Project Instructions for Introduction to Machine Learning (CSC2515) (2024 Fall)

Rahul G. Krishnan
Department of Computer Science, University of Toronto
Vector Institute, Toronto, Canada

1 Introduction

The goal of the project is to give you an opportunity to gain experience in doing research in machine learning (ML).[1] The range of acceptable research topics is wide. Your work can be application-oriented, algorithmic, empirical, or a combination of them. This is discussed more in Section 2.

The project contributes to 30% of your final mark. It has two components. The percentage of each component and the deadline for each of them are (subject to minor changes)

- Proposal (5%): October 25, 2024 (Section 3)
- Report (and Source Code) (25%): November 29, 2024 (Section 4)

Collaboration. The project should be conducted collaboratively as a team of 3 or 4 (*preferred*) members. Everyone should contribute to the project in a meaningful way, and they should be clear about their contributions.

Some frequently asked questions are answered in Section 5.

2 Types of Projects

You may choose any topic of your choice, as long as it is related to ML. Below I describe a few general directions you may want to pursue, and how you can choose a topic within them. Your project might be a combination of them.

You do not need to invent a new algorithm, reach a state of the art performance, or completely solve a new application domain to be successful in this project. I realize that there is not much time in a semester to learn enough about a new research area to innovate a new idea and execute it completely (though sometimes your fresh perspective might lead to an idea that others have not thought before. In that case, you have to cherish that opportunity and pursue it). The goal is to give you a taste of what research in ML might look like. If you feel delighted enough after the end of this course, you have the option to continue working on the project with your team. This being said, you have to spend a good amount of time on this project and you must follow rigorous scientific methodology in pursuing your research.

Before providing specific suggestions for each type of project, I have a general advice on how you can start if you do not have any idea already: You can consult the proceedings of top AI/ML venues such as NeurIPS, ICML, ICLR, COLT, AAAI, IJCAI, AISTATS, JMLR, MLJ, and JAIR, to find many interesting papers. A paper may catch your attention. Read it carefully. Afterwards, read some

[1] This document may be slightly changed in the coming weeks in order to clarify any questions that you have. This version: October 15, 2024.

of the papers that are cited within that paper, and backtrack. You often find the same set of papers referred to again and again. It is also very helpful to see what other more recent papers have cited your originally selected paper. This helps you figure out what advances has been make since that original paper.[2]

2.1 Application

If your main research topic is in an application area, broadly defined, you can investigate whether you can formulate it as an ML problem and solve it using ML algorithms.

For example, if you work on healthcare, user modelling, fraud detection, product recommendation, speech recognition, you can probably formulate your problem as an ML problem. Of course, there are many application domains that can potentially fit well within the ML framework, but have not been explored much yet. Discovering this possibility is an exciting endeavour.

If you decide to go through this path, you should collect appropriate dataset, compare several ML algorithms that we have covered during the course (or some new ones that you find in research papers), and compare them with each other. You need to follow high standards of empirical evaluations.

2.2 Empirical Study

Empirically investigating an already existing algorithm is a reasonable project. Think of yourself as an experimentalist who wants to understand the behaviour of an algorithm through careful design and conduct of experiments.

You want to know when the algorithm works and when it does not. The original paper that introduced the algorithm might have not explored all relevant questions. It is possible that their authors only reported successful results. Your goal is to empirically investigate the range of problems and conditions that result in success and failure of an algorithm.

You need to evaluate the algorithm(s) on several problem domains to see how widely applicable the idea is. In this process, you may want to explore the effect of different hyper-parameters on the performance of the algorithm, and study issues such as sensitivity of the performance on the hyper-parameter.

When you perform an empirical study, you need to follow good statistical practices. For example, if there is any randomness in the algorithm (e.g., random initial weights of neural network), you need to run the algorithm multiple times in order to compute the mean performance as well as its confidence interval.

You also need to be careful in making sure that you separate your training/validation/test sets properly. Otherwise, your results would not be meaningful.

2.3 Algorithm Design

You may decide to design a new algorithm by varying certain components of an already existing algorithm, and effectively search in the space of "adjacent possible". For example, what happens if you learn a suitable distance function for the K-NN algorithm, perhaps using an auto-encoder? Or what happens if you use both ℓ_2 and ℓ_1 penalty as the regularizer for linear models? What about other choices of regularizers? Or what happens if instead of using scalar step size in an online algorithm such as TD, you use a matrix step size? And perhaps design an adaptive mechanism to change the gain?

[2] You may also consult the following good set of suggestions by Csaba Szepesvári at https://rltheory.github.io/pages/assignments/, especially if you want to work on theoretical work. In addition to that, I consulted Animesh Garg's course on 3D and Geometric Deep Learning http://www.pair.toronto.edu/csc2547-w21/project/, and Roger Grosse's project instructions https://www.cs.toronto.edu/~rgrosse/courses/csc2541_2021/assignments/project.pdf for his Neural Net Training Dynamics. You may find good advice there too. Note that their evaluation criteria and what is acceptable or not is not the same as this course, so do not rely on that.

The space of all possible algorithms is combinatorially large. Maybe we can use a computer to automatically explore it. For this course, however, we want you to explore it based on the insight gained in this course as well as your other courses and research experience.

You do not want to randomly wander in the space of all algorithms. Any new algorithm should be justified. You do not need to rigorously prove that the algorithm works before trying it empirically, but it is good to have a theoretical insight before your start implementing.

2.4 Theoretical Analysis

Although this course is not focused on the theoretical analysis of ML algorithm, you may still decide to work on the theoretical understanding of an ML algorithm or problem. Some example research directions are (by no means comprehensive):

- Understanding the convergence properties of different ML algorithms (SVM, Boosting, Lasso, etc.) under various assumptions
- Investigating the effect of relaxing the i.i.d. assumption on the properties of ML algorithms.

You should understand an aspect of theory literature very well and figure out

- What are interesting questions to ask? For example, is the convergence of the algorithm the main concern? Or the sample complexity is?
- What are known and unknown about the topic? For example, do we have any upper bound for the sample complexity? What about a lower bound? Do they match?
- What assumptions are required to make the analysis work? Is there any assumption that can be relaxed? For example, do we need an i.i.d. assumption in the proofs? What changes if we relax that assumption? Or is boundedness of some quantities assumed? Is that necessary?
- Is there any part of the theoretical analysis that can be improved? Is there any tool that the authors of the paper use that is known to be improvable? For example, if they use Hoeffding's inequality, can we use Bernstein's inequality instead to get improved results?

3 Proposal

Your proposal should be a maximum of two (2) page summary of your intended research direction (references excluded in the page limit). You need to clearly

- define the problem,
- provide a brief summary of prior work, and
- what you intend to achieve.

The instruction team will provide you feedback on this. We also provide some office hours before the proposal deadline, in case you want to bounce ideas back and forth before submitting them in the written form.

4 Report (and Source Code)

You should write a 6-8 page report summarizing your work. I encourage you to use LaTeX.[3] You can use the NeurIPS style file https://nips.cc/Conferences/2022/PaperInformation/StyleFiles, though you are not required to use this specific style. Whatever style you use, make sure it has large margins, so I can leave you handwritten comments.

At a high-level, your report should include

- Problem definition and motivation: Clearly state what problem you are tackling and why we should care about it.

[3] If you do not know how to use LaTeX, this is a great opportunity to learn.

- Summary of prior work: What other attempts have been done in order to address this problem.
- Your contributions: Statement and proof of a new result or summary and critique of prior results (Theoretical); clear description of the algorithm and evidence (Theoretical or Empirical) supporting how it works (Algorithmic); the description of the algorithm, the experimental design to evaluate them, and the empirical results (Empirical); how you formulated your application, the description of algorithms you have tried, and the performance you achieved in comparison with other baselines (Application).
- Conclusions: What have you learned and what is remained to be done or figured out?

Your report will be evaluated based on its quality of writing and explanations, how well you cover the prior work, precision of your statements, your contributions (which depends on the type of research you have conducted), and following the good scientific methodology.

It is common in research that each co-authors contribute to different aspects of the project. Some may come up with the high-level ideas, some design the algorithm, some study the idea theoretically, some design and conduct the experiments, and some others write the paper. I'd like to acknowledge that this is how modern science works and let you have different contributions. That being said, *you need to have a section describing the rule of each team member in the whole project in some detail*. The only requirement is that *all team members must be involved in writing the paper*. If you are not good in writing yet, grad school is the right place to practice.[4]

If your project has a source code, which most projects do, you should submit it too.

5 FAQ

Q: Is it acceptable to have a project that overlaps with my thesis project?

A: Yes, and I encourage it. But you should be clear about what part has been done before this project, and what contributions are new. The basic idea is that you should not reuse your prior work for this project; you have to spend a significant amount of time during this semester to work on this project, but you can use it for your thesis (of course, if your supervisor is OK with it).

Q: Is it acceptable to have a project that overlaps with another course project?

A: Try to avoid it! If there is a good reason to have a project that spans more than one course, that can be discussed. You need to get the permission of all instructors for this. Which means that you need to send an email to me and the other instructor(s) and get a joint permission. Since your research is done within a team and you may have different teams in different courses, this makes the credit assignment difficult, hence the discouragement.

Q: Can I extend the project from a previous course?

A: Yes! You should mention it in your proposal, include the report from the previous project in your submission, and be explicit about the new contributions specific to this course. In other words, be clear about the Δ.

Q: Can I have a team size of 5+ or 1 or 2?

A: Teams should consist of three or four members, with four being preferred. Any other arrangement requires a clear reason and my prior permission.

Q: What if I discover that someone else has done a very similar thing to what I am doing in this project?

A: That is OK. It is a part of science. Make sure to cite those paper(s) in your prior work, and provide a detailed comparison.

[4] If there is any reason that you cannot participate in writing the paper, for example a medical reason, you should discuss it with me beforehand.