Machine Learning for Social Science

Summer Institute
July 2023

Time: July 12-19, 2023
Synchronous Zoom Classes Monday, Wednesday, Friday 10am - 12pm

Instructor: Dr. Brian Kim, PhD
kimbrian@umd.edu

Overview:

This course provides an introduction to supervised statistical learning techniques such as decision trees, random forests and boosting and discusses their potential application in the social sciences. These methods focus on predicting an outcome Y based on some learned function f(X) and therefore facilitate new research perspectives in comparison with traditional regression models, which primarily focus on causation. Predictive methods also provide a valuable extension to the empirical social scientists' toolkit as new data sources become more prominent. In addition to introducing supervised learning methods, the course will include practical sessions to demonstrate how to tune and evaluate prediction models using the statistical programming language R.

Learning Outcomes:

After taking this course, students will be able to:

- Implement the full machine learning workflow, including data preparation, model building, validation, and testing.
- Describe the bias-variance tradeoff and how it affects the machine learning process.
- Use R to fit a wide variety of machine learning models for both regression and classification.
- Critically evaluate models based on performance metrics as well as bias and fairness measures.

Class Meetings:

This course uses a flipped classroom design. In this course, you are responsible for watching video recorded lectures and going through the readings, and then attending class meetings where students have the chance to discuss the materials from a unit with the instructor. In general, the class time will be used to answer questions as well as walking through R code examples.

Prerequisites:

Basic knowledge of R is required for this course. Students who are not proficient in R are encouraged to work through one or more tutorials prior to this class. Some resources can be found here:

https://www.rstudio.com/online-learning/#R
https://rstudio.cloud/learn/primers
http://www.statmethods.net/
https://swirlstats.com/

Course Schedule

Video Lectures should be viewed in between class meetings.

Recommended Video Lectures:

- Classification with Trees
- Training and Testing Models
- Evaluating Models
- Overview of Machine Learning Workflow

Wednesday, July 12: Introduction to Machine Learning

Lecture: Machine Learning Workflow, Bias-Variance Tradeoff

Video Lectures:

- K- Nearest Neighbors
- Performance Measures
- Introduction to Regularized Regression
- Lasso and Ridge Regression
- Elastic Net and Group Lasso
- Regularized Tuning and Cross-Validation

Assignment 1 Assigned

Friday, July 14: Fitting Models

Lecture: Fitting Models in R

Video Lectures:

- Introduction to Ensemble Methods
- Bagging
- Random Forests
- AdaBoost
- Gradient Boosting
- XGBoost

Assignment 1 Due. Assignment 2 Assigned.

Monday, July 17: Ensemble Methods

Lecture: Ensemble Methods in R

Video Lectures:

- Variable Importance
- PDP, ICE, and ALE
- Surrogate Models
- ML Bias
- ML Fairness

Assignment 2 Due. Assignment 3 Assigned.

Wednesday, July 19: Machine Learning in Context

Lecture: ML in Context, Advanced ML methods, Unsupervised Learning

Assignment 3 Due.