COS424/SML302: Fundamentals of Machine Learning

Spring 2018

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.

Administration

Lectures: Tuesdays and Thursdays, 11:00PM-12:20PM
Friend Center 101

Precepts: Fridays, 11:00PM-11:50AM or 1:30-2:20PM
Friend Center 101 or Computer Science 105

Instructor: Prof. Barbara Engelhardt
Office hours: Tuesdays 2:30PM-3:30PM in COS 302
Email: bee@princeton.edu

Lecturer: Dr. Xiaoyan Li
Office hours: See Piazza for times, location at 221 Nassau Street, Room 104
Email: xiaoyan@princeton.edu

Teaching assistants:
Brian Jo, Ariel Gewirtz, Greg Gunderson, Archit Verma. See Piazza for contact information and office hours.

Piazza

We will use Piazza to host all communication.

1. Sign up for the Piazza site at `piazza.com/princeton/spring2018/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.
6. Use it to access weekly quizzes.

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

For the first time, we will 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, classification, cross validation, etc.).

For visualization and downstream analysis of the results, 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. There are a number of excellent packages for data visualization in R, such as `ggplot2`.

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 five page write up of the data, analyses, one page of methods, and results (see Piazza 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. If you choose to pair with another classmate, you may not pair with that same classmate for more than one of the homework assignments. Because of the nature of the team structure, late days are given at the discretion of the professor.

- 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.
- 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 15th by 5pm; on May 14th, 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

- 6-Feb HW 1 out
- 27-Feb HW 1 due; HW 2 out
- 27-Mar HW 2 due; HW 3 out
- 17-Apr HW 3 due
- 19-Apr Project proposals due
- 14-May 9am-4pm Final project poster session
- 15-May Final project due (Dean's Date)

Grading and Collaboration Policies

As mentioned above, each homework assignment may be done alone or in pairs, although no more than two people may work on a homework assignment together. If individuals or groups discuss important details of their approach outside of these groups, please acknowledge this discussion in the Acknowledgements section of the homework. No pair of students may re-pair to work on a later homework assignment without the instructor's permission beforehand.

The final project may be done with zero to three partners (i.e., the largest teams may be of size four); we scale our expectations in the deliverables with the size of the groups. You are allowed to collaborate with people with whom you worked on a previous homework assignment.

The late day policy is that, across the three assignments, you may take one week of additional time to hand in one of the assignments. The week may not be split among assignments, but must be used on only one. Notice that the additional week will be taken on a specific assignment must be requested before the due date of that assignment. For every day that an assignment is late, 5% of the grade (out of 100%) will be taken off. When one person in a pair has used their late days, and another has not, we cannot extend the late days to both partners, and the partners may receive different grades on the assignment – this is something to consider when finding assignment partners. No late days will be given for any of the 12 reading quizzes – due every Thursday during the semester before class – nor for the final project – due Dean's Date at 5pm. However, only 10 of the 12 quizzes are required – you may skip any two you choose with no penalty.

Note that we have software for examining the similarity of code and homework/project write ups, and we will be running these on everything that is handed in. If we find that there is substantial replication of code or of text in the write up, we will give that assignment a 0 grade and send the assignment and evidence of the violation to the appropriate honor code committee.

If you feel your work has been graded unfairly, please bring the full set of graded work for the semester — not just the graded work in question — to both Profs Li and Engelhardt. We will work with you to make sure the grades are fair and appropriate, but, to be clear, this does not always mean that the grades will improve. Often problems in work that the graders missed will be identified as well as the opposite.

Syllabus and Readings

Most readings come from: Murphy, K. *Machine Learning: A Probabilistic Perspective.* MIT, in press. (MLAPP); the e-book is posted online on Blackboard. Additional resources for material will be posted on Slack, as well as the slides from lecture.

- 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)

Readings below are tentative. Please see Piazza for any changes to the reading.

Lecture	Date	Subject	Reading
01	T 06 Feb	Introduction	MLAPP Ch 1
02	R 08 Feb	Probability and statistics review	MLAPP Ch 2; [Opt] MLAPP Ch 3.1-3.4
03	T 13 Feb	Graphical models	MLAPP Ch 10.1-10.2, 10.4
04	R 17 Feb	Probabilistic classification	MLAPP Ch 3.5
05	T 20 Feb	Features and kernels	MLAPP 14.1-14.2
06	R 22 Feb	Kernel classifiers	MLAPP 14.3-14.5
07	T 27 Feb	Linear regression	MLAPP 7.1-7.3; [Opt] ESL Ch 3.1-3.2
08	R 01 Mar	Regularized linear regression	MLAPP 7.5.1,7.6.1,7.6.2; [Opt] ESL Ch 3.4
09	T 06 Mar	Logistic regression	MLAPP 8.1-8.2
10	R 08 Mar	Generalized linear models	MLAPP 9.1-9.3.2; [Opt] McCullagh and Nelder, Ch 2
11	T 13 Mar	K-Means	MLAPP 11.1-11.3
12	R 15 Mar	Mixture models	
	T 21 Mar	Spring break	
	R 23 Mar	Spring break	
13	T 27 Mar	Optimization	MLAPP 8.3 & 8.5
14	R 29 Mar	Expectation-maximization	MLAPP 11.4-11.6
15	T 03 Apr	Hidden Markov models	MLAPP 17.1-17.2
16	R 05 Apr	Dimension reduction and PCA	MLAPP Ch 12.1-12.2
17	T 10 Apr	Factor analysis	
18	R 12 Apr	Probabilistic topic models	Blei (2011)
19	T 17 Apr	Communities in networks	Airoldi et al. (2008)
20	R 19 Apr	Dirichlet processes	MLAPP 25.2
21	T 24 Apr	Gaussian process regression	Roberts et al. 2013
22	R 26 Apr	Markov chain Monte Carlo	MLAPP 23 (optional), 24.1-24.3.5
23	T 01 May	Scalable machine learning	MLAPP 21.1-21.5 (not 21.4)
24	R 03 May	Summary and discussion	
	M 14 May	Poster session	
	T 15 May	Dean's Date	Projects due (5pm)