Predictive and Causal Machine Learning in R
(taught by Prof. Martin Huber)

This course provides an introduction to predictive and causal machine learning based on the software "R".

Predictive machine learning aims at forecasting the value of an outcome of interest, e.g. sales or turnover, based on observing specific patterns of potentially relevant factors (or "predictors") like price, quality, weather, advertisement campaigns etc. That is, predictive machine learning "learns" from patterns among predictors in (past) data to forecast the value of the outcome in the future.

Causal machine learning aims at assessing the causal effect of some intervention or treatment, like offering or not offering a training program to jobseekers, on an outcome of interest, like employment. The assessment of a causal effect requires that groups receiving and not receiving the treatment are comparable in background characteristics which also affect their outcome (e.g. previous labor market history, education etc.). Causal machine learning can be used to generate such comparable groups in a data-driven way by estimating two separate models for how the characteristics affect the intervention and the outcome. Such approaches also permit detecting subgroups for whom the treatment effect is particularly large as a function of their observed characteristics (effect heterogeneity analysis). This is useful for optimally targeting specific subgroups by the treatment (optimal policy learning). Finally, by repeatedly assigning alternative treatments over time in an appropriate way, one may learn and converge to the assignment of the most effective treatment (reinforcement learning).

This course first discusses the underlying assumptions, intuition and usefulness of machine learning for forecasting and causal analysis. It then introduces various machine learning algorithms and discusses their application for prediction/forecasting and causal analysis. Using the statistical software "R" and its interface "R Studio", these methods are applied to various real-world data sets.

Objectives

- To understand the ideas, goals, and differences of machine learning for prediction and for causal analysis
- To understand the intuition, advantages, and disadvantages of alternative methods
- To be able to apply predictive and causal machine learning to real world data using the software "R" and its interface "R Studio"

Content

- Introduction to the concepts and purposes of predictive and causal machine learning
- Basics of predictive machine learning: Model tuning (cross-validation) and performance evaluation (out-of-sample testing)
- Prediction based on penalized regression (lasso and ridge regression)
- Prediction based on tree-based approaches (trees, bagging, random forests)
- Further predictive machine learners: boosting, support vector machines, neural networks (deep learning), and ensemble methods
- Causal analysis based on penalized regression (lasso and ridge regression)
- Causal analysis using tree-based approaches (causal trees and causal forests)
- Causal analysis based on double machine learning
- Assessing effect heterogeneity across subgroups
- Optimal policy learning to maximize treatment effectiveness using tree-based approaches
- Reinforcement learning to learn the most effective treatment (among several alternatives) by repeated treatment assignment over time
- Application of all methods to real world data using the statistical software "R" and its interface "R Studio"

Prerequisites

- Introductory statistics (probability theory, conditional means, linear regression), basic command of the statistical software "R".

Daily schedule:

Monday:

9:00-10:30 First lecture

10:30-11:00 Coffee break

11:00-12:30 Second lecture

12:30-14:00 Lunch

14:00-15:30 Third lecture

15:30-16:00 Coffee break

16:00-17:30 PC lab and problem sets

19:00 Dinner

Tuesday-Thursday:

9:00-10:30 First lecture

10:30-11:00 Coffee break

11:00-12:30 Second lecture

12:30-14:00 Lunch

14:00-15:30 PC lab and problem sets

15:30-16:00 Coffee break

16:00-17:30 Third lecture and/or discussion of PC lab and problem sets

19:00 Dinner

Friday:

9:00-10:30 First lecture

10:30-11:00 Coffee Break

11:00-12:30 Second lecture

12:30-14:00 Lunch

14:00-15:00 PC lab and problem sets

Please note that the course will start on Sunday, August 23, in the evening with a welcome meeting at 19:00 followed by dinner. The course will finish at 15:00 on Friday, August 28.

Approximate schedule in terms of topics covered:

Day 1

- Purpose of statistical modelling and machine learning for predictive analysis (in contrast to causal analysis).

- Basic concepts underlying all machine learning algorithms like performance measurement, the variance-bias tradeoff, sample splitting (into training, validation, and test data), and cross-validation.

- Machine learning algorithms based on penalized regression or shrinkage methods, which shrink the importance of weak predictors, namely lasso regression, ridge regression, and elastic nets.

Day 2

- Machine learning algorithms that based on decision trees, namely regression/classification trees, bagged trees, and random forests.

- Further machine learning algorithms: Boosting, support vector machines, neural networks, and ensemble methods.

Day 3

- Purpose of causal analysis, potential outcome notation, identifying assumptions for treatment evaluation (like the selection-on-observables assumption).

- Causal analysis based on lasso estimation (penalized regression): double selection or partialling out procedures to condition on important control variables, estimator's properties with and without sample splitting/cross-fitting.

Day 4

- Causal analysis based on random forests: effect interpretation in nonparametric models with binary and non-binary treatments, average treatment effect estimation, estimator's properties, and heterogeneity analysis.

- Causal analysis based on double machine learning: concept of doubly robust estimation (can be combined with a range of machine learning algorithms), average treatment effect estimation, estimator's properties, and heterogeneity analysis.

Day 5

- Optimal policy learning: Combining double machine learning and tree based-methods for (1) a data-driven segmentation into subgroups and (2) an optimal treatment assignment across subgroups for maximizing treatment effectiveness.

- Reinforcement learning: Repeated random assignment of alternative treatments over time to iteratively converge to the treatment assignment which is (on average) most effective.

Course Material:

Lecture slides, R code, and data files will be made available to the course participants.

Textbooks:

For predictive machine learning: G. James, D. Witten, T. Hastie, and R. Tibshirani (2021): An Introduction to Statistical Learning with Applications in R, Springer, New York. Freely available at: https://www.statlearning.com/

For causal machine learning: M. Huber (2023): Causal analysis - Impact evaluation and causal machine learning with applications in R, MIT Press, Cambridge. Free online version available at: https://mitpress.ublish.com/ebook/causal-analysis-impact-evaluation-and-causal-machine-learning-with-applications-in-r-preview/12759/162