Machine Learning for Social Sciences

Jorge Cimentada[*]

20 November, 2019

Course description

With the increasing amounts of data being collected on a daily basis, Machine Learning strategies have gained mainstream attention for their predictive power. Shifting away from focusing on inference, Machine Learning is a field at the intersection of statistics and computer science that is focused on maximizing predictive performance. That is, the goal of Machine Learning is to predict something – and predict it very well, regardless of whether you understand it. The field can be intimidating as it is vast and growing every year. On this course we will introduce students to the basic ideas of the Machine Learning framework. Moreover, we will introduce the basic algorithms used for prediction and discuss the potentials that Machine Learning can have in the social sciences.

We will introduce the main concepts in Machine Learning, such as regularized regressions, classification trees and clustering techniques, through basic examples and discuss their advantages and disadvantages. We will also pay great attention to discussing how it's been used in research and the potential it has for being applied in new fields. Although many social scientists do not see how predictive models can help explain social phenomena, we will also discuss how machine learning can play a role as a tool for discovery, improving causal inference and generalizing our classical models through cross validation.

We will end the course with a prediction challenge that will put to test all of your acquired knowledge. Start with a discussion on the role of predictive challenges such as the Fragile Families Challenge in social sciences, our predictive challenge will require the student to run machine learning algorithms, test it's out-of-sample error rate and publish their results. This will give the class a real hands-on example of how to incorporate Machine Learning into their research right away. Below is a detailed description of the syllabus.

Schedule

Session 1
July 6th 09:00h-10:45h

- Introduction to the Machine Learning Framework
  - Inference vs Prediction
  - Can inference and prediction complement each other?
  - "The Fragile Families Challenge"
  - Bias-variance / Interpretability-prediction tradeoffs
  - Resampling methods: validation, k-fold CV

Readings:

- Sections 2.1 and 2.2 from James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013
- Molina, M., & Garip, F. (2019). Machine Learning for Sociology. Annual Review of Sociology, 45.

[*]Laboratory of Digital and Computational Demography, Max Planck Institute of Demographic Research

- Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.
- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

Break 10:45h-11:15h

Session 2
July 6th 11:15h-13:00h

- Linear regression and regularization
  - Continuous predictions and loss functions
  - Lasso
    - Advantages/Disadvantages
    - R example
  - Ridge regression
    - Advantages/Disadvantages
    - R example
  - Elastic Net
    - Advantages/Disadvantages
    - R example
- Exercises

Readings:

- For a theoretical introduction to Lasso/Ridge, sections 6.1, 6.2 and 6.6 from *James, Gareth, et al. (2013) An Introduction To Statistical Learning. Vol. 112. New York: springer*
- For hands-on examples, Chapter 6 of *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*

Session 3
July 7th 09:00h-10:45h

- Supervised Regression
  - Introduction to supervised regression
  - Classification
    - Confusion matrices
    - ROC Curves
  - Classification Trees
    - Advantages/Disadvantages

        * R example

- Exercises

Readings:

- For an introduction to classification trees, Section 8.1, 8.3.1 and 8.3.2 from *James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013*
- For hands-on examples, chapter 9 from *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*
- For real-world applications of Classification Trees:
  - Billari, F. C., Fürnkranz, J., & Prskawetz, A. (2006). Timing, sequencing, and quantum of life course events: A machine learning approach. European Journal of Population/Revue Européenne de Démographie, 22(1), 37-65.
  - Chapter 3 of Nolan, D., & Lang, D. T. (2015). Data science in R: a case studies approach to computational reasoning and problem solving. CRC Press.

Break 10:45h-11:15h

Session 4
July 7th 11:15h-13:00h

- Supervised Regression
  - Bagging
    * Advantages/Disadvantages
    * R example
  - Random Forest
    * Advantages/Disadvantages
    * R example
  - Gradient Boosting
    * Advantages/Disadvantages
    * R example
- Exercises

Readings:

- For an introduction to bagging/random forests/boosting, Chapter 8 from *James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013*

- For hands-on examples, chapter 10, 11 and 12 from *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*
- For real-world applications of Random Forests:
  - Perry, C. (2013). Machine learning and conflict prediction: a use case. Stability: International Journal of Security and Development, 2(3), 56.
  - Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. Journal of Empirical Legal Studies, 13(1), 94-115.

Session 5
July 8th 09h-10:45h

- Unsupervised Regression
  - Introduction to unsupervised learning
  - Principal Component Analysis (PCA)
    - Advantages/Disadvantages
    - R example
  - K-Means clustering
    - Advantages/Disadvantages
    - R example
- Exercises

Readings:

- For an introduction to unsupervised learning, Section 10.1 from *James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013*
- For an introduction to PCA
  - Section 10.2 and 10.4 from *James, Gareth, et al. An Introduction To Statistical Learning. 112. New York: springer, 2013*
  - For hands-on examples, chapter 17 from *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*
- For an introduction to K-Means clustering
  - Section 10.5 from *James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013*
  - For hands-on examples, chapter 20 from *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*
- For real-world applications of K-means clustering:
  - Garip, F. (2012). Discovering diverse mechanisms of migration: The Mexico–US Stream 1970–2000. Population and Development Review, 38(3), 393-433.

  - Bail, C. A. (2008). The configuration of symbolic boundaries against immigrants in Europe. American Sociological Review, 73(1), 37-59.

Break 10:45h-11:15h

Session 6
July 8th 11:15h-13:00h

- Unsupervised Regression
  - Hierarchical clustering
    * Advantages/Disadvantages
    * R example

- Final challenge: Prediction competition
  - Explanation of strategies
  - No free lunch theorem
  - Presentation of results

Readings:

- For an introduction to hierarchical clustering, sections 10.3.2, 10.3.3, 10.5.2 from *James, Gareth, et al. An Introduction To Statistical Learning. Vol. 112. New York: springer, 2013*
- For hands-on examples, chapter 21 from *Boehmke & Greenwell (2019) Hands-On Machine Learning with R, 1st Edition, Chapman & Hall/CRC The R Series. Accessible at: https://bradleyboehmke.github.io/HOML/*
- For examples on prediction competitions:
  - Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. (2016). Crowdsourcing city government: Using tournaments to improve inspection accuracy. American Economic Review, 106(5), 114-18.
  - Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Introduction to the Special Collection on the Fragile Families Challenge. Socius, 5, 2378023119871580. Accessible at https://www.researchgate.net/publication/335733962_Introduction_to_the_Special_Collection_on_the_Fragile_Families_Challenge

Software:

We will be using the R software together with the Rstudio interface. No laptop is required as the seminars will take place in the RECSM facilities. Any packages we plan to use will be already downloaded previous to the session.

Prerequisites:

- The course assumes that the student is familiar with R and should be familiar with reading, manipulating and cleaning data frames. Ideally, the student has conducted some type of research using the software.
- Students should have solid knowledge of basic statistics such as linear and logistic regression, ideally with more advanced concepts such as multilevel modelling.

Instructor:

Jorge Cimentada has a PhD in Sociology from Pompeu Fabra University and is currently a Research Scientist at the Laboratory of Digital and Computational Demography at the Max Planck Institute for Demographic Research. His research is mainly focused on the study of educational inequality, inequality in spatial mobility and computational social science. He has worked on data science projects both in the private sector and in academic research and is interested in merging cutting edge machine learning techniques with classical social statistics. You can check out his blog at cimentadaj.github.io or contact him through twitter at @cimentadaj.