MA-M-320 - Quantitative Methods II, Machine Learning (90568 S, 6 ECTS)

Lecturer

Marica Valente, PhD (marica.valente@uibk.ac.at)

https://sites.google.com/view/maricavalente

Purpose and Content

This course provides a broad introduction to microeconometric empirical methods for economists, including traditional econometric methods and machine learning techniques. The target audience are master students interested in learning how to perform data analysis and solve prediction problems. Students will learn how to use the statistical software R. Completing the course will enable students to conduct independent empirical research in their master thesis as well as future jobs (e.g. public policy institutions, consulting firms, and doctoral programs).

Machine learning (ML) defines a set of modern empirical tools used in fields like statistics, computer science, AI and, more recently, economics. ML in economics is often viewed as a black-box: this course aims to make ML less obscure and more accessible. In this course, we will walk through the basics of ML with a focus on supervised learning such as regularized linear regression and tree-based methods. In addition, I will show R codes to familiarize with the algorithms' implementation. Existing statistical packages make it trivial to do ML in practice. However, I will show how economic intuition still plays a crucial role in improving the algorithms' performance. At the end of the course, students will know how to use ML methods to solve problems that traditional econometrics cannot.

Requirements

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

Format

The exercises in the course will be conducted in R. I recommend that participants familiarize themselves with the software using free online tools, e.g. https://www.datacamp.com/courses/free-introduction-tor (sign up and start the free course on Introduction to R) or https://swirlstats.com/ The first R session will also serve as an introductory session into R.

Readings: Venables, W. N., Smith, D. M. and the R Core Team (2018): An Introduction to R. https://cran.rproject.org/doc/manuals/r-release/R-intro.pdf

You will use your own laptops for exercises in R. 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/.

Syllabus

Three sessions of lectures (2022-2023):

- December 5 (Monday) and December 9 (Friday)
- December 15-16 (Thursday-Friday)
- January 12-13 (Thursday-Friday)

Time: 9-18h at the latest, ECTS: 6 Credit Points

SESSION I (10 SWS)

1. **Statistics, econometrics and machine learning.**
   How does econometrics handle low- versus high-dimensional problems?
   Starting from the basics of econometrics and 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?

4 SWS – 2 lectures (session I, Monday, December 5, 2022)

3. **How to use machine learning methods for prediction?**
   Alternative algorithms to 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.
- Nonparametric methods. Tree-Based Methods: Classification and Regression Trees, and Random Forests

4SWS – 2 lectures (session I, Friday, December 9, 2022)

4. **Q&A session / Office hours**

2SWS – 1 lecture (session I, Friday, December 9, 2022)

SESSION II (10 SWS)

Prior experience with the software R is not a prerequisite, however, it is certainly advantageous.

5. **How to use machine learning tools in R?**
   After revising the programming basics in R, we will learn which are the most important functions enabling empirical analysis in high-dimensions, how they work concretely, and how to interpret their output.
- **Analyze the basic concepts of ML in R**

6. **Tree-based methods in R**. Explain homework assignment

4SWS – 2 lectures (session II, Thursday, December 15, 2022)

7. **Parametric methods.** Regression-Based Methods: Lasso (Ridge, Bridge, and Elastic Nets)

8. **Analyze regression-based methods in R**

4SWS – 2 lectures (session II, Friday, December 16, 2022)

9. **How to conduct empirical research**. How to present research idea / description of available high-dimensional data sources

2SWS – 1 lecture (session II, Friday, December 16, 2022)

Readings:

- 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.
- 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

SESSION III (10 SWS)

10. **How to write an empirical paper?**
    How to show that what you do matters? This part of the course will focus on how to (i) motivate a paper from both the political scientists' and the economists' view; (ii) improve on the literature stressing how your work differs from past relevant studies; (iii) discuss the paper's identifying assumptions and provide evidence supporting them; (iv) lay out the empirical analysis and results while being upfront on the paper's limitations; (v) decide what to show, how to show it, and why to show it; (vi) summarize main contributions and results without being repetitive and stressing the bottom line; (vii) give qualitative and quantitative takeaways; (viii) conclude with lessons and implications both for policy and future research.

2SWS – 1 lecture (session III, Thursday, January 12, 2023)

11. Office hours / Q&A

2SWS – 1 lecture (session III, Thursday, January 12, 2023)

12. Student oral presentation of research ideas + Feedback

4 SWS – 1 lectures (session III, Friday, January 13, 2023)

13. Office hours / Q&A

2SWS – 1 lecture (session III, Friday, January 13, 2023)

Exam:

- **50% oral exam:** presentation of an *original* empirical research question using the methods learned in the course (in groups of more than two people).

  The empirical question should be policy relevant, and the motivation should be supported with both anecdotal evidence from the real world and from previous gaps in the literature. I prefer motivations coming from real world's problems and stylized facts, rather than from gaps in the literature. Use graphs when striking.

  Select a question that you *can* answer empirically using the data at hand. Empirical analyses should be based on a dataset provided by the instructor. Under certain circumstances, the data can be freely chosen from publicly available datasets associated to published papers. I also welcome original research ideas using high-dimensional dataset from *Kaggle*.

  Presentation style and slides matter for the grade.

- **50% term paper:** This written article should contain the motivation, economic and econometric mechanisms, data description, and full analysis of the empirical research question. R codes used for the analysis need to be attached and will be evaluated. The paper is due by the end of the semester.