COMP9417 Project

June 21, 2021

Aims

Learning objectives of this assignment:

- a self-selected task to extend aspects of the course material
- involves practical aspects of the machine learning problem, i.e.
  - implementing or modifying algorithms and/or
  - experimental evaluation of algorithms on data set(s)
- exercise written communication skills in motivating, recording and summarising work done on a specified task

Submission

The hand-in for this assignment has two parts:

- files containing program code to do something interesting with data set(s) and/or results of running programs on data set(s)
  - compressed archive of files
  - any programming language can be used
  - must be combined into a single tar or zip archive.
- a report on what you did.
  - must be a single document in PDF format.
  - must include names and zIDs of **ALL** team members

Note: **ONLY ONE** person on the team submits both parts of the assignment.

Note: **Be sure to submit a single PDF file containing your report, and a single tar/zip file containing other files. DO NOT combine your PDF with your tar file.**

Marking

Total: 30 marks available

- **Part 1:** [15 marks]
  - 8 marks: solving the basic problem as described in the topic
  - 7 marks: extra features, or 1 person solving most or all of a > 1 person problem *[at grader's discretion]*
- **Part 2:** [10 marks]
  - 6 marks: describing the problem and your solution
  - 4 marks: good presentation and communication of results
- **Achievement:** [5 marks] At the discretion of the grader - basically a score based on how impressive the work is based on the group size, difficulty of the topic, depth of the analysis, etc.

Part 1

Marks will be gained by:

- evidence of good design or planning by breaking down the problem into sub-components
- rigorous collection of results
- use of comments and notes to record decisions taken and reasons for them in the process of the work
- Motivating the choice of project and your approach (e.g. why was the project interesting? has it been done before? what is different about your approach?)

Part 1

Marks will be lost by:

- programs failing to compile or run
- missing results files
- no clear information on contents of files submitted (e.g. in README)
- evidence of plagiarism (including submissions that are very similar to existing implementations online). This includes recycling work done for other courses, for example, COMP9444.

Part 2

Marks will be gained by:

- evidence of thorough testing of an idea
- good presentation and summary of key results using tables, graphs, etc.
- simple, clear and relevant explanations
- well-formatted, well-organised, spell-checked and grammar-checked documents

Part 2

Marks will be lost by:

- inappropriate length (aim for length of $3 + 2.5x$ where x is the number of group members. Extra figures, tables, etc. can go in an appendix of *reasonable* length). This is not a hard cut-off, and longer reports are OK so long as the length is justified (i.e., the content is crucial to the report, use your best judgement).
- digression, rambling or waffling to fill space unnecessarily
- errors or inconsistencies in presentation, such as
  - incorrect description of algorithms or their properties
  - poor algorithm selection for a task
  - errors in evaluation, like not using an independent test set or cross-validation if this is required
  - statements or conclusions not based either on your experimental results or referenced sources
  - incorrect or inappropriate use of statistical tests
- evidence of plagiarism

Group Configuration

Each team must be configured with 1-5 students currently enrolled in the course

- Teams can be made up of students from different tutorials, and groups can consist of both PG and UG students.
- Larger teams are expected to do more (achievement grade will be affected by this)
- Teams should submit a summary of work completed by each member. If missing, we will assume that all members contributed equally.
- You can use the Moodle Group Project (finding a group) Forum to find group members if needed.
- **Add your group to the 'Group Project - Member Selection' object on Moodle. Deadline to do this is Friday 2 July 2021. You can only join a group if you have the permission of the other group members!**

Member Contributions

- All group members should contribute equally to any work submitted.
- In the case the group feels that one or more of the students have not contributed sufficiently, we will take steps to re-distribute the marks accordingly.
- Some good advice: Keep a record of your contributions throughout the project. Keep a record of all communications with other group members (emails/chat), etc. In the event of a group dispute, we will request evidence from all group members about contribution. Failure to produce evidence means that all group members will receive the same grade.

Report Structure

Giving a very strict set of guidelines to the format of the report for the project is difficult since the different projects are very varied. However, some things to keep in mind are:

- **Length:** Keep it concise. Include a README file with the code so you don't have to put that type of information in the report.
- **Introduction:** You must explain the problem you have tackled, the basic approach taken to solving it, why you chose it, and any important aspects of that approach in terms of machine learning.
- **Implementation:** If your work was mostly implementation, focus on that. Otherwise briefly describe what you did.

Report Structure

- **Experimentation:** All methods must be tested on some data, so these results should be included. Additionally, if this was a major focus, you will need to explain the work done and what was accomplished, for example on setting up the learning task, choice of evaluation, and so on. Detailed statistical analyses are probably outwith the scope of the project, so don't include these unless you are already very familiar with this kind of thing.
- **References:** Should be there for algorithms used or other aspects of the work.
- **Appendix:** Should be used if you have a lot of experimental results. However, consider plotting graphs or using other visualizations like histograms to summarize a lot of results concisely.

Deadline

Sunday August 1st, 2021 23:59:59

Topics: Topic 0 - Propose your own

The objective of this topic is to propose a machine learning problem, source the dataset(s) and implement a method to solve it. This will typically come from an area of work or research of which you have some previous experience.

- it must involve some practical work with some implementation of machine learning
- you must send an email to the course admin (use the class account) with a description of what you are planning (a couple of paragraphs should be enough) that needs to be approved in an emailed reply before you start
- it must not involve double-dipping, i.e., be part of project for another course, or for research postgrads it must include a statement to the effect that it is not part of the main work planned for the thesis (although it can be related)
- If you choose to do topic 0, the deadline to propose a project is 9th July.

Topics: Topic 1 - Machine Learning Paper

The objective of this topic is to choose a journal or conference paper, summarise its findings, and implement the proposed algorithm on a new or simulated dataset.

- Good sources for papers are: NeurIPS, ICML, JMLR, JAIR, ICLR, or ArXiv
- You may also choose a series of papers and compare various approaches to the same problem.
- Email the course admin before you get started on this one too. If you choose to do topic 1, the deadline to propose a project is 9th July.

Topics: Topic 2 - Competitions & Challenges - Kaggle

- Kaggle competitions are hosted here. You may only work on competitions that are labelled either **Featured** or **Research** or **Analytics**. You can select one from either Active or Completed competitions to work on.
- assess carefully the time you will need to understand the competition requirements, get familiar with the data and run the algorithm(s) you plan to use
- for live competitions you can include your submission's placing on the leaderboard at submission time! Note however, that your grade will not be determined solely by your leaderboard ranking. Of course it will be great to do well in the competition, but we are mainly grading you based on your approach and final report.
- You do not need admin approval for this topic. You **must** include a link to the competition on the first page of your report. Failure to do so will result in a 2 mark immediate penalty.

Topics: Other Considerations

- Do not choose a project that needs a significant amount of data processing, or 'create' a dataset, as we are primarily interested in machine learning in this course, not data cleaning. Of course most tasks will require *some* preprocessing.
- A larger group is expected to achieve more, and group size will be taken into consideration when assigning marks for achievement and extra features.
- Choose a topic that interests you, but be pragmatic when it comes to time requirements and difficulty of the project.
- Use common sense when choosing competitions/datasets/models. Do not expect a good grade if you choose a very simple task.
- Before using advanced machine learning techniques, always use a simple baseline such as a decision tree or logistic regression.

Examples: Project Reports

- Every project is different.
- However, if you follow the guidelines above your group should be able to produce a good report.
- We have provided two recent reports as examples.
- The first is from last year on a Kaggle competition.
- The second is from a few years ago on an application of Reinforcement Learning (this topic is no longer available, but it could give you some ideas for an original topic of your own).
- **These are available on the course Moodle page under the Project object in the 'Project Examples' folder**.

Examples: Topic 1 - Machine Learning Paper

You don't have to pick any of the suggestions below, they are just to give you an idea. They range from quite theoretical to practical.

- A comprehensive look at evaluation.
- Proposal for a new technique called stagewise regression.
- Mathematical analysis of neural network approximation.
- A new Python library for tensor processing.
- A Python library for the important task of outlier detection.

Please discuss carefully with your group if you want to do this topic, and also search the sources in the Topic 1 slide above for more options before you make your selection.

Examples: Topic 2 - Competitions and Challenges

Kaggle is a go-to site for machine learning problems and datasets. These provide an excellent opportunity to acquire the essential skills for applied machine learning. You don't have to pick any of the suggestions below, they are just to give you an idea.

Note: some of the datasets on Kaggle are big — you can sample a subset of the data for your project, just make sure that how you do this is detailed in your group's report.

Some typical prediction tasks:

- Fraud detection.
- Malware prediction.
- Hourly rainfall.

Examples: Topic 2 - Competitions and Challenges (continued)

Some image analysis tasks:

- Classify cloud organization patterns from satellite images.
- Global wheat detection.
- Cassava leaf disease classification.
- Right whale recognition.

Some natural langage processing (NLP) tasks:

- Help end gender bias in pronoun resolution.
- English text normalization challenge.

Please discuss carefully with your group if you want to do this topic, and also search Kaggle for more options before you make your selection.