COS424: Interacting with Data

Spring, 2015

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 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: Tuesday 3:00PM-4:00PM; sign up at http://wass.princeton.edu/
Email: bee@cs.princeton.edu

Lecturer: Dr. Xiaoyan Li
Office hours: Thursday 3:00PM-5:00PM at 221 Nassau Street, Room 104
Email: xiaoyan@princeton.edu

Teaching assistant: Pingmei Xu
Office hours: Friday 11:00AM-12:00PM
Email: pingmeix@princeton.edu
Teaching assistant: Bianca Dumitrascu  
*Office hours:* Monday 10:00AM-11:00AM  
*Email:* biancad@princeton.edu

Teaching assistant: Wei Hao  
*Office hours:* Wednesday 11:00AM-12:00AM  
*Email:* whao@princeton.edu

**Piazza**

We will use Piazza to host all communication.

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. 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/](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.). Cython also may be used to substantially speed up code for large data sets.

**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 $\LaTeX$ file among collaborators, consider using *ShareLaTeX, WriteLaTeX*, or *Git* (all free).

**Course requirements**

There are three kinds of work required for the course.
• Homeworks. (60%) There are three homeworks 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.

• In-class quizzes. (10%) There will be occasional five minute quizzes in class to evaluate whether or not you have done the reading assignments.

• 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 11th, 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 three; 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

• 3-Feb HW 1 out
• 26-Feb HW 1 due; HW 2 out
• 2-Apr HW 2 due; HW 3 out
• 30-Apr HW 3 due
• 11-May 10am-2pm Final project poster session
• 12-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)


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

The readings will be posted to Blackboard or Piazza.
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Subject</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>T 03 Feb</td>
<td>Introduction</td>
<td>MLAPA Ch 1</td>
</tr>
<tr>
<td>02</td>
<td>R 05 Feb</td>
<td>Probability and statistics review</td>
<td>MLAPA Ch 2; [Opt] MLAPA Ch 3.1-3.4</td>
</tr>
<tr>
<td>03</td>
<td>T 10 Feb</td>
<td>Graphical models</td>
<td>MLAPA Ch 10.1-10.2, 10.4</td>
</tr>
<tr>
<td>04</td>
<td>R 12 Feb</td>
<td>Probabilistic classification</td>
<td>MLAPA Ch 3.5</td>
</tr>
<tr>
<td>05</td>
<td>T 17 Feb</td>
<td>Features and kernels</td>
<td>MLAPA 14.1-14.2</td>
</tr>
<tr>
<td>06</td>
<td>R 19 Feb</td>
<td>Kernel classifiers</td>
<td>MLAPA 14.3-14.5</td>
</tr>
<tr>
<td>07</td>
<td>T 24 Feb</td>
<td>Topics in machine learning</td>
<td>MLAPA Ch 6.2.1</td>
</tr>
<tr>
<td>08</td>
<td>R 26 Feb</td>
<td>Linear regression</td>
<td>ESL Ch 3.1-3.2</td>
</tr>
<tr>
<td>09</td>
<td>T 03 Mar</td>
<td>Regularized linear regression</td>
<td>ESL Ch 3.4</td>
</tr>
<tr>
<td>10</td>
<td>R 05 Mar</td>
<td>Logistic regression</td>
<td>MLAPA 8.1-8.2</td>
</tr>
<tr>
<td>11</td>
<td>T 10 Mar</td>
<td>Generalized linear models</td>
<td>McCullagh and Nelder, Ch 2</td>
</tr>
<tr>
<td>12</td>
<td>R 12 Mar</td>
<td>Optimization</td>
<td>MLAPA 8.3 &amp; 8.5</td>
</tr>
<tr>
<td></td>
<td>T 17 Mar</td>
<td>Spring break</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R 19 Mar</td>
<td>Spring break</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>R 26 Mar</td>
<td>Mixture models</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>T 31 Mar</td>
<td>Expectation-maximization I</td>
<td>MLAPA 11.4-11.6</td>
</tr>
<tr>
<td>16</td>
<td>R 02 Apr</td>
<td>Expectation-maximization II</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>T 07 Apr</td>
<td>Hidden Markov models I</td>
<td>MLAPA 17.1-17.2</td>
</tr>
<tr>
<td>18</td>
<td>R 09 Apr</td>
<td>Baum Welch and Viterbi</td>
<td>MLAPA 17.4</td>
</tr>
<tr>
<td>19</td>
<td>T 14 Apr</td>
<td>Dimension reduction and PCA</td>
<td>MLAPA Ch 12.1-12.2</td>
</tr>
<tr>
<td>20</td>
<td>R 16 Apr</td>
<td>Factor analysis</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>T 21 Apr</td>
<td>Probabilistic topic models</td>
<td>Blei (2011)</td>
</tr>
<tr>
<td>22</td>
<td>R 23 Apr</td>
<td>Active learning</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>T 28 Apr</td>
<td>Scalable machine learning</td>
<td>MLAPA 21.1-21.5 (not 21.4)</td>
</tr>
<tr>
<td>24</td>
<td>R 30 Apr</td>
<td>Summary and discussion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M 11 May</td>
<td>Poster session</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T 12 May</td>
<td>Dean’s Date</td>
<td>Projects due (5pm)</td>
</tr>
</tbody>
</table>