CSCI547: Machine Learning Final Project

October 22, 2020

Due date

The project will be due **Wednesday, November 25th by 5:00 PM**.

Overview

One of the primary goals for CSCI547 is to give you some experience in applying the algorithms that we learn about in Machine Learning to a variety of problems, perhaps in your particular area of specialization. The final project for this course is designed to give you an opportunity to apply the machine learning skills you've learned to a non-trivial problem of interest.

Project Scope

Your project will entail the application of a machine learning technique to a dataset. You may apply a technique that we learned about in class to a dataset that we haven't seen before, or you may apply an algorithm that you would like to research on your own to one of the datasets we've seen in class, or you may try something entirely new. For obvious reasons, it will not be allowed to apply an algorithm we learned about in class to a dataset we used in class.

Deliverables

Once you've identified the scope of your project, you should write a code base that represents a working machine learning model. This code base should not only deal with training the model, but also include elements of validation and model selection indicating a deep exploration of the chosen method and dataset. **In addition to the code, you will need to prepare a paper consisting of a literature review framing the problem that you are trying to solve (you should read and cite at least 2 papers, but probably more), methods, results, discussion, and conclusions**. There is no specified length for this paper, but it does need to be long enough to justify the work that you did! As with all scientific writing, well-made figures are a highly welcome inclusion.

Grading Rubric

The best project will be similar in quality to a short scientific paper. That means that the figures and format should be polished, the language clear and complete, the work fully cited, the implications thoroughly considered, and the accompanying code well-commented and easily accessible. Papers that exhibit all of the above characteristics will surely receive an A. Papers exhibiting some of these qualities will receive some fraction of an A grade. Papers that do not exhibit any of these characteristics will receive an F. A very important note is that you will *not* be penalized for a negative result: if you apply an interesting algorithm to a non-trivial dataset and things just don't quite work out to a satisfying conclusion, that is okay. It is much better to be bold, do something interesting, and have it not work that to do something certain but uninspired.

Examples of previous projects

- Using CNNs for cancer detection in mammograms
- Using a naive Bayes classifier to identify intersection crash probability in Missoula
- Using linear regression to understand factors influencing transportation choice
- Using a recurrent neural network to try and predict the stock market
- Identifying grouse from thermal camera images
- Classifying irrigated land from Landsat images
- Development of a recommender system for movies based on reviews

Dataset sources

The following links have many good ML-suitable datasets (but feel free to use your own!).

Kaggle kaggle.com

UCI Machine Learning Repository https://archive.ics.uci.edu/ml/index.php

University of Florida Statistics http://users.stat.ufl.edu/ winner/datasets.html