Hong Kong University of Science and Technology
COMP 4211: Machine Learning
Spring 2020

Project
Due: 19 May 2020, 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 participating in an ongoing competition or pursuing a proposed study using a provided dataset. The anticipated difficulty of each competition or dataset is indicated as: [*] for easy; [**] for medium; [***] for challenging.

Unlike the two 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 80-90% ($\approx 30/35$) of the total workload of the two programming assignments which are worth 15% and 20%, respectively. Consequently, as a two-person group project, its total workload is expected to be about 1.7 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 Kaggle Competitions

You may choose to participate in one of these three ongoing Kaggle competitions:

- [**] **Plant Pathology 2020 - FGVC7** (end date: 11 May 2020)
  `https://www.kaggle.com/c/plant-pathology-2020-fgvc7`
- [**] **M5 Forecasting - Accuracy** (end date: 30 June 2020)
  `https://www.kaggle.com/c/m5-forecasting-accuracy/overview/timeline`
- [***] **Abstraction and Reasoning Challenge** (end date: 27 May 2020)
  `https://www.kaggle.com/c/abstraction-and-reasoning-challenge`

If you wish, you may enroll in this online course titled "How to Win a Data Science Competition: Learn from Top Kagglers":

`https://www.coursera.org/learn/competitive-data-science`

You should provide evidence to show that the team members of your group for this course project are exactly the same as those in the team joining the Kaggle competition. The Kaggle team should not contain additional members.

3 Kaggle Datasets

Alternatively, you may choose to propose and work on a machine learning task using one of the following Kaggle datasets:

- [*] **200+ Financial Indicators of US Stocks (2014-2018)**
  `https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018`
- [*] **Air Tickets between Shanghai and Beijing**
  `https://www.kaggle.com/lpisallerl/air-tickets-between-shanghai-and-beijing`
- [*] **Income Classification**
  `https://www.kaggle.com/lodetomasi1995/income-classification`
- [*] **Logistic Regression to Predict Heart Disease**
  `https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression`
- [*] **Mobile Price Classification**
  `https://www.kaggle.com/iabhishekofficial/mobile-price-classification`
- [*] **Password Strength Classifier Dataset**
  `https://www.kaggle.com/bhavikbb/password-strength-classifier-dataset`
- [*] **Predicting Student Grades**
  `https://www.kaggle.com/daviddraper1518/predicting-student-grades`
- [*] **Used Cars Catalog**
  `https://www.kaggle.com/lepchenkov/usedcarscatalog`
- [**] **Best Artworks of All Time**
  `https://www.kaggle.com/ikarus777/best-artworks-of-all-time`
- [**] **Novel Corona Virus 2019 Dataset**
  `https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`
- [**] **Predict Click Through Rate (CTR) for a Website**
  `https://www.kaggle.com/animeshgoyal9/predict-click-through-rate-ctr-for-a-website`
- [**] **Real and Fake News Dataset**
  `https://www.kaggle.com/nopdev/real-and-fake-news-dataset`
- [**] **Spam and Ham Dataset**
  `https://www.kaggle.com/rushirdx/spam-and-ham-dataset`

For inspiration, you may take a look at the Kaggle website (`https://www.kaggle.com/`) and many 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 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.

4 Other Datasets

Instead of using a Kaggle dataset, you may also choose one from the following:

- [***] **MVTec Anomaly Detection (MVTec AD) Dataset**
  `https://www.cse.ust.hk/faculty/dyyeung/mvtec_ad.html`
- [***] **Omniglot Handwritten Character Dataset**
  `https://github.com/brendenlake/omniglot`
- [***] **CORe50 Continual Object Recognition Dataset**
  `https://vlomonaco.github.io/core50`

5 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 may use the GPU servers provided by the Department of Computer Science and Engineering (`https://cssystem.cse.ust.hk/Facilities/ug_cluster/gpu.html`) or the Colaboratory, or called Colab, provided by Google (`https://colab.research.google.com/`).

6 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 cannot be used for earning credits in a different course.

6.1 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. You should try your best to ensure that the workload is shared in a fair manner.

6.2 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.

6.3 Video Presentation

You are required to prepare an oral presentation of your project in the form of a video. 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.

6.4 Submission

The project report and source code will be due on **19 May 2019, Tuesday, 11:59pm**.

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.

7 Grading Guidelines

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

- Difficulty level of the project [**10 points**]
- 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 [**25 points**]

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

If you join a Kaggle competition, you are recommended to include a screenshot of the leaderboard as proof.

The video presentation will also be assessed according to the corresponding aspects above.

Please note again that this project cannot 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.

8 Academic Integrity

Please read carefully the relevant web pages linked from the course website.