Cardiff School of Computer Science and Informatics

Coursework Assessment Pro-forma

Module Code: CMT307
Module Title: Applied Machine Learning
Lecturer: Yuhua Li, Yukun Lai
Assessment Title: Coursework 2 Machine learning project
Assessment Number: 2
Date Set: 28 February 2022
Submission Date and Time: 25 April 2022 at 9:30am
Return Date: 27 May 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.

This will apply to any of the two parts to be submitted as part of this assignment.

Your individual 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 submission consists of a **group submission** (note: group submission refers to the submission of the two files for Part 1 hereafter) and an **individual submission**. The group submission will be submitted in Learning Central by a nominated team member, and the individual submission will be submitted in Learning Central by individuals.

The **group submission** (from Part 1 of this assignment) consists of two files:

1. A single PDF file for your **group report** (up to 4500 words) on a specific machine learning project.
2. A zip file containing **all source code** of your group project.

All group members must have seen and agreed to the final version of the submission.

The **individual submission** (from Part 2 of this assignment) consists of a single PDF file for the self-reflection and peer assessment proforma.

You must submit to Learning Central the following files using a naming convention as required in the table below, replacing [group number] by your group number and [student

number] by your student ID number, e.g., *groupReport_G1.pdf* for the report of Group 1 and *peerAssessment_C1234567*.pdf for student C1234567.

Description		Type	Name
Part 1 (group submission)	**Compulsory**	One PDF (.pdf) file for group report (up to 4500 words)	groupReport_G[group number].pdf
	Compulsory	One ZIP (.zip) file containing the Python code	groupCode_G[group number].zip
Part 2 (individual submission)	**Compulsory**	One PDF (.pdf) file for the individual peer assessment proforma	peerAssessment_[student number].pdf
	Compulsory	One PDF (.pdf) file for Cover sheet (to be individually submitted with peer assessment proforma)	[student number].pdf

Note: This coursework consists of two parts: Part 1 is for group project code and group report. Part 2 is for individual work.

Part 1: **Group report** and **project code**. The deliverable includes a zip file containing all file(s) of the source code, and a single PDF file for the written report (up to 4500 words) describing solutions, design choices, evaluation and a reflection on the main challenges faced during development and insights gained throughout the process. Prior to handing in make sure all documentation has been collected. Additional supporting material, such as sources or data may also be submitted if appropriate along with the code zip file. Any code submitted will be run in Python 3 and must be submitted as stipulated in the instructions. Make sure the report clearly mentions your group number, the project title, the name of supervisor and a list of student ID numbers of all members of the group on the title page of your report.

Part 2: **Peer assessment**. Part 2 consists of a peer assessment proforma where students reflect on individual contributions and assign marks to other members in your group. Each individual will submit a cover sheet together with the peer assessment proforma in Learning Central by the deadline.

Any deviation from the submission instructions above (including the number and types of files submitted) will result in a reduction in marks for this assessment of 20%.

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

Assignment

In this coursework, students demonstrate their knowledge and skills with the topics covered in the module via a group project. This coursework consists of two parts: Part 1 is for group report and project code. Part 2 is for individual work.

Marks will be awarded to the individual student based on the quality of the group report, the project code and their contribution.

Part 1: Group report (80%)

In Part 1, students will be allocated in groups to design a machine learning project in one specific topic. The list of all topics along with their descriptions is available in Appendix A.

Each group is given a specific dataset and a supervisor. The task of each group consists of developing a whole machine learning pipeline that attempts to solve the task. The usage of neural networks as methods/baselines is not mandatory but will be positively assessed; the non-usage of neural methods should be properly justified.

Throughout the course the groups will have several milestones and should present their progress to their supervisor in each session. Finally, the group will write a report summarizing the steps followed and the main insights gained as part of the process.

As part of the group decisions, each student will be allocated to one of the following tasks:

- Descriptive analysis of the dataset + Error analysis
- Preprocessing + Literature review
- Implementation + Results

Each of these tasks will have a minimum of two students involved (except in exceptional cases when this is not possible), who will work together in the specific task and as part of the group. The structure of the report will be decided by the group members. In Appendix B, students can find some guidelines to write the report, including some of the common sections that groups may want to include in their report.

Note: These are just guidelines and students are not forced to follow this structure. New sections may be added or adjusted if necessary.

Each student will also be involved in all group activities/tasks and will be responsible for the well-functioning and coordination of the team members.

Deliverables

The deliverables for this part include **a report of no more than 4500 words** and a **zip file with all the Python code and a README file**. The group report must have the first page from Appendix C. The code and README should contain these contents:

(1) Code to get the statistics used to complement the descriptive analysis of the dataset.
(2) Code to train models on the training set and evaluate them on the test set. This code should also include all steps for preprocessing the original dataset, if it were necessary.
(3) A README file explaining how to run the code for each of the two parts.

The code will not be marked separately and will only be used as a complement to assess specific parts of the report.

Assessment

The final mark for this part (80% of the total marks) will result from the following items:

- Descriptive analysis of the dataset + result analysis (20%)
- Preprocessing + Literature review (20%)
- Implementation + Results (20%)
- Group report as a whole, including its coherence and structure (20%)

Note: Normally every member of the group will receive the same mark for the group report and project code, but in some cases marks might be weighted by the individual contribution in the project. This would be based on peer assessment for which instructions will be given in Part 2.

Credit will be awarded against the following criteria.

Criteria	Fail (0-49%)	Pass (50-59%)	Merit (60-69%)	Distinction(>=70%)
Descriptive analysis of the dataset + result analysis (20%)	No or arbitrary data exploration. No or little meaningful result analysis and discussion.	Suitable but limited data exploration. General result analysis and discussion.	Good data exploration but miss some insightful analysis. Good result analysis and discussion but lack of depth.	Thorough and insightful data exploration. Insightful result analysis and discussion.
Preprocessing + Literature review (20%)	No or very little data pre-processing and literature review.	Some necessary pre-processing and basic literature review are conducted.	Adequate pre-processing to prepare the data for model development. Adequate literature review.	Extensive pre-processing to deal with all aspects of non-ideal characteristics of the data with an aim to achieve a best classification performance. Extensive and insightful literature review.
Implementation + Results (20%)	Unsuitable ML method is chosen. The models are not correctly implemented and optimised. Little/improper performance evaluation.	The models are implemented but not properly/sufficiently trained. Performance evaluation using metrics without considering data characteristics.	The implemented models are properly trained and optimised Good performance evaluation with suitable metrics.	All models are excellently implemented, properly trained and systematically optimised. Comprehensive model evaluation for best results.
Group report as a whole, including its coherence and structure (20%)	The report is poorly presented. No or little meaningful discussion.	The report is acceptable in terms of technical contents and structure. General discussion and vague conclusion.	The report well written. Good discussion but lack of depth.	The report is professionally and cohesively presented. Insightful discussion and clear conclusions.

Part 2: Self-reflection and peer assessment (20%)

In Part 2, students are asked to do a self-reflection and peer assessment using the proforma in Appendix D. In the proforma, you must discuss your and each member's contribution to the group project and to the overall group report. You must show that you contributed to the group report and code. Discuss what tasks you have performed and provide evidence of your work (you may refer to the group report for the actual work/results). Discuss how you approached these tasks and how you interacted with other members, both in sharing your results and in organising the team's activities. Consider how well your existing skills were utilised and what new skills you have learned. Then reflect on your overall performance and role in the team and suggest what went well and what changes you will be making to improve (1) your performance in particular, and (2) the performance and results of methods and analyses performed as part of the project. You may also reflect on how your perspective and approach changed over time and adapted to improve your work.

Note: Please indicate the information about your group (group number, project name) in the proforma.

This part weighs 20% of the total marks.

Contribution of group members

This is a team project and this assignment is assessed as a team, apart from the individual mark for the supporting evidence for each team members' service. Each team member is expected to contribute to the project for the tasks that are agreed in the group.

You will also be asked to submit a peer assessment form. You will evaluate the contribution of each group member and your own contribution to the deliverables. Normally every member of the group will receive the same mark for the group components except in the case where a group member's contribution and/or quality of work falls significantly below that of the majority of the group, in which case the marks will be adjusted accordingly.

Please inform the module leader if there are any problems with any group members not engaging in group tasks or missing group meetings. The teaching team will check on the engagement of the group members in the contact session and review meetings. Students should therefore inform the module leader (and the other team members, if appropriate) if circumstances arise that are likely to affect their engagement with their work and/or attendance at weekly meetings with the rest of the team.

The teaching team will provide formative feedback during contact sessions. Your team should also meet regularly outside of these.

Learning Outcomes Assessed

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.
5. Critically appraise the ethical implications and societal risks associated with the deployment of machine learning methods.
6. Explain the nature, strengths and limitations of an implemented machine learning technique.

Criteria for assessment

Criteria for each individual part is provided separately as in previous sections. The final mark will be obtained from a weighted sum of the two parts: Part 1 - 80%; Part 2 - 20%.

The grade range is divided in:

Distinction	(70-100%)
Merit	(60-69%)
Pass	(50-59%)
Fail	(0-50)

Feedback

Feedback on your coursework will address the given criteria. Feedback and marks will be returned by 27 May via Learning Central. There will be opportunity for individual and group feedback during an agreed time.

Appendix A: CMT307 Group Projects

Note 1: Datasets are provided for all projects
Note 2: Not all datasets contain train/dev/test splits. It is up to the group members to decide a suitable split in those cases (or cross-validation).

1. **Emoji prediction**

The goal of this task is to predict an emoji (e.g. 😉) given a tweet. Dataset based on the SemEval 2018 task on emoji prediction (https://competitions.codalab.org/competitions/17344).
Direct link to download the dataset: emoji_prediction.zip

2. **Text categorization**

Text categorization (also referred to as text classification) consists of associating a document with a given topic (e.g. sports, politics, etc.). 20 Newsgroups dataset (http://qwone.com/~jason/20Newsgroups/).
Direct link for download the dataset (bydate version):
http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz

3. **Hate speech detection**

Given a tweet or a piece of comment, the task of hate speech consists of predicting whether the given text represents hate speech or not, and classify it accordingly. This task is based on the SemEval 2019 task on detection of hate speech against immigrants and women in Twitter (https://competitions.codalab.org/competitions/19935).
Direct link to download the dataset:https://drive.google.com/file/d/1Cn60H0klYNRNI_q5SeghzFlu6eMYmMOO/view?usp=sharing

4. **Opinion vs. factual news stories**

When building models to analyse news data at scale, it is important to distinguish between articles where the author is reporting "facts" (e.g. that something happened, or that a person said something) and those where the author is reporting their own opinion (e.g. that they think something will happen or that something happening is a good thing). Using a dataset of financial news articles, the task consists of building a classifier which labels an article as being fact or opinion. (Note: this dataset may contain noisy labels as they were obtained semi-automatically).
Direct link to download the dataset (dataset provided by AYLIEN):
https://drive.google.com/file/d/1Mqoh7gG-g3Sc3Zh6zEO8o2sKBpdSeFiJ/view?usp=sharing

5. **Urban Sound Classification**

Dataset from here: https://urbansounddataset.weebly.com/urbansound8k.html . The goal of this task is to classify different urban sounds (e.g. dog bark or street horn) into their correct classes.
Direct link to download the dataset:
https://drive.google.com/file/d/15fojQ3xKcPMLwIm6s8hoyM0xJjZ6SOZU/view?usp=sharing
Note: the size of this dataset is over 6GB (a portion of it may be used if hard to process). Therefore an important part of this project is the preprocessing and handling of the data. The following repository includes a machine learning model that can be used as a guide/reference at the beginning: https://github.com/AmritK10/Urban-Sound-Classification

6. **Fine-grained image classification**

The goal of this task is to develop an algorithm to learn to classify images containing objects of the same category (e.g. birds, dogs) into specific sub-categories, i.e. specific species.
Datasets are available:
Caltech-UCSD Birds-200-2011: 200 categories of birds
http://www.vision.caltech.edu/visipedia/CUB-200-2011.html
Stanford dogs: 120 categories of dogs
http://vision.stanford.edu/aditya86/ImageNetDogs/

7. **Detection and recognition of traffic signs**

Detection and recognition of traffic signs is an important task for autonomous driving. The task is to identify traffic signs in images and recognise them.
Dataset available at
https://sid.erda.dk/public/archives/daaeac0d7ce1152aea9b61d9f1e19370/published-archive.html (German Traffic Sign Recognition Benchmark)

8. **Object localisation**

It is straightforward for people to locate objects in an image, but can you develop a system to learn to do this? Given an image, the task is to find all the instances of relevant objects (as bounding boxes).

Dataset available at http://host.robots.ox.ac.uk/pascal/VOC/voc2012/index.html

It contains multiple datasets, and you have the flexibility to choose the ones that satisfy your project needs. So the following are just suggestions, and you may come up with your own approach, as long as it is reasonable and well justified.

For object detection/localisation, the most relevant dataset is the one used for

> Detection: Predicting the bounding box and label of each object from the twenty target classes in the test image.

You can see example images for this task at
http://host.robots.ox.ac.uk/pascal/VOC/voc2012/examples/index.html

9. Energy usage prediction

Prediction of building energy consumption is important to electricity distribution, management and environment. This project will develop a machine learning model to predict energy usage based on historic usage rates and observed weather. Dataset available at https://www.kaggle.com/c/ashrae-energy-prediction/overview.

10. Stock price prediction

Machine learning can be useful for algorithmic trading. It can help to develop a profitable trading strategy by predicting a stock price in the future based on historical daily OHLC price data. Data can be readily downloaded from Yahoo Finance using Python package yfinance: https://pypi.org/project/yfinance/ .

For groups working on image-related tasks (Projects 10-13):

Some tutorials for basic image processing with Keras/Tensorflow:

https://stackabuse.com/image-recognition-in-python-with-tensorflow-and-keras/
https://developer.ibm.com/technologies/artificial-intelligence/articles/image-recognition-challenge-with-tensorflow-and-keras-pt1
https://www.tensorflow.org/tutorials/keras/classification

Appendix B: Group report guidelines

In this document we detail some of the common sections involved in a machine learning project report. These sections are only presented as a guideline, but the report may have a different structure both in terms of sections and order.

1. **Introduction**

Summary of the task and main goals/contributions/insights of the project.

2. **Literature review / Related work**

Overview of the related work most connected to the methods and tasks of the projects. Explain the differences and the connection between works in the literature with respect to the employed method (e.g. advantages/disadvantages, ideas you exploited, etc.).
Tip: Google Scholar is a good resource to find relevant articles to any of the topics.

3. **Description of the task/dataset**

Description of the task and dataset, including relevant statistics of dataset splits.

4. **Methodology**

Description of the machine learning methods used in the project.

5. **Experimental setting**

Description of the specific details of the evaluation (e.g. parameter tuning, usage of the development set).

6. **Results**

Final results of the experiments, including baselines and table/s with precision/recall/accuracy/f1, etc.

7. **Analysis**

Analysis of the results, error analysis (investigate the type of error the system makes, etc.).

8. **Conclusion and future work**

Summary of the main conclusions and takeaways from the experiments. Explain ways to investigate or improve the method in the future.

Appendix C: Group report front page

You will include the following information on the first page of your group report.

CMT307 Coursework 2 Group Project

Group number	
Project title	
Supervisor	
	[list student IDs of all group members here]

Appendix D: Peer Assessment and Self-reflection

Notes for completing the peer assessment form

1. Each row of the form is for one member
2. In the first column, add full names of all other group members (excluding yourself) in alphabetical order by surname.
3. In the second column, evaluate contributions of each member of the group to the project (up to 150 words for each member). This evaluation should be based on evidences, such as tasks completed, interaction with other members, participation in the group activities, team working skills, etc.
4. In the third column, fairly and honestly allocate marks to each member based on your evaluation in the second column. **You have a total of 20 marks to be allocated to all other members**. For example, for a group with 4 other members, an allocation of the 20 marks to those 4 members could be 7, 3, 10, 0, respectively. A 0 mark indicates that member has done nothing to the group project. If contributions are felt to be fair, you can score the same marks to all other students in your group.
5. Total peer assessment (TPA) marks of a member will be obtained by aggregating marks of all peer assessment forms in the group, and a member can have a maximum peer assessment mark of 20 (even if her/his aggregated TPA mark is greater than 20).
6. The final peer assessment (FPA) marks of a member will be weighted by the total group mark from Part 1. For example, if the total group mark is 57 out of 80 marks and a member called Boris has a TPA mark of 16, then Boris's FPA will be (57/80) x 16 = 11 (round to nearest integer). Finally Boris's overall mark for this coursework will be 57+11=68%.
7. Normally every member of the group will receive the same mark for the group components except in the case where a group member's contribution and/or quality of work falls significantly below that of the majority of the group, in which case his/her marks will be adjusted down accordingly.
8. If you failed to submit Peer Assessment and Self-reflection Proforma or the submitted proforma was not usable (e.g., clearly unfair mark allocation), your FPA would be zero.

Peer Assessment and Self-reflection Proforma

Student ID:
Group number:
Project title:
Supervisor:

Peer assessment

Member name	Contribution and justification	Marks

Self-reflection

[write your self-reflection here, up to 300 words. Your self-reflection will be used to cross-check if you have been fairly assessed by your peers]