CS5785 / ORIE5750 / ECE5414 Final Project

This final project is due on **Wednesday, December 13th, 2023 at 11:59PM ET**, uploaded to Gradescope (Canvas->Gradescope). The final project *proposal* is due on **Friday, November 17th, 2023 at 11:59PM ET**, uploaded to Gradescope (Canvas->Gradescope). Your submission will have two parts:

1. A write-up as a single `.pdf` file. Submit this under the `final-report` assignment in Gradescope.
2. Source code and *data* (if your project involves datasets of your own choice) files for all of your experiments (AND figures) in `.ipynb` files (file format for IPython Jupyter Notebook). These files should be placed in a folder titled `final` and uploaded to the `final-code` assignment in Gradescope.

The instruction on the write-up can be found below. On the cover page, include the class name, and the names of the entire team. In addition, briefly describe the responsibility of each team member. You could use online LaTeX templates from Overleaf, under "Academic Jounal" and "Project / Lab Report".

Please include all relevant information for your project, including text response, equations, figures, graphs, output, etc. If you include graphs, be sure to include the source code that generated them. Please pay attention to Canvas for announcements, policy changes, etc. and Piazza for final project related questions.

You are allowed to use well known libraries such as `scikit-learn`, `scikit-image`, `numpy`, `scipy`. When required, we only ask that you develop your algorithms in `PyTorch` instead of `Keras`, `TensorFlow`, etc.

PROJECT GUIDELINES

Becoming a data scientist or machine learning professional requires many concrete skills, but typically (if you're successful), there usually won't be someone constantly monitoring your every move and providing explicit instructions every step of the way. Doing a good job requires (1) identifying interesting problems where machine learning can make an impact; (2) identifying and collecting relevant data; (3) documenting precisely how this data relates (and how it doesn't) to the present application; (4) analyzing the data and summarizing its properties; (5) implementing, running, and evaluating relevant algorithms; and (6) convincing your stakeholders of your findings, through solid technical work and experiments, conveyed via clear writing, quantitative analysis, and qualitative analysis (including visualizations).

In this open-ended assignment, you will have an opportunity to combine your creativity, coding know-how, knowledge of algorithms, and writing skills towards an adventure of your choosing. In groups of up to 3, you must identify an interesting problem, relevant sources of data (perhaps by scraping from the internet), and the appropriate machine learning techniques that can be applied. *Although individual submissions are allowed, you are highly encouraged to work in teams!* A great final report will convince the reader that you have identified an interesting problem, collected compelling data, conducted thorough analysis, and explored the usefulness of the algorithms that you've been taught (and perhaps some that you haven't been taught yet) for making predictions and/or making sense of the data. The final deliverable will consist primarily of a write-up (complete with quantitative results and visualizations), with supplementary materials including code and data. There is no strict page requirement or limit, but reports will be judged on insight and technical depth more than volume.

PROJECT PROPOSAL For the project proposal deadline, you need to briefly describe the objective and overview of your project on a Canvas quiz, and indicate whether you would like early feedback on the feasibility of the idea, identifying potential sources of data, and so on. If you want to work on one of the following projects, but are not sure how to formulate the problem, come to my office hour!

Here, we offer a selection of project suggestions that you may consider working on. (You are also encouraged to devise your own project ideas.)

1 EXAMPLE: CREATING INSTRUCTIONS DURING A SURGICAL PROCEDURE

A Queens-based eye surgery center is exploring the development of real-time surgical guidance during eye procedures by utilizing historical recordings. This guidance can be delivered in the form of either video or text. The concept revolves around harnessing past surgical data to anticipate potential issues that may arise during the ongoing procedures, offering real-time recommendations for the best course of action.

The above describe problem is a reinforcement learning problem by nature – the end goal is to learn the optimal/best surgical recommendation. A natural way to approach the problem is to break it into two parts: 1) extract labels (the name of the surgical steps) from the videos, and 2) learn the optimal surgical action given the current history using some reinforcement learning algorithm.

There are broadly two directs that you can pursue under this topic:

1. Write a comprehensive literature review on how past methods could be combined to produce a solution in this problem, and demonstrate a subset of your proposed solution on appropriate dataset(s). (This route emphasis more heavily on the literature review side, and less on the coding side)
2. identify a subarea of this problem, and clearly state how your solution would be used as a subroutine in the overall framework. You should demonstrate the feasibility of your ideas on appropriate dataset(s). (This route emphasis more heavily on the coding part. You should focusing on solving the subproblem that you propose.)

2 Example: Abnormality Detection in Healthcare

Hospital A has many locations in NYC, and each site has their own management team overseeing the entire hospital operations. By looking through the data, one senior doctor noticed that the revenue at one location is abnormally low. By examining the data more closely, this doctor discovered that patients at this location were accidentally forced to purchase the prescriptions rather than it being optional. After correcting for this error, the revenue at this site in the subsequent year went back to normal. Motivated by the above incident, please propose a method that can detect abnormality in medical data. Please identify the type of abnormality that you would like to address, propose an appropriate method that solves this problem, and demonstrate your idea on the appropriate dataset(s).

3 Example: Does Consumption of Meat Increase Breast Cancer Incidents?

It is known that certain hormone replacement therapy could lead to increased breast cancer rates [1]. On the other side, it is known that the breast cancer rate among women in the U.S. is 13%, while the breast cancer rate in China, for example, is 3.5% in 2015. One observation that we have is that beef cattle and sheep in the U.S. are prevalent grown using growth promoting hormones [2]. Please identify appropriate datasets and methods to answer the questions *whether the consumption of meat increase the number of breast cancer incidents*?

PROJECT REPORT WRITEUP

Each group should submit a writeup of their project work, exceeding no more than 6 pages including figures. References are excluded from the 6 pages (e.g., they may overflow onto a 6th page and be numerous). It is fine to be below the page limit; this is the maximum. We appreciate conciseness. We will also ask you to submit your **code** and **dataset**, so that we can validate your results.

The writeup must adhere to the following template, accessible via this link – so that we can judge all writeups in the same manner without having to worry about different font sizes, etc.

Your report writeup should contain the following sections:

1. **Abstract** A brief description of what you did in the project and the results observed.
2. **Introduction** Introduce the motivation and the setup of your problem: why is the problem important (if applicable, how does your method can be used for the downstream task that we care about). Describe your dataset.
3. **Related Work** If you decide to take the coding route, this section should be brief and discusses prior works that have explored the same problem / dataset as yours.

   If you decide to focus on the literature review for the final project, you can merge this section with methods to give a comprehensive overview of past methods and how they can combined to produce a feasible solution.
4. **Methods** If you take the coding route, state the details about your models. This section should summarize all the methods that you have tested. Specifically, explain the details about 1) your model and 2) the training (e.g., how to initialize the model parameters).
5. **Results** State the results of your experiments, and insights.

   This section should include a table summarizing the performance of your models (and different combinations of them). If you explored different hyperparameters, you are encouraged to add figures that plot performance vs. hyperparameter values. You should make sure that your figures are readable and not take up so much space (you have the page limit).
6. **Discussion** Summarize what you have done in this project, and highlight the limitations. Discuss future work. You can combine this section with the Results section if needed.
7. **References** List references used in your writeup. We expect you to cite all papers, books and websites used for ideas, code and phrasing. Please review the Canvas course syllabus for our policy on Academic Integrity.

REFERENCES

[1] N. R. Shah and T. Wong, "Current breast cancer risks of hormone replacement therapy in postmenopausal women," *Expert Opinion on Pharmacotherapy*, vol. 7, pp. 2455 – 2463, 2006.

[2] U. Food and D. Administration, "Steroid hormone implants used for growth in food-producing animals."