STOR 565: Introduction to Machine Learning (Due: 04/24/20)

Final Project

Instructor: Andrew Nobel *TA:* Alexander Murph

Project Overview: The overall goal of the project is to apply supervised and unsupervised statistical learning methods to a dataset of your choice to answer a research question. The formulation of this question, which methods you will use, and how you interpret the results, are entirely up to you and your team. You will additionally be expected to research and implement a supervised learning technique not covered in class, write a report on this technique, and use this technique in the analysis of your data. The following submissions will be expected.

1. Project Proposal: Due 04/01/2020; no more than 1 page in length. This is expected to be a brief of your project plan, outlining your teams choice of data, research question(s), and learning methods (both old and new, both supervised and unsupervised). Besides explicitly stating each of these points of interest, this proposal should include a summary of your data that discusses its origin, who collected it, why it was collected, and what predictors are available. If you cannot explain what a given predictor measures then you may not use it in your analysis. You should also include a top-level explanation of whatever new supervised learning method you have chosen.

2. Presentation Dates Document: Due 04/15/2020; no more than 1 page in length. A list of dates & times that work for every member in your group to do your Final Project Presentation over Zoom. Your team should decided on *at least* 8 dates & times that work between 04/20/2020 and 05/05/2020. Presentations are expected to be 15 minutes in length, with a 5 minute Q&A session, so please plan on having at least 25 minutes available.

3. Presentation: Due before the last day of finals; all group members must be present. Presentations are expected to be 15 minutes in length and should include well-thought-out slides. They should be rehearsed extensively beforehand to ensure they last for the proper amount of time. Each member of your group should speak for an approximately equal amount. You will be expected to take questions from the instructors following your presentation. Grading will be based on clarity, thoroughness, the degree to which contributions seem equal, and ability to answer questions accurately and concisely.

4. Final Report: Due 05/01/2020; 10-12 pages in length (including graphics). This should include the following sections:

    - *Introduction*: Should include a detailed discussion of your data and your predictors, why your team chose this data, and what research question(s) you will investigate. You should also include a discussion of any data cleaning that was necessary, and any pre-processing you did.
    - *Learning Methods*: Should outline which supervised and unsupervised methods you plan to use and discuss why they are relevant to your project aims. This section should include a very detailed report of the supervised learning technique not covered in class. You should be able to discuss specifically what this new method does, how it does it and how it can be used for your classification aims.
    - *Results*: Should include detailed, informative visuals with a discussion of how you got them.
    - *Discussion*: Should interpret the visuals you created in the context of your research question.
    - *Conclusion*: Based on your analysis, what is your conclusion?

Submission Item	*Weight*
Project Proposal	10%
Presentation Dates	5%
Presentation	40%
Final Report	45%

Data Suggestions	*New Learning Method Suggestions*
Enron Emails	SVM
Twitter Sentiment	Decision Trees
Wine Quality	Neural Network
Pima Indians Diabetes	Naive Bayes

If there are any data or supervised learning techniques that you would like to use for this project that are not listed above, you may do so with prior approval from Murph. Email him at least 24 hours before the Project Proposal is due.