CS 4232/5232 – Fall 2023 – Project v2 (Additions in red)
Project Requirements and Proposal Information
Distributed: 09/07/2023

I. About the Project

This document provides instructions for the semester-long group project of the course. The purpose of this project is for each group to pick an area in machine learning and data mining that you find interesting so that you can explore it experimentally throughout the semester.

This project will involve programming; therefore, you should not postpone working on this project until the day before an item is due. This is because waiting until the last minute will almost certainly show in the quality of your project and will not result in a high grade. For this reason, I very strongly encourage you to work hard on your project throughout the semester.

II. Submission and Grading

Items	Due	Grade
Submit Group Members	Sept 7	-
Oral Proposal Presentation	Sept 21	10 %
Written Proposal	Sept 28	10 %
Progress Report	Oct 24	20 %
Preliminary Writeup	Nov 16	15 %
Final Oral Presentation	Dec 5–7	15 %
Final Report	Dec 7	30 %
	TOTAL:	100 %

III. Group Requirements

The following are the requirements regarding the groups for this project:

- This will be a group project with each group consisting of 4 students.
- Since this is a slash-listed course, ideally there should not be groups with both undergraduate and graduate students.
- By September 7th, 2023, **send an email to the TA, CC'ing me, listing all the members of your group**. If your group has fewer than 4 students (this includes the situation of a group having only 1 student), then also submit an email letting us know about this.

After we receive all your emails, we will post the final group membership list. To do this, we might have to reorganize the groups to ensure the above group requirements, giving higher preference to (i.e., trying not to reorganize) those groups that sent their emails earlier.

IV. Project Requirements and Specifications

Your group project must satisfy the following requirements:

- Each *undergraduate group must experimentally compare the performance of at least 3 different machine learning/data mining algorithms*. *Graduate groups must compare 4 different machine learning/data mining algorithms*.

  For example, if your goal is to design a classifier to predict how late a flight will arrive at its destination, then 3 different ML algorithms are: i) use a random forest algorithm to predict the time of arrival; ii) use an elasticnet model; and iii) use a deep network. If your goal is to design an agent that learns how to play chess using reinforcement learning, then 3 different algorithms would be the following: i) use q-learning (reinforcement learning) to learn a good policy; ii) use SARSA (reinforcement learning) to learn a good policy; and iii) use the A3C algorithm.

The performance measure for your learning agent is highly dependent on the topic of your project. However, some examples of performance measures are the following: training time, prediction time, accuracy, precision, recall, F1 score, memory consumption, etc.

- In your project you need to *implement at least one of your machine learning/data mining algorithms from scratch*. For this latter machine learning/data mining algorithm, *you may not use packages like caret, scikit-learn, or any other that already implements it.*

  You can, however, use packages that support other functionalities (different from ML algorithms) like matrix-vector operations, plotting, optimization packages, data frame operations, etc. Examples of packages that do not implement ML algorithms are the following: R's basic utilities, R's tidyverse, numpy, and pandas. The idea of this constraint is for you to learn how to program the algorithms. You can consult me if you have any doubts about the packages that you can use.
- You must analyze at least 4 papers or book chapters related to the topic of your project. This analysis must compare and contrast your proposed project against existing techniques.
- To implement your project, you can use any programming language that you like. I recommend using either Python or R, but you can also use Matlab, SAS, etc.
- Your project cannot be about handwritten digit recognition, Boston housing price prediction, or any other problem that has been solved everywhere.

V. Oral Project Proposal

For your oral project proposal, you must prepare slides and then make a 4-minute-long presentation in front of the class. Your presentation must **contain one slide for each of the following items**:

1) Project title and the names of the authors
2) Objectives and justification of the project

   State concisely what the project is intended to accomplish overall using a bullet format to describe your individual objectives (one bullet for each individual objective). To justify your project, explain the reasons why you chose to do the project you are proposing and why it is worth doing.

   This slide **must contain two things**: **a bulleted list of objectives** and the justification of the project.
3) Contributions of the project

   State what the differences are between your project and what has already been done in the area. In other words, explain what is new about your project that has not been done before.

   This slide **must be simply a list of bullet points** with each bullet point presenting a contribution.
4) Machine learning algorithms that will be employed

   Mention the machine learning/data mining algorithms that you will employ to solve the problem in your project. You need to briefly explain why you suspect that those algorithms can solve your problem.

There should be a bulleted list of algorithms in this slide. Each **bullet presents an algorithm** and is followed by an explanation of why the algorithm solves your problem.

5) Description of the dataset
   Describe the dataset(s) that you will use for your project, listing and explaining its attributes. If your dataset(s) consist(s) of record data, then you need to mention the number of rows, the number of columns, and the size occupied by the totality of the file(s) containing the dataset.
6) Description of the learning environment
   If your project involves reinforcement learning, then describe the learning environment, i.e., the states, actions, and rewards, and a link to the simulator that you will use (in case you are not implementing your own simulator).

On the day of the proposal, each group **must upload to Canvas a PDF document** containing their slides. Only one member of each group will upload this PDF to Canvas.

The oral project proposal, like all other items due for this project, is a joint effort of all members in a group. For this reason, every member of a group must present a part of the oral project proposal in front of the class.

VI. Written Project Proposal

Your written project proposal must consist of 2 typed, 12-point font, double-spaced, 1-inch margin pages. In addition to all the items of your oral project proposal, the written proposal must contain the following ones:

1) Related work
   Provide a bullet list of at least 4 references, which can be papers/books/websites. For each paper or book reference, you need to provide the name of the author, title, publisher or Journal name, and year; for websites, you must provide the name of the author, title, web URL, publication date, and last accessed date. After each bullet, explain in detail what the reference does and the difference between what you will do and what was done by the reference.

   This section aims to explain existing knowledge in the area, identifying the gaps your project intends to fill. The group must compare their approach to the papers discussed in this section. These 4 references must be different from the ones mentioned in this course's syllabus.
2) Research methodology and timetable
   This must be a table outlining a general plan of work including a description of the ML approaches methods and procedures to be used, and *how you plan to evaluate the performance of your approaches*. This plan must specify clearly tasks to be accomplished, the deliverables of each task, the starting and ending dates of each task, and the person(s) in charge of each task.

3) Addressing comments of Prof. Leal

   You need to provide a response to each comment I made to your oral proposal, indicating how you are addressing it. This needs to be a well-reasoned response.

The written report must follow the requirements for bulleted lists for the following sections:

- objectives (one bullet per objective),
- contributions (one bullet per contribution),
- machine learning algorithms (one bullet per algorithm)

described in the oral proposal requirements. The only **difference between the written report and the oral presentation** is that in the former you will explain each objective, contribution, and algorithm in writing, while you do not need to do so in the slides.

Each group must upload to Canvas **a PDF document (<u>it cannot be Word</u>)** containing the written project proposal on the day that the written proposal is due. Only one member of the group will upload this PDF to Canvas.

VII. Project Suggestions

Some suggestions for projects are the following:

1) Improve upon the work of students who have taken this course before. The advantages of this route are that there is code already implemented and that you have a baseline against which to compare your own project. Examples of projects from past semesters are facial recognition in an image, designing a chess player, flappy bird, etc. If you are interested in this route, I can inform you more about this.
2) Find a Kaggle competition and participate in it.
3) Find big dataset(s) online (e.g., from kdnuggets.com or check 50 Best Public Datasets) and then make an agent that uses them (or it) to predict stock prices, or the amount of crude produced, etc.
4) Make an agent that is a very good player of some complex game like chess, go, othello (reversi), backgammon, snake, minesweeper, FlappyBird, etc. It cannot be a trivial game like tic-tac-toe. If you are choosing this path, check open-gym (open-ai) for simulators so that, if possible, you do not end up investing a significant portion of your time in this course implementing a simulator.
5) Implement and compare algorithms for image processing (like those on smartphones): noise reduction, image super-resolution, facial recognition, etc.
6) Implement and compare algorithms for synthetic image generation or synthetic image recognition.
7) Experimentally compare two published machine learning/data mining algorithms, or two similar machine learning (e.g., Tensorflow vs. Theano). This comparison needs to be rigorous and comprehensive, so it cannot simply be a discussion of the advantages/disadvantages of each.
8) Select a ML/Data Mining paper in a topic paper that interests you, and either replicate their results or propose an extension to the paper. The extension can be a theoretical (a new idea or technique, eliminate a parameter of the technique, make it work on a different type of data, etc.) or a practical extension (make the technique work on multiple cores, GPUs, etc.). To find ML papers you can search in conferences like the following: KDD, ICML, NIPS, ICDE, SIGMOD, AAAI, VLDB, etc., and in journals like the following: the Machine

Learning Journal, the VLDB journal, etc. Your papers must be from peer-reviewed journals or conferences.

9) For graduate students, you can use data from your own research and extend what you have done so far. You need to state very clearly in your proposal the work that you have done previously and the new work that you intend to do in this course.
10) If you have other ideas that you would like to try, you can talk to me.

Last Modification: 9/7/23 6:52:00 AM