Class Project

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 datasets 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.

The 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 as R studio notebooks, using LaTeX, Jupyter, or any other typesetting environment. Projects can be done individually or in groups of 1-5 people.

Some sources of 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.

I prefer that you do not just download 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.

You may choose to continue working on the dataset from the mini-project and try to improve the methods that you have developed in the first part of the semester. I have some additional datasets available if you are interested. Some examples are invasive species, log from greenhouses, or NH and NY sweet corn,

The report should be about $3 * \sqrt{\text{group size}}$ pages long. Reports for projects that apply ML to a new(-ish) dataset should contain roughly the following sections. Please feel free to improve on this structure as you see fit.

- **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?
- **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.
- **Evaluation methodology**: You should answer questions like:
    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?

- **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?

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.