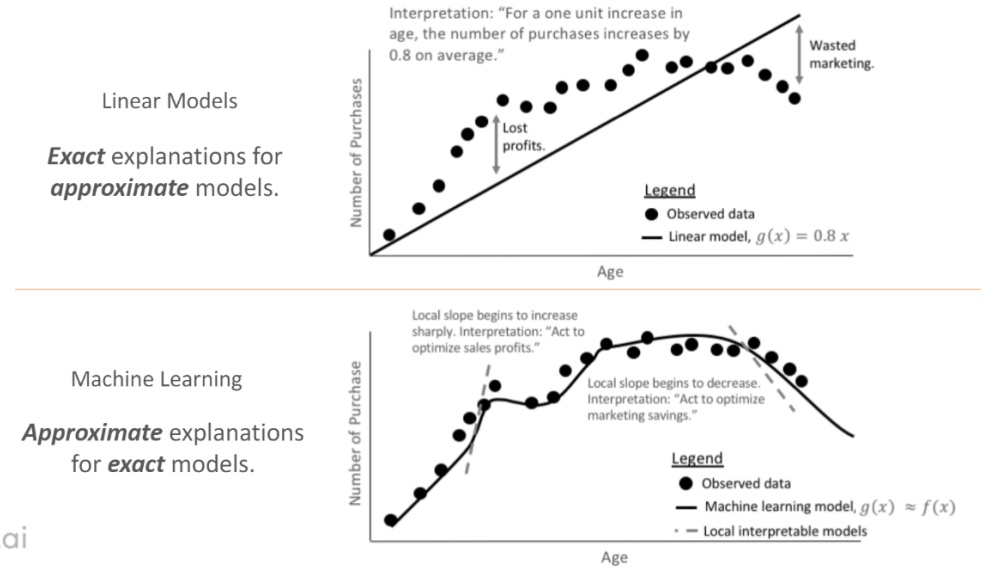
Machine Learning: Applications in Social Science Research
ICPSR Summer Program in Quantitative Methods of Social Research
22 June – 17 July, 2019 (3-5pm EST)



Instructor:
Christopher Hare
Assistant Professor
Department of Political Science
University of California, Davis
cdhare@ucdavis.edu
Office Hours: 10am-12pm EST

TAs:
Sam Fuller
PhD Student
Department of Political Science
University of California, Davis
sjfuller@ucdavis.edu
Office Hours: 12-2pm EST

Patrick Wu
PhD Student
Department of Political Science
University of Michigan
pywu@umich.edu
Office Hours: 5-7pm EST

Course Description:

A growing number of social scientists are taking advantage of machine learning methods to uncover hidden structure in their data, improve model predictive power, and gain a better understanding of complex relationships between variables. This workshops 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, the workshop 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 workshop will demonstrate how these models can be estimated in `R` (and, time permitting, `Python`).

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

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. Boyd, Stephen and Lieven Vandenberghe. 2018. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares.* New York: Cambridge University Press.

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

13. 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.

14. Sechidis, Konstantinos and Gavin Brown. 2018. "Simple Strategies for Semi-supervised Feature Selection." *Machine Learning* 107 (2): 357-395.

15. **R**

    (a) James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R.* New York: Springer.

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

16. **Python**

(a) VanderPlas, Jake. 2017. *Python Data Science Handbook: Essential Tools for Working with Data*. Sebastapol, CA: O'Reilly.

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 highly 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. Chris recommends the line editor Atom, which is open-source and free (`https://atom.io/`). Sublime and Notepad++ are other popular options.

Many people prefer to use `RStudio` to run `R`, which is perfectly fine.

You'll want to install the following packages in `R`:

```
install.packages(c("BMS","caret","ClassDiscovery","corrplot",
"doParallel","dplyr","extraTrees","fastICA","foreach","foreign","gbm",
"GenAlgo","ggfortify","ggplot2","kernlab","lavaan","MASS","MCMCpack",
"mlbench","nnet","pROC","quadprog","randomForest","RANN"))
```

Course materials: Course materials (including slides, code, and problem sets) will be available on the course Canvas page. Four (± 1) assignments will be given over the workshop.

Tentative Schedule:
This schedule is subject to change:

- **Week One: Machine Learning: Theory and Concepts**

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

  Computational Learning Theory and the Development of Machine Learning

  The Bias-Variance Tradeoff and Error Rates

  Model Validation and Tuning

  Resampling Techniques

  Predictions and Counterfactuals

  Quick Review of Linear Regression Models

  Programming in `R`

  Computing Performance and Practical Tips

- **Week Two: Supervised and Semi-Supervised Learning**

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

  Generalized Linear Models and Extensions

  Shrinkage/Regularization Methods and the Lasso

  Regression Splines and Generalized Additive Models

  Linear and Flexible Discriminant Analysis

  Naive Bayes

  Bayesian Model Averaging

  Neural Networks and Generative Adversarial Networks

  Graphical Models

  Support Vector Machines and Relevance Vector Machines

  k-nearest Neighbors

- **Week Three: Tree-Based Methods and Learning Ensembles**

  → Zhao, Qingyuan and Trevor Hastie. 2019. "Causal Interpretations of Black-Box Models." *Journal of Business & Economic Statistics*, forthcoming.

  → 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.

  Classification and Regression Trees

  Ensemble Methods: Random Forests and Boosting

  Assessing Variable Importance and Effects

  Partial Dependency Plots and Model Visualization

  Ensemble Modeling and Heterogeneous Treatment Effects

- **Week Four: Unsupervised Learning**

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

  k-means Clustering

  Principal Components Analysis

  Manifold Learning and Multidimensional Scaling

  Self-Organizing Maps

  Deep Learning

  Mixture Models and Latent Class Analysis

  Novelty/Outlier Detection