Cos 524 syllabus

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 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.

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

We require the code for the data analysis homework assignments to be done in Python. Python 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, classi cation, cross validation, etc.).

For visualization and downstream analysis of the results, R is a powerful open-source platform for statis-tical 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. There are a number of excellent packages for data visualization in R, such as ggplot2.

Writing with LATEX

We will use LATEX to 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 LATEX document among collaborators, consider using Overleaf or Git (all free through Princeton).

Course requirements

There are three kinds of work required for the course.

1. *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 five page write up of the data, analyses, one page of methods, and results (see Canvas page for the write-up template and an example write-up), and the Python code used to analyze the data. All homework assignments may be done alone or in pairs. Because of the nature of the team structure, late days are given at the discretion of the professor; at the start, each person will have exactly one week-long extension available, so please plan accordingly.
2. *Reading quizzes online. (10%)* There will be weekly multiple choice question quizzes online about the reading that week. Each quiz will close before the start of class on Thursdays. These will consist of a few short questions about the assigned reading material for the week. There are 12 weeks of class, and you are expected to complete 10 of these reading quizzes (i.e., you are excused from two of them with no penalty). There are no extensions and no late days. There is no extra credit for completing more than ten.
3. *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 5th by 5pm EST; on May 3rd, 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.

** *For COS 524 only*: The 10% of your grade devoted to reading quizzes will be changed to be participation in precepts, where the readings are discussed. There are 12 weeks of class,
and you are expected to complete 10 of these reading responses (i.e., you are excused from two of them with no penalty). There are no extensions and no late days. There is no extra credit for completing more than ten.

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

Syllabus and Readings

Most readings come from: Murphy, K. Machine Learning: A Probabilistic Approach. MIT, in press. (MLAPA); the e-book is posted online in the reserve section on Canvas.

-- Hastie, T., Tibshirani, R. and Freedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition, Springer, 2009. (ESL); the e-book is posted online in the reserve section on Canvas.
-- Bishop, C. Pattern Recognition and Machine Learning. Springer-Verlag, 2006. (PRML) the e-book is available on the web: https://www.microsoft.com/en-us/research/people/cmbishop/

Modifications for COS 524

COS 524 will require lectures and readings, and an additional set of readings related to the lecture material will be discussed in the precepts for graduate students. Graduate students are expected to prepare short responses to the papers assigned for precepts. The quizzes will no longer be required, and instead 10% of your grade will be for participation in the precepts and completion of the graduate-level reading assignments and responses. The additional readings for COS524 will be placed in the COS524 Reading folder in the Files section on Canvas.

Schedule

Lecture/ Precept	Week	Subject	Reading
L01 L02 P01	01 Feb 01 Feb	Introduction Probability and statistics review Homework1 get started	MLAPP Ch 1 MLAPP Ch 2; [Opt] MLAPP Ch 3.1-3.4 Reading for COS524: 50 Years of Test (Un)fairness: Lessons for Machine Learning
L03 L04 P02	08 Feb 08 Feb	Graphical models Probabilistic classification Writing a good report & Cross-validation	MLAPP Ch 10.1-10.2, 10.4 MLAPP Ch 3.5 Reading for COS524: On Discriminative vs. Generative Classifier: A Comparison of Logistic Regression and Naive Bayes
L05 L06 P03	15 Feb 15 Feb	Features and kernels Kernel classifiers Evaluation metrics and feature selection for classification	MLAPP 14.1-14.2 MLAPP 14.3-14.5 Reading for COS524: Support Vector Networks

L07 L08 P04	22 Feb 22 Feb	Linear regression Regularized linear regression Homework2 get started	MLAPP 7.1-7.3; [Opt] ESL Ch 3.1-3.2 MLAPP 7.5.1,7.6.1,7.6.2; [Opt] ESL Ch 3.4 Reading for COS524: Statistical Modeling: The Two Cultures
L09 L10 P05	01 Mar 01 Mar	Logistic regression Generalized linear models Regularization in linear models and Hyperparameter tuning using corss-validation	MLAPP 8.1-8.2 MLAPP 9.1-9.3.2; [Opt] McCullagh and Nelder, Ch 2 Reading for COS524: Wide and Deep Learning for Recommender Systems
L11 L12 P06	08 Mar 08 Mar	K-Means Mixture models Imputation methods and Bootstrapping	MLAPP 11.1-11.3 Reading for COS524: Refining Initial Points for K-Means Clustering
	15 Mar	Spring break	
L13	17 Mar	Optimization No precept this week.	MLAPP 8.3 & 8.5 Reading for COS524: Measuring the predictability of life outcomes with a scientific mass collaboration

L14 L15 P07	22 Mar 22 Mar	Expectation-maximization Hidden Markov models Homework3 get started	MLAPP 11.4-11.6 MLAPP 17.1-17.2 Reading for COS524: Maximum Likelihood from Incomplete Data via the EM Algorithm
L16 L17 P08	29 Mar 29 Mar	Dimension reduction and PCA Factor analysis PCA, SVD, and NMF, Suggestions on HW3	MLAPP Ch 12.1-12.2 Reading for COS524: EM Algorithms for PCA and SPCA
L18 L19 P09	05 Apr 05 Apr	Probabilistic topic models Communities in networks LDA, graph/network properties and analysis	Blei (2011) Airoldi et al. (2008) Reading for COS524: Inference of Population Structure Using Multilocus Genotype Data
L20 L21 No precept	12 Apr 12 Apr	Dirichlet process Gaussian process regression	MLAPP 25.2 Roberts et al. 2013 Reading for COS524:Bayesian nonparametric Models
L22 L23 no precept	19 Apr 19 Apr	Markov chain Monte Carlo Scalable machine learning	MLAPP 23 (optional), 24.1-24.3.5 MLAPP 21.1-21.5 (not 21.4) Reading for COS524: Variational Inference: A Review for Statisticians

L24 no precept	26 Apr	Summary and discussion	
	M 03 May W 05 May	Poster session Dean's Date	 Project due (5pm EST)