RMIT UNIVERSITY

Machine Learning Assignment 2

COSC2673/COSC2793 Semester 1 2023
Machine Learning & Computational Machine Learning

Weight: 50% of the final course mark
Type: group of 2
Due Date: 9:00 am, 16th of May 2023 (Week 11)
Learning Outcomes: This assignment contributes to CLOs: 1,2, 3, 4,6
Note: Marks will be awarded for meeting requirements as close as possible. Clarifications/Updates may be made via announcements / relevant discussion forums, you are required to check them regularly.

1   Overview

In this assignment you will design and create an end-to-end machine learning system for a real-world problem. This assignment is designed for you to apply and practice skills of critical analysis and evaluation to circumstances similar to those found in real-world problems. This is a group project (maximum of two students). In this assignment you will:

- Design and create an end-to-end machine learning system.

- Apply multiple algorithms to a real-world machine learning problem.

- Analyse and evaluate the output of the algorithms.

- Research into extending techniques that are taught in class.

- Provide an ultimate judgement of the final trained model(s) that you would use in a real-world setting.

This assignment has the following deliverables:

1. A report (of no more than 3 pages, plus up to 2 pages for appendices) critically analysing your approach and ultimate judgement.

2. An independent evaluation of your model and ultimate judgement (to be included in the report).

3. Your Python scripts, Jupyter notebooks, and software used to build your learning system and produce the models and results.

To complete this assignment, you will require skills and knowledge from lecture and lab material for Weeks 1 to 12 (inclusive). You may find that you will be unable to complete some of the activities until you have completed the relevant lab work. However, you will be able to commence work on some sections. Thus, do the work you can initially, and continue to build in new features as you learn the relevant skills. *A machine learning model cannot be developed within a day or two. Therefore, start early.*

1.1 Learning Outcomes

This assignment contributes to the following course CLOs:

- **CLO 1:** Understand the fundamental concepts and algorithms of machine learning and applications.

- **CLO 2:** Understand a range of machine learning methods and the kinds of problem to which they are suited.

- **CLO 3:** Set up a machine learning configuration, including processing data and performing feature engineering, for a range of applications.

- **CLO 4:** Apply machine learning software and toolkits for diverse applications.

- **CLO 6:** understand the ethical considerations involved in the application of machine learning.

1.2 Group Work

The group work must be completed in groups of 2.

1. You may form groups with any student in the course.

2. We strongly recommend that you form groups from within your labs, because:

   - Your tutor will help you form groups, but only within your lab.
   - You will have plenty of opportunity to discuss your group's progress and get help from your tutor during the rest of the course. It will be extremely helpful for your whole group to be present, but this can't happen if you have group members outside the lab.

Groups for Assignment 2 must be **registered with your tutor by week 9 lab**[1]. Your tutor "register" your group on Canvas during the lab when you request it. you should not Email the tutor for registration as they will not consider it. If you are unable to find a group, discuss this with your tutor as soon as possible. *If you do not register the group by week 9, the course coordinator will randomly assign you to a group*.

If at any point you have problems working with your group, inform your tutor immediately, so that issues may be resolved. This is especially important with the online delivery of the course. We will do our best to help manage group issues, so that everybody receives a fair grade for their contributions. To help with managing your group work we will be requiring your group to use particular tools. These are detailed in Section 5.

---

[1]If your group spans multiple labs, have one of your tutors register the group.

2 Assessment details

2.1 Task

Using machine learning in real-world settings involves more than just running a data set through a particular algorithm. In this assignment, you will design, analyse and evaluate a complete machine learning system.

The key aspect of this assignment is the ***design*** , ***analysis***, and ***evaluation*** of your methodology, investigation, and results. This assignment focuses on both the accuracy of your model, *and* your understanding of your approach and model.

For this assignment you have a choice of your project. You may select this project from the list in Section 3, or you may negotiate a project with the course coordinator. Regardless of the problem you choose, you must conduct the following tasks:

1. You need to come up with an ***approach*** , where each element of the system is *justified* using data analysis, performance analysis and/or knowledge from relevant literature.

2. Investigate various Machine Learning solutions to the problem.

3. Make an ultimate judgement.

4. Independent evaluate of your ultimate judgement.

5. Present your design, investigation, evaluation and findings in report format.

2.2 Investigation

Your investigation will require you to design, use, analyse and evaluate an end-to-end machine learning system. You should consider a variety of techniques that have been discussed in class, and techniques you have researched. Your end-to-end system may consist of elements such as:

- A well justified evaluation framework.

- Pre-processing the data set to make it suitable for providing to various machine learning algorithms.

- Carefully selected and justified baseline model(s).

- Hyper-parameter setting and tuning to refine the model.

- Analysing model and outputs & interpreting the trained models.

Each project features many of these above aspects. Each project also has unique aspects which cover a sample of issues from across machine learning. Additionally, each project has unique mandatory requirement(s), detailed for each project. The details of each project are listed in Section 3.

The details in this spec are the **minimum requirements**. *A thorough investigation must consider more than the minimum to receive high grades.*

2.3 Ultimate Judgement

You must make an ***ultimate judgement*** of the "best" model that you would use and recommend for your particular project. It is up to you to determine the criteria by which you evaluate your model and determine what is means to be "the best model".

*For higher grades you must use techniques that goes **beyond simple performance metric analysis** when making the ultimate judgment.*

2.3.1 Independent Evaluation of your Ultimate Judgement

You may conduct an independent evaluation of your ultimate judgement. This can be conducted where possible by:

1. Using data collected completely outside of the scope of your original training and evaluation.

2. Comparing your performance to other works in literature that use same/similar

data. This evaluation simulates how your ultimate judgement would perform if it were

deployed
in a real-world setting, where you are unable to re-train and adjust the model.

Each project describes a method to conduct this independent evaluation, which you may extend.

2.4 Approach, Critical Analysis & Report

You must compile a report analysing the approach you have taken in your investigation. Your report:

- Must be no longer that 3 pages of text

- May contain ***an additional*** 2 pages for appendices

- Use a *single-column* layout with no less than size 11pt font

- The appendices may only contain citations, figures, diagrams, or data tables that provide evidence to support the statements in your report.

- Include the name(s) and student ID's of the student(s) who wrote the report.

Any over length content, or content outside of these requirements will not be marked. For example, if you report is too long, ONLY the first 3 pages pages of text will be read and marked.

In this report you should analyse elements such as:

- Machine learning algorithms that you considered

- Why you selected these approaches

- Evaluations of the performance of trained model(s)

- Your ultimate judgement with supporting analysis and evidence

This will allow us to understand your rationale. We encourage you to explore this problem and not just focus on maximising a single performance metric. By the end of your report, we should be convinced that of your ultimate judgement and that you have considered all reasonable aspects in investigating your chosen problem.

The key aspect of this assignment isn't your code or model, but the thought process behind your work.

Remember that good analysis provides factual statements, evidence and justifications for conclusions that you draw. A statements such as:

"I did <xyz> because I felt that it was good"

is not analysis. This is an unjustified opinion. Instead, you should aim for statements such as:

"I did <xyz> because it is more efficient. It is more efficient because . . . "

3 Projects

For your project:

1. You may choose from one of the below 2 suggested projects. Each project has unique aspects and will allow you to explore different aspects of the ML field.

2. You may negotiate your own project. Note the special requirements and timeline for negotiating a project.

3.1 Suggested Projects

Each project has different requirements, so ensure you are aware of these differences.

Project 1: Classify Images of Cancer

Assume you are a team of machine learning engineers working for a biomedical startup company. Your task is to develop a machine learning system that can classify histopathelogy images of cancerous cells. A basic description of histopathelogy images can be found here.

You will be using a modified version of the "*CRCHistoPhenotypes*" dataset for this task. The data set for you to use in this assignment has been specifically prepared for you, and is provided on Canvas. The dataset consists of 27x27 RGB images of cells from 99 different patients and you are expected to use the dataset to perform two tasks:

- Classify images according to whether given cell image represents cancerous cells or not (isCancerous).

- Classify images according to cell-type, such as: fibroblast, inflammatory, epithelial or others.

The correct classification of the images is given by the "data_labels_mainData.csv" and "data_labels_extraData.csv". For the first 60 patients, the medical experts have provided labels isCancerous and cell-type. However, for the remaining 39 patients, the medical experts have only provided labels for isCancerous.
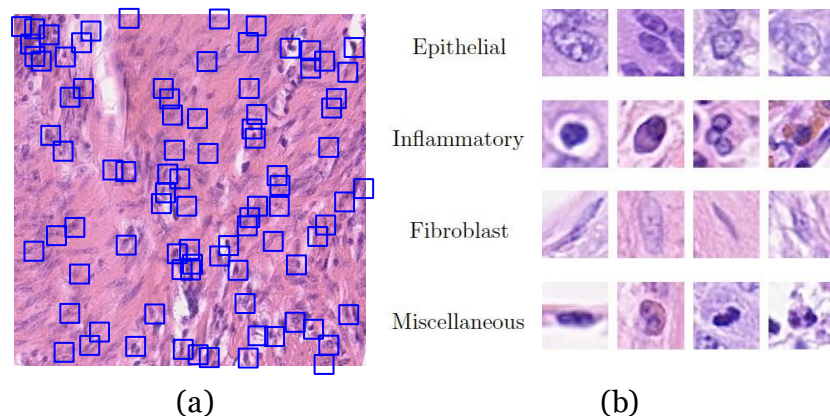
$$(a) \qquad\qquad (b)$$

Figure 1: (a) Example histopathelogy image of the with individual cells marked with "blue" rectangles of size 27x27. (b) Example of different cell types present in histopathpelogy images of the.

Requirements

- You must investigate at least one supervised machine learning algorithms *for each* of the two categories (Tasks). That is, you must build at least one model capable of predicting isCancerous, and at least one model capable of classifying the cell-type. One model with post processing label is not considered adequate.

- You are not required to use separate type(s) of machine learning algorithms, however, a thorough instigations should consider different types of algorithms.

- You are required to *fully train* your own algorithms. You may not use pre-trained systems which are trained on other datasets (not given to you as part of this assignment).

- For higher grades (HD/DI) you must explore how the data in "Data_labels_extraData.csv" can improve the cell-type classification model and use it accordingly.

- Your final report must conduct an analysis and comparison between classifying the two categories.

Independent Evaluation

- As you don't have access to histopathelogy images outside the dataset, the independent evaluation of this project is different to the other projects. A fundamental feature of the scientific process is reproducing existing work, and comparing new results against exist work.

- The original data set was published publicly as part of the following paper (available via RMIT library):

    K. Sirinukunwattana, S. E. A. Raza, Y. Tsang, D. R. J. Snead, I. A. Cree and N. M. Rajpoot,*"Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images,"* in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1196-1206, May 2016, doi: 10.1109/TMI.2016.2525803.

Since publication, a number of papers, and online resources, have been published that use this data set.

- Your independent evaluation is to *research* a number of these published works. Then you must *compare and contrast* your results to those other works.

Project 2: Packet Scheduling in Routers

Assume you are a team of machine learning engineers working for a technology company that is manufacturing routers. Your team is tasked with developing a *reinforcement learning* based scheduling algorithm for a router that is being designed currently by your company.

A major challenge in designing packet scheduling mechanisms is, how to support multiple classes of traffic with different quality of service (QoS) requirements. Services such as video, voice and data each have different service requirements in terms of delay and delay variations. A common approach used in routers is to provide separate queues for each class of traffic with guarantees on expected delay (will also have a dedicated queue for "best-effort" traffic where no guarantees need to be given). A classifier is used to assign incoming traffic to each queue. A scheduler will then select a packet from a queue at each time-slot to be transmitted via the outgoing interface. An example router interface is shown in Figure 2.
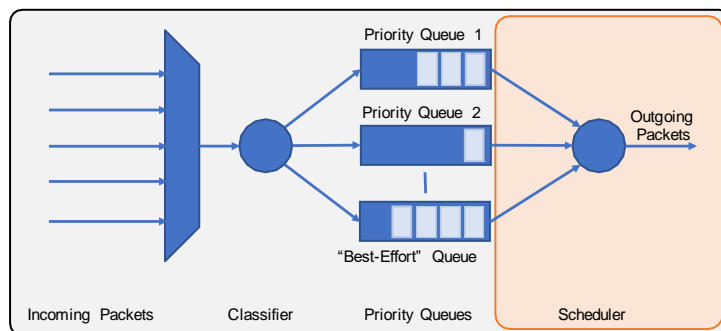

Figure 2: An example router interface

Your task is to implement a suitable scheduling policy (in scheduler) that satisfies the QoS constraints of each traffic class while still ensuring reasonable performance for "best-effort" traffic (minimise latency - mean delay - of the "best-effort" queue. You should consider two different scenarios:

Scenario 1: For each timeslot the scheduler selects a queue and transmits the first packet in the queue (ignore time taken to switch queues).

Scenario 2: For each timeslot the scheduler selects to send a packet from the current queue or switch the queue (i.e. one timeslot is taken to switch queues).

To simplify the problem, you may **assume** that:

1. There are only three traffic classes (or 3 queues): Video (Priority Queue 1), Voice (Priority Queue 2) and data (Best-effort Queue).

2. All packets in the system has constant fixed length.

Queue	Arrival Rate [Packets per timeslot]	Mean Delay Requirement [timeslots]
Priority Queue 1	0.3	6
Priority Queue 2	0.25	4
Best-effort	0.4	best-effort

Table 1: Arrival statistics and mean delay requirements for initial testing

3. A fixed length timeslot is required to transmit a packet, and at most one packet can be serviced at each timeslot.

4. For initial testing, you may use the arrival statistics and mean delay requirements listed in Table 1. A through investigation will explore performance under varying conditions.

Requirements

- Need to identify how best to setup the reinforcement learning problem(s). i.e. choosing a good representation for the state, a good reward function, and the independent performance measure used to compare against baseline schedulers.

- This project will **require** you to build a simple simulation to conduct the reinforcement learning. It is advisable to build a simulation system that has configurable parameters for the flow of packets.

- The simulation **should be** based on the *OpenAI gym*. You need to code your own environment.

- The simulated routing interface *must include* at least two priority queues and one "best-effort" queue.

- Due to above work, you only need to use a single form of reinforcement learning to train an optimal policy.

- For higher grades (HD/DI), you must test your learned system under varying arrival rate conditions that are reasonable and discuss the limitations.

- For higher grades (HD/DI), you must test your learned system under both scenarios mentioned above.

Independent Evaluation

- You will need to compare the performance of the system under varying conditions to fixed scheduling policies like:

    1. First-in-First-out (FIFO): The scheduler will select the packet that arrived the earliest irrespective of the queue type to transmit next.

    2. Earliest deadline first (EDF): Scheduler selects the packet with the smallest time difference between its delay requirement and its accumulated delay to transmit next.

    3. Sequential Priority (SP): selects a packet from the highest priority queue that has packets waiting to transmit next.

    4. Any other scheduling policy that you think is suitable for the task.

3.2 Negotiated Project

You may propose and negotiate a project and machine learning problem to investigate, with the course coordinator. This project must meet a number of constraints:

- The project must be of a suitable complexity and challenge that is similar to the suggested projects. As part of the negotiation, the scope and deliverables of the project will be set.

- If the project is using an existing data set, the problem should be phrased in a manner that can be solved by multiple machine learning methods, of which at least two methods will be investigated.

- If the project requires a data set to be generated, devised, or collected, this collection should require sufficient effort. In these cases, especially for reinforcement learning tasks, only one machine learning method may need to be investigated.

- The proposed project must be independent of previously or concurrently assessed work. You may not conduct a project if you have already been assessed on the work, or are concurrently being assessed on the work.

- *You may only conduct a negotiated project if you are working in a group.*

In general, negotiations will take place via email, during consultation hours, or by appointment. To start negotiations you should fill in the "Project Negotiation Template" on canvas and forward it to the course coordinator. Please note, that the course coordinator is not available outside of business hours.

All negotiated projects must be finalised by no later than **5pm Friday (Week 9)**. This is the absolute deadline. If you wish to conduct a negotiated project, **begin the negotiation process early** . A negotiated project may be denied *before* the deadline if there is insufficient time for the negotiation process.

4 Submission

You have to submit all the relevant material as listed below via Canvas.

1. A report (of no more than 3 pages , plus up to 2 pages for appendices) critically analysing your approach and ultimate judgement

2. An independent evaluation of your model and ultimate judgement (included in the report).

3. Your Python scripts, Jupyter notebooks, and software used to build your learning system and produce the models and results.

The submission portal on canvas consists of two sub-pages. First page for report submission, the second page for code submission. More information is provided on canvas. Include only source code in a zip file containing your name. We strongly recommend you to attach a README file with instructions on how to run your application. Make sure that your assignment can run only with the code included in your zip file!

After the due date, you will have 5 days to submit your assignment as a late submission. Late submissions will incur a penalty of 10% per day. After these five days, Canvas

will be closed and you will lose ALL the assignment marks.

Assessment declaration:
When you submit work electronically, you agree to the assessment declaration - https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration

5 Teams

This group assignment can be conducted entirely online, without you ever meeting your group members face-to-face. This isn't a problem, with the available online tools. The challenge for you will be using these tools *effectively*. You will need to **make extra efforts** and be **very dedicated and diligent** in working with your team members. This will include setting up dedicated times for meetings and group programming sessions.

5.1 Group Work Tools

To help manage your group work, and demonstrate that you are consistently contributing to your group, we are going to require you to use a set of tools.

5.1.1 MS Teams

Each group will be required to create a team on the RMIT MS Teams platform. Your group must **add your tutor(s)** to your MS team[2]. Your MS team will be the *only official* communication platform for the assignment. This means you must:

- Only use the MS Team channels for group chats

- Hold all team meeting through MS Teams and record all team meetings (except for the conversations in the lab sessions).

- Store any group files (not in Git) in the MS Team

If there are disputes, we will use the record on MS Teams as the source of evidence. You may not use other platforms (including Discord) as we cannot verify the identify of the users.

5.1.2 Git Repository

Your group must have a **private Git repository** that hosts your group's code. This may be on BitBucket or Github. Your group must **add your tutor(s)** to this repository[3]. This git repository will be used as the *evidence of your individual contribution* to your group.

---

[2]If your group is from multiple labs with different tutors, add *all* of your tutors to the MS Team.
[3]If your group is from multiple labs with different tutors, add *all* of your tutors to the git repository.

5.2 Notifying of Issues

If there are **any issues** that affect your work, such as being sick, you **must keep your group informed** in a timely manner. Your final grade is determined by you (and your group's) work over the entire period of the assignment. We will treat everybody fairly, and account for times when issues arise when marking your group work and contributions. However, you must be upfront and communicate with us.

If you fail to inform us of issues in a **timely fashion** and we deem that your actions significantly harm your group's performance, we may award you a reduced grade. It is academic misconduct if you are *deliberately dishonest*, *lie to*, or *mislead* your group or teaching staff in a way that harms your group.

6 Marking guidelines

A detailed rubric is attached on canvas. In summary:

- Approach 60%;

- Ultimate Judgment & Analysis (Independent Evaluation) 20%;

- Report Presentation 20%;

7 Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods

- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviors, including:

- Failure to properly document a source

- Copyright material from the internet or databases

- Collusion between students

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity.