POL 290G: Statistical Learning in the Social Sciences

W, 11am–1:50pm, Kerr 594

`https://canvas.ucdavis.edu/courses/876614`

Instructor

Christopher Hare
cdhare@ucdavis.edu
Office Location: Kerr 574
Office Hours: Mondays 11am–1pm

Course description

A growing number of social scientists are taking advantage of machine learning methods to enhance prediction *and* inference when analyzing social science data. This course covers the mechanics underlying machine learning methods and discusses how these techniques can be leveraged by social scientists to gain new insight from their data. Specifically, we will cover both supervised and unsupervised methods: decision trees, random forests, boosting, support vector machines, neural networks, deep and adversarial learning, ensemble learning, principal components analysis, factor analysis, and manifold learning/ multidimensional scaling. We will also discuss best practices in fitting and interpreting these models, including cross-validation techniques, bootstrapping, and presenting output. The course will demonstrate how these models can be estimated in `R`.

Texts

I include **many** books, articles, and other sources in this syllabus. I do not expect (or even necessarily advise) you to read all of them, but a big part of graduate school is simply compiling and organizing sources that may be useful for specific research purposes now or in the future. The following are "required" texts, but "strongly recommended" would be a better description.

1. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning with Applications in R*. 2nd ed. New York: Springer. `https://www.statlearning.com/`

2. Boehmke, Bradley and Brandon Greenwell. 2019. *Hands-On Machine Learning with R*. Boca Raton, FL: CRC Press. `https://koalaverse.github.io/homlr/`

3. Chollet, Francois, J.J. Allaire, and Tomasz Kalinowski. 2022. *Deep Learning with R*. 2nd ed. Shelter Island, NY: Manning.

Recommended texts and readings

1. Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition. New York: Springer.

2. Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16 (3): 199-231.

3. Efron, Bradley and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. New York: Cambridge University Press.

4. Watt, Jeremy, Reza Borhani, and Aggelos K. Katsaggelos. 2020. *Machine Learning Refined: Foundations, Algorithms, and Applications*, second edition. New York: Cambridge University Press.

5. Izenman, Alan Julian. 2013. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York: Springer

6. Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling.* New York: Springer.

7. Duboue, Pablo. 2020. *The Art of Feature Engineering: Essentials for Machine Learning.* New York: Cambridge University Press.

8. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* Cambridge, MA: MIT Press.

9. Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: MIT Press.

10. Berk, Richard A. 2016. *Statistical Learning from a Regression Perspective*, second edition. New York: Springer.

11. Buduma, Nikhil and Nicholas Locascio. 2017. *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms.* Sebastopol, CA: O'Reilly.

12. Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67.

13. Zhao, Qingyuan and Trevor Hastie. 2021. "Causal Interpretations of Black-Box Models." *Journal of Business & Economic Statistics* 39 (1): 272-281.

14. **Social science guides and applications**

    (a) Athey, Susan and Guido W. Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11 (1): 685-725.

    (b) Green, Jon and Mark H. White II. 2023. *Machine Learning for Experiments in the Social Sciences.* Cambridge Elements in Experimental Political Science. New York: Cambridge University Press.

    (c) Justin Grimmer, Margaret E. Roberts, Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24: 1-20.

(d) Mullainathan, Susan and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87-106.

(e) Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhill Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. 2021. "Integrating Explanation and Prediction in Computational Social Science." *Nature* 595: 181–188.

(f) Grimmer, Justin, Solomon Messing, and Sean J. Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25 (4): 413-434.

(g) Ramirez, Christina M., Marisa A. Abrajano, and R. Michael Alvarez. 2019. "Using Machine Learning to Uncover Hidden Heterogeneities in Survey Data." *Scientific Reports* 9 (1): 160-61.

(h) Shatte, Adrian B. R., Delyse M. Hutchinson, and Samantha J. Teague. 2019. "Machine Learning in Mental Health: a Scoping Review of Methods and Applications." *Psychological Medicine* 49 (9): 1426–1448.

(i) Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris. 2018. "Machine Learning Approaches for Clinical Psychology and Psychiatry." *Annual Review of Clinical Psychology* 14 (1): 91-118 .

(j) Athey, Susan and Stefan Wager. 2021. "Policy Learning With Observational Data." *Econometrica* 89 (1): 133-161.

(k) Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47 (2): 1148–1178.

(l) Wager, Stefan and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228-1242.

(m) Jang, Jaewon, and David B. Hitchcock. 2012. "Model-Based Cluster Analysis of Democracies." *Journal of Data Science* 10: 297-319.

(n) Alvarez, R. Michael, Ines Levin, and Lucas Núñez. 2017. "The Four Faces of Political Participation in Argentina: Using Latent Class Analysis to Study Political Behavior." *Journal of Politics* 79 (4): 1386-1402.

(o) Ludwig, Jens and Sendhil Mullainathan. 2023. "Machine Learning as a Tool for Hypothesis Generation." *National Bureau of Economic Research Working Paper Series.* `http://www.nber.org/papers/w31017`

(p) Wankmüller, Sandra. 2022. "Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis." *Sociological Methods & Research.* In press.

(q) Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19-42.

15. **R**

(a) Ismay, Chester and Albert Y. Kim. 2019. *Statistical Inference via Data Science: A ModernDive into R and the Tidyverse.* Boca Raton, FL: CRC Press.

(b) Kuhn, Max and Julia Silge. 2022. *Tidy Modeling with R: A Framework for Modeling in the Tidyverse.* Sebastapol, CA: O'Reilly.

Grade components

Assignments/problem sets	30%
Presentation	25%
Final project	25%
Participation	20%

Grading system

≥ 97	A+	87 - 89	B+	77 - 79	C+	67 - 69	D+	< 60	F
93 - 96	A	83 - 86	B	73 - 76	C	63 - 66	D		
90 - 92	A-	80 - 82	B-	70 - 72	C-	60 - 62	D-		

Assignments/problem sets

Assignments will be given out throughout the quarter. You can pick any three to complete (worth 10 points each).

Presentation

Find a a substantive (ideally in political science [or whatever your field of study is]) application of a machine learning method and provide a 10-15 minute presentation in class.

Final project

Complete a project (ideally a term paper) that uses one or more of the methods covered in class. The goal is to apply these methods to data **you** care about!

Course materials

Course materials (including slides, `R` code, and assignments) will be available on Canvas.

Academic honesty

I strictly adhere to the UC Davis Code of Academic Conduct (`http://sja.ucdavis.edu/cac.html`). I will report any suspected academic misconduct to Student Judicial Affairs.

Disabilities

Students with special needs that require accommodation should notify the Student Disability Center (54 Cowell Building; `530.752.3184`; `https://sdc.ucdavis.edu`) as soon as possible so that arrangements detailed in the Letter of Accommodation can be made.

Required software

`R` is available for download from CRAN (Comprehensive R Archive Network): `https://cran.r-project.org/`. You'll want to install the most recent (compatible) version. If installing on Windows, I recommending also downloading and installing Rtools (this is optional, but comes in handy if you ever need to compile or test a package).

I generally recommend the use of a text (or line) editor. Line editors are designed for writing and modifying programming code, and have useful functionality (e.g., macros) for programmers. Sublime and Notepad++ are both good choices that are open-source and free.

Many people prefer to use `RStudio` to run `R`, which is perfectly fine. Indeed, this is the approach we will take in the class!

Course materials

Course materials (including slides, code, and problem sets) will be available on the course Canvas page.

Tentative schedule

- **3 April: NO CLASS (Midwest meetings)**
- **10 April: Introduction**
- **17 April: Tuning and extensions of the regression model**

  Introduction to Statistical Learning with R, chapters 2 and 5.
  Hands-on Machine Learning with R, chapter 7.

  The bias-variance tradeoff and error rates

  Cross-validation and tuning

  Multivariate Adaptive Regression Splines (MARS)

  Shrinkage/regularization methods and the LASSO

- **24 April: Trees I**

  Introduction to Statistical Learning with R, chapter 8.
  Hands-on Machine Learning with R, chapters 9–10.

  Single decision trees

  Bagging

- **1 May: Trees II**

  Hands-on Machine Learning with R, chapters 11–12.
  Montgomery, Jacob M. and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729-744.

  Random forests

  Boosting

- **8 May: Interpretable machine learning**

  Hands-on Machine Learning with R, chapter 16.
  Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, second edition. `christophm.github.io/interpretable-ml-book`

  Assessing variable importance and effects

  Partial dependency plots and model visualization

- **15 May: Neural nets/deep learning I**

  Introduction to Statistical Learning with R, chapter 10.
  Deep Learning with R, chapters 1–3.
  Hands-on Machine Learning with R, chapter 13.

  Foundations

- **22 May: Neural nets/deep learning II**

  Deep Learning with R, chapters 4–6.
  LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521: 436-444.

  Keras/Tensorflow

  Text data

- **29 May: Neural nets/deep learning III**

  Deep Learning with R, chapters 10–11.

  Text data (continued)

  Time series data

- **5 June: Unsupervised learning**

  Introduction to Statistical Learning with R, chapter 12.
  Hands-on Machine Learning with R, chapter 17.
  Blaydes, Lisa and Justin Grimmer. 2020. "Political Cultures: Measuring Values Heterogeneity." *Political Science Research and Methods* 8(3): 571-579.

  k-means clustering

  Principal components analysis

  Manifold learning/multidimensional scaling