Machine Learning: Applications and Opportunities in Social Science Research

ICPSR Summer Program in Quantitative Methods of Social Research
9-13 July, 2018, Ann Arbor, MI

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

Course Description:
The field of machine learning is most commonly associated with "big data": how we can use massive
datasets to make better predictions about things like credit card fraud, Netflix recommendations,
and the like. Though machine learning has been most influential in its commercial and medical
applications, a growing number of social scientists are taking advantage of these methods to: (1)
uncover patterns and structure embedded in data, (2) test and improve model specification and
predictions, and (3) perform data reduction. 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, the course will cover: decision trees, random forests,
boosting, k-means clustering and nearest neighbors, support vector machines, kernels, neural
networks, and ensemble learning. We will also discuss topics related to best practices, including
error rates, cross-validation, and the use of bootstrapping methods to develop uncertainty estimates.
The course will demonstrate how to estimate these models in both R and Python, as well as methods
for interpreting and presenting model output.

Recommended Texts/Readings:


   MA: MIT Press.

   MIT Press.


   New York: Springer.


   for Estimating Heterogeneous Treatment Effects using Machine Learning.”


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:

- **Monday, July 9: 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

- **Tuesday, July 10: Supervised and Semi-Supervised Learning I**
  - 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
• **Wednesday, July 11: Supervised and Semi-Supervised Learning II**  
  Neural Networks and Generative Adversarial Networks  
  Graphical Models  
  Support Vector Machines and Relevance Vector Machines  
  \( k \)-nearest Neighbors

• **Thursday, July 12: Tree-Based Methods and Heterogeneous Treatment Effects**  
  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

• **Friday, July 13: Unsupervised Learning**  
  \( k \)-means Clustering  
  PCA and Factor Analysis  
  Manifold Learning and Multidimensional Scaling  
  Deep Learning  
  Mixture Models and Latent Class Analysis  
  Programming in Python (*time permitting*)