Term Project - Guidelines

DASC 5420: Theoretical Machine Learning

Summary

The term project for this course will be an in-depth report on a machine learning experiment. It should match the format and structure of an academic report on research.

The objectives of the term project include:

- Practice your written skills.
- Practice your critical thinking skills.
- Foster curiosity.

Deadline

Friday, **April 12th, 2024** at 11:59 pm PT.

Rules

This is a group (2 students) assignment. You may discuss your ideas and get advice from other students, but all writing and content must be your own. **No plagiarism** is acceptable: use citations to mark the sources of the facts and ideas mentioned in your report.

Submissions

1. **A summary 5-6 pages PDF report of your work.**
    a. The report should be a complete summary of your project. You should concisely write your report with most of the information and try to avoid too many details.
    b. Make sure you include a working link to the dataset you use and a link to the **GitHub** repository of your project.
    Note: *Create a project repository in **GitHub** (we will discuss it in class) which should contain all of your code (R/Python) associated with the project and a full PDF report consisting of a details description of the project. You can add more details/more information if needed in the GitHub report. Your GitHub report may not necessarily be the same as your final submitted report. I expect you to provide more details in the GitHub one. A copy of the R/python file associated with your code should be in the GitHub repository. Your code should be well documented.*
    c. An evaluation report (one page) of your group member mentioning his/her contribution to the project. Be specific about what he/she did and what you did. Be professional and be honest when you evaluate. Your group mate shouldn't see the evaluation report at any cost. You need to submit the report separately (we will discuss it).
    d. A copy of the R/Python file associated with your code.

Data

You will need to select a data set to use for this project. You can come up with your preferred data or you can use the link provided in Moodle to choose your preferred data. You can also find data from the **UCI Machine Learning Repository (https://archive.ics.uci.edu/).** You must validate your choice of data with me before starting your project. The deadline for finding your data is **March 25th**.

Written report

The written report should be between 5-6 pages excluding references. The PDF should be clearly formatted as a report. You may use the Word Template to guide your formatting, or download another IEEE template (e.g., in Latex) (https://www.ieee.org/conferences/publishing/templates.html). The key formatting requirements you need to meet (regardless of your use of any template):

a. Font: Times New Roman. The size of the font for the main body text must be 11 pt.
b. One column.
c. Single line spacing.
d. Citations/references are to be done in Vancouver style. Citations are marked by a number in square brackets [1], which refers to a numbered reference in the references section.
e. Sections should be clearly titled and numbered.

The paper should begin with a title (in large text, centred) followed by your name (centered). These are followed by the following sections:

Abstract (one paragraph, centered, **in bold**, 150-250 words)
1. Introduction (why your work is important/motivation of your work)
2. Data
3. Method
4. Results
5. Discussion and/or Conclusion

The written report needs to be submitted electronically on Moodle by Friday, **April 12th, 2024** at 11:59 pm PT.

Abstract

This is a short summary of your work. It does not need to be colorful or contain complex equations or explanations. The key points to cover are what your work is about, and what you have achieved (e.g., main results).

Introduction

This section should explain the problem you are focusing on in this report. This can be a domain problem (e.g., detecting a disease), a technical innovation (e.g., a new machine learning method), or a combination of both. This section should provide any information a general reader will need to understand the rest of the work. Most importantly, it should motivate your work, i.e., it should answer the question "Who cares?" (in formal language, of course).

Data

This section should include a description of the data, including its name, contents, where you acquired it, size, etc. You can include general statistical analysis or other comments about the data to help the reader understand the problem.

Method

Describe your method for this experiment. **Be specific.** List the key parts of your process, hyperparameters, algorithm choices, etc. You do not need to mention everything that is obvious, but you need to include every detail necessary to replicate your research. You need to provide the **GitHub link** to the project repository which should include all of your code with associated files at the end of this section. If your GitHub link doesn't work properly, your project will be marked as ZERO.

Results

Display and describe your results. You should use figures (for data visualizations) and tables (for showing results from different configurations or experiments) as necessary. Be sure to label all figures and tables and refer to them appropriately within the text. Make sure you are using cross-validation for your primary score to demonstrate its reliability. What do the

results mean? Do they tell you something about the data or the problem? Do they show the effectiveness of your method, your process, or some other aspect of your work?

Discussion and/or Conclusions

A discussion section describes key questions and observations raised by the research. A conclusion section describes what you have concluded from your work. Either can include suggestions for future research or practical applications of your work.

References

Include references for any website or text you refer to in your report. Use Vancouver style: this is the standard for medicine, mathematics and the technical sciences.

Academic Integrity

Don't cheat or copy text without providing a citation. Claims that require support should also be cited (e.g. this method was developed by X; this method is 250% more efficient, etc.). As we've seen in this course, many of the most common lines of machine learning are the same across programs, and it is fine to use those. If you want to include a function, file, or library from another author, that is fine as long as it is referenced. However, you cannot use other people's unique code without providing a reference.

Academic dishonesty is a serious offense. The TRU policy can be found at http://www.tru.ca/__shared/assets/ed05-05657.pdf. Plagiarism or cheating will not be tolerated and will be reported as per the policy. Note that this may result in zero grade, course failure, or expulsion from TRU.