17.835: Machine Learning and Data Science in Politics

Spring 2020

Instructor: Professor In Song Kim
TAs: Matias Giannoni, Zeyu Chris Peng, Guillermo Toral

Department of Political Science
MIT

1 Contact Information

<table>
<thead>
<tr>
<th></th>
<th>In Song Kim</th>
<th>Matias Giannoni</th>
<th>Zeyu Chris Peng</th>
<th>Guillermo Toral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email:</td>
<td><a href="mailto:insong@mit.edu">insong@mit.edu</a></td>
<td><a href="mailto:giannoni@mit.edu">giannoni@mit.edu</a></td>
<td><a href="mailto:pzy0337@mit.edu">pzy0337@mit.edu</a></td>
<td><a href="mailto:gtoral@mit.edu">gtoral@mit.edu</a></td>
</tr>
<tr>
<td>Office Hours:</td>
<td>Fri 3:30pm-5:30pm</td>
<td>Tues 2pm-3pm</td>
<td>Mon 10am-11am</td>
<td>Wed 2pm-3pm</td>
</tr>
</tbody>
</table>

2 Logistics

- Lectures: Tuesdays and Thursdays 3:30pm –5:00pm, 56-114
- Recitations: Fridays 10-11; Fridays 11-12; Fridays 1-2. Rooms TBA.
  - There will be three weekly recitation sessions. We will cover a review of the theoretical material and also provide help with computing issues. The teaching assistants will run the sessions and can give more details. Attendance is strongly encouraged.

3 Course Description

Empirical studies in political science is entering a new era of “Big Data” where a diverse range of data sources have become available to researchers. Examples include network data from political campaigns, data from social media generated by individuals, campaign contribution and lobbying expenditure made by firms and individuals, and massive amount of international trade flows data. How can we take advantage of these new data sources and improve our understanding of politics? This course introduces various machine learning methods and their applications in political science research. Students will

1. Be introduced to various quantitative political science research topics in its four subfields: American Politics, International Relations, Comparative Politics, and Political Methodology.

2. Learn basic machine learning algorithms and data science tools that are applied in political science research.

3. Apply data analysis tools using R programming language through problem sets.
4. Collect and analyze data to learn substantive topics of own interest.

5. Learn how to communicate data-driven findings and insights.

*Note:* the topics covered in this class represent only a very small subset of political science research. If you enjoy this class, please consider a HASS concentration in Political Science. We also offer a major and a minor in Political Science, as well as a minor in Public Policy and a minor in Applied International Studies. Internships and research opportunities too. Check out these programs and more at: [https://polisci.mit.edu/undergraduate](https://polisci.mit.edu/undergraduate)

4 Prerequisites

This class will assume that you do not have any prior exposure to political science and machine learning. One prerequisite for this course is basic programming skills in at least one language (e.g., Python). Students who have taken **6.0001: Introduction to Computer Science and Programming in Python** or the equivalent are ready to take this course. If you have any questions about whether you are prepared for this course, please talk to the instructor.

5 Notes on Computing

In this course we use \( R \), an open-source statistical computing environment that is very widely used in statistics and data science. (If you are already well versed in another statistical software, you are free to use it, but you will be on your own). We will begin the course with an introduction to \( R \) and no prior exposure to the programming language is required. 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.

To use \( R \), install it by visiting [https://cran.r-project.org/](https://cran.r-project.org/) and clicking the appropriate link for your operating system. *After installing R*, we strongly recommend you also install \( RStudio \), a tremendously useful interface to work with \( R \). To install the free RStudio Desktop, visit [this webpage](https://cran.r-project.org/).

6 Course Requirements

The final grades are based on the following items:

- **Problem sets (35%)**: Five 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 (We will use Letter Grade grading scheme with \(+/-\) modifiers, e.g., \(A^+, A, A^-, B^+, \ldots\)). The following instructions will apply to all problem sets unless otherwise noted.

  - *Late submission will not 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.
– All answers should be typed. Students are strongly encouraged to use a typesetting system such as \LaTeX.

• **Final problem set** (15%): The last problems set will be a special problem set, which will be weighted more heavily toward the calculation of the final grade. You will *not be allowed to collaborate with anybody on the final problem set*. This is to check if you have developed sufficient experience to work through problems on your own.

• **Final group project** (40%): Students are expected to form a group. Each group will apply methods they learned in this course to an empirical problem of own substantive interest. The group project will be evaluated based on the performance across the following three tasks.

  − **Three Tasks**
    1. **Data Collection and Cleaning** (5%): Collecting political science data. Final output: (1) dataset and (2) replication codes.
       * Your group will either engage in your own data collection (e.g., using web-scraping) or utilize multiple existing datasets in political science. The dataset should be submitted in a standard data format (e.g., .csv, .sql, .json) as an output of this task.
       · Collecting new data: Your team will choose a topic of interest related to various subfields in political science such as American Politics, International Relations, Comparative Politics, and Political Economy. The instructor will provide guidance to identify potential sources for novel data collection.
       · Utilizing existing datasets: You should merge various datasets available in political science research. The instructor will also make two of his own databases available: (1) Money in Politics Database (see [www.LobbyView.org](http://www.LobbyView.org)), and (2) International Trade data.
    2. **Descriptive Analysis** (15%): Your group will then conduct descriptive data analysis. You should submit tables and figures that effectively illustrate key patterns in your data. A five-page report should be submitted as an output of this task by April 14.
    3. **Data Analysis/Presentation** (10%): You will utilize various tools that you learn from the course to conduct an in-depth data analysis. Each group will give 10 minutes in-class presentation of their main findings in the last weeks of the semester.
    4. **Final Poster** (10%): A poster should be submitted as a final output of the project on the last day of the class.

− **Deadlines**: Please be aware of the following deadlines. Late submission will not be accepted. You are welcome to arrange a meeting (during the office hours) with the instructor and the TAs as you make a progress over the semester.
  * **February 20**: By this date, please form your team. We will create a forum so that students in the same recitation group can find members of their final group project. A group should have at least three students and consist of no more than 5 students. Your team should arrange a meeting with the recitation TA within a week after this date.
* **March 3:** By this date, your team should identify the dataset to analyze. Please submit one-page description of your project that explains (a) the specific dataset that your team is going to collect/analyze, (b) the main puzzle/problem that your team plans to study.

* **April 14:** By this date, your team should submit a five-page long report summarizing the results from your descriptive data analysis. The report should have 1-inch margins with double-spaced 12 point font text. Please submit a document with at most 5 figures or tables that summarize your data with informative caption for each.

* **May 12:** By this date, your team should submit the final poster of your project. There will be in-class poster session on this date.

- **Participation & Attendance** (10%): Students are expected to attend all classes and actively participate in their group project. Students are strongly encouraged to ask questions and participate in discussions during lectures and recitation sessions. We will take attendance on randomly chosen dates across the semester. In addition, there will be required readings for each section of the course which students are expected to complete prior to the lectures.

## 7 Course Website

You can find the Stellar website for this course at:


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

## 8 Questions about Course Materials

In this course, we will utilize an online discussion board called Piazza. 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/spring2020/17835](https://piazza.com/mit/spring2020/17835)

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!

## 9 Notes on Poster

Poster presentation is an efficient way to get valuable feedback from a large number of people. A poster should effectively summarize your key findings. Here are some notes.
1. **Use keywords and bullet points:** You should not use full sentences—you're 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

--

10 **Books**

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

11 **Tentative Course Outline**

11.1 **Introduction**

- Machine Learning and Data Science in Political Science

  *Optional Reading:*

11.2 **Data Analysis using R**

- Dataframe

- Descriptive data analysis

- Visualization

  *Optional Reading:*
  - QSS: Chapter 1.3

11.3 **Causality**

- Causal Inference

- Average Treatment Effect (ATE) and Average Treatment Effect for the Treated (ATT)

  *Required Reading:*
• QSS: Chapter 2

*Recommended Reading:*


11.4 Linear Regression

• OLS (Ordinary Least Squares)
• Difference in means estimator
• Regression and Causation

*Required Reading:*

• Chapter 4.2 (First Week)
• Chapter 4.3 (Second Week)

*Optional Reading:*


11.5 Supervised Learning

• Introduction to Supervised Learning
• K-Nearest-Neighbor (KNN) Classifier
• Support Vector Machine (SVM)
• Over fitting
• Ridge Regression
• Least Absolute Shrinkage and Selection Operator (LASSO)

*Optional Reading:*

11.6 Unsupervised Learning Methods

- Principal Component Analysis (PCA)
- Clustering Algorithm

**Reading:**
- QSS: Chapter 3.7

- Mixture Models and EM Algorithm

**Reading:**
- Bishop Ch.9

11.7 Text Analysis

- Introduction to Text Analysis

**Required Reading:**
- QSS: Chapter 5.1

**Recommended Reading:**

- Latent Dirichlet Analysis (LDA)

**Reading:**

- Word Embeddings

**Recommended Reading:**
11.8 Network Analysis

- Network Analysis

  Recommended Reading:
  - QSS: Chapter 5.2

11.9 Applications in Political Science

- International Trade with Big Data

  Reading:

- Lobbying and Campaign Contribution

  Reading:

- Identifying Behavioral Patterns using Massive Data

  Reading:

- Measuring Ideological and Political Preferences using Social Network Data

  Reading:
  - Pablo Barberá “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23.1 (2014): 76-91

- What do Politicians Do?
Reading:

• Big Administrative Data: Promises and Pitfalls

  Reading:

• Machine Learning Algorithms in Society

  Reading: