STAT 2270: Data Mining
University of Pittsburgh

Project Ideas and Possibilities

The project in this course should constitute a deep dive into a particular topic or collection of topics related to data mining and statistical learning. These projects can take a number of different forms, several of which are described below. Students are in no way limited to the possibilities described below, nor to the particular topics/papers listed, but please discuss any substantial deviations with me prior to submitting your proposal. Projects will culminate with a 25-30 minute in-class presentation near the end of the course. Depending on the particular project you choose, a written component may also be required. Further details on the projects and project proposals will be discussed in class.

1. General Topic Not Covered in Class: There are an enormous number of topics, methods, and entire fields of study related to statistical learning; we'll cover only a select handful of these in class. Your project may consist of reviewing a topic of interest that we did not cover. You should provide a readable introduction to the topic as well as important results and in many cases, simulations and/or code. Think of this as a chapter of a textbook. Some interesting topics that we won't cover include:

- Causal Inference
- Tree-related ideas outside of CART (e.g. c-trees, BART)
- Explainable AI
- Density Estimation
- Graphical Models
- Post-selection Inference
- Clustering
- Further topics in high-dimensional regression
- Further topics in deep learning
- High-dimensional testing and random projections
- Bayesian ideas in statistical learning
- Statistical learning methods related to survival analysis (e.g. survival trees)
- Reinforcement Learning

2. Summaries of Recent Paper(s): Statistical learning is a very active area of current research; dozens of papers (at least) are published daily. Many of these papers discuss new and interesting ideas. You may select one such recent paper (or a recent series of papers) and summarize these. You should highlight the main results, discuss why they are important, and attempt to replicate the results. You may also want to experiment with additional simulations and if you're covering a new method, apply that method to real data.

A very few select examples of recent papers that I've found interesting are included as references here – see [1–9]. Certainly you are absolutely by no means limited to these examples. Feel free to ask friends, senior statistics graduate students, or other professors for interesting recent papers that they might suggest. You should also keep up with recent work being done by following uploads to `arxiv.org` – nearly every new and meaningful contribution to statistics is uploaded here, often years prior to its appearance in a journal. You can sign up for an account and also elect to receive daily emails on papers uploaded on topics of your choice.

3. Paper Criticism: As part of the course, we'll discuss a number of important issues that are important to consider when performing various analyses. Unfortunately, the vast majority of these are rarely done in practice, especially in scientific applications. You are welcome to find relatively recent scientific papers which perform subpar analyses and explain their shortcomings. You'll need to find papers in which the data utilized is publicly accessible (which is more difficult than you may think), replicate the work done, demonstrate what *should* have been done, experiment with other approaches, and explain how your findings may affect the stated results of that paper.

4. Data Analysis: Finally, you're welcome to apply the methods and procedures discussed in class to a dataset of your choice. If you choose this route, this should be a very careful and thorough analysis where essentially every reasonable method we discussed in class is applied and every guideline followed. The dataset should be large enough to constitute a substantial challenge and relate to an interesting topic.

References

[1] Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

[2] Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *arXiv preprint arXiv:1807.11408*, 2018.

[3] Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.

[4] Jing Lei, Max G?Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, pages 1–18, 2018.

[5] Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

[6] Scott Powers, Trevor Hastie, Robert Tibshirani, et al. Customized training with an application to mass spectrometric imaging of cancer tissue. *The Annals of Applied Statistics*, 9(4):1709–1725, 2015.

[7] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

[8] Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.

[9] Wenhao Hu, Eric Laber, and Leonard Stefanski. Variable selection using pseudo-variables. *arXiv preprint arXiv:1804.01201*, 2018.