PS279: Special Topics in Methodology: Graphical Models and Statistical Learning

Prerequisite: PS271

Fall 2010
Fridays 12:00-2:50, SSB 353

Langche Zeng
lazeng@ucsd.edu
Office Hours: F 3-5 (wks 1&2), Tu 2-4 (rest of quarter), SSB399

Course Description

This course builds on PS271, and covers some methods and models essential for improving causal inference and prediction, central goals of empirical research. Topics include graphical models, in particular causal Bayesian networks, causal inference aided by causal graphs, and aspects of statistical learning theory and methods, focusing on nonlinear models for supervised learning and model ensemble and selection methods. The class format is a combination of lecture/seminar/workshop.

Course Requirements

Evaluation of course work is based on class participation (such as problem solving, paper presentation/discussion, tool use demonstration), peer feedback/evaluation, and a collaborative research project. It is of vital importance that you make every effort to attend every class meeting. The project should apply some methods/models relevant to the course to some substantive problem in the field of your interest (or develop some new methods.) Your team must consist of members of this class and all papers must be coauthored. The project should be finished, in the form of (mainly the data/methods/results/analysis part of) a potentially publishable research paper, by the last class meeting (12/3), when you will submit the paper and present your work to the class. All class work submissions should be in hard copies unless otherwise indicated.

Software

You are required to use R for this course, and you are strongly encouraged to use LaTeX for typesetting and Emacs (or XEmacs; also useful is ESS) for text editing. All are available on the web for free download.

Books and Internet Resources

--Hastie, Trevor, Robert Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer, 2009. website

--Pearl, Judea. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 2nd ed. 2009. website

These are the main references. We will also draw on other books, papers, some materials at the authors' websites, as well as some excellent tutorials here on some topics. Specific recommendations/assignments will be made along the way and papers for class use will be sent to you a week in advance.

The following book, a good reference for R, can also be found in the campus bookstore:

--Venables, William and Brian D. Ripley. *Modern Applied Statistics with S*. 4th Edition. Springer, 2002. website

Other useful references include:

Statistical learning:

--Bishop, C. *Pattern Recognition and Machine Learning*. Springer, 2006. website

--Ripley, B. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

Causal Inference:

--Morgan, S. and Winship, C. *Counterfactuals and Causal Inference*. Cambridge University Press, 2007.

--Spirtes, P., C. Glymour, and R. Scheines. *Causation, Prediction, and Search* . The MIT Press, 200?

R and Stats "in a nutshell":

--Adler, J. *R in a Nutshell*. O'Reilly, 2009.

--Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

Course Plan

The research project: You should find your coauthor(s) and locate your data set (if you don't already have one) as soon as possible. For comprehensive data resources, see for example the Social Science Data Services on campus, Micah Altman's The Impoverished Social Scientist's Guide to Data Resources , Paul Hensel's International Relations Data Site , and the Comparative Politics Data at U. Michigan. Look for data that likely have a relatively high signal-to-noise ratio. By the end of the 5th week (10/22), I expect a 2 page research proposal outlining the motivation of your project, the methods/models you plan to use, the status of data work, and expected contributions to the literature. Two weeks before the last class meeting (11/19), I expect a concise draft including key findings in tables/figures. We'll discuss the class projects along the way, but at these two points, each team will be paired with another for guaranteed peer feedback. The "final" paper is due by the last class meeting (12/3). Given the nature of the course the term paper should focus on the methods part (data/coding/models/results..), with a couple of additional pages for providing the context. (Along with the final paper, you'll also submit a sheet of paper rating your experience with your teammates and the helpfulness of the peer feedback you have received. This is for my info only, as one factor for the course grades.)

Papers for class use: You are encouraged to suggest 1-3 papers that are particularly relevant to your project to use in class. Please send the suggestions to me in pdf. You will take turn presenting/leading discussions on the papers (focusing on the methods).

Topics: The following is a tentative plan for the course content. We may adjust the plan according to the needs and pace of the class. Updates to the syllabus will be posted here or announced in class. Suggested readings from the main texts are indicated below. Papers for class use will be sent a week in advance.

1. Methodological challenges in prediction and causal inference; Overview of course content; R packages for the course.

Pearl, Epilogue; Hastie et al. chap.1

2. Introduction to graphical models; Causal Bayesian networks.

Pearl, chap.1; Bishop, 8.1-8.3 (online); Hastie et al. 17.1-17.2

3. Overview of supervised learning.

Hastie et al., chap.2, 3.4.1-3.4.3, 3.6, 4.4.5, 5.1.

4. Identification of causal effects; Direct and indirect effects; Causality and structural models; Structure learning.

Pearl, 2, 3, 4.1, 4.5, 5, 6.2.

5. Model assessment and selection: model complexity, regularization, bias-variance trade-off, cross validation; Committee/ensemble methods, model averaging, bagging, boosting.

Hastie, et al. 7.1-7.7, 7.10-7.11, 8.7, 8.8, 10.1-10.3. Bishop, 14.1-14.4

6. Class projects proposal discussion; Tool demonstrations

7. Neural networks

Hastie et al., 11. Bishop, 5.

8. Support vector machines; Kernel methods, Prototypes and Nearest Neighbors

Hastie et al. 4.5.2, 12.1-12.3, 6.6-6.9, 13.1, 13.2.1, 13.2.3, 13.3

9. GAMs; Trees/Random forests; More on boosting and ensemble learning

Hastie et al., 9.1, 9.2, 15.1-15.3, 10.9-10.14, 16.1, 16.3

10. Class project presentations

Communication

I have a profound high frequency hearing loss that's not effectively helped by hearing aids, and I rely on speech reading and email for communication. This will certainly cause some inconvenience to you (e.g., I may ask you to repeat what you say, sometimes more than once; I cannot follow group discussion; In general you cannot reach me by phone; etc.) I appreciate your understanding and accommodation. I will be compiling a class email list. Please send me your email address if it differs from the official UCSD address in the class roster.
