**Hong Kong University of Science and Technology**
**COMP 4211: Machine Learning**
**Spring 2021**

**Project 1**
Due: 27 April 2021, Tuesday, 11:59pm

# 1 Preamble

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.

Unlike the programming assignments, this 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.

The project is expected to be substantial and hence will be a group project, with each project group consisting of two students. Since the project is worth 30% of the final course grade, its workload per group member is expected to be about the total workload of Programming Assignments 2 and 3. Consequently, as a two-person group project, its total workload is expected to be about two times the total workload of the two programming assignments. This comparison is by no means exact but serves to give you some ideas about the expected workload.

# 2 Datasets

You are asked to choose a dataset from the following list and propose a machine learning task to work on using the chosen dataset:

- **COVID-19 Dataset**
  `https://github.com/GoogleCloudPlatform/covid-19-open-data`
- **Face Mask Detection**
  `https://www.kaggle.com/andrewmvd/face-mask-detection`
- **Influenza Dataset**
  `https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`
- **MIND: Microsoft News Recommendation Dataset**
  `https://www.kaggle.com/arashnic/mind-news-dataset`
- **Text Classification Dataset**
  `https://www.kaggle.com/kaushiksuresh147/the-social-dilemma-tweets`

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.

Depending on the machine learning task you propose to work on, it may use only a subset of a dataset above (either a subset of the features or a subset of the instances).

Note that like many real-world datasets, the datasets above may have 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 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.

Depending on the computational demand of your project, you are recommended to use either the GPU servers provided by the Department of Computer Science and Engineering (`https://cssystem.cse.ust.hk/Facilities/ug_cluster/gpu.html`) or Google Colab (`https://colab.research.google.com`).

## 4 Project Group Formation

You may form your own project group. If you need help, we can also form a group for you. More information about this will be announced in due course separately.

## 5 Assessment Components and Submission

There are three assessment components:

- Project report
- Source code
- Video presentation

Only one member of each project group will submit all the assessment components on behalf of the group, but the names of both members should be listed clearly in all the assessment components.

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

### 5.1 Project Report

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

- Project title
- Students with full names, student IDs, and HKUST email addresses
- Description of the dataset and any preprocessing
- Description of the machine learning task performed on the dataset
- Machine learning methods used for solving the task
- Experiments and results
- Division of labor
- Hyperlink to YouTube video

You should state clearly the division of labor between the two group members by listing the main duties and contributions of each member. The overall contribution of each member to the project should also be given in percentage (e.g., 55% by A and 45% by B). You should try your best to ensure that the workload is shared evenly (i.e., 50% each). Grading will be done individually according to the workload distribution.

### 5.2 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.

### 5.3 Video Presentation

You are required to prepare an oral presentation of your project in the form of a video using informative slides that summarize the key aspects of the project. The video should be no longer than 15 minutes.

Note that the video is not a movie for entertainment or an advertisement. Instead, it is for a technical presentation of your project. You should pay attention to both the technical content and the quality of your video.

When your video is ready, upload it to YouTube as an 'unlisted' (not 'private' or 'public') video and include its hyperlink in your report. The video should be ready by the time you submit the report and no change should be made to it after the deadline.

### 5.4 Submission

Like the programming assignments, submission of the project report and source code should be done electronically using the Course Assignment Submission System (CASS):

`https://cssystem.cse.ust.hk/UGuides/cass/student.html`

Your submission should contain two files: report (`<StudentID>_report.pdf`) and compressed source code (`<StudentID>_code.zip` or `<StudentID>_code.rar`).

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.

# 6 Grading Guidelines

This project will be counted towards 30% of your final course grade. The breakdown is as follows:

- Description of the dataset and any preprocessing [**10 points**]
- Description of the machine learning task performed on the dataset [**10 points**]
- Description of the hardware and software computing environment, machine learning methods, and parameter settings [**10 points**]
- Source code adhering to good programming practices [**10 points**]
- Description of the experiments [**25 points**]
- Visualization and discussion of the results obtained [**20 points**]
- Video presentation (expected to cover all of the above) [**15 points**]

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. At most one NQA coupon may be used to entitle you to submit this project late for one day without grade penalty.

# 7 Academic Integrity

Please refer to the regulations for student conduct and academic integrity on this webpage: `https://acadreg.ust.hk/generalreg`.