Class Project: Overview

The goal of the machine learning project is to get hands-on experience in independently defining, analyzing, and executing a machine learning (or data science) project. It is not necessary (but of course allowed) that the project conducts original research in machine learning.

You can either take an interesting dataset and try to make predictions/inference from it using machine learning techniques that we have covered or, even better, ones that we have not covered. Another option is to take a machine learning method and analyze its behavior, or propose an improvement. At least one aspect of the project should be novel and creative.

Some sources of relevant datasets are:

- https://archive.ics.uci.edu/ml/datasets.php
- https://github.com/nytimes/covid-19-data
- https://data.gov
- https://github.com/CSSEGISandData/COVID-19
- https://www.eddmaps.org/
- https://kaggle.com

Feel free to use piazza to solicit ideas on where to get other datasets.

For more inspiration and prior work, you may also want to look at some recent machine learning conferences and workshops:

- https://neurips.cc/Conferences/2020/Schedule?type=Workshop
- https://icml.cc/virtual/2020/workshops

A good project will be creative and not be just based on a dataset from Kaggle or a similar site. You may use datasets and problems on Kaggle as an inspiration, but then please try to go beyond the problem definition stated on the site. The goal is to try a machine learning project that goes beyond simply running a standard algorithm using a standard dataset.

Project Report

The overall project deliverable is a brief report that describes the results and provides the appropriate evidence that supports them. Please do not simply include a deluge of plots. Be brief and to the point. Only include the most relevant evidence. The reports can be prepared using a quarto notebook, LaTeX, Jupyter, or any other typesetting environment (or even MS Word if you'd like). Projects can be done individually or in *groups of 1-5 people*.

The completed report should be *up to 4 pages long* (excluding references) and *succinctly* describe the problem, the motivation, and the results. The expectations on the quality and length of the work grows linearly with the number of students involved in the group. A short

2-page report is better than a 4-page report that does not highlight the importance of the findings.

There are many ways to structure the report, but a great report will have content that resembles a workshop paper. The following paper is a good example:

https://www.climatechange.ai/papers/neurips2020/6/paper.pdf

Deliverables

To make it easier to make incremental progress and to get feedback on your ideas, the project will include a sequence of deliverables culminating with the final report and a short presentation. The intermediate deliverable will be graded as pass fail, with any good faith effort counting for pass. You may change your to any extent that you would like based on the feedback that you get from the intermediate deliverables.

1. Motivation/Introduction

Which addresses these issues:

1. What is the problem?
2. Is it prediction or inference?
3. Is it classification or regression?
4. Why is the problem important?
5. What does success look like?
6. What are the data sources that will be used. Is it likely that they will suffice to achieve the goals?

The report should be provided in a form of a free flowing text and not just as answers to the questions above.

2. Related work

Describe most relevant methods that have been used to solve the problem you are tackling. If the focus of the project is on an application, describe previous work addressing the application. Describe what methods and data sets were used previously. The relevant work should be based in peer-reviewed research papers, books.

A good resource is the Google Scholar search engine:

http://scholar.google.com

3. Methodology

You should answer questions such as:

1. What is the right metric for success?
2. How good does it need to be for the project to succeed? For example, does the prediction error needs to be at most 5%? What about the area under the curve. Argue why.
3. Use a test set? Bootstrapping to understand parameter variability?
4. How to make sure that the results are valid?
5. What kind of methods will be using?
6. Will there be any theoretical analysis or improvements that you are proposing?

The report should be provided in a form of a free flowing text and not just as answers to the questions above.

4. Results

Describe the results of the method. Describe how well the method did in the evaluation and compare with prior work (if applicable). Discuss what the results mean in the context of the problem definition. Is there anything that can be done to improve the results, or are they good enough? What about confidence in the results?

5. Project presentations and final version

If you are analyzing/improving a ML method, make sure you motivate your analysis, describe the method you chose, and present a clear analysis of your results. The final version will be graded based on the quality of the results and the clarity of the report. The final presentation will either be slides or perhaps in form of posters. The decision will depend on the number of project in the class and other ligistical considerations.