SOCIOL 273L: Computational Social Science, Part A

UC Berkeley

Fall 2021

Instructors: David Harding (Sociology) and Kelly Quinn (Sociology)

Lecture: Tuesdays 10am-noon (Berkeley Institute for Data Science, 190 Doe Library)
Lab: Thursdays 10am-noon (Berkeley Institute for Data Science, 190 Doe Library)
Instructor Office Hours:

Harding (dharding@berkeley.edu): Tuesdays 1-3pm (sign-up: https://tinyurl.com/HardingOH), or by appointment

Quinn (kellyquinn@berkeley.edu): TBD (see bcourses)

Course bCourses Site: https://bcourses.berkeley.edu/courses/1507109
Course Github repository: https://github.com/dlab-berkeley/Computational-Social-Science-Training-Program

Course description: This is the first semester of a two-semester course that provides a rigorous introduction to methods and tools in advanced data analytics for social science doctoral students. The goal of the course is to provide students with a strong foundation of knowledge of core methods, thereby preparing them to contribute to research teams, to conduct their own research, and to enroll in more advanced courses. The course will cover research reproducibility (fall), machine learning (fall), natural language processing (spring), and causal inference (spring). In contrast to other courses currently offered on campus, this course's intended audience is applied researchers, typically social science doctoral students in their second or third year of graduate school. *This is a required course for students in the Computational Social Science Training Program (website forthcoming). Enrollment is open to doctoral students from any department.*

The course is divided into modules, each lasting 3-5 weeks. Each module will include lectures, discussion of example research articles, lab exercises, and a group project involving Python (fall and spring) or R programming (spring only). Projects, typically done in groups of 3-4 students, will also provide the opportunity to practice reproducibility techniques, data manipulation and transformation, and data science workflows.

Course objectives (Fall semester):

- Proficiency with tools for reproducibility of research
- Conceptual understanding of machine learning methods, including strengths and weaknesses of various algorithms and their appropriate application to different kinds of prediction and classification problems
- Ability to apply these concepts and execute relevant methodologies on social science data in Python and correctly interpret results
- Familiarity with key empirical papers that apply computational social science methods to research

Prerequisites: A year-long course in statistical methods for social science graduate students (or equivalent prior experience) will generally be sufficient preparation. Students should have a background in multivariate regression (both linear and non-linear models), maximum likelihood estimation, and introductory causal inference (omitted variable bias, potential outcomes, average treatment effects, causal graphs). Students may consult the instructor about readings on these topics to ensure adequate preparation. In addition, this course will be taught in Python. Students without a background in introductory Python programming should take the D-Lab Python Fundamentals Workshop series, which is usually offered in the week before the fall semester begins. Those who need a Python refresher may wish to review the Jupyter Notebooks for D-Lab Python Fundamentals here: https://github.com/dlab-berkeley/python-fundamentals

Instructional technology: Examples and student projects will occur in Python using Jupyter Notebooks. Python is a general purpose open-source programming language that is known for its ease of use and readability and is a strong data science tool. Jupyter Notebooks is an open-source web application for documents that contain live code, equations, visualizations and narrative text, and it can be used for data cleaning, statistical modeling, data visualization, machine learning, and more. Students should install Anaconda before the first lab.

Instructional Resilience and the possibility of Remote Instruction: All "lecture" and lab meetings will be held in-person by we will need to be prepared to move to Zoom if conditions require (see bcourses site for links).

How we will use class time: During most weeks, we will have pre-recorded videos, typically in 10-15 minute segments. Students should view these lectures, take the self quizzes, and do the readings BEFORE each week's lecture session. Class "lecture" meeting times will be used to review key points, answer questions and discuss the readings. Occasionally we may use half of the lecture section for lab. Each student will submit a one-page weekly reflection memo by 5pm the night before lecture (students may skip 5 weeks during the semester). Lab times will be used to work through Jupyter Notebooks applying the week's concepts, tools, and models to data. Group projects with rotating group membership will provide students with opportunities to build a course community with fellow students.

Grading:

- Lecture and Lab Participation: 25%
- Weekly reflection memos (graded credit/no credit): 20%
- Project #1 (graded credit/no credit): 5%
- Project #2: 15%
- Project #3: 15%
- Project #4: 20%

Course Schedule
(note: links to readings are on bcourses; some require UC-Berkeley CalNet login)

Module 1: Introduction and Reproducibility

Week 1 (Aug 31/ Sept 2): Introduction to Computational Social Science (Python/Statistics Refresher in Lab, *No Reflection Memo this week*)
Readings:

- Henry Brady. 2019. "The Challenge of Big Data and Data Science." *Annual Review of Political Science*
- Matt Salganik. 2019. *Bit by Bit: Social Research in the Digital Age*. Princeton UP (Preface Chapters 1 and 2)
- Breiman, L. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science,* 16(3), 199–231.

Week 2 (Sept 7/9): Reproducibility and Transparency, Part 1
Readings:

- Christensen, Freese and Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science.* UC Press (Chapters 1-6)

Week 3 (Sept 14/16): Reproducibility and Transparency, Part 2
Readings:

- Christensen, Freese and Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science.* UC Press (Chapters 7-12)

*** Project #1 (Reproducibility) Due Sept 24 ***

Module 2: Introduction to Machine Learning

Week 4 (Sept 21/23): Ethics and Machine Learning (Math Review in Lab)

Readings:

- Brian Christian. 2020. *The Alignment Problem* (Chapters 1-3). Norton. (on bcourses)
- Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel. 2016. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *The Washington Post*. October 17, 2016.

Week 5 (Sept 28/30): Introduction to Machine Learning

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer (Sections 1, 2.1-2.2, 5.1-5.2)
- Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *Annals of the American Academy of Political and Social Science*, 659(1), 48–62. http://doi.org/10.1177/0002716215570279
- Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review: Papers & Proceedings* 2015, 105(5): 491–495

Week 6 (Oct 5/7): Regression (Ridge and LASSO)

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer (Sections 6.1-6.2, skim 3.1-3.5 for review)
- Mario Molina and Filiz Garip. 2019."Machine Learning for Sociology." *Annual Review of Sociology*. 45:27–45

*** Project #2 (Regression) Due Oct 19 ***

Module 3: Supervised Machine Learning

Week 7 (Oct 12/14): Classification, Part 1

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer (Sections 4.1, 9.1-9.6, skim 4.2-4.3 for review)
- Vito D'Orazio, Steven T. Landis, Glenn Palmer and Philip Schrodt. 2014. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines. *Political Analysis* 22 (2): 224-242

Week 8 (Oct 19/21): Classification, Part 2

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer (Sections 4.4-4.6)
- Choose ONE of the following two papers:

- Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan. Human Decisions and Machine Predictions. NBER Working Paper No. 23180
- Sendhil Mullainathan, Ziad Obermeyer. A Machine Learning Approach to Low-Value Health Care: Wasted Tests, Missed Heart Attacks and Mis-Predictions. NBER Working Paper No. 26168

Week 9 (Oct 26/28): Tree-Based Learning and Ensemble Learning

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*..Springer (Sections 8.1-8.2)
- Hastie, Tibshirani, and Friedman. 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed)* . Springer (Chapter 16)
- Elliott Ash, Sergio Galletta, Tommaso Giommoni. A Machine Learning Approach to Analyze and Support Anti-Corruption Policy. Working Paper (https://elliottash.com/wp-content/uploads/2020/04/Ash-Galletta-Giommoni-paper-2020-04-28.pdf)

Week 10 (Nov 2/4): Deep Learning

Readings:

- Hastie, Tibshirani, and Friedman. 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Ed)* . Springer (Sections 11.1-11.8)
- EECS 189 Notes: Note 14, Note 15.
- Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *PNAS* December 12, 2017 114 (50) 13108-13113

**** Project #3 (Classification) Due Nov 19 ****

Module 4: Unsupervised Machine Learning

Week 11 (Nov 9/11): Clustering *(no Lab on Thursday, Veteran's Day holiday)*

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*.Springer (Sections 10.1, 10.3)
- Jakulin, A., Buntine, W., Pira, T., & Brasher, H. (2009). Analyzing the U.S. Senate in 2003: Similarities, Clusters, and Blocs. *Political Analysis*, 17(3), 291-310. doi:10.1093/pan/mpp006

Week 12 (Nov 16/18): Principal Components Analysis/Dimensionality Reduction

Readings:

- James, Witten, Hastie, and Tibshirani. 2013. *An Introduction to Statistical Learning*.Springer (Section 10.2, review 4.4)
- Danyi Qi and Brian E. Roe. Household Food Waste: Multivariate Regression and Principal Components Analyses of Awareness and Attitudes among U.S. Consumers. *PLOS One.*

Week 13 (Nov 23): Lab only this week, on Tuesday

(No class November 26 -- Thanksgiving Break)

Module 5: Wrap Up

Week 14 (Nov 30/Dec 2): Machine Learning Applications in the Social Sciences

Readings:

- Blumenstock, J. E., Cadamuro, G., & On, R. 2015. Supplementary materials for "Predicting poverty and wealth from mobile phone metadata." *Science*, 350(6264), 1073–1076. http://doi.org/10.1126/science.aac4420
- Mullainathan, S., & Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives,* 31(2—Spring), 87–106. http://doi.org/10.1257/jep.31.2.87
- Wang, Q., Phillips, N. E., Small, M. L., & Sampson, R. J. 2018. "Urban mobility and neighborhood isolation in America's 50 largest cities". *Proceedings of the National Academy of Sciences*, 115(30), 7735–7740. http://doi.org/10.1073/pnas.1802537115
- Barberá, P., Boydstun, A., Linn, S., McMahon, R., & Nagler, J. (2020). Automated Text Classification of News Articles: A Practical Guide. *Political Analysis*, 1-24. doi:10.1017/pan.2020.8

**** Project #4 (Unsupervised Machine Learning) Due Dec 17 ****