upf. Universitat Pompeu Fabra Barcelona

RECSM Research and Expertise Centre for Survey Methodology

Machine Learning for Social Science

Week 3 (Regular course)

Duration: 12 hours

Format: In person and online

Instructor: Christoph Kern

Course description: 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 based on some learned function 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 (high dimensional) data sources become more prominent. In addition to introducing supervised learning methods, the course will include practical sessions to exemplify how to tune and evaluate prediction models using the statistical programming language R. The course aims to illustrate the covered concepts and methods from a social science perspective by discussing typical applications and social science research problems that may benefit from machine learning tools.

Learning schedule:

Day 1		Machine Learning Foundations Training error, test error, Bias-Variance Trade-Off Data splitting and Cross-Validation Applications in Social Science
		Break
		Regularized Regression (w. lab session) Variable Selection Ridge regression, Lasso regression, Elastic net
Day 2		Performance evaluation Performance metrics for regression Performance metrics for classification
		Break
		Classification and Regression Trees (CART) (w. lab session) Classification and Regression Trees (CART) CTREE and MOB
Day 3		Tree-based Ensemble Methods Bagging Random Forests, Extra Trees
		Break
		Boosting Methods & Interpretability (w. lab session) Adaboost, Gradient Boosting (GBM) ML Interpretability

Prerequisites: It is assumed that students have solid knowledge of basic statistics, including linear and logistic regression. Familiarity with the statistical programming language R is recommended but not strictly necessary. Students may work through one or more R tutorials prior to the first class meeting.

Software: R, including packages tidyverse, mlbench, glmnet, rpart, mlr3verse

Readings:

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning. New York, NY: Springer

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (Eds.). (2017). Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press Taylor & Francis Group.

Kuhn, M. and Johnson, K. (2013). Applied Predictive Modeling. New York, NY: Springer.

Instructor short bio:

Christoph Kern is Junior Professor of Social Data Science and Statistical Learning at the Ludwig-Maximilians-University of Munich and Project Director at the Mannheim Centre for European Social Research (MZES). He received his PhD in social science (Dr. rer. pol.) from the University of Duisburg-Essen in 2016. Before joining LMU Munich, he was a Post-Doctoral Researcher at the Professorship for Statistics and Methodology at the University of Mannheim and Research Assistant Professor at the Joint Program in Survey Methodology (JPSM) at the University of Maryland. His work focuses on the reliable use of machine learning methods and new data sources in social science, survey research, and algorithmic fairness.