COS 597D: Data driven statistical genetics

Syllabus (9/10/14)

Lecturer: Barbara Engelhardt

Overview

In this course, you will arrive with a specific, data-driven statistical question in mind. Throughout the semester, we will walk through all of the necessary steps to analyze these data and produce independent, original scientific analyses of those data for your project in small groups of classmates with complementary scientific backgrounds. We will focus on reproducible statistical and machine learning methods after selecting, for each project, a data set and analytic problem to drive the plan of study.

I assume that everyone in the course has a background that includes linear algebra and introductory probability and statistics. Programming experience is heavily encouraged, in particular in Python, C, or R. I do not assume any knowledge of biological or genomic data, or more sophisticated machine learning background. I expect homework and reading to take approximately 10 hours per week per student, and I also expect much of this time to be spent among project group members.

Course logistics

The course instructor is Barbara Engelhardt (bee@princeton.edu)

The course meets Mondays and Wednesdays 11-12:20 in CS 302

Office hours will be Wednesdays 2-3:30 and (occasionally) Thursdays 2-3:30

We will use a Google groups email list for out-of-class discussions (please sign up to get on this list): `princetoncos597@googlegroups.com`

We will have a course Dropbox to share materials, data, results, and course information.

Grading

The final grade for the course will consist of 50% class participation and 50% final project. You will find it essential to perform the weekly homework assignments in order to get full marks for both class participation and the final project.

Weekly homework will include reading a scientific paper describing application of a specific statistical method to genomic data, occasionally developing a 30 minute technical presentation for the class, writing down, implementing, or applying statistical methods to data, and presenting results and observations from the application of these methods.

There are no auditors. Those that cannot enroll for a grade (e.g., postdocs) must still complete all of the assignments.

Purpose of the course

Classes will be structured so that you will gain experience in:

- Analysis of large scale data
- Independent and original thinking and research
- Scientific and methodological writing
- Oral presentations of technical material
- Critical reading and technical reviews
- Scientific collaboration and interdisciplinary research

Each class will include the following three components: 1) a presentation from each project group and a class discussion on analysis of results, possible directions, and changes to the current analysis; 2) a student-lead presentation on a statistical or machine learning approach that is driven by the current status of each of the projects based on the weekly reading, and 3) a short discussion of the homework assignment.

The first homework will be for each student to identify a specific data set that they are interested in studying. The second class will include brief presentations from each student, and the identification of project groups with collaborative synergies. Throughout the course, we will step through each part of each project together in order to develop a statistical method for application and analysis of these data. We will adapt the lecture portion of the course to cover specific areas of interest identified from the projects. The course will culminate with drafts of each scientific manuscript being distributed and peer-reviewed among classmates, and the final project manuscript being submitted to journals for peer review.

I expect each project group to coordinate the project closely within the group. I also expect the project manuscripts to be submission-ready at the time of the project due date (January 19, 2015).

Lecture outline (subject to change)

Week	Topic	Reading	Homework
9/10	Introduction	N/A	N/A
9/15	Data Imputation	[Troyanskaya *et al.* 2001]	Identify data, project ideas
9/22	Data exploration: clustering	[Eisen *et al.* 1998]	Clean data
9/29	Data exploration: factor analysis and PCA	[Engelhardt & Stephens 2010]	Visualize data
10/6	Hypothesis testing	[Storey & Tibshirani 2003]	Define model
10/13	Linear models & association testing	[Stephens & Balding 2009]	Related work
10/20	Latent Dirichlet allocation	[Pritchard, Stephens, Donnelly 2000]	Parameter estimation
11/3	Model selection and sparsity	[Tibshirani 1996]	Simulation and validation
11/10	Discrete time series, HMMs	[Cawley & Pachter 2003]	Application to data
11/17	Time series and Gaussian Processes	[Roberts *et al.* 2013]	Model validation and replication
11/24	Undirected networks and Markov random fields	[Schafer and Strimmer 2005]	Results presentation
12/1	Coalescent processes	[Zollner & Pritchard 2005]	Bigger picture
12/8	Dirichlet processes	[Huelsenbeck & Andolfatto 2007]	Manuscript reviews
1/19			Final projects due

Possible projects

- Online variational methods for HDP, incorporating multi-scale learning
- Kernels for GPs that encode genomic regulatory elements to facilitate association mapping
- Test for variance eQTLs
- Tensor regression models with GP prior for time series data
- Subspace models for population structure
- MCMC sampling to compute posterior distributions of selection pressure at a specific genomic locus.

Resources

Software

Reproducibility and open access in research and methods development is essential. In this course, we will be developing our code and writing our papers with GitHub and version control. We will be able to make these methods publicly available when the manuscripts are submitted. We will write documents and perform some data analysis and visualization with LaTeX, KnitR, and iPython in order to produce reproducible, statistically clear manuscripts and analysis pipelines. We will post the submitted manuscripts on a preprint server to allow you to reference them in your work before they are peer-reviewed and published in journals.

- GitHub and Git version control
- KnitR
- iPython
- LaTeX
- Preprint servers: BioRxiv, arXiv

Textbooks

- Murphy "Machine learning: A Probabilistic Perspective"
- Bishop "Pattern Recognition and Machine Learning"
- Hastie, Tibshirani, Friedman "Elements of Statistical Learning"

Help with writing

- Silvia "How to write a lot"
- Belcher "Writing your Journal Article in 12 Weeks"
- Zimmer "The index of banned words" (Discover Magazine blog post)

The Princeton Writing Program has an initiative called "Writing in Science & Engineering" that includes half-term courses for graduate students and one-on-one consultations to allow you to craft your manuscript and language. I encourage you and your group to sign up for a session to get feedback on your project manuscript.