Project Final Paper

Purpose:

The primary purpose of the project is to provide you with the opportunity to explore, deploy and analyze machine learning techniques in a meaningful way. Students should be able to demonstrate an understanding of the interaction of data, learning algorithm, and outcome reliability/generalizability.

A secondary purpose is to provide students with an opportunity to write a technical paper, in the style of a conference proceedings.

Final Paper

The final paper is the primary deliverable for your project. It should demonstrate that you have applied machine learning technique(s) to real data, and that you understand what you have done. The application does not have to be entirely successful. After all, this is a course project, and not a real publication.

You should demonstrate your knowledge of how the algorithm(s) works, and explain the performance of your application, whether it's a good, happy result, or whether it's a little disappointing. You will not be explicitly graded on the success of your investigation. To be sure, it is easier to write a paper describing a successful application of knowledge than an unsuccessful one.

Requirements

Your paper should meet the following requirement for format:

- ACM or IEEE format.
    - PDF only. Please convert to PDF for me. Thanks!
    - The filename for your report should include your names. One group put all family names. Or all first names also works. But please, nothing generic like "CMPT 820 Final Report".
    - Please continue to have one group member submit the report. It is no longer necessary to have other group members submit anything.
- Maximum length is 10 pages, subject to:
    - The 10 page limit does not include your references.
    - You may include an appendix of supplementary plots and graphs submitted in earlier deliverables. Use this when you have many plots or figures showing different aspects of your data or results. You can refer to these in your main report.
- With references and an appendix, your report may be longer than 10 pages, but no longer than 20-25 pages.

This is arbitrary, but conference proceedings always include a page limit, and fitting your work into the limit is part of the work of writing a paper. Use this opportunity to practice prioritizing and making the most out of the space you have.

Your paper should have the following **content.**

The final paper should have the following sections, which you started on your progress report. Other sections can be added if appropriate, or the overall organization can be changed to facilitate readability, as appropriate. However, all the information described below must be covered.

- **Abstract**:
    - Write a 200-400 word abstract describing the paper.
    - An abstract may be the only part of your paper most scientists read, because only a small fraction of scientists are working in your area. You have to give your reader the information they need to decide whether to look more carefully at your paper. And you cannot waste their time with details that do not assist this decision.
    - Your abstract must say what problem you are solving, how you solved it, and how well your solution works. There is no room for motivations, long descriptions, introductions, explanations, or anything else.

- **Introduction**:
  - Some readers, particularly senior scientists like your supervisors, whose expertise is in the area that you are writing about; they might read your introduction and conclusion sections, skipping over the rest. So these sections have to give your reader the content of your paper in broad strokes. You can assume that the person reading your introduction is a scientist familiar with some aspect of your problem or academic area.
  - Your introduction must briefly describe the problem and why it is important. Briefly describe the approach and why you think it might be novel/valid. Briefly describe your results.
  - Do not waste any space by describing the organization of your paper, unless your organization is unusual.
- **Brief Literature Review:**
  - This part of the paper is useful for scientists whose expertise n the application area is not deep. This could be students like yourselves, post-docs, or senior scientists reading outside their area.
  - The topic of the literature review is the application of computational methods to solve a problem related to your project. Usually, the computational methods will be applications of machine learning, but other statistical or ad hoc methods might be relevant.
  - The topic of the literature review is not machine learning in general, nor is it your application area in general.
  - The literature review serves two purposes. As part of the course, doing the research for this review gives you depth and breadth. The research you review might have given you some ideas on how to approach your problem.
  - As part of a paper (in general, not necessarily a course project), this review puts your work into context. Your research might build on this literature, or apply the same or similar technique to a slightly different related problem. Your research might improve on the work of others, broaden its applicability, or you might have a new (radically, or slightly) idea that you are proposing. In all these cases, your reader needs to know enough about the relevant research and current studies to evaluate your work.
  - Pay special attention to research that's related to your project. Mention what techniques were used, and how well they worked, in broad strokes. You want to compare your results to this literature. In general, a publication presents an advancement or improvement in some way. For a course project, you don't need an advancement or improvement, but we still need the context.
- **Algorithm Description**
  - This part of the paper describes the background in machine learning necessary to understand your results.
  - In general, for a conference or journal paper, what you write here really depends on your research area. If your expected audience is familiar with machine learning, then you only need to describe enough detail to explain what you have done, and you will skip the basics. If you are bringing machine learning to a new area, then you may need to explain a little more of the basics.
  - **For us, in CMPT 820, this is the core of your paper.** This is where you demonstrate what you learned about machine learning algorithms. You must present your algorithms as if to another student in the course who wants to learn about the methods you used. This requires more detail than simple overview, but less than a textbook. Basically, you want to demonstrate to your instructor that you understand what the techniques are intended to do, and how it works, using algorithms, equations, diagrams. This is more detail that you would put into a publication, unless you were writing about a new learning method you developed.
  - This section should include other background material related to your application. For example, you might talk about real estate if you're applying machine learning to real estate.
  - This section is not specific to your application of the techniques. If the techniques you are using have hyper-parameters to tune, or options to select, you describe them, but you don't say what values you used.
  - You can assume that your reader will understand the following terms and concepts, and therefore you will not need to define, explain, or describe them:
    - Over-fitting, under-fitting, training set, test set, validation set, cross-validation
    - Performance metrics such as accuracy, error, precision, recall, confusion matrix
    - Bayes Rule, basic probability
- **Data Description**:
  - Provide an empirical summary of the dataset you used. You should provide high level details such as the number of records, classes, and dimensionality, but also more targeted summary statistics such as the distribution of features in labeled classes, or the number and distribution of missing data if known. This kind of high level statistical summary can provide an important sanity check for the performance of the machine learning algorithm. If you did this well in the first deliverable, you can reproduce it, with minor edits, here.
  - Keep this relatively short, because we've seen it before, but a summary is always required. It will help your readers understand your problem, and it will remind your instructor about the details. Feel free to use some pages in the appendix for this purpose.

- **Results**
  - The first part of your results section is methodology. This section describes the way you treated or pre-processed your data, how you chose hyper-parameters, your test-train strategy, and other details specific to your project. This part of the paper should provide enough detail that another researcher could replicate your work. You have to be judicious in the amount of detail you put here. Too much detail is unwelcome, because it leads to papers that are too long. But a researcher should be able to replicate your results because you provided the important considerations.
  - You should also describe why you decided on these details. In a conference or journal paper, you will be expected to have made the right choices, and your rationale for choosing them will be part of the evaluation of your method. In CMPT 820, you might not have had time to select optimal values, and that's perfectly fine. If you chose value arbitrarily, say so, and perhaps, in your conclusions, explain how different choices might have affected your results.
  - The second part of the results section is where you describe the results of your application. This part of the section simply presents the data objectively. Use tables and plots. Your written text should use the tables and plots as support. Describe the important aspects of the results, accuracies, plots and tables, stating the facts, and interpreting them objectively.
  - Make comparisons between different techniques here. Describe your comparison objectively, using quantitative language, not subjective language.
  - This is not the place where you evaluate the quality of your results, or assess whether you have been successful in this project. A scientist replicating your work should be able to make similar observations, plots, and tables, doing exactly what you did.
- **Discussion**
  - Here is where a paper will interpret the results, subjectively. If the results were good or bad, you can explain or hypothesize why. Compare your results to the work reviewed in the literature review (better than, not as good as, faster than, etc). Be fair, but don't be afraid if your course project is not as good as you had hoped.
  - Here also is where you should comment about what you've learned from doing this CMPT 820 project. Discuss how you'd change your methodology, or how you'd determine better hyper-parameters. Focus your discussion here on your experience.
- **Conclusions and future work**
  - In this section, you summarize your paper, focussing on your results, and their comparison to the state of the art. This is one of the sections that a senior scientist will reader. You will present your results in a little more detail that the introduction section, and the comparison will assist the reader in understanding how the state of the art has advanced.
  - Also discuss future work. Here you describe ideas that you have for pushing this work forward. This discussion has to be deeper than modifying hyper-parameters. Are there other experiments to do? More data to collect? Another step in a larger problem. Don't use this to discuss trivialities.
- **References:**
  - A list of papers and textbooks you consulted.
  - The format should be neat, and standard according to some standard, but any standard is fine.

Important Notes:

- **Every sentence in your paper must be your own words (you and your project team members). A few exceptions are discussed below.**
  - Writing is hard work, but you must practice it to get better at it. Please use the remaining time to focus on your writing, and help each other recognize and fix writing problems.
  - An exception to the requirement is that **you may quote a source appropriately**. However, you should use very limited quotes in the Literature Review and Algorithm Description sections, as this is where you show what you have learned and researched. Writing about what you learned is a good way to demonstrate it.
  - **You do not have to quote equations or formulae.** If there is a special result that you need to draw attention to, you can cite the source of the definition or formula. For example, if you are explaining EM, and you are working from some sources (a textbook, say), you may cite the sources near the beginning of the description. Equations and formulae and definitions from the source can be presented without quoting, and is not considered plagiarism.
  - For the purposes of the CMPT 820 report, you may also borrow images and diagrams, provided you cite them in the caption or the text. However, in a conference or journal paper, borrowing images and diagrams may not be appropriate, and you should discuss doing so with your co-authors, especially you supervisor.
  - I will be reading your papers on my computer, and I will be selecting sentences at random and using Google search. **If I discover even one single sentence in your paper that has been lifted, copied, modified in minor trivial ways, from a website or a textbook, or any other source, I will deduct 15/35 marks from every member in the project group, regardless of who is responsible.** When my marks are submitted and approved, I will then proceed to initiate formal allegations of Academic Misconduct to the College. For grad students, the consequences are severe.

- **If your paper is not ACM or IEEE format, your paper will be rejected, I will deduct 15/35 marks from every member in the project group.**

  Really, there's no reason not to use the common format. Conferences often reject papers written in non-standard format without review.
- **If your paper exceeds 10 pages (not including references and supporting material appendices), I will deduct 15/35 marks from every member in the project group.**

  Part of your experience is to prioritize and make your ideas concise. Conferences often reject papers that are too long, or charge researchers hundreds of dollars per page that exceeds the limit.
- A paper that does a good lit-review and presents the algorithms well, but has poor results, is worth more than a paper that treats these weakly but has good results.

Grading

Grading will consider how well you present your literature, your algorithm, and your results, and not how good your results are. The grading scheme below is out of 70 marks, but the report is worth 35 marks towards your final grade.

- 6 marks: **Abstract**
  - 6/6: Excellent: concise, descriptive, helpful
  - 5/6: Good: helpful, but a little long
  - 4/6: Satisfactory: a little long, maybe not helpful
  - 2/6: Unsatisfactory: too long, too much detail, not descriptive
- 6 marks: **Introduction**
  - 6/6: Excellent: to the point, introducing the topic, presenting the essence of the results
  - 5/6: Good: introduced the topic, and results, but maybe a little long
  - 4/6: Satisfactory: introduced the topic, but didn't discuss results, or didn't avoid irrelevant discussion
  - 2/6: Unsatisfactory: did not introduce the topic, did not consider the reader
- 12 marks: **Literature revie**w
  - 12/12: Excellent: papers summarized concisely, relation to project highlighted, and key aspects presented
  - 10/12: Good: papers and key aspects described, but maybe not briefly enough
  - 8/12: Satisfactory: papers described dismissively, or without meaningful discussion of relationship to your project
  - 4/12: Unsatisfactory: too brief, not enough detail, too much detail, concepts or ideas dismissively presented
- 12 marks: **Algorithm description**
  - 12/12: Excellent: demonstrated superior understanding, presented concisely
  - 10/12: Good: principles and algorithms presented; demonstration of understanding (minor errors are insignificant)
  - 8/12: Satisfactory: principles and algorithms presented, with significant factual errors, or without meaningful interpretation
  - 4/12: Unsatisfactory: no demonstration of understanding of the algorithms or principles
- 6 marks: **Data description**
  - 6/6: Excellent: key points highlighted concisely
  - 5/6: Good: thorough discussion, maybe not concisely
  - 4/6: Satisfactory: a lot of presentation, but without focus on key aspects
  - 2/6: Unsatisfactory: a lot of text and tables, but not much else
- 6 marks: **Results**
  - 6/6: Excellent: key aspects discussed, plots and graphs tidy, good coordination between texts and graphs
  - 5/6: Good: key aspects discussed, plots and graphs tidy, loose coordination between texts and graphs
  - 4/6: Satisfactory: text description does not focus on key aspects, plots and graphs not integrated to discussion
  - 2/6: Unsatisfactory: weak text description; graphs unlabelled, plots confusing or too small to read
- 6 marks: **Discussion**
  - 6/6: Excellent: demonstrated understanding of the results and implications, comparisons, and good reflection
  - 5/6: Good: compared results, made reflections, suggested improvements or opportunities
  - 4/6: Satisfactory: compared results to expectations only, little thought about implications
  - 2/6: Unsatisfactory: little discussion, little reflection
- 6 marks: **Conclusions**
  - 6/6: Excellent: concise summary of the paper, including results, comparisons, and future work
  - 5/6: Good: summarized the paper, including results, but trivial future work
  - 4/6: Satisfactory: casual summary of the paper, without highlighting key results or comparisons.
  - 2/6: Unsatisfactory: no meaningful summary, no highlight of key results, no comparison or future work

- 10 marks: **Presentation**
    - 10/10: Excellent. Very few errors, very coherent writing style, well-structured sections
    - 8/10: Good: Very coherent, well-structured, but some grammatical errors
    - 6/10: Satisfactory: Grammatical problems, and structural flaws in the structure of the sections
    - 4/10: Unsatisfactory: grammar and structural flaws that impede reading
- **Deductions**:
    - -30/70: Not ACM or IEEE format.
    - -30/70: Exceeded the 10 page limit not including references or supplementary material.
    - -30/70: Not in your own words.