PSC 8120
Special Topics: Machine Learning for Social Science
Spring 2021

PSC 8120: Machine Learning for Social Science
Spring 2021
Sync Session: Mondays 12:45pm-3:15PM ET

Instructor: Iris Malone
Phone: 202-994-0992
Email: irismalone@gwu.edu
Virtual Office Hours: 3pm-4pm ET Wednesday, 3pm-4pm ET Thursday, or by appointment
Office Hour Sign-Ups: https://www.wejoinin.com/irismalone@gwu.edu

Overview

This course provides an overview of machine learning techniques for use in social science. Machine learning involves the use of non-linear and even non-parametric modeling for use in pattern recognition, classification, and prediction. The growth in big data creates huge opportunities for social scientists to leverage this data in new and exciting ways. The course will survey supervised and unsupervised machine learning techniques, show students how to use these tools in the programming language R, and apply these methods to different social science topics, including election forecasting, threat assessment, and text analysis.

Course Prerequisites

An introductory statistics course and familiarity with the programming language R is recommended.

Learning Outcomes

By the end of the course, students will be able to:

- compare and contrast parametric vs non-parametric modeling approaches
- use R to train different machine learning models
- know which algorithms to use for different classification or prediction problems
- evaluate the predictive performance of different classification algorithm
- visualize and interpret machine learning results

Assignments and Evaluation

- **Problem Sets (70%):** There will be seven problem sets due approximately every two weeks. Students are encouraged to collaborate on problem sets together, but must write up their problem sets on their own.
- **Final Project (30%):** Students have two options for their final project. First, students can do a replication study of an existing study and apply a machine learning algorithm to see whether the results change using non-parametric approach. Alternatively, students can use machine learning skills for an original project and write a short research note on the results. For either option, students will create a virtual poster summarizing their findings for the final session and submit a write-up (3000-4000 words) to the professor by **5pm ET May 6**.

Course Materials and Technology Requirements

Books

- Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* (ISLR) The book can be accessed online and downloaded here: http://www-bcf.usc.edu/~gareth/ISL/index.html

Software

In this course, we will be using R, a statistical software that can be downloaded and installed for free (https://cran.r-project.org/). To write and edit R code for problem sets, I recommend using R Studio. We will briefly use Python when we turn to web-scraping. I recommend using Jupyter notebook.

Technology Requirements

Blackboard will be used for posting course files, modules, and assignments and for communicating with the class. We will use Zoom for synchronous sessions. You are already enrolled for this course on Blackboard if you have completed registration for the course. It is your responsibility to periodically check the course site (log in at http://blackboard.gwu.edu/ Using your gwu.edu address) for updates to the syllabus/readings.

As an online student, it is necessary to possess baseline technology skills in order to participate fully in the course. Please consult online.gwu.edu/student-support for further information about recommended configurations and support.

Course Format

Workload Expectations

In this 3 credit graduate course students are expected to work for 7.5 hours per week (this includes 2.5 hours of synchronous class time per week) totaling to 112.5 hours of work over the duration of this 15-week semester.

Methods of Instruction

What is Synchronous?

1. **Weekly Zoom Discussions** (Mondays 12:45pm-3:00pm ET): We will have a *semi-flipped classroom.* The first hour of the course will be a lecture. The second hour of the course will walk through different applied examples and coding exercises in R. Zoom is a collaborative meeting tool that allows for real-time video meetings and sharing computer content over the web. You should familiarize yourself with the topics and cases for discussion before coming to class. **Please adjust for the time zone difference if you are not in ET.**

What is Asynchronous?

1. **Independent Readings:** Please read the ISLR text and come prepared to discuss before our weekly discussion. Applied example readings are optional.

Submission of Assignments and Late Work

Submission of Papers

- Assignments are to be submitted to the professor via email before the deadline. The assignments are intended to expand upon the lecture material and to help students develop the actual skills that will be useful for applied work. Portions of the homework completed in R should be submitted using R markdown, a markup language for producing well-formatted HTML documents with embedded R code and outputs. R mark- down requires installation of the *knitr* package. We recommend using Rstudio, a user interface for R, which is set up well for the creation of R markdown documents. An assignment is considered late if the professor cannot successfully open the document before the deadline.
- Please name your files using the header LastName-Assignment, e.g. Malone-PS1.pdf

Late Work and Extensions

- Late papers will be penalized at a rate of one-half letter grade (e.g. A to A-) for every 12 hours past the deadline. They are not accepted after 48 hours.
- Extensions for late work will not be accepted except in extenuating circumstances (e.g. illness, family emergency).

Contact and Virtual Office Hours

Email is the best way to contact me. I promise to respond to your emails within 24 hours. I will hold office hours via Zoom on Wednesdays and Thursdays. Please sign-up in advance using wejoinin.com/irismalone@gwu.edu. If those hours do not work for you, please email for an individual appointment.

University Policies

- **Plagiarism, Cheating, and Academic Integrity:** Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information. For details and complete code, see studentconduct.gwu.edu/code-academicintegrity
- **Sharing of Course Content** Unauthorized downloading, distributing, or sharing of any part of a recorded lecture or course materials, as well as using provided information for purposes other than the student's own learning may be deemed a violation of GW's Student Conduct Code.
- **Observance of Religious Holidays:** In accordance with University policy, students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance. For details and policy, see: provost.gwu.edu/policies-procedures-and-guidelines
- **Security and Safety:** In an emergency: call GWPD 202-994-6111 or 911. For situation-specific actions: review the Emergency Response Handbook: safety.gwu.edu/emergencyresponse-handbook In an active violence situation: Get Out, Hide Out or Take Out: For more info, see go.gwu.edu/shooterprep and safety.gwu.edu/stay-informed
- **Alert DC:** Alert DC provides free notification by e-mail or text message during an emergency. Visit GW Campus Advisories for a link and instructions on how to sign up for alerts pertaining to GW. If you receive an Alert DC notification during class, you are encouraged to share the information immediately.
- **GW Alert:** GW Alert provides popup notification to desktop and laptop computers during an emergency. In the event that we receive an alert to the computer, we will follow the instructions given. You are also encouraged to download this application to your personal computer. Visit GW Campus Advisories to learn how.

Student Support and Resources

- **Virtual Academic Support** A full range of academic support is offered virtually in spring 2021. See coronavirus.gwu.edu/top-faqs for updates. Tutoring and

course review sessions are offered through Academic Commons in an online format. See academiccommons.gwu.edu/tutoring Writing and research consultations are available online. See academiccommons.gwu.edu/writing-research-help. Coaching, offered through the Office of Student Success, is available in a virtual format. See studentsuccess.gwu.edu/academic-program-support

- **Commitment to Inclusive Teaching:** Higher education works best when it encourages a vigorous exchange of ideas in which all points of view are heard. Free expression in the classroom is an integral part of the process. At the same time, this process is most effective when all approach the enterprise with empathy and respect for others, irrespective of their ideology, views, or identity. I encourage you to report bias incidents here: https://diversity.gwu.edu/bias-incident-response.

- **Disabilities and Accommodations:** If you need disability accommodations, please register with Disability Support Services (DSS). If you have questions about disability accommodations, contact DSS at 202-994-8250 or dss@gwu.edu or visit them in person in Rome Hall, Suite 102. For information about how the course technology is accessible to all learners, see the following resources:

  - https://www.blackboard.com/blackboard-accessibility-commitment
  - https://corp.kaltura.com/products/video-accessibility/
  - https://voicethread.com/about/features/accessibility/

- **Counseling and Psychological Services:** The University's Mental Health Services offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include: crisis and emergency mental health consultations confidential assessment, counseling services (individual and small group), and referrals. For additional information see: counselingcenter.gwu.edu/

Course Calendar

1. **January 11: Overview**

   - Topics:
     - Supervised vs Unsupervised Methods
     - Prediction vs Inference
     - Bias-Variance Trade-Off
   - Readings
     - ISLR Chp 2.
     - Henry E. Brady. 2019. "The Challenge of Big Data and Data Science." *Annual Review of Political Science* 22 (1): 297–323
     - Justin Grimmer, Molly Roberts, and Brandon Stewart. "Machine Learning for Social Science." *Annual Review of Political Science.* Forthcoming.
     - Michael D. Ward, Brian D. Greenhill, and Kristin M. Bakke. 2010. "The perils of policy by p-value: Predicting civil conflicts." *Journal of peace research* 47 (4): 363–375

2. **January 18: MLK Jr Day (No Class)**

Unit 1: Supervised Methods

3. **January 25: Regression and Classification**

   - Topics:
     - Review: Linear Regression
     - Logistic
     - Linear Discriminant Analysis
     - K-Nearest Neighbors
   - Readings
     - ISLR Chp. 2, p. 39-42
     - ISLR Chp. 3-4
   - Applied Examples
     - Adam Bonica. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" [in en]. *American Journal of Political Science* 62 (4): 830–848
     - Azusa Katagiri and Eric Min. 2019. "The Credibility of Public and Private Signals: A Document-Based Approach." *American Political Science Review* 113 (1): 156–172

  - Tamar Mitts. 2019a. "From isolation to radicalization: anti-Muslim hostility and support for ISIS in the West." *American Political Science Review* 113 (1): 173–194

4. **February 1: Non-Linear Models**

   - Topics
     - Splines
     - Generalized Additive Models
     - Local Regression
     - Polynomial Regression
   - Readings
     - ISLR Chp. 7
     - Nathaniel Beck and Simon Jackman. 1998. "Beyond Linearity by Default: Generalized Additive Models." *American Journal of Political Science* 42 (2): 596–627
     - Nathaniel Beck, Jonathan N. Katz, and Richard Tucker. 1998. "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable." *American Journal of Political Science* 42 (4): 1260–1288
     - David B. Carter and Curtis S. Signorino. 2010. "Back to the future: Modeling time dependence in binary data." *Political Analysis* 18 (3): 271–292

5. **February 8: Model Selection and Assessment**

   - **Problem Set 1 Due 12pm ET**
   - Topics
     - Performance Metrics (AUC, Accuracy, F-Score, Kappa Score)
     - Resampling
     - Bootstrap
     - Cross-Validation
       * K-Fold CV
       * LOOCV
   - Readings
     - ISLR Chp. 5
   - Applied Examples
     - Erik J. Engstrom. 2012. "The Rise and Decline of Turnout in Congressional Elections: Electoral Institutions, Competition, and Strategic Mobilization" [in en]. *American Journal of Political Science* 56 (2): 373–386

  - David Siroky et al. 2016. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data." *Political Analysis* 24 (1): 87–103
  - Marcel Neunhoeffer and Sebastian Sternberg. 2019. "How Cross-Validation Can Go Wrong and What to Do About It." *Political Analysis* 27, no. 1 (January): 101–106
  - Yu Wang. 2019. "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data: A comment." *Political Analysis* 27 (1): 107–110

6. **February 15: President's Day (No Class)**

7. **February 22: Variable Selection and Assessment**

   - **Problem Set 2 Due 12pm ET**
   - Topics
     - Curse of Dimensionality
     - Variable Selection
     - Shrinkage Methods
       * LASSO
       * Ridge Regression
     - Dimension Reduction
   - Readings
     - ISLR Chp. 6
     - Applied Examples
       * Sonali Singh and Christopher R. Way. 2004. "The correlates of nuclear proliferation: A quantitative test." *Journal of Conflict Resolution* 48 (6): 859–885
       * Mark S. Bell. 2016. "Examining explanations for nuclear proliferation." *International Studies Quarterly* 60 (3): 520–529

8. **March 1: Tree-Based Methods I**

   - **Problem Set 3 Due 12pm ET**
   - Topics
     - CART
     - Decision Tree
     - Bagging
   - Readings

  - ISLR Chp. 8, p. 303-316
- Applied Examples
  - Daniel W. Hill and Zachary M. Jones. 2014. "An empirical evaluation of explanations for state repression." *American Political Science Review* 108 (3): 661–687
  - Zachary M. Jones and Yonatan Lupu. 2018. "Is there more violence in the middle?" *American Journal of Political Science* 62 (3): 652–667
  - Shea Streeter. 2019. "Lethal force in black and white: Assessing racial disparities in the circumstances of police killings." *The Journal of Politics* 81 (3): 1124–1132

9. **March 8: Tree-Based Methods II**
   - Topics
     - Random Forests
     - Gradient Boosting (GBM)
     - BART (Bayesian Additive Regression Trees)
   - Readings
     - ISLR Chp 8, p. 316-330
     - Jacob M. Montgomery and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–744
   - Applied Examples
     - Iris Malone. "Group-Level Predictors of Civil War." Working Paper.

10. **March 16: Spring Break (No Class)**

11. **March 22: Support Vector Machines**
    - **Problem Set 4 Due 12pm ET**
    - Topics
      - Hyperplanes
      - Support Vector Classifiers
      - Naive Bayes Classifier
    - Readings
      - ISLR Chp 9
    - Applied Examples

  - Vito D'Orazio et al. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22 (2): 224–242
  - Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. "Detecting opinion spams and fake news using text classification" [in en]. *Security and Privacy* 1 (1): e9

12. **March 29: Neural Networks and Deep Learning**

    - Topics
      - Sequential Data
      - Convolutional Neural Networks
      - Recurrent Neural Networks
    - Readings
      - Trevor Hastie, Robert Tibshirani, and Jerome Friedman. "'Elements of Statistic Learning." *Springer Series in Statistics.* 2013. Chp 11.
    - Applied Examples
      - Nathaniel Beck, Gary King, and Langche Zeng. 2000. "Improving quantitative studies of international conflict: A conjecture." *American Political science review,* 21–35
      - Scott De Marchi, Christopher Gelpi, and Jeffrey D. Grynaviski. 2004. "Untangling neural nets." *American Political Science Review,* 371–378
      - Nathaniel Beck, Gary King, and Langche Zeng. 2004. "Theory and evidence in international conflict: a response to de Marchi, Gelpi, and Grynaviski." *American Political Science Review,* 379–389
      - Dean Knox and Christopher Lucas. 2019. *A Dynamic Model of Speech for the Social Sciences.* SSRN Scholarly Paper ID 3490753. Rochester, NY: Social Science Research Network

Unit 2: Unsupervised Methods

13. **April 5: Principal Component Analysis and Clustering**

    - **Problem Set 5 Due 12pm ET**
    - Topics
      - Principal Component Analysis
      - K-Means Clustering
      - Hierarchical Clustering
      - Factor Analysis

- Readings
  - ISLR Chp. 10
- Applied Examples
  - Cullen S. Hendrix. 2010. "Measuring state capacity: Theoretical and empirical implications for the study of civil conflict." *Journal of Peace Research* 47 (3): 273–285
  - Iris Malone. 2019. "Unmasking Militants: Organizational Trends in Armed Groups, 1970-2012." *Working Paper,* George Washington University

14. **April 12: Text as Data/Web-scraping**

- Topics
  - Quick Intro to HTML/Python
  - Beautiful Soup
  - Web-Scraping
  - Pre-Processing (Tokenize, Stemmers)
- Readings
  - Justin Grimmer and Brandon M. Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21 (3): 267–297

15. **April 19: Text Analysis I**

- **Problem Set 6 Due 12pm ET**
- Topics
  - Dictionary-Based Methods
  - Sentiment Analysis
  - Distinctive Words
- Applied Examples
  - Arthur Spirling. 2012. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." *American Journal of Political Science* 56 (1): 84–97
  - Amaney A. Jamal et al. 2015. "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses" [in en]. *Perspectives on Politics* 13, no. 1 (March): 55–73
  - Ruowei Liu et al. 2020. "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach." *Annals of GIS* 0, no. 0 (October): 1–14
  - Erin Baggott Carter and Brett L. Carter. 2020. "Propaganda and Protest in Autocracies." *Journal of Conflict Resolution*

16. **April 26: Text Analysis II**

    - Topics
      - Distances
      - Clustering
      - Topic Models (LDA, STM)
    - Applied Examples
      - Margaret E. Roberts et al. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58 (4): 1064–1082
      - Hannes Mueller and Christopher Rauh. 2018. "Reading between the lines: Prediction of political violence using newspaper text." *American Political Science Review* 112 (2): 358–375
      - Tamar Mitts. 2019b. "Terrorism and the Rise of Right-Wing Content in Israeli Books." *International Organization* 73 (1): 203–224
      - Erin Rossiter. 2020. "Measuring Agenda Setting in Interactive Political Communications." *Working Paper,* Washington University St Louis

17. **April 29: Poster Session (Designated Monday)**

    - **Problem Set 7 Due 12pm ET**
    - **Final Project Due May 6 5pm ET**