Statistics Senior Project Outline

The data: <https://www.kaggle.com/sammy123/lower-back-pain-symptoms-dataset>

The objective of this project is to model lower back pain data using Logistic Regression, Support Vector Machines (SVM), and Random Forest, and compare how the models perform. Upon initial inspection, the last six variables in the data appear to be random noise. After modeling the data as is (or without the random noise variables), I plan to incrementally add in more random noise and compare how they perform as the signal becomes more buried in noise - the goal here would be to have a plot for each model showing the number of random variables added and the predictability/accuracy of the model. The analysis will be implemented in R.

Tentative: If the above goes well and I want to do something more, I will attempt to implement GUIDE. From the UW Madison website - “GUIDE is a multi-purpose machine learning algorithm for constructing classification and regression trees. It is designed and maintained by Wei-Yin Loh at the University of Wisconsin, Madison. GUIDE stands for *Generalized, Unbiased, Interaction Detection and Estimation.*” <http://www.stat.wisc.edu/~loh/guide.html>

Timeline: (subject to change to comply with Math Department deadlines as needed)

* Dates listed is when item should be completed

September 17: Complete Initial Exploration and Models (LR, SVM, Random Forest) and Outline

September 28: Add in more random noise and compare initial 3 models

(October 12-16: Fall Break)

October 17: Learn GUIDE and observe how it handles increasing noise

October 26: Finalize First Draft of Analysis

November 5: Complete First Draft of Paper

November 16: Finish edits to first draft

(November 21-25: Thanksgiving Break)

November 26: Finish PowerPoint/Practice Presentation

...