit area in the test set. The problem should be framed as both regression and classification. For regression, the house price of a unit area is given; for classification, there would be two labels (expensive and not-expensive) depending on the house price of a unit area: expensive if it is higher or equal to 30, and not-expensive if it is lower than 30.

The task is therefore to train two machine learning models (one regression and another one classification) and check their performance. The student can choose the models to solve this problem. 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 instructions in the pdf) and the performance in the test set in terms of Root Mean Squared Error (regression) and accuracy (classification). While you will need to write the full code to get to the results, only these instructions are required in the pdf. **(25 points)**

Part 2 (60%)

In Part 2, students are provided with a text classification dataset (named "bbc_news"). The dataset contains news articles split into five categories: tech, business, sport, politics and entertainment. Based on this dataset, students are asked to preprocess the data, select features and train and evaluate a machine learning model of their choice for classifying news articles. 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.

Note: Training, development and test sets are not provided. It is up to the student to decide the evaluation protocol and partition (e.g., cross-validation or pre-defining a training, development and test set). This should be explained in the report.

Deliverables for this part are the Python code including all steps and a report of up to 1200 words. The Python code should include the Python scripts and a small 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 dataset set as input, and output the results according to the chosen evaluation protocol. The code will consist of 25% of the marks for this part **(15 points)** and the report the remaining 75% **(45 points)**. 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 report should include:

1) Description of all steps taken in the process (preprocessing, choice of features, feature selection and training and testing of the model). This description should be

such that one could understand all steps without looking at the code **(15 points - 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 can be included to perform additional experiments. **(10 points - 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 (accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1) of the trained model in the dataset. **(10 points - 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. **(10 points - The depth and correctness of insights related to your deliverable will be assessed)**

The report may include tables and/or figures.

Extra credit (optional) - 10% extra marks in the second part (6 points): 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. text classification/categorization). 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.

Note: The maximum marks for the second part will be 60 in any case.

Learning Outcomes Assessed

This coursework covers the six 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 is based on the correctness of the answers, for which the process is also required. Full marks will be given for answers including both the correct answer and a correct justification or methodology.

- **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 broadly divided in:
Distinction (70-100%) - Full understanding of all the concepts, correct answers and methodology, well-documented and working code, accurate justification and description of all steps and critical analysis.
Merit (60-69%) - Good understanding of all the concepts, working code, justification and description of steps and analysis.
Pass (50-59%) - Few errors in questions and code, methodology with issues and not detailed description of steps and justification or with issues
Fail (0-50%) - Code with errors, flawed methodology, incorrect solutions, and no clear description of justification of steps.

Feedback and suggestion for future learning

Feedback on your coursework will address the above criteria. Feedback and marks will be returned between May 10th and May 14th 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 part of the module).