Homework 6 - Stat 495

Cassandra Jin

Due Wednesday, October 25th by midnight

Homework 6 - Idea Brainstorm for Project

This homework counts towards the project as a project component (not towards homework).

The idea here is to list several potential statistical topics for your project, as well as your motivation for choosing them, and a little bit about what you'd like to do for the project (e.g. application or simulation, at least, your preliminary thoughts on this). This should include where you've heard about the topic, why you want to learn about each topic, why you are interested in a particular analysis, etc. Remember that the topics you are brainstorming should be relatively new to you. You should brainstorm at least 3 topics - feel free to add more following the outline below.

Some of you may have data sets that you'd like to perform analysis on. That's fine, and you can choose topics that align with the data set analysis, so long as the other requirements for the project are met. Feel free to list such connections below. However, the statistical topic must be the focus here. This allows the project to align with our learning goals for the statistics major.

The feedback provided to you from this brainstorm will help you to prepare a proposal for the project.

Topic Idea 1

- Time series methods, maybe specifically forecasting

We've reached this topic in my Time Series Analysis and Applications class but the material focuses on the theoretical fundamentals of the technique. We haven't explored many applications of the models besides in financial data analysis contexts, and I'd be more interested in both understanding the math beneath the process and studying its applications in fields like astronomy, earthquake prediction, weather forecasting, etc. if possible. I think time series are particularly interesting, as time is a constant axis and variable that everything measures against (constant in the ever-present sense). Change between groups is much of what I've learned in Statistics, rather than change over time, so I'd love to dive deeper into the latter.

Topic Idea 2

- Optimization

I'm considering taking the Optimization class listed in the Math department next semester, but as I slowly dip further into Math and Statistics, I see that the two become more and more inseparable and mutually fundamental. I think it'd be fun to see the topic in the statistical context and application before diving into the math, and I would likely come to have a sufficient understanding of the math building blocks through this project anyway. I've also previously applied for jobs in the DoE which capitalize on optimization and carry the theory into practical, widespread applications, like battery and resource optimization. I'm not sure how much my research for this project might intersect with those fields, but I'd love to find out! Is the topic of just "optimization" way too broad?

Topic Idea 3

- Gaussian mixture model (GMM)

I ran into this topic in a statistical paper that I found for a STAT-225 homework assignment. Upon looking into it a little further, I've found awesome packages allowing one to learn, sample, and estimate GMM's from data. I understand that mixture models are a generalization of k-means clustering to incorporate information about the distributional structure of the data, which I think would be helpful in strengthening my understanding and affinity for the uses of underlying distributions (I'm not very good at working with distributions, I'd say). Also, I haven't done much statistical method research, so seeing the dimensional representation of LASSO in Tibshirani's paper was eye-opening. I've run into visuals of GMM's, and rather than confidence intervals, the technique generates confidence ellipsoids for multivariate models, which just seems very cool at the moment!

Topic Idea 4

- Linear discriminant analysis (LDA) or support vector machine (SVM)

I looked at the survival analysis section of CASI hoping that just based on the name it'd be settled in the field of biostatistics, but that isn't quite what it is. So I searched up methods in biostatistics and found these two. Both pertain to classification and work to either separate the data into effective classes or capture the performance of them. I'd love to try them out on an interesting dataset and understand how they're working on a deeper level.