17.806: Quantitative Research Methods IV

Spring 2016

Instructor: In Song Kim
TA: Yiqing Xu

Department of Political Science
MIT

1 Contact Information

<table>
<thead>
<tr>
<th></th>
<th>In Song</th>
<th>Yiqing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office:</td>
<td>E53–407</td>
<td>E53–408</td>
</tr>
<tr>
<td>Email:</td>
<td><a href="mailto:insong@mit.edu">insong@mit.edu</a></td>
<td><a href="mailto:xyq@mit.edu">xyq@mit.edu</a></td>
</tr>
<tr>
<td>Phone:</td>
<td>617–253–3138</td>
<td>617-866-0547</td>
</tr>
</tbody>
</table>

2 Logistics

- Lectures: Tuesdays and Thursdays 4:00–5:30pm, E53–485
- Recitations: Fridays 10:30–11:30am, E53–438
- In Song’s office hours: Make an appointment
- Yiqing’s office hours: Friday 11:30am–12:30pm

Note that the first class meets on February 2. We will hold no class on February 16 (Presidents’ Day) and on April 19 (Patriots Day). Last day of class is May 12.

3 Course Description

This course is the fourth and final course in the quantitative methods sequence at the MIT political science department. The course covers various advanced topics in applied statistics, including those that have only recently been developed in the methodological literature and are yet to be widely applied in political science. The topics for this year are organized into three broad areas: (1) research computing, where we introduce various techniques for automated data collection, visualization, and analysis of massive datasets; (2) statistical learning, where we provide an overview of machine learning algorithms for predictive and descriptive inference; and (3) finite mixture models (e.g., Latent Dirichlet allocation for text analysis), as well as a variety of estimation techniques such as EM Algorithm and Variational Inference.
4 Prerequisites

There are three prerequisites for this course:


2. Probability and statistics covered in 17.800, 17.802 and 17.804, including linear regression, Bayesian statistics

3. Statistical computing: proficiency with at least one statistical software. We will use \texttt{R} in this course (more on this below).

For 1, refer to this year’s math camp materials to see the minimum you need to know; see

- **Math Camp 1**: [https://stellar.mit.edu/S/project/mathprefresher/](https://stellar.mit.edu/S/project/mathprefresher/)
- **Math Camp 2**: [https://stellar.mit.edu/S/project/mathcamp2/](https://stellar.mit.edu/S/project/mathcamp2/)

This class will assume that you have already had some prior exposure to the material covered and go through many concepts relatively quickly.

5 Course Requirements

The final grades are based on the following items:

- **Problem sets** (45%): Seven problem sets will be given throughout the semester. Problem sets will contain analytical, computational, and data analysis questions. Each problem set will contribute equally toward the calculation of the final grade. The following instructions will apply to all problem sets unless otherwise noted.
  - All answers should be typed. Students are strongly encouraged to use \texttt{LaTeX}, a typesetting system that has become popular in the field. Please make sure that your code follows Google’s \texttt{R} Style Guide rules (here is the [URL](https://www.google.com)).
  - Neither late submission nor electronic submission will be accepted unless you ask for special permission from the instructor in advance. (Permission may be granted or not granted, with or without penalty, depending on the specific circumstances.)
  - Working in groups is encouraged, but each student must submit their own writeup of the solutions. In particular, you should not copy someone else’s answers or computer code. We also ask you to write down the names of the other students with whom you solved the problems together on the first sheet of your solutions.
  - For analytical questions, you should include your intermediate steps, as well as comments on those steps when appropriate. For data analysis questions, include annotated code as part of your answers. All results should be presented so that they can be easily understood.

- **Final project** (50%): The final project will be a paper which applies methods learned in this course to an empirical problem of your substantive interest.
  
  1. **Data** (10%)  
     - Students are expected to collect their own data related to an empirical problem of own interest. To help this, we are going to cover (1) the basics of Python, and (2) web-scraping at the beginning of the semester.
– Students who do not have particular target data sources should consult with the instructor by February 13th.
– Replication papers are allowed, but you must go beyond the original analysis in some significant way by collecting additional data and applying techniques learned in the course. If you have any doubts, please consult with the instructor and TA.

2. **Paper** (35%)

   – Title
   – Abstract (150 words)
   – Introduction (2 pages max): Introduction must contain the following.
     (a) The problem/puzzle to be solved
     (b) Explain why previous work and methods leave the problem unresolved
     (c) Your contribution, i.e., the solution to the problem/puzzle. You need to give the reader a clear sense of how you will solve the problem.
     (d) Brief summary of your findings
   – Data section (2 pages max)
   – Figures and tables with informative captions

3. **Poster** (5%)

**Collaboration:** We encourage you to collaborate with another student (a group should not consist of more than 2 students). Note that most cutting-edge research is collaborative (see any recent issue of *APSR* or *AJPS*), and collaboration is more likely result in a good, potentially publishable paper (multiple brains are usually better than one).

**Deadlines:** Please be aware of the following deadlines. Late submission will be penalized.

– **March 29 (Descriptive data analysis):** By this date, you should acquire the data to be analyzed and finish descriptive data analysis. Please upload a brief memo to the [Stellar webpage](#) with the following components. You will be giving in-class presentations on data on this date.
  * Data description (why better than previous data)
  * Main theoretical/empirical contributions/motivations
  * Figures/tables

– **April 26 (Initial analysis):** By this date, you should finish initial data analysis. Meet with the instructor to get feedback on your analysis (schedule a meeting with the instructor in the week of April 11).

– **May 10 (Poster due):** The poster should summarize the theoretical/empirical contributions, the methods you utilized, and the results (figures and tables). The poster should be posted to the [Stellar webpage](#) by midnight. The size of the poster should be A0 (33.1 × 46.8 inches).

– **May 12 (Final Paper):** By this date, you should submit your final paper to the [Stellar webpage](#) by midnight.

– **May 13 (Poster Presentation) 10:30AM**
• **Participation and presentation** (5%): Students are strongly encouraged to ask questions and actively participate in discussions during lectures and recitation sessions.

In addition, there will be recommended readings for each section of the course which students are strongly encouraged to complete prior to the lectures in order to get the most out of them.

### 6 Course Website

You can find the Stellar website for this course at:  

http://stellar.mit.edu/S/course/17/sp16/17.806/

We will distribute course materials, including readings, lecture slides and problem sets, on this website.

### 7 Questions about Course Materials

In this course, we will utilize an online discussion board called *Piazza*. Below is an official blurb from the Piazza team:

Piazza is a question-and-answer platform specifically designed to get you answers fast. They support LaTeX, code formatting, embedding of images, and attaching of files. The quicker you begin asking questions on Piazza (rather than via individual emails to a classmate or one of us), the quicker you’ll benefit from the collective knowledge of your classmates and instructors. We encourage you to ask questions when you’re struggling to understand a concept ... 

See this New York Times article to learn more about their founder’s story:  


In addition to recitation sessions and office hours, please use the Piazza Q&A board when asking questions about lectures, problem sets, and other course materials. You can access the Piazza course page either directly from the below address or the link posted on the Stellar course website:

https://piazza.com/mit/spring2016/17806

Using Piazza will allow students to see other students’ questions and learn from them. Both the TA and the instructor will regularly check the board and answer questions posted, although everyone else is also encouraged to contribute to the discussion. A student’s respectful and constructive participation on the forum will count toward his/her class participation grade. *Do not email your questions directly to the instructor or TA* (unless they are of a personal nature)— we will not answer them!

### 8 Recitation Sessions

Weekly recitation sessions will be held in **E53-438** on Fridays 10:30–11:30am. Sessions will cover a review of the theoretical material and also provide help with computing issues. The teaching assistant will run the sessions and can give more details. Attendance is strongly encouraged.
9 Notes on Auditing

In order to audit this course, one must

- Obtain the course instructor’s permission
- Register officially as a listener
- Complete all problem sets
- Submit comments on each project’s descriptive data analysis by April 5.

10 Notes on Poster

Poster presentation is an efficient way to get valuable feedback from a large number of people. A poster should follow the structure of your paper, and thus it is a helpful way to think about the organization of your paper before writing it. Here are some notes.

1. **Use keywords and bullet points:** You should not use full sentences—your audience will never read them. Try to use keywords (or half sentences when needed), and make sure that you use only one line to deliver each point.

2. **Use \LaTeX:** There are many online templates to help you make posters easily, e.g., http://www-i6.informatik.rwth-aachen.de/dreuw/latexbeamerposter.php

3. **Examples:** You may find it helpful to look at some of the posters presented at Political Methods conferences. It’s available here: http://polmeth.wustl.edu/media.php.

11 Notes on Computing

- In this course we use \texttt{R}, an open-source statistical computing environment that is very widely used in statistics and political science. (If you are already well versed in another statistical software, you are free to use it, but you will be on your own.) Each problem set will contain computing and/or data analysis exercises which can be solved with \texttt{R} but often require going beyond canned functions to write your own program.

- If your project requires large computational resources, I recommend using Research Computing Environment (RCE) available through the Harvard-MIT Data Center (HMDC).

12 Books

- Recommended books: We will read chapters from these books throughout the course. We strongly recommend that you at least purchase Bishop. These books will be available for purchase at COOP and online bookstores (e.g. Amazon) and on reserve in the library.

– Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014 An Introduction to Statistical Learning. Springer.

13 Tentative Course Outline

13.1 Research Computing

1. The Basics of Python

2. Web-scraping
   • Regular expression
   • Getting Data from the Web
   
   *Recommended:

3. Repp, Armadillo

13.2 Supervised Learning

1. Support Vector Machine (SVM)

   *Recommended:
   • Bishop Appendix E. Lagrange Multipliers
   • Bishop 7.1 (7.1.3, 7.1.4 optional)
   • Murphy Ch.14 (optional)

2. Over-fitting (Model Selection), Cross-validation

   *Recommended:
   • Bishop 1.1

3. Variable Selection (Ridge Regression, LASSO)

   *Recommended:
   • Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. Ch 3.1–3.4


*Recommended:*

• Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 9, 15, 16

• Bishop 14

• Murphy Ch.16

5. Machine learning for Causal Inference

*Recommended:*


### 13.3 Dimension Reduction

1. Principal Component Analysis

• Bishop Ch. 12 (towards 12.2.1)

• Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 14.5

2. Factor Analysis


### 13.4 Mixture Models

1. Probability Distributions

*Recommended:*

• Bishop Ch.2, Appendix B
2. EM Algorithm

Recommended:

- Bishop Ch.9
- Murphy Ch.11

3. Variational Inference

Recommended:

- Bishop Ch.10
- Murphy Ch.21

13.5 Text Analysis

1. Text as Data: regular expression, stemming

Recommended:


2. Latent Dirichlet Analysis

Recommended:


3. Correlated Topic Models

Recommended:

4. Structural Topic Models

*Recommended:*


5. Words and Votes: Scaling with Text

*Recommended:*


13.6 Network as Topics: Network Block Models

*Recommended:*


13.7 Sequential Data

1. Hidden Markov Models

*Recommended:*

• Bishop Ch.13
• Murphy Ch.17