Sta-395 Intro to Machine Learning - Final Project (Fall '25)

Overview: The overarching aim of the final project is for you to engage in a capstone experience that demonstrates your ability to thoughtfully apply modern machine learning methods to novel data of your choosing.

To begin, you must pose a research question involving *complex data* with a format/structure that goes beyond the standard spreadsheet format you've worked with in lower level statistics courses. Some examples include:

- Images or videos
- Text
- Time-series

Your project should then apply machine learning methods from two different perspectives:

- Feature engineering with conventional models/pipelines
- Deep learning using a model specialized for your task

Depending upon the application, it is acceptable if one of these is a "straw man". As an example, feature engineering isn't easy for image data, so you might come up with some features based upon the distribution of pixel intensities, the prevalence of certain colors, etc.; however, it would be okay if these features aren't overly comprehensive and don't perform well so long as you put forth substantial effort into your deep learning approach.

Your project will be evaluated on the follow components:

1. Code Repository - you are to provide an organized and appropriated documented set of Python files that recreate your main results
2. Presentation - a 10-minute presentation providing a high-level overview of your research question, data, and main results
3. Report - a 3-5 page report (not including figures or graphics) written in a formal scientific tone providing a detailed overview of your methods and results

Additionally, you will receive a small amount of completion credit for the following:

1. Submitting a topic proposal containing an appropriate level of detail by the stated deadline (Friday 10/31)
2. Presenting a short, informal progress update to the class (Tuesday 12/2)

You are free to pursue any topic that allows you to demonstrate your knowledge of machine learning methods within the basic constraints described above. This might include working on a Kaggle competition, reanalyzing data from a scientific paper, or any other application you're interested in.

The expectation is that you form a project team containing 2-3 members (including yourself), but if there are special circumstances regarding your topic/plans you may work alone.

Timeline:

- Tuesday 10/28 - One member of your intended project team provides the names of all team members via email. Any students who are not part of a team and who have not indicated a preference to work alone will be assigned to a team.
- Friday 10/31 - One member of your team must submit a short project proposal via email. This must include your research question, your data source, and a paragraph explaining your planned methods.

- Tuesday 11/25 - We will have a built-in work day where you can collaborate with your team during class and ask questions of myself and others
- Tuesday 12/2 - You will give a short ($\leq$ 5-minute) progress briefing where you share your project topic and current status with the class. The purpose of this step is for your team to solicit feedback and ideas from your peers.
- Tuesday 12/9 *or* Thursday 12/11 - Formal presentations during class
- Friday 12/19 - Final reports are due by 5:00 PM

Presentation Details:

During class on either Tuesday 12/9 or Thursday 12/11 your team will give a roughly 10-minute presentation. Your presentation should summarize the key components of your group's work, including the project's goals, data, methods, and results. You should treat the presentation as if you were at a scientific conference, meaning you may assume some familiarity with basic machine learning methods and concepts, but you should thoroughly cover more advanced methods or non-standard methods. Your presentation will be scored on the following criteria:

1. *Comprehensiveness* - Did you communicate all of the necessary information that someone would need to understand your work? Were any important details not adequately explained? Could an audience member provide an accurate overview of everything you did?
2. *Coherence* - Was the presentation organized in a logical order? Did it flow smoothly? Was it easy for the audience to follow along?
3. *Professionalism* - Were your presentation materials appropriately prepared? Did you adhere to stated 10-minute time limit and speak using appropriate volume and pacing? Did each group member contribute to the presentation in a manner that didn't interrupt its flow?

Report Details:

Final reports, code, and data are to be submitted via P-web no later than 5pm on Friday at the end of final's week. Your paper should be 3-5 pages in length (not including figures, tables, references, and any supplemental material). You are encouraged to embed a link to a repository containing your code and data (rather than submitting them directly). Your report will be scored according to the following criteria:

1. *Structure* - Does your report follow a proper scientific structure (Intro, Methods, Results, Discussion, References)? Does the report include appropriate figures and tables that help communicate your main results? Does information appear in the proper section? (ie: results are not given in the Methods section, etc.)
2. *Technical Correctness* - Are your methods and results described correctly? Or there any errors, misleading statements, or important information that is omitted?
3. *Level of Detail* - Are your methods described in sufficient detail? Are your results thorough and comprehensive?
4. *Coherence* - Can a reader of your paper fully understand your work? How easily could a reader summarize the main steps of your project? Are there any gaps, awkward transitions, or omitted sections?
5. *Professionalism* - Are your visualizations and tables professional in appearance? Do you cite scientific sources when appropriate? Is the report written in a scientific tone and free of grammatical/spelling errors?

For additional guidance on scientific writing, I recommend this article, which provides some general advice on how to structure and prepare scientific writing.

Deep Learning Models:

An expectation of the project is that you consider an appropriate deep learning model. This model doesn't need to be something you develop from scratch, and you may opt to use or fine-tune a model built by someone else for a related machine learning task. A few places where you could identify such a model are:

1. Hugging Face - https://huggingface.co/models
2. Kaggle - https://www.kaggle.com/models
3. PyTorch - https://pytorch.org/hub/
4. TSAI (for time-series data) - https://timeseriesai.github.io/tsai/