17.806: Quantitative Research Methods IV

Spring 2020

Instructor: In Song Kim
TA: Zach Markovich

Department of Political Science
MIT

1 Contact Information

<table>
<thead>
<tr>
<th></th>
<th>In Song</th>
<th>Zach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office</td>
<td>E53–407</td>
<td>E53–458</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:insong@mit.edu">insong@mit.edu</a></td>
<td><a href="mailto:zmarko@mit.edu">zmarko@mit.edu</a></td>
</tr>
<tr>
<td>Phone</td>
<td>617–253–3138</td>
<td></td>
</tr>
</tbody>
</table>

2 Logistics

- Lectures: Tuesdays and Thursdays, 9:30am–11:00am, E53–438
- Recitations: Friday, 1pm-2pm (tentative), location tbd
- In Song’s office hours: Friday 3:30pm–5:00pm
- Zach’s office hours: Monday 1pm-3:30pm (tentative).

Note that the first class meets on February 4. Class will begin at 10AM on February 13. No class will be held on February 18 (President’s 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 as well as their applications in causal inference methods; and (3) finite mixture models (e.g., Latent Dirichlet allocation for text analysis), as well as a variety of estimation techniques such as the 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 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%): Six bi-weekly 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 LaTeX, a typesetting system that has become popular in the field. Please make sure that your code follows Google’s R Style Guide rules (here is the URL).
  
  - 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 in a single document so that they can be easily understood. RMarkdown is strongly encouraged.

- **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 and Initial Analysis** (15%)
     
     - Students are expected to collect their own data related to an empirical problem of own interest.
A one page

Students who do not have particular target data sources should consult with the instructor by February 20.

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.

March 20: A one-page memo due (see below).

April 14: A five-page report due (see below).

2. **Paper** (35%): The paper should be *maximum 10 pages* of double-spaced 12-point font text (including references and appendix) with 1-inch margins.

   - Title
   - Abstract (150 words)
   - Introduction: 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
   - Empirical analysis: Figures and tables with informative captions

**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 20 (Data Collection):** By this date, you should acquire the data to be analyzed and start preliminary descriptive data analysis. Please upload one-page memo to the Stellar webpage with the following components.
  * Main theoretical/empirical contributions/motivations
  * Data description (source, collection methods, and why better than previous data)

- **April 14 (Initial descriptive analysis):** By this date, you should submit a five-page report summarizing your data collection and descriptive data analysis to the Stellar webpage.

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

**Participation** (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/sp20/17.806/

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

7 Questions about Course Materials

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/spring2020/17806

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 (time tbd). 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
- Complete all problem sets
- Submit written comments on each project’s descriptive data analysis

10 Notes on Computing

- In this course we use 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 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).
11 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.
  
  
  - Daniel Jurafsky and James Martin. 2018. *Speech and Language Processing*. Prentice Hall. [PDF]

12 Course Outline

12.1 Introduction

1. Big Data in Political Science

  Recommended Reading:

  
  

12.2 Automated Data Collection

1. Web Scraping, Regular Expressions

  Recommended Reading:

  - Jurafsky and Martin 2.1
  
  - For a basic tutorial on HTML, consult 3 sources linked from this blog post: [Three great places to start learning HTML](https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/atheyimpactmlecon.pdf)
  
  
  - Data Camp Course: [Working with Web Data in R](https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/atheyimpactmlecon.pdf)
12.3 Supervised Learning

1. Support Vector Machine (SVM)

*Recommended Reading:*
- 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

*Required Reading:*

*Recommended Reading:*
- Bishop 1.1

3. Variable Selection (Ridge Regression, LASSO)

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

*Recommended Reading:*


*Recommended Reading:*
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Ch 9, 15, 16
- Bishop 14
- Murphy Ch.16
12.4 Machine Learning for Causal Inference

1. Machine learning for Causal Inference

**Required Reading:**


**Recommended Reading:**

- Susan Athey and Guido Imbens. 2016. “Recursive Partitioning for heterogeneous causal effects.” *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360

12.5 Dimension Reduction

1. Principal Component Analysis, Factor Analysis

**Recommended Reading:**

- Bishop Ch. 12 (towards 12.2.1)

2. T-SNE

*Recommended Reading:*


12.6 Mixture Models

1. Probability Distributions

*Required Reading:*

• Bishop Ch.2, Appendix B

2. EM Algorithm

*Required Reading:*

• Bishop Ch.9

*Recommended Reading:*

• Murphy Ch.11

3. Variational Inference

*Required Reading:*


*Recommended Reading:*

• Bishop Ch.10
• Murphy Ch.21
12.7 Text Analysis

1. Text as Data: regular expression, stemming

   **Recommended Reading:**


   **Recommended Reading:**


3. Words and Votes: Scaling with Text

   **Recommended Reading:**


4. Word Embeddings

   **Recommended Reading:**

   - Jurafsky and Martin 6
12.8 Causal Inference with Time-Series Cross-Section Data

Recommended Reading:


12.9 Network Models

Recommended Reading:


