Hong Kong University of Science and Technology
COMP 4211: Machine Learning
Fall 2022

Project
Due: 25 November 2022, Friday, 11:59pm

1 Preamble

Unlike the programming assignments and problem set, this individual project is intended to be more open-ended like many other course projects or final year projects. As such, much room is left for you to explore. Consequently, there will only be grading guidelines but not a detailed marking scheme.

Since the project is worth 20% of the final course grade, its workload is expected to be about two times the workload of the first programming assignment. This comparison is by no means exact but serves to give you some ideas about the expected workload.

You have a choice between two types of project:

- A *software project* that involves hands-on experience of solving a machine learning task
- A *presentation project* that involves learning and presenting an extended topic beyond the topics covered in class.

These two types of project are described in detail in two separate sections below.

2 Academic Integrity

Please refer to the regulations for student conduct and academic integrity on this webpage: `https://registry.hkust.edu.hk/resource-library/academic-standards`.

3 Software Project

3.1 Objective

The objective of this project is to gain and practise the hands-on skills needed for solving more realistic machine learning tasks through pursuing a proposed study using one of the datasets provided.

3.2 Datasets

You are asked to choose a dataset from the following list and propose a machine learning project to work on using the chosen dataset (click on the name of a dataset to access the corresponding web page):

- Credit Card Classification
- FaceMask Dataset
- Top 10000 Movies Based On Ratings
- Traffic Signs 1 Million Images for Classification
- UK MET Office Weather Data
- Wonders of the World Image Dataset

For inspiration, you may take a look at the Kaggle website (`https://www.kaggle.com`) and other online resources. Sometimes a data set originally used for one task may be used in a very different way for another task that has not been studied by others before. This will make your project novel.

Depending on the machine learning task(s) you propose to work on, you may use only a subset of a dataset above (either a subset of the features or a subset of the instances). However, the subset used should not be too small to make the project trivial.

Note that like many real-world datasets, the datasets above may contain missing values for some instances. Excluding those instances may not be the best treatment. Instead, you are recommended to explore the use of imputation methods for estimating and filling in the missing values before use.

3.3 Machine Learning Models and Computing Facilities

The machine learning tasks based on the datasets above will more likely involve supervised and unsupervised learning techniques than reinforcement learning techniques.

In case you plan to use some more advanced machine learning methods not covered in the course for your project, please make sure that you also include the related methods covered in the course as baselines for comparison. Among other things, including the baselines will help to justify using more advanced methods if they can indeed outperform the simpler baselines significantly.

You should estimate the computational demand of your proposed project. In case the computing resources provided by the basic Colab plan cannot meet your need, you may consider a short-

term subscription of the Colab Pro/Pro+ plan or other cloud computing services.

3.4 Assessment Components and Submission

There are two assessment components:

- Project report
- Source code

Note that this project should not be used for earning credits in a different course to avoid double-dipping.

3.4.1 Project Report

The report should cover at least the following aspects of the project:

- Project title
- Student information with full name, student ID, and HKUST email address
- Description of the dataset and any preprocessing
- Description of the machine learning task(s) performed on the dataset
- Machine learning methods used for solving the task(s)
- Experiments and results

3.4.2 Source Code

All the source code that you have written for this project should be submitted for grading. In case your code is modified from another source, you are expected to acknowledge it clearly in your report. Failure to do so is considered plagiarism.

Data files should not be submitted to keep the submission file size small.

3.4.3 Submission

Assignment submission should only be done electronically in the Canvas course site.

Your submission should contain two files: report (`report.pdf`) and compressed source code (`code.zip`). When multiple versions with the same filename are submitted, only the latest version according to the timestamp will be used for grading. Files not adhering to the naming convention above will be ignored.

3.5 Grading Guidelines

This project will be counted towards 20% of your final course grade. The breakdown is as follows:

- Description of the dataset and any preprocessing [**15 points**]
- Description of the machine learning task(s) performed on the dataset [**10 points**]
- Description of the hardware and software computing environment, machine learning methods, and parameter settings [**10 points**]
- Good programming practices in source code [**10 points**]
- Description of the experiments [**30 points**]
- Visualization and discussion of the results obtained [**25 points**]

Grading will be based on rubrics with five levels of achievement (excellent, good, satisfactory, unsatisfactory, poor) for each of the items above.

An important general criterion is clarity, to the extent that others can replicate your experiments based on the information provided in the report.

Please note again that this project should not be used for another course.

Late submission will be accepted but with penalty. The late penalty is deduction of one point (out of a maximum of 100 points) for every minute late after 11:59pm. Being late for a fraction of a minute is considered a full minute. For example, two points will be deducted if the submission time is 00:00:34.

In case the bonus part you did for either one of the programming assignments allows you to submit this project late for up to 24 hours without grade penalty, the late submission policy above will no longer be applicable. In other words, no submission 24 hours after the original deadline will be accepted. Please also note that late submission should also be done in Canvas.

4 Presentation Project

4.1 Objective

Since this is an introductory machine learning course, the topics covered are mostly elementary though they are fundamental to understanding more advanced topics. To strike a balance between learning elementary and more advanced/recent machine learning topics, you are asked to self-learn an extended topic which is related in some ways to one or more of the basic topics covered in the course. The objective of this project is *learning to learn*, or called *meta-learning* in machine learning. Through the course, you do not just learn the topics covered but also the learning skills to learn new topics yourself. This is arguably a more important reason for taking the course.

4.2 Extended Topics

You are asked to choose an extended topic from the following list of topics which are ordered alphabetically:

- Anomaly detection
- Deep reinforcement learning
- Generative adversarial networks
- Gradient boosting trees
- Graph convolutional networks
- Neural architecture search
- Self-supervised learning
- Transformer networks

Each of these extended topics is related in some ways to the topics covered. No other topics outside this list will be allowed.

The easiest way for you to learn these topics yourself would be to look for relevant materials online. If you happen to know a topic, I suggest that you do not choose it but a topic that is new to you. After all, the whole purpose is for you to learn something new to expand your machine learning knowledge.

4.3 Assessment Components and Submission

There are two assessment components:

- Presentation slides
- Video presentation

Note that this project should not be used for earning credits in a different course to avoid double-dipping.

4.3.1 Video Presentation and Slides

You are asked to prepare an oral presentation of your chosen topic in the form of a video that is about 30 minutes long. Your face should be visible throughout the whole presentation. As a general guideline, you are advised not to have more than one slide per minute, i.e., no more than 30 slides in total, excluding the slides for the title and section headings. The target audience of your presentation should be students immediately after taking COMP 4211, i.e., students who have learned the basic machine learning topics. In other words, in your presentation, you

may refer to concepts and techniques covered in the course without having to review them again.

Here are some guidelines which may help you prepare your slides:

- Motivate the topic using intuitive examples
- Discuss how the chosen topic is related to some basic topics covered in the course
- Choose a representative model or algorithm for the topic to present
- Discuss the strengths and weaknesses of the model or algorithm
- Briefly mention or list some variants without going into details
- List the references that you have read to help you prepare the slides

Please note that this presentation flow is by no means the only possibility. It is just a suggestion for your reference. Note that presenting a broad topic at too high a level is generally discouraged. Rather, it is better to spend more time on a specific model or algorithm although it would be useful to mention about variants of your choice later in the presentation.

If some of your slides are modified from another source, you are expected to acknowledge it clearly at the end of your presentation. Failure to do so is considered plagiarism.

When your video is ready, upload it to YouTube as an 'unlisted' (not 'private' or 'public') video and include its hyperlink clearly in the first slide. The video should be ready by the time you submit the slides and no change should be made to it after the deadline. Note that it takes time to upload a video of 30 minutes long to YouTube, so you should manage your time well instead of rushing towards the end. No request for extension of deadline will be entertained for failing to complete video upload by the deadline.

4.3.2 Submission

Assignment submission should only be done electronically in the Canvas course site.

Your submission should be named `slides.pdf` or `slides.pptx`. When multiple versions with the same filename are submitted, only the latest version according to the timestamp will be used for grading. Files not adhering to the naming convention above will be ignored.

4.4 Grading Guidelines

This project will be counted towards 20% of your final course grade. The breakdown is as follows:

- Good time management in presentation [**10 points**]
- Clarity of content in slides [**20 points**]
- Clarity of oral presentation [**20 points**]
- Effective use of examples and visual aids [**20 points**]
- Appropriate content for illustrating the topic [**30 points**]

Grading will be based on rubrics with five levels of achievement (excellent, good, satisfactory, unsatisfactory, poor) for each of the items above.

Please note again that this project should not be used for another course.

Late submission will be accepted but with penalty. The late penalty is deduction of one point (out of a maximum of 100 points) for every minute late after 11:59pm. Being late for a fraction of a minute is considered a full minute. For example, two points will be deducted if the submission

time is 00:00:34.

In case the bonus part you did for either one of the programming assignments allows you to submit this project late for up to 24 hours without grade penalty, the late submission policy above will no longer be applicable. In other words, no submission 24 hours after the original deadline will be accepted. Please also note that late submission should also be done in Canvas.