COS 513: Foundations of Probabilistic Modeling

Syllabus (9/10/15)

Lecturer: Barbara Engelhardt

Overview

In this course, you will dive deep into the exploration of machine learning with a specific question in mind, with the idea that a carefully selected (or crafted) statistical model and associated method paired with an existing data set will be able to answer that question. 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 models and 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 a more sophisticated machine learning background. I expect homework and reading to take approximately 10 hours per week per student, 90 minutes of which is reading the assigned paper and the rest as part of the project development.

Course logistics

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

The course meets Mondays and Wednesdays 1:30 - 2:50pm in CS 302.

Office hours will be Wednesdays 3-4pm in COS 322.

Greg Darnell (gdarnell@princeton.edu) is the course TA. His office hours will be Tuesdays 1:30-3:30pm.

We will use Piazza for out-of-class discussions (please sign up).

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

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 machine learning paper, developing (with your group) a 20 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

The primary purpose of the class is to learn about more advanced statistical machine learning models and methods than are taught at the undergraduate level. Classes will be structured so that you will gain additional 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

Monday's class will have a 40 minute presentation by a pair of students about the assigned paper for that week. The remaining 40 minutes of Monday's class, and class on Wednesday, will consist of a presentation from each project group and a class discussion on analysis of results, possible directions, and changes to the current analysis. Wednesday's class will conclude with 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 course readings to cover specific areas of interest identified from the projects. The course will culminate with drafts of each 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 12, 2015). There are opportunities to submit these manuscripts to conferences, and, if they are accepted, I will support your registration and travel. In particular, ICML 2016 submission deadline is February 15th 2016, and ISMB 2016 will be around that time as well.

Lecture outline (subject to change)

Resources

Software

Reproducibility and open access in research and methods development is essential to making science and technology useful. In this course, I strongly recommend the following (types of) approaches to your research. For example, it is useful to develop code and write papers with GitHub and version control. You will be able

Table 1: Syllabus. Asterisk indicates written work due.

Week	Topic	Reading	Homework
9/14	Introduction	N/A	N/A
9/21	Linear models	[Cunningham & Ghahramani 2015]	Identify data, project ideas
9/28	Nonlinear dimension reduction	[Roweis & Saul 2000]	Visualize data; define problem
10/5	Infinite mixture models	[Rasmussen 1999]	Define modeling approach
10/12	Hierarchical Dirichlet processes	[Teh, Jordan, Beal, Blei 2004]	Parameter estimation
10/19	Structured Dirichlet processes	[Rao & Teh 2009]	Related work
10/26	Indian buffet process	[Griffiths & Ghahramani 2005]	Simulation and validation*
11/9	Gaussian processes	[Williams & Rasmussen 1996]	Application to data
11/16	Determinantal point processes	[Kulesza & Taskar 2010]	Analyze results
11/23	Bayesian optimization	[Snoek, Larochelle, Adams 2012]	Model validation and replication
11/30	Heavy-tailed process priors	[Wauthier & Jordan 2010]	Results presentation*
12/7	Multiscale Gaussian processes	[Fox & Dunson 2012]	Bigger picture
12/14	Exchangeable random graphs	[Lloyd, Orbanz, Ghahramani, Roy 2012]	Manuscript reviews*
1/12			Dean's Date*

to make these methods publicly available when the manuscripts are submitted. Some proportion of the data analysis and all of the visualization can be performed 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 or conferences.

- 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.

Learning more about Machine Learning

- Independent work!
- Local, relevant talks at Princeton (stat-ml-talks): sign up at `https://lists.cs.princeton.edu/mailman/listinfo/ml-stat-talks`
- CSML reading group: send an email to listserv@princeton.edu, and in the body of the email include this line only:

  SUB CSML-reading
- Read NIPS, ICML, UAI, AI-Stats, JMLR and other conference proceedings and journals