Introduction to Machine Learning [CS-3410-1]

Midterm Project

This grading component carries 10% weightage and should be the most engaging part of the course. You will be expected to submit a small implementation-based machine learning (ML) project as a part of your midterm. You will have the option of extending this project as a part of your final project, given you have confirmed this with the teaching staff.

Topic Choice: You have the flexibility to either come up with your own project idea or choose from the options provided below. Once you have a project idea, you must get it approved by the teaching staff.

Group Policy: You may work individually or in pairs of two - however, the expected output and rigor exhibited by pairs of students will be higher.

Rules:

1. You must code all algorithms manually. The use of machine learning libraries such as scikit-learn, TensorFlow, Keras, PyTorch, etc., is strictly prohibited. You may use libraries for basic data handling and mathematical computations like NumPy and Pandas.

2. You cannot use preprocessed data for this project. You need to use raw data and perform data profiling, cleaning, transformation, etc. Make sure to include all these steps in your code and project report.

For any queries related to the project, please email anyone from the teaching staff (and cc the professor) with the subject line "Final Project - [Your Topic]". Only one member of the group should send the email.

Project Options

You have two options for your final project:

1. Your Own Project:

Come up with your own idea and implement it. This could be an application, algorithm, or theoretical exploration.

Your project can fall into one of the following categories:

i. Application Project: Apply machine learning algorithms to solve a real-world problem
ii. Algorithmic Project: Develop a new algorithm or improve an existing one to solve a specific problem
iii. Theoretical Project: Prove or analyze properties of a non-trivial machine learning algorithm (note: This is more challenging and less common)

Please note that these categories are not mutually exclusive, and that you will be expected to demonstrate at least some understanding of theory and application in your project. Many projects combine elements of applications, algorithms, and theory.

Datasets: You can use datasets from platforms like Kaggle (https://www.kaggle.com/), UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php), or collect your own data. Use academic search engines like Google Scholar(http://scholar.google.com) to find relevant research papers and datasets.

2. Choose from Suggested Ideas:

If you're unsure about what to do, you can pick one of the following project ideas:

i. Spam Email Detection (Classification): Build a model to classify emails as spam or not spam based on their content. https://www.kaggle.com/datasets/wcukierski/enron-email-dataset
ii. Antibiotic Resistance Prediction (Classification): Build a model to understand antibiotic resistance on the basis of available genomic information. https://www.kaggle.com/datasets/nwheeler443/gono-unitigs
iii. Air Quality Prediction (Regression): Build a model to predict air quality, given air quality data from 5 major cities in the US (2014-2024). kaggle.com/datasets/prajapatirishabh/air-quality-data2012-to-2024

For inspiration, you might also look at some recent machine learning research papers. Two of the main machine learning conferences are ICML and NeurIPS. You can find papers from the recent ICML https://icml.cc/Conferences/2023 and NeurIPS conference https://neurips.cc/Conferences/2023.

Project Components:

1. ***Project Report :*** The final writeup needs to be as comprehensive as possible. It should justifiably highlight the research and effort you've put into this final project. The main components expected are:
    a. Introduction
    b. Literature Review (Optional)
    c. Dataset(s) Source and Description
    d. Data Exploration and Important Features
    e. Methods
    f. Experimentation
    g. Final Results
    h. Conclusion
    i. References
2. ***Project Code:*** A detailed, organized GitHub repository that contains all datasets and codefiles. We recommend consistently committing changes to your repository, both as good practice and from a grading point of view.

3. ***Midterm Presentation:*** Everyone will have to present their projects to the teaching staff. It will be a 10-minute presentation with snippets from your working model or just the interface you would have designed for your project. An accompanying, short slide deck is expected.

Additional Resources

1. Here are links to past year student projects from Professor Andrew Ng's CS229 course-

					2023 (Autumn)	2022 (Autumn)
2022 (Spring)	2021 (Spring)	2020	2019 (Autumn)	2019 (Spring)	2018	2017
2016	2016 (Spring)	2015	2014	2013	2012	2011
2010	2009	2008	2007	2006	2005	2004

Try using this only for inspiration/ideas. Cite any and every resource you pick from these. They are just to guide you with potential ideas.

Good luck, and have fun working on your project!