Machine Learning: Applications and Opportunities in the Social Sciences
ICPSR Summer Program in Quantitative Methods of Social Research
24 June-19 July, 2019, Ann Arbor, MI

Instructor: Christopher Hare
Assistant Professor
Department of Political Science
University of California, Davis
cdhare@ucdavis.edu

TA: Sam Fuller
PhD Student
Department of Political Science
University of California, Davis
sjfuller@ucdavis.edu

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/Readings:


11. R

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

12. Python


Course materials: Course materials (including slides, code, and problem sets) will be available on a private Dropbox folder.

Tentative Schedule:
This schedule is subject to change:

- **Week One: Machine Learning: Theory and Concepts**
  - 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**
  - 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
  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
  $k$-means Clustering
  Principal Components Analysis
  Manifold Learning and Multidimensional Scaling
  Self-Organizing Maps
  Deep Learning
  Mixture Models and Latent Class Analysis
  Novelty/Outlier Detection