COS424/SML302: Fundamentals of Machine Learning

Spring 2016

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 search 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. Can she make a hypothesis about which genotypes regulate 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 user?

Data analysis is central to many modern problems in science, industry, and culture. In science and and engineering, it is essential to be fluent in solving 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 approach the analysis of large, complex data sets. In particular, you will be able to, given a data set, define the data analysis problem, learn about new methods, apply these methods to data, and understand the meaning of the results.

Administration

Lectures: Tuesdays and Thursdays, 1:30PM-2:50PM
Friend Center 101

Instructor: Prof. Barbara Engelhardt
Office hours: Thursday 3:00PM-4:00PM
Email: bee@princeton.edu

Lecturer: Dr. Xiaoyan Li
Office hours: Tuesday 10:00 - 12:00 noon at 221 Nassau Street, Room 104
Email: xiaoyan@princeton.edu

Teaching assistants: TBA

Piazza

We will use Piazza to host all communication.

1. Sign up for the Piazza site at `piazza.com/princeton/spring2016/cos424sml302`.
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.
5. Use it to download course notes, programming assignments, and reading materials.

Prerequisites

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

Course programming

Although you may use whatever programming language you choose for these problems in data analysis, we suggest R or Python. R is a powerful open-source platform for statistical computing and visualization. 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.

KnitR files that illustrate how to generate many of the visualizations in R presented in class will accompany the lecture slides on the Piazza course website.

Python is another good option for a programming language for the programming assignments and the final project, as it has emerged as an easy and fast platform to develop many machine learning methods. In particular, the library SciKit-Learn has a large number of ML methods and approaches for use (including regression, classification, cross validation, etc.).

Writing with LaTeX

We will use LaTeXto write the homework assignments and the final project. We will post templates for the homework assignments and the final project on the website. To jointly edit a single LaTeXfile among collaborators, consider using *ShareLaTeX*, *Overleaf*, or *Git* (all free).

Course requirements

There are three kinds of work required for the course.

- Homework assignments. (60%) There are three homework assignments due throughout the semester. These will all be the analysis of a specific data set, disseminated with the homework description, using methods discussed in class; the deliverables will be a four page write up of the data, analyses, and results (see Piazza page for the write-up template and an example write-up). All homeworks may be done alone or in pairs. If you choose to pair with another classmate, you may not pair with that same classmate for more than one of the homeworks. Because of the nature of the team structure, late days are given at the discretion of the professor.
- Reading responses. (10%) There will be weekly responses to the reading due in class printed out on the LaTeXreading response template we provide. These will consist of a paragraph of your thoughts about and reactions to the assigned reading material for the week. There are 12 weeks of class, and you are expected to hand in 10 of these reading responses (i.e., you are excused from writing two of them with no penalty). There are no extensions and no late days.

- Final project. (30%) The class project will be either a dramatic extension of one of the three homework projects in the course, or your own work on the development or application of machine learning methods to a large data set. You will turn in an eight-page write-up of your project on Dean's date; on May 6th, you will present your work at a poster session for the Princeton community. You may work alone on your project, but we encourage you to work in groups of up to four; you may pair with a classmate that you worked with on a previous assignment for the project.

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

Important Dates

- 2-Feb HW 1 out
- Precept 1: Feb 8 at 7:30PM COS105
- 23-Feb HW 1 due; HW 2 out
- Precept 2: Feb 29 at 7:30PM COS105
- 22-Mar HW 2 due; HW 3 out
- Precept 3: Mar 28 at 7:30PM COS105
- 12-Apr HW 3 due
- 9-May 9am-4pm Final project poster session
- 10-May Final project due (Dean's Date)

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)

The readings will be posted to Blackboard.

Readings below are tentative. Please see Blackboard for updated information.

Lecture	Date	Subject	Reading
01	T 02 Feb	Introduction	MLAPA Ch 1
02	R 04 Feb	Probability and statistics review	MLAPA Ch 2; [Opt] MLAPA Ch 3.1-3.4
03	T 09 Feb	Graphical models	MLAPA Ch 10.1-10.2, 10.4
04	R 11 Feb	Probabilistic classification	MLAPA Ch 3.5
05	T 16 Feb	Features and kernels	MLAPA 14.1-14.2
06	R 18 Feb	Kernel classifiers	MLAPA 14.3-14.5
07	T 23 Feb	Topics in machine learning	MLAPA Ch 6.2.1
08	R 25 Feb	Linear regression	ESL Ch 3.1-3.2
09	T 01 Mar	Regularized linear regression	ESL Ch 3.4
10	R 03 Mar	Logistic regression	MLAPA 8.1-8.2
11	T 08 Mar	Generalized linear models	McCullagh and Nelder, Ch 2
12	R 10 Mar	Optimization	MLAPA 8.3 & 8.5
	T 16 Mar	Spring break	
	R 18 Mar	Spring break	
13	T 22 Mar	K-Means	MLAPA 11.1-11.3
14	R 24 Mar	Mixture models	
15	T 29 Mar	Expectation-maximization	MLAPA 11.4-11.6
16	R 31 Mar	Hidden Markov models	MLAPA 17.1-17.2
17	T 05 Apr	Dimension reduction and PCA	MLAPA Ch 12.1-12.2
18	R 07 Apr	Factor analysis	
19	T 12 Apr	Probabilistic topic models	Blei (2011)
20	R 14 Apr	Nonparametric Bayesian distributions	
21	T 19 Apr	Gaussian process regression	
22	R 21 Apr	Markov chain Monte Carlo	
23	T 26 Apr	Scalable machine learning	MLAPA 21.1-21.5 (not 21.4)
24	R 28 Apr	Summary and discussion	
	M 09 May	Poster session	
	T 10 May	Dean's Date	Projects due (5pm)