Machine learning for social sciences

Week 3 (Regular course)

Duration: 12 hours/3 days

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 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 (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.

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. Some resources can be found here: https://rstudio.cloud/learn/primers

Course Objectives

At the completion of this course, students will have a profound understanding of tree-based prediction methods and the machine learning perspective on statistical modeling. Students will learn the computational skills to apply and evaluate these methods using R.

References

• 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 & Fr

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

Additional readings

• Kern, C., Klausch, T., and Kreuter, F.(2019). Tree-based Machine Learning Methods for Survey Research. Survey Research Methods 13(1), 73–93. Https://doi.org/10.18148/srm/2019.v1i1.7395

• Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. https://arxiv.org/abs/1908.09635.

• Molnar, C. (2020). Interpretable machine learning. A Guide for Making Black Box Models Explainable.https://christophm.github.io/interpretable-ml-book/

Short biography

Christoph Kern is 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. He also is a Project Director at the Mannheim Centre for European Social Research (MZES). He received his PhD (Dr. rer. pol.) in social science from the University of Duisburg-Essen (UDE) in 2016. His research focuses on the usage of machine learning methods in the social sciences and survey research.