CS182 - Introduction to Machine Learning, Fall 2021-22
Course Projects

Prof. Ziping Zhao

One of main goals of CS182 is to prepare you to utilize machine learning techniques to solve real-world problems, and this course project provides a good opportunity for you to start in this direction.

I. Specifications

As a part of evaluation of CS182, you are required to complete a course project based on this instruction. Your project must be closely related to what you have learnt in this course, and you cannot use the ideas or results in existing materials or in your previous taken courses. Also, you should not submit a project that is largely collaborated with people outside this course.

In short, a typical project consists of picking an interesting dataset or application, applying one or more well-known machine learning algorithms as baselines, and extending these baselines in creative and innovative ways. The general guidelines are listed as follows.

- Projects should be completed in groups, each of which is composed of **1-2** students.
- Each project consists of two major parts: one final writeup and the source code.
  - **Final writeup**: You are expected to submit a final writeup summarizing your findings, ideas, and contributions. Final writeup must be in the form of a NeurIPS paper and **no longer than 5 pages** in total: 4 pages for the body of the writeup, and an additional page for references. Only the hard copy in **pdf**, rather than the source latex code, should be submitted.
  - **Source code**: For the sake of convenience, it is highly recommended to use Python or Matlab to implement your ideas and algorithms in the project. However, any other programming languages are allowed. It is your responsibility to make sure the source code is executable and contains no severe bugs. Please submit the code in a separate **zip** file.

II. Project Types

Basically speaking, there are three types of projects:

1. **Application project**. This is the easiest and most common one. Select one application that interests you, and explore how best to apply existing learning algorithms to solve it. If you want to choose this project type, please make sure the application is somewhat new, and compare a sufficient number (≥ 5) of cutting-edge algorithms in your projects.
2. **Algorithmic project**. It aims to solve a problem or a family of problems by developing a new learning algorithm, or a novel variant of an existing algorithm. The proposed algorithm is expected to be comparable with state-of-the-art algorithms on some datasets or specific problem settings, and should not be the exactly the same with any existing algorithms.
3. **Theoretical project**. This is purely theoretical, and is usually the most difficult one. Prove some interesting or non-trivial properties of a new or an existing algorithm. If you decide to challenge this type of project, please let me know before you start your project.

Of course, it will be also fine if you would like to complete your project by combing the three elements of applications, algorithms, and theoretical analysis.

III. Schedule and Submission

Please follow the deadlines below. They are strict deadlines and there will be penalties for not respecting them. In particular, the final reports late by 1 day will be penalized with 20% of the grade, late by 2 days will be penalized with 40% of the grade, and late by 3 days is most likely a Fail.

1) *Group member and Topic:* By **Dec. ~~13th~~20th, 2021 (CST)**, you need to form your group (if you want), choose a topic (either inspired on the list of topics below or not, preferably the student will come up with a topic of his/her interest), and report these information to the link:
https://docs.qq.com/sheet/DVWtuenhPUlllTWtU?tab=BB08J2
2) *Final report:* By **Jan. 10th 11pm, 2022 (CST)**, submit your final report with filename

   stu_name_1-stu_name_2-project_name.pdf

   with the source codes and all the cited references (optional) with filename

   stu_name_1-stu_name_2-project_name.zip

   to the link:
   http://pan.shanghaitech.edu.cn/cloudservice/outerLink/decode?c3Vnb24xNjM4Nzg2MDY0Mzgxc3Vnb24=
   Only one group member is supposed to submit the project, tag the rest of group members, and make sure the member-specific contributions.

IV. Project Topics

The first task for you is to pick one interesting project topic. There are many avenues that you may pursue for this project, and we encourage you to be brave and creative even if you don't think you'll necessarily get "good" results. Here are some preliminary ideas[1]:

- Extend classical supervised learning algorithms in the setting of semi-supervised or active or reinforcement learning.
- Develop algorithms that stores both global/linear and local/non-linear information during learning.
- Propose methods that overcome the limitations of existing methods in terms of classification accuracy or computational complexity or theoretical analysis.
- Extend existing binary classification/regression models to handle multi-class or multi-task or multi-label or multi-view/modal/source or multi-instance problems.
- Propose a variant of top methods to address real-world problems in modern applications, such as missing feature values, imbalanced labels, large-scale sample space, high-dimensional feature space, and so on.
- Solve clustering or dimensionality reduction problems by revising current unsupervised methods to handle semi-supervised or supervised applications.
- Improve the popular optimization algorithms to prevent them from converging to the saddle point or local minimum.

In addition to the ideas listed above, you might also refer to some recent machine learning research papers. Three top-tier conferences in machine learning are NeurIPS, ICML, and NeurlIPS.

V. After CS182

An excellent CS182 project will be publishable or nearly-publishable piece of work. After completing CS182, if you would like to continue working on your project along this direction as your graduation project, or submit your work to a machine learning (or other appropriate non-machine learning) conference or journal, please feel free to talk with me. I am happy to give you some guidance and support your further work.

[1] Note that this list is by no means comprehensive, and you can pick any topic that is related to our course and interests you.

APPENDIX A
SOME SUGGESTIONS FOR TECHNICAL WRITEUPS

The final writeup is an important part of the project, as well as the course learning exercise. Please spend enough time on completing your writeup so that it is well motivated, precisely described, and clearly presented. Typically, a technical report/paper comprises of four sections: introduction, methodology, experiment, and conclusion. When writing these sections, it is better for you to keep the following ideas in mind.

- The **introduction** section should set up the problem (e.g., why is it interesting? important?) and provide some context for what work has been done in the past (e.g., what is known? what is the open problem? what are the deficiencies of existing methods?).
- The **methodology** and **experiment** sections should describe clearly what you did (e.g., enough details for someone to re-implement your algorithms, if needed.), and provide some summary tables or figures that illustrate your experimental results, along with some descriptive interpretations.
- The **conclusion** section should concisely summarize your answers for the following questions:
    - What is your problem?
    - What have you done?
    - Why did it work well? (or why didn't?)
    - What might you do if you want to continue working on this project?

APPENDIX B
SOME USEFUL DATA REPOSITORIES

- Kaggle Datasets
- UCL ML Repository
- UCL KDD Repository
- CIFAR Dataset
- Caltech 101 Dataset
- NUS-Wide Dataset
- MirFlickr Dataset
- Amazon Dataset
- MovieLens Dataset
- Multi-Label Repository