Machine Learning Methods for Prediction and Treatment Effect Estimation

Course lecturer: Marica Valente

Overview of class purpose and content:
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, and this course aims to make such a box 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 regression and tree-based methods for both prediction and causal effect estimation. At the end of the course you will know how to use ML methods to solve problems that standard econometrics cannot. In addition, there will be two R sessions to familiarize with the algorithms' implementation. Existing statistical packages make it trivial to do ML in practice. However, we will show how economic intuition still plays a crucial role in improving the algorithms' performance.

Requirements:
No previous knowledge of ML is required since this is an introductory class. The course requires some basic knowledge of econometrics, and R coding. Please make sure to have RStudio installed before the beginning of the class.

In total 9 sessions over 3 days:

Day	Time	Title
WED. 10 June	3-4.30 p.m. (I)	**Introduction to High-Dimensional Predictive Problems**
THU. 11 June	9.30-11.00 a.m. (II)	**ML for Prediction and Causal Inference**
	Coffee break 30 min.	(II) Parametric: Regression-based methods
	11.30-12.30 a.m. (III)	(III) Nonparametric: Tree-based methods
	Lunch break 1h	(IV) R-session
	1.30-3.00 p.m. (IV)	(V) Causal inference and high-dimensional data
	Coffee break 30 min.	
	3.30-4.30 p.m. (V)	
FRI. 12 June	9.30-11.00 a.m. (VI)	**ML for Treatment Effect Estimation**
	Coffee break 30 min.	(VI) Causal and Generalized random forests
	11.30-12.30 a.m. (VII)	(VII) Double ML
	Lunch break 1h	(VIII) ML application for policy evaluation and continuous TE
	1.30-3.00 p.m. (VIII)	(IX) R-session
	Coffee break 30 min.	
	3.30-4.30 p.m. (IX)	

Syllabus

I) Introduction to High-Dimensional Predictive Problems

- Operational definition(s), motivating empirical facts, the key concepts of ML
- 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?

II-IV) Machine Learning Methods for Prediction

- Parametric models for high-dimensional data:
  Regression-Based Methods: LASSO, Ridge, Bridge, and Elastic Nets
- Nonparametric models for high-dimensional data:
  Tree-Based Methods: Classification and Regression Trees, and Random Forests
- R-session

Readings:

- Breiman, L. (1996) Heuristics of instability and stabilization in model selection. Ann. Statist., 24, 2350–2383.
- 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.
- Giraud, C. (2014): Introduction to High-Dimensional Statistics, Monographs on Statistics & Applied Probability, Chapman & Hall CRC (mathematical foundations of high-dimensional statistics)
- Hoerl, A. and Kennard, R. (1988) Ridge regression. In Encyclopedia of Statistical Sciences, vol. 8, pp. 129–136. New York: Wiley.
- James, G., Witten, D., Hastie, T., and R. Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R. Springer.
- Jones, Z., and Linder, F. (2015): Exploratory Data Analysis using Random Forests.
- Frank, I. E. and Friedman, J.H. (1993) A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109-148.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008): The Elements of Statistical Learning (Downloadable on Tibshirani website)
- Fu, W. (1998) Penalized regression: the bridge versus the lasso. J. Computnl Graph. Statist., 7, 397–416.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B, 58, 267–288
- Varian, H. (2014): Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28(2), pp. 3-28.

V-IX) Machine Learning Methods for Treatment Effect Estimation

- Causal inference and big data: Synthetic controls and diff-in-diff in low- and high-dimensions
- Analysis of Causal and Generalized Random Forests
- Double Machine Learning
- ML application for policy evaluation with continuous treatment
- R-session

Readings:

- Arkhangelsky, D., Athey, S., Hirshberg, D.A., Imbens, G.W. and Wager, S. (2019). Synthetic Difference in Differences. Working Papers wp2019_1907, CEMFI.
- 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
- Doudchenko, N., and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. National Bureau of Economic Research.
- Kinn, D. (2018): Synthetic controls and high-dimensional data. Working Paper arXiv:1803.00096
- 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.