COS424: Interacting with Data
Spring, 2012

(This document was updated on April 26, 2012.)

Course description

Problems about data abound. Here are some examples:

- Netflix collects ratings about movies from millions of its users. From these ratings, how can they predict which movies a user will like?
- JSTOR scans and runs OCR software on millions of scholarly articles. Scholars want to browse and explore their collection. How should JSTOR organize it?
- A biologist has collected hundreds of thousands of measurements about the genotypes and traits of a large population. She would like to run a new experiment. Can she make a hypothesis about which genes are associated with which traits?
- Google sends and receives hundreds of millions of email messages each day. Are some of them spam? Which advertisements should they show next to each one?

Data analysis is central to many modern problems in science, industry and culture. Scientists and engineers have to be fluent in thinking about how to solve modern data analysis problems. This class puts you on the path towards that fluency.

In this course, we will learn about a suite of tools in modern data analysis: when to use them, the kinds of assumptions they make about data, their capabilities, and their limitations. More importantly, we will learn about the language for and process of solving data analysis problems. On completing the course, you will be able to learn about a new tool, apply it data, and understand the meaning of the result.

Administration

Lectures
Tuesdays and Thursdays, 1:30PM-2:50PM
CS Room 104 (Large auditorium)

Recitation Office Hour
Thursdays, 4:40PM-5:30PM
CS Room 104 (Large auditorium)

Instructor
Prof. David Blei
Office hours: Tuesday 3:00PM-4:00PM; sign up at *http://wass.princeton.edu/*
blei@cs.princeton.edu

Lecturer
Dr. Xiaoyan Li
Office hours: Wednesday 10:00AM-12:00PM
xiaoyan@princeton.edu

Teaching assistant
Yuhui Luo
Office hours: Friday 12:30PM-1:30PM
yuhuiluo@princeton.edu

Piazza
We will use Piazza to host all communication.

1. Sign up for the Piazza site at *http://piazza.com/class#spring2012/cos424*.
2. Use it to ask and answer questions about the course.
3. Use it to communicate with the instructors privately.
4. Use it to receive important announcements from the instructors.

Prerequisites

The prerequisite knowledge is calculus, linear algebra, computer programming, and some exposure to probability and/or statistics. Contact Prof. Blei if you have concerns about your prerequisite coursework.

Programming with R

We will use R, which is a powerful open-source platform for statistical computing and visualization. We will hold a special session about learning R in the beginning of the semester.

You can download R for many platforms at *http://www.r-project.org/*.

To get started with R, see *Introductory Statistics with R* by Peter Daalgard. It is available as a PDF from the Princeton Library.

Course requirements

There are three kinds of work required for the course.

- **Homeworks.** (50%) There are three homeworks due throughout the semester. These contain written questions and programming and data analysis in R. Throughout the semester, students can use seven late days of automatic extension. (Weekends are fully counted.) After using all seven days, late homeworks are not accepted.

- **Reading responses.** (10%) You are to hand in a response to the reading each week. They should be typed, and can be from one paragraph to one page long. We read them, but they are not graded. Each student must hand in his or her response at the beginning of lecture on Thursday. No reading response will be accepted late or at another time.

- **Final project.** (40%) The class project constitutes about one month of work. For the project, we expect you to undertake a thorough piece of applied data analysis and clearly report your findings in a written report and poster presentation. You can work alone, but we encourage you to work in groups of two or three.

Failure to complete any significant component of the course may result in a D or F.

Important Dates

14-Feb	HW 1 out
28-Feb	HW 1 due; HW 2 out
13-Mar	HW 2 due
29-Mar	HW 3 out
05-Apr	Final project proposal due
12-Apr	HW 3 due
15-May	Final project due

Syllabus and Readings

Most readings come from

- Murphy, K. *Machine Learning: A Probabilistic Approach.* MIT, in press. (MLAPA)

- Hastie, T., Tibshirani, R. and Freedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Edition, Springer, 2009. (ESL)

- Bishop, C. *Pattern Recognition and Machine Learning.* Springer-Verlag, 2006. (PRML)

When available, readings will be posted on the course website. Otherwise, they will be handed out in class or posted to blackboard. (The Hastie et al. book is available online as a PDF.)

#	Date			Subject	Reading
01	T	07	Feb	Introduction	MLAPA Ch 1; PRML Ch 2
02	R	09	Feb	Probability and statistics review I	[Opt] PRML Ch 1.2; MLAPA Ch 2; PRML Ch 2
03	T	14	Feb	Probability and statistics review II	MLAPA Ch 5.1-5.2
04	R	16	Feb	Probabilistic classification	
05	T	21	Feb	Probabilistic classification	PRML Ch 9.1-9.2
06	R	23	Feb	Naive Bayes	[Opt] ESL Ch 14.3.4-14.3.11
07	T	28	Feb	K-Means	PRML Ch 9.3-9.4
08	R	01	Mar	Mixture models	
09	T	06	Mar	Expectation-maximization I	PRML Ch 13.1-13.2
10	R	08	Mar	Expectation-maximization II	
11	T	13	Mar	Hidden Markov models I	[No reading]
12	R	15	Mar	Hidden Markov models II	
—	T	20	Mar	*Spring break*	
—	R	22	Mar	*Spring break*	
13	T	27	Mar	Linear regression I	ESL Ch 3.1-3.2
14	R	29	Mar	Linear regression II	
15	T	03	Apr	Regularized linear regression	MLAPA 1.2.9, 7.4.1-7.4.3; ESL Ch 3.4
16	R	05	Apr	Logistic regression	
17	T	10	Apr	Generalized linear models I	McCullagh and Nelder, Ch 2
18	R	12	Apr	Generalized linear models II	
19	T	17	Apr	Dimension reduction I	PRML Ch 12.1-12.2, 13.3
20	R	19	Apr	Dimension reduction II	
21	T	24	Apr	The Kalman Filter and NMF	Lee and Seung (1999); Spall (2003) Ch 4
22	R	26	Apr	Scalable machine learning	
23	T	01	May	Probabilistic topic models	Blei (2011)
24	R	03	May	Summary and discussion	