POLS 6394:
Machine Learning for Social Sciences
Fall 2020

Instructor: Ryan Kennedy
Office: 447 Philip G. Hoffman (PGH)
Phone: 713-743-1663
Email: rkennedy@uh.edu
Class Time: 13:00-16:00 Th
Office Hours: Tuesday 11:00 – 13:00

Teaching Assistant: Myriam Shiran
Email: myriam.shiran@gmail.com
Office Hours: Friday 10:00 – 12:00

Course Description:
Driven by an explosion of data availability and computational power, the past twenty years have seen a resurgence of interest in machine learning – a field of study that allows machines to learn without being explicitly programmed (Samuel 1959). Growing out of artificial intelligence (AI) research, machine learning encompasses both an approach to learning from data and techniques for statistical estimation that differ in important ways from traditional approaches to statistical modeling in the social sciences.

The social sciences have not been left out of this resurgence. Indeed, machine learning approaches have made their way into many social science studies – usually in order to develop forecasting models, estimate highly non-linear and/or interactive patterns, or produce unique data for analysis. These application have opened new vistas for analysis.

This course is an introduction to the philosophy and methods of machine learning. We will be covering foundational aspects of machine learning, basic models, and applications in statistical software (R in this case). We will also be covering some of the applications of machine learning in the social sciences and methods that have become more popular within the social science context (versus, say, in computer science).

At the end of this course, students should be able to deploy machine learning solutions to social science research problems and have the foundational understanding to build on this knowledge.

Assignments and Grading:
The grading for this course will be based on bi-weekly problem sets, class participation, and a final research project.

Problem sets will require students to apply the methods and techniques covered in the course. Some will require essay-style answers to explain important concepts. Others will require students to work with a dataset in R and provide outputs for their answers. Students will, unless otherwise noted, complete these assignments in R Markdown. It is ok for students to work on these assignments together, but the work you submit should be your own. If plagiarism is suspected, the student will be confronted, will likely receive a zero for the assignment, and may be reported for further action.

Class participation is based on several components. First, each week there will be journal articles assigned and students will be expected to present on these articles to the rest of the class. These presentations should not just include the basic information about the article (after all, we are all going to read them), but also critical evaluation. Second, each week, there will be conceptual questions for the week. Students should evaluate these questions, write answers to them, and be ready to discuss them in the class sessions. Finally, when we get to the final projects, students will present their project to the class and should provide constructive comments on each other's presentations and projects.

Finally, students will be asked to write a final paper for the course. If the student already has a project in mind that they would like to pursue, this can be the topic for the final project. Otherwise, students should find an article or dataset that has been previously published and pursue a replication project. The replication should include both a reproduction of the results from the article and a modification of the article's approach using a machine learning technique. In either case, it is strongly recommended that the student discuss their final project with the instructor early in the semester and continue this discussion regularly throughout the semester.

Readings:
We will largely be following the structure of *An Introduction to Statistical Learning with Application in R*. This book will provide the foundations for the course and generally covers the core materials with which the student should be familiar. It is also available freely online – the link can be found on Blackboard. If students would like a print version of the book, it can be purchased on Amazon or from the publisher.

The other required readings will take the form of articles that are posted on Blackboard.

Students should familiarize themselves with the Blackboard site as soon as possible.

Required Readings:
James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2017. *An Introduction to Statistical Learning with Applications in R* (ISLR), corrected 8th printing. New York: Springer.

Recommended Readings:
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2017. *Elements of Statistical Learning* (ESL), corrected 12th printing. New York: Springer. [Available as an online pdf – see Blackboard. This is widely cited as a classic in the field and covers the mathematical foundations of the models and approaches we will discuss in much greater detail.]

Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. 2012. *Learning from Data: A Short Course*. AMLbook. [For those really interested in digging into the philosophy and mathematical proofs underlying machine learning, this book is a gem.

Wickham, Hadley and Garrett Grolemund. 2017. *R for Data Science*. New York: O'Reilly. [Available as an online book – see Blackboard. This is an essential reference for the Tidyverse – a set of more modern programming tools in R. While it is not necessary that your write your code in a Tidy format (and ISLR does not), it is a good idea for you to become familiar with these tools. Much of the information you will find online will use these tools.]

Silge, Julia and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. New York: O'Reilly. [Available as an online book – see Blackboard. For those interested in using text mining for their projects, this is a good resource to get you into basic analysis like sentiment analysis and topic models.]

Grus, Joel. 2019. *Data Science from Scratch: First Principles with Python*, 2nd ed. New York: O'Reilly. [For those interested in learning how to do this in Python, this book provides both a discussion of how to do machine learning in Python and a lot of detail about programming the models themselves. Warning, this is not a book for learning Python itself – it does cover some Python basics, but you might be better served by a general introduction.]

Students with Disabilities:
The College of Liberal Arts and Social Sciences, in accordance with 504/ADA guidelines, is committed to providing reasonable academic accommodations to students who request them. Students seeking accommodation must register with the Center for Students with Disabilities (CSD) 713-743-5400 and present approved documentation to me as soon as possible.

Academic Honesty:
To cultivate an environment of academic integrity, the University of Houston expects students to abide by the University's Undergraduate Academic Honesty Policy, found in the Undergraduate Catalog. http://www.uh.edu/academic-honesty-undergraduate

Counseling and Psychological Services:
Counseling and Psychological Services (CAPS)--www.uh.edu/caps--are available for students having difficulties managing stress, adjusting to college, or feeling sad and hopeless. You can reach CAPS) by calling 713-743-5454 during and after business hours for routine appointments or if you or somebody you know is in crisis. The "Let's Talk" program provides a drop-in consultation service at convenient locations and hours around campus. http://www.uh.edu/caps/outreach/lets_talk.html

Classroom Conduct:
As noted above, attendance is expected on days that are not noted as asynchronous. Since this is a participation-based course, it is very difficult to actively participate in the class when you are not online during the simulation sessions. I understand that there may be some circumstances in which you cannot make it to class. You have three allowed absences this semester. Any absences beyond this must be justified with documentation. I also expect you to show up to class on time. If you are late, you may miss important announcements about the class, as well as important happenings in the simulation. If you show up more than halfway through the class, or leave when more than half of the class has is left, you will be counted with a ½ absence. Finally, I strongly discourage using a mobile device during the class. If you are on your phone, you are likely going to miss important information and be a less effective negotiator in the simulation. They should be turned off (not turned to vibrate).

Online Modifications:
We will be conducting both online asynchronous and online synchronous parts. The asynchronous parts are lectures that can be viewed on your own time, but the asynchronous lectures for a particular week should be viewed before our synchronous class meetings. You will need to view these, pay attention, and take notes to be able to understand key elements of the class.

The synchronous class sessions will take place on Microsoft Teams. You should download and install Microsoft Teams as soon as possible (https://teams.microsoft.com/downloads). You will need to have it

installed for the first week of class. The instructor will be in touch through email to give you the code to access our team.

Student Conduct Policy:
CLASS students are expected to abide by the University of Houston's Code of Student Conduct: http://www.uh.edu/dos/behavior-conduct/student-code-of-conduct/

Additional Information Regarding Classes During COVID-19:

Excused Absence Policy

Regular class attendance, participation, and engagement in coursework are important contributors to student success. Absences may be excused as provided in the University of Houston Undergraduate Excused Absence Policy and Graduate Excused Absence Policy for reasons including: medical illness of student or close relative, death of a close family member, legal or government proceeding that a student is obligated to attend, recognized professional and educational activities where the student is presenting, and University-sponsored activity or athletic competition. Additional policies address absences related to military service, religious holy days, pregnancy and related conditions, and disability.

Recording of Class

Students may not record all or part of class, livestream all or part of class, or make/distribute screen captures, without advanced written consent of the instructor. If you have or think you may have a disability such that you need to record class-related activities, please contact the Center for Students with DisABILITIES. If you have an accommodation to record class-related activities, those recordings may not be shared with any other student, whether in this course or not, or with any other person or on any other platform. Classes may be recorded by the instructor. Students may use instructor's recordings for their own studying and notetaking. Instructor's recordings are not authorized to be shared with *anyone* without the prior written approval of the instructor. Failure to comply with requirements regarding recordings will result in a disciplinary referral to the Dean of Students Office and may result in disciplinary action.

Syllabus Changes

Due to the changing nature of the COVID-19 pandemic, please note that the instructor may need to make modifications to the course syllabus and may do so at any time. Notice of such changes will be announced as quickly as possible through (*specify how students will be notified of changes*).

Resources for Online Learning

The University of Houston is committed to student success, and provides information to optimize the online learning experience through our Power-On website. Please visit this website for a comprehensive set of resources, tools, and tips including: obtaining access to the internet, AccessUH, and Blackboard; requesting a laptop through the Laptop Loaner Program; using your smartphone as a webcam; and downloading Microsoft Office 365 at no cost. For questions or assistance contact UHOnline@uh.edu.

UH Email

Email communications related to this course will be sent to your Exchange email account which each University of Houston student receives. The Exchange mail server can be accessed via Outlook, which provides a single location for organizing and managing day-to-day information, from email and calendars to contacts and task lists. Exchange email accounts can be accessed by logging into Office 365 with your Cougarnet credentials or through Acccess UH. They can also be configured on IOS and Android mobile devices. Additional assistance can be found at the Get Help page.

Webcams

Access to a webcam is strongly encouraged for students participating remotely in this course. Webcams should be turned on when students are doing class presentations. Webcams can be turned off during other discussion.

Helpful Information

COVID-19 Updates: https://uh.edu/covid-19/

Coogs Care: https://www.uh.edu/dsaes/coogscare/

Laptop Checkout Requests: https://www.uh.edu/infotech/about/planning/off-campus/index.php#do-you-need-a-laptop

Health FAQs: https://uh.edu/covid-19/faq/health-wellness-prevention-faqs/

Student Health Center: https://uh.edu/class/english/lcc/current-students/student-health-center/index.php

Class Schedule

Week 1: Course Introduction (8/27) [Last day to add course, 8/31]

- Reviewed Readings:
  - King, G., 1995. Replication, replication. *PS: Political Science and Politics*, *28*(3), pp.444-452.
  - ISLR, chapter 1.
  - Introduction to R Markdown (Recommended)

Week 2: Review of Basic Mathematical Tools & Foundations of ML I (9/3)

- Reviewed Readings:
  - ISLR, chapter 2.
- Student Readings:
  - Molina, M. and Garip, F., 2019. Machine learning for sociology. *Annual Review of Sociology*.
  - Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), pp.199-231.
  - Athey, S. and Imbens, G.W., 2019. Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, pp.685-725.

Week 3: Foundations of ML II (9/10) [Last day to drop course without grade, 9/9]

- Reviewed Readings:
  - Kennedy, R., 2015. Making useful conflict predictions: Methods for addressing skewed classes and implementing cost-sensitive learning in the study of state failure. *Journal of Peace Research*, *52*(5), pp.649-664.
- Student Readings:
  - Clark, W.R. and Golder, M., 2015. Big data, causal inference, and formal theory: Contradictory trends in political science?. *PS, Political Science & Politics*, *48*(1), p.65.
  - Grimmer, J., 2015. We are all social scientists now: How big data, machine learning, and causal inference work together. *PS, Political Science & Politics*, *48*(1), p.80.
  - Cowgill, B. and Tucker, C., 2017, December. Algorithmic bias: A counterfactual perspective. In *Workshop on Trustworthy Algorithmic Decision-Making*.
  - Lazer, D. and Radford, J., 2017. Data ex machina: introduction to big data. *Annual Review of Sociology*, *43*, pp.19-39.

Week 4: Linear and KNN Regression (9/17)

- Reviewed Readings:
  - ISLR, chapter 3.
- Student Readings:
  - Hindman, M., 2015. Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), pp.48-62.
  - Dietrich, B.J., Hayes, M. and O'BRIEN, D.Z., 2019. Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech. *American Political Science Review*, *113*(4), pp.941-962.
  - Schneider, G., Gleditsch, N.P. and Carey, S., 2011. Forecasting in international relations: One quest, three approaches. *Conflict Management and Peace Science*, *28*(1), pp.5-14.

Week 5: Basic Classification Models (9/24)

- Reviewed Readings:
  - ISLR, chapter 4.
- Student Readings:
  - King, G., Pan, J. and Roberts, M.E., 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, pp.326-343.
  - Cantú, F. and Saiegh, S.M., 2011. Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning. *Political Analysis*, *19*(4), pp.409-433.
  - Caughlin, T.T., Ruktanonchai, N., Acevedo, M.A., Lopiano, K.K., Prosper, O., Eagle, N. and Tatem, A.J., 2013. Place-based attributes predict community membership in a mobile phone communication network. *PloS one*, *8*(2), p.e56057.

Week 6: Cross-Validation and Bootstrapping (10/1)

- Reviewed Readings:
  - ISLR, chapter 5.
- Student Readings:
  - Neunhoeffer, M. and Sternberg, S., 2019. How cross-validation can go wrong and what to do about it. *Political Analysis*, *27*(1), pp.101-106.
  - Ward, M.D., Greenhill, B.D. and Bakke, K.M., 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of peace research*, *47*(4), pp.363-375.

Week 7: Model Selection (10/8)

- Reviewed Readings:
  - ISLR, chapter 6.
- Student Readings:
  - Waggoner, P., and Macmillen A. 2020. Pursuing Open-Source Development of Predictive Algorithms: The Case of Criminal Sentencing Algorithms. University of Chicago: Working Paper.
  - Greene, K.T., Park, B. and Colaresi, M., 2019. Machine learning human rights and wrongs: How the successes and failures of supervised learning algorithms can inform the debate about information effects. *Political Analysis*, *27*(2), pp.223-230.
  - Bloniarz, A., Liu, H., Zhang, C.H., Sekhon, J.S. and Yu, B., 2016. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, *113*(27), pp.7383-7390.

Week 8: Non-Linear Models (10/15)

- Reviewed Readings:
  - ISLR, chapter 7.
- Student Readings:
  - Zhao, Q. and Hastie, T., 2019. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, pp.1-10.
  - Kaufman, A., King, G. and Komisarchik, M., 2017. How to measure legislative district compactness if you only know it when you see it. *American Journal of Political Science*.
  - Keele, L., 2006. How to be smooth: automated smoothing in political science. *Unpublished Manuscript, Ohio State University*.
  - Beck, N. and Jackman, S., 1998. Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, pp.596-627.

Week 9: Tree-Based Methods and Ensembles (10/22)

- Reviewed Readings:
  - ISLR, chapter 8.
- Student Readings:
  - Montgomery, J.M. and Olivella, S., 201[illegible]ree-Based Models for Political Science Data. *American Journal of Political Science*, *62*(3), pp.729-744.
  - Green, D.P. and Kern, H.L., 2012. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive re[illegible]ression trees. *Public opinion quarterly*, *76*(3), pp.491-511.
  - Kaufman, A.R., Kraft, P. and Sen, M., 2019. Improving Supreme Court Forecasting Using Boosted Decision Trees. *Political Analysis*, *27*(3), pp.381-387.
  - Bisbee, J., 2019. BARP: Improving Mister P Using Bayesian Additive Regression Trees. *American Political Science Review*, *113*(4), pp.1060-1065.

Week 10: Neural Networks (10/29)

- Reviewed Readings:
  - ESL, chapter 11.
- Student Readings:
  - Rodriguez, P.L., and Spirling, A. Forthcoming. Word Embeddings: What Works, What Doesn't, and How to Tell the Difference in Applied Research. *Journal of Politic*
  - Cantú, F., 2019. The fingerprints of fraud: Evidence from Mexico's 1988 presidential election. *American Political Science Review*, *113*(3), pp.710-726.
  - Beck, N., King, G. and Zeng, L., 2000. Improving quantitative studies of international conflict: A conjecture. *American Political science review*, pp.21-35. [Note: might also want to look at De Marchi, S., Gelpi, C. and Grynaviski, J.D., 2004. Untangling neural nets. *American Political Science Review*, pp.371-378; and Beck, N., King, G. and Zeng, L., 2004. Theory and evidence in international conflict: a response to de Marchi, Gel[illegible]and Grynaviski. *American Political Science Review*, pp.379-389; for more commentary.]
  - Torres, M., 2018. *Framing a P[illegible]est: Determinants and Effects of Visual Frames*. Working Paper.

Week 11: Support Vector Machines (11/5) [Last day to drop with a W, 11/3]

- Reviewed Readings:
  - ISLR, chapter 9.
- Student Readings:
  - Hainmueller, J. and Hazlett, C., 2014. Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, pp.143-168.
  - Rheault, L., Rayment, E. and Musulan, A., 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, *6*(1), p.2053168018816228.
  - Dube, A., Jacobs, J., Naidu, S. and Suri, S., 2020. Monopsony in online labor markets. *American Economic Review: Insights*, *2*(1), pp.33-46.
  - Hager, A. and Hilbig, H., 2020. Does Public Opinion Affect Political Speech?. *American Journal of Political Science*.

Week 12: Unsupervised Learning (11/12)

- Reviewed Readings:
  - ISLR, chapter 10.
- Student Readings:
  - Jang, J. and Hitchcock, D.B., 2012. Model-Based Cluster Analysis of Democracies. *Journal of Data Science*, *10*.
  - Denny, M.J. and Spirling, A., 2018. Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, *26*(2), pp.168-189.
  - Slapin, J.B. and Proksch, S.O., 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, *52*(3), pp.705-721.
  - Grimmer, J. and King, G., 2011. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), pp.2643-2650.

Week 13: ML Tips and Tricks (11/19)

- Reviewed Readings:
  - None
- Student Readings:
  - None

Week 14: No class. Happy Thanksgiving

Week 15: Student Presentations (12/3)

- Student slides for presentations should be turned in by 12/1.
- All students should read other students' slides prior to class and come ready to discuss and critique.

Final Papers Due 12/10 on Blackboard Turnitin Link.