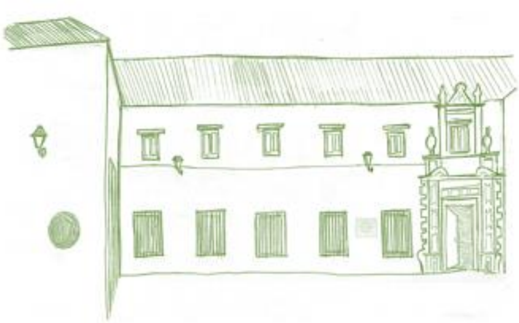
Machine Learning for Prediction and Causal Analysis



PhD in Economics, Business, Finance & Computer Science
University of Huelva
1-5 July 2024

1, 3 y 5 de julio 2024 UNIA-UHU
Sala virtual:
https://eu.bbcollab.com/guest/f83ed8d7cbc44341b0843a5a0410f5ab

Machine Learning for Prediction and Causal Analysis

Dates: 1, 3 and 5 July 2024
Time schedule: 14:00-17:00; 14:00-17:00; 14:00-18:00 (GTM+2, UCT+1)
Location: Online learning thorough UNIA's virtual campus. This comprehensive webinar is hosted through Blackboard and runs over a total of three days. Link: https://eu.bbcollab.com/guest/f83ed8d7cbc44341b0843a5a0410f5ab
Delivered by: Ass. Prof. Marica Valente, Dr. (University of Innsbruck)

Overview

How does statistics handle low- versus high-dimensional problems? The logical starting point is understanding linear regression, as many machine learning techniques are based on this simple method. The course then shows how linear regression can become flexible and be adjusted to explore high-dimensional associations and causal problems. In addition, I provide an introduction to highly flexible, nonparametric machine learning techniques such as tree- based methods. I present their applicability to both predictive and causal problems and draw contrasts with traditional regression approaches.

Estimation of average treatment effects (ATEs) accounting for a large number of variables such as individuals' socio-economic attributes plays an essential role in modern economics, medicine, and other disciplines by informing policy decision-making or physicians on the effects of their interventions. Several machine learning algorithms have been proposed recently to estimate ATEs in an effective and flexible way by re-purposing predictive machine learning models for causal estimation. In this course I summarize the literature on predictive algorithms and provide concrete guidance for their application for causal effect estimation in high dimensions.

Moreover, targeting policy or treatment interventions to specific subgroups of the population requires the understanding of heterogeneous causal effects. Heterogeneous causal effect analysis focuses on examining individualized treatment effects for individuals or subgroups in a population. Understanding heterogeneous treatment effects can critically guide, for instance, policymakers to identify socioeconomic groups of individuals for which the policy causes the largest effects, design effective policies, and tailor information campaigns for the least- responsive groups. In this course I present the recent advancements in the treatment effect and machine learning literature on the estimation of conditional average treatment effects (CATEs) from observational data with binary or continuous treatments.

The guidance I provide is supported by *comprehensive R tutorials* in which I will carefully explain codes piece by piece and provide tools for autonomous work. This course provides well-documented implementations of different ATE and CATE estimation strategies used in the literature as well as R codes to allow easy use of these methods as well as reproduction of the case studies analysed in class.

Contents

The course will cover the following topics:

- Draw differences between Statistics, Econometrics and Machine Learning
- Linear Regression, Assumptions, and Flexibility
- Machine Learning Methods for Prediction
- Non-parametric methods: CART and Random Forests
- Parametric methods: LASSO and other regression-based methods
- Machine Learning Methods for Causal Analysis
- Double/Debiased Machine Learning
- Causal Forests
- Applications in Social Sciences and Economics (e.g. Environmental Policy Evaluation, Crime Detection in the Shadow Economy, Drug Costs in the Health Sector)

The content is divided into two Modules: I and II.

Learning objectives

Module I: Predictive vs. Causal Problems

- Distinguish prediction problems from causal problems
- Describe and justify that correlation is not causation
- Describe why standard statistical methods, such as linear regression, fail in high dimension
- Perform data analysis in R and apply simple predictive machine learning algorithms
- Analyze algorithms' outputs and compare outputs across methods

Module II: Causal Problems with Standard and Machine Learning Methods

- Describe how causal models fit into machine learning
- Distinguish how to use machine learning for prediction vs. for causal effect estimation
- Describe the justification behind double machine learning methods
- Describe the justification behind causal forests
- Perform causal effect analysis with machine learning in R
- Analyze applications of causal problems in social science

Literature

Lecture notes (slides).

Inspirational readings (recommended):

- Varian, H. (2014): Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28(2), pp. 3-28 https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3
- Athey, S. (2018): The Impact of Machine Learning on Economics. The Economics of Artificial Intelligence: An Agenda. University of Chicago Press https://www.gsb.stanford.edu/faculty-research/publications/impact-machine-learning- economics
- Frey, S. & Savage, A. & Torgler, B. (2011). Behavior under Extreme Conditions: The Titanic Disaster. Journal of Economic Perspectives 25(1), pp. 209-22 https://www.aeaweb.org/articles?id=10.1257/jep.25.1.209

The full list of references will be provided in the Syllabus below. Selected references:

- (textbook) James, G., Witten, D., Hastie, T., and R. Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R. Springer.
- (paper) Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics. Journal of Economic Perspectives 28(2), pp. 29-50 https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29

- (paper) Wager, S., and Athey, S. (2018): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association 113(523), pp. 1228-1242 https://www.tandfonline.com/doi/abs/10.1080/01621459.2017.1319839?journalCode=uasa20

Methods

The course consists of lectures complemented by practical sessions.

The exercises in the course will be conducted in R. If you want to replicate the codes during or after the lecture, please make sure that R and RStudio are installed on your laptops. To download R, go to https://www.r-project.org/, for RStudio, go to https://www.rstudio.com/products/rstudio/download/.

Readings:

- [Introductory] Stauffer, R., Chimiak-Opoka, J. Rodríguez-R, L. M., Thorsten, S. Zeileis, A. Introduction to Programming with R. https://discdown.org/rprogramming/
- [Technical] Venables, W. N., Smith, D. M. and the R Core Team (2018): An Introduction to R. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Free online resources to learn R:

If you want to familiarize with the software R, you can use free online tools, e.g.

- https://www.datacamp.com/courses/free-introduction-to-r (sign up and start the free course on Introduction to R)
- https://swirlstats.com/

Prerequisites

No previous knowledge of machine learning is required since this is an introductory class. I expect that students have completed a graduate course of statistics or econometrics. The course requires basic knowledge of the linear OLS regression method. Prior experience with R is not a prerequisite, however, it is certainly advantageous.

Syllabus

MONDAY JULY 1st
2 pm – 5 pm

MODUL I – Predictive vs. Causal Problems

1. Statistics, econometrics and machine learning.

   - How does econometrics handle low- versus high-dimensional problems? Starting from the basics of linear regression (OLS), this part of the course will introduce students to high-dimensional predictive problems.
   - Operational definition(s), motivating empirical facts, the key concepts of ML

2. Draw contrasts with traditional approaches (OLS in classical statistics)

   - The curse of dimensionality for local average estimators and linear regression
   - High-dimensional data: Curse or blessing?

3. How to use machine learning methods for prediction?

   - Alternative algorithms to linear regression (OLS) that are better suited for prediction are now easily available: This part of the course introduces some of the machine learning algorithms that are most commonly adopted by economists.

4. Nonparametric methods. Tree-Based Methods

   - Classification and Regression Trees, Random Forests (R packages rpart, randomForest, etc.)

WEDNESDAY JULY 3rd
2 pm – 5 pm

5. Parametric methods. Variable Selection Techniques

   - Regression-Based Methods: Lasso.
   - Other methods, only scratch surface: Forward Selection, Ridge, Bridge, and Elastic Nets (R packages glmnet, caret, etc.)

6. R Session: Machine Learning Methods for Prediction

   - Prediction of survival in the Titanic Disaster
   - For social scientists, evidence about how people behaved as the Titanic sunk offers a

quasi-natural field experiment to explore behavior under extreme conditions of life and death

- Using our intuition and individual data on the Titanic Disaster will allow us to predict survival rates and answer some questions: Was it favourable for survival to travel alone or in company? Does one's role or function (being a crew member or a passenger) affect the probability of survival? Do social norms, such as "Women and children first!" have any effect? Does nationality affect the chance of survival?

<u>Readings:</u>

- Breiman, L. (1996) Heuristics of instability and stabilization in model selection. Ann. Statist., 24, 2350–2383.
- Hoerl, A. and Kennard, R. (1988) Ridge regression. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136. New York: Wiley.
- Flom, P. L. and Cassell, D. L. (2007): Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NESUG 2007.
- Varian, H. (2014): Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28(2), pp. 3-28.
- Giraud, C. (2014): Introduction to High-Dimensional Statistics, Monographs on Statistics & Applied Probability, Chapman & Hall CRC (mathematical foundations of high- dimensional statistics)
- Jones, Z., and Linder, F. (2015): Exploratory Data Analysis using Random Forests.
- Frey, S. & Savage, A. & Torgler, B. (2011). Behavior under Extreme Conditions: The Titanic Disaster. Journal of Economic Perspectives 25(1), pp. 209-22 https://www.aeaweb.org/articles?id=10.1257/jep.25.1.209
- Friedman, J., Hastie, T., and Tibshirani, R. (2008): The Elements of Statistical Learning (Downloadable on Tibshirani website)
- James, G., Witten, D., Hastie, T., and R. Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R. Springer.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288

FRIDAY JULY 5th
2 pm – 6 pm

MODUL II – Causal Models and Machine Learning Methods

7. Linear Regression Methods for Causal Analysis

- Estimation of ATEs using Linear Regression

8. Machine Learning Methods for Causal Analysis

- Estimation of ATEs: Double/debiased Machine Learning

9. R Session: Machine Learning Methods for Causal Analysis

- Lasso for double machine learning and causal effect estimation (package hdm)

10. Machine Learning Methods for Heterogeneous Causal Effects

- Intuition behind CATEs and Causal Random Forests

11. (optional) R Session: Machine Learning Methods for Heterogeneous Causal Effects

- Causal forests for CATE estimation (package grf)

12. Applications

Application fields:

- Applications from Social Sciences and Economics (broadly defined)
- From my research (see https://sites.google.com/view/maricavalente/research): this includes e.g. quantifying crime and irregular migration in the shadow economy, estimating environmental policy effects, and the effects of pharmaceutical payments to physicians in the health sector

Readings:

- Athey, S., Tibshirani, J., and Wager, S. (2018): Generalized Random Forests. Annals of Statistics 47(2), pp. 1148-1178.
- Belloni, A., Chernozhukov, V. (2013). Least squares after model selection in high- dimensional sparse models. Bernoulli 19(2), 521-547.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Treatment and Structural Effects in Economics. Journal of Economic Perspectives 28(2), pp. 29-50.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," Review of Economic Studies, 81, 608–650.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018), Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21: C1-C68.
- Chernozhukov, V., Hansen, C. and Spindler, M. (2016). High-dimensional Metrics in R. arXiv:1603.01700
- Wager, S., and Athey, S. (2018): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association 113(523), pp. 1228-1242.

Registration

Deadline: Registration closes 1 calendar day prior to the start of the course
Fee: Gratis
Registration: mecofin.uhu@gmail.com

Instructor

Marica Valente is Assistant Professor in the Department of Economics at the University of Innsbruck (Austria). Marica has MSc in Economics from the Toulouse School of Economics, and holds a PhD from The Berlin School of Economics. Before joining the University of Innsbruck, she was a Post-doctoral Research Fellow at the ETH Zurich. Marica is an Environmental economist with a passion for causal inference, big data, and machine learning. Her research also includes works in the fields of environment, health, labor, migration, conflicts and gender.
Email: marica.valente@uibk.ac.at
Website: https://sites.google.com/view/maricavalente

Academic Coordination

Concepción Román	Universidad de Huelva
Mónica Carmona	Universidad de Huelva
Contact	mecofin.uhu@gmail.com

Organized by

MSc and PhD in Economics, Finance and Computer Science
University of Huelva (Escuela de Doctorado)
Spanish Labour Economics Association (AEET)
International University of Andalusia