

SYLLABUS

Halle Institute for Economic Research (IWH)

Machine Learning for Microeconometrics

by

Prof. A. Colin Cameron, Ph.D.

(University of California at Davis)

February 17-19, 2020

The course is taught in 24 sessions of 45 Minutes starting

Monday February 17: Sessions 1-8

Tuesday February 18: Sessions 9-16

Wednesday February 19: Sessions 17-24.

course outline

Monday 1-4: Machine learning: Overview, terminology, selection of regressors using goodness-of-fit or cross-validation.

Monday 5-8. Regression: shrinkage methods (ridge, lasso, elastic net), Dimension reduction (principal components)

Tuesday 9-12: Regression: Nonlinear models: local regression, neural networks, regression trees, bagging, random forests and boosting.

Tuesday 13-16: Nonparametric density estimation, nonparametric and semiparametric regression, bootstrap.

Wednesday 17-20: Causal inference with machine learning: IV estimation with many instruments; partial linear model with many controls; ATE with heterogeneous effects and many controls.

Wednesday 21-24: Classification (categorical y): logit, k-nn, LDA, SVM; Unsupervised learning (no y): PCA, cluster analysis.

The material will cover applications using Stata Version 16. R will not be used.

Material posted at Course Website

All slides will be posted.

All programs and datasets generating the slides will be posted.

Most papers should be accessible e.g. through JSTOR.

I strongly suggest getting either a pdf or hardcopy of James et al. *An Introduction to Statistical Learning: with Applications in R* – see below.

The course uses Stata version 16. But most of the basic methods of machine learning are well explained in *An Introduction to Statistical Learning: with Applications in R*, and there is much more machine learning code in R than in Stata.

Slides (posted)

| | |
|------------------|--|
| Machlearn2020_1A | Basics – selection and cross validation |
| Machlearn2020_1B | Shrinkage estimators |
| Machlearn2020_2A | Other Estimators |
| Machlearn2020_2B | Nonparametrics and semiparametrics |
| Machlearn2020_2C | Bootstrap |
| Machlearn2020_3A | Causal inference |
| Machlearn2020_3B | Classification and unsupervised learning |

Programs and Data (posted)

| | |
|----------------------------|-------------------------------------|
| mus228prediction_halle.do | Stata program |
| mus228prediction_halle.txt | Output file from Stata program |
| mus203mepsmedexp.dta | Stata data set |
| mus228ajr.dta | Stata data set |
| mus228generateddata.dta | Stata dataset |
| mus228generateddata.csv | Comma separated file with same data |
| bootstrap2020.do | Stata program |
| bootstrap2020.txt | Output file from Stata program |
| bootdata.dta | Stata dataset |
| nonparametric2020.do | Stata program |
| nonparametrics.txt | Output file from Stata program |
| nonparametric.dta | Stata dataset |

Key readings (only the first posted)

Chapter 28 "Machine Learning for prediction and inference" in A. Colin Cameron and Pravin K. Trivedi, *Microeconometrics using Stata*, Second edition, forthcoming.
Posted as Cameron_Trivedi_MUS2_chapter_28.pdf

ISL: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibsharani (2013), *An Introduction to Statistical Learning: with Applications in R*, Springer.
A free legal pdf is at <http://www-bcf.usc.edu/~gareth/ISL/> and if you have access a cheap hardcopy can be obtained via <http://www.springer.com/gp/products/books/mycopy>

ESL: Trevor Hastie, Robert Tibsharani and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.
A free legal pdf is at <http://statweb.stanford.edu/~tibs/ElemStatLearn/index.html> and a \$25 hardcopy can be obtained via <http://www.springer.com/gp/products/books/mycopy>

Next most important readings (not posted)

Alex Belloni, Victor Chernozhukov and Christian Hansen (2014), "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, Spring, 29-50.

Sendhil Mullainathan and J. Spiess: "Machine Learning: An Applied Econometric Approach", *Journal of Economic Perspectives*, Spring 2017, 87-106.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins (2018), "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1-C68.

Other suggested readings (not posted)

Bradley Efron and Trevor Hastie (2016), *Computer Age Statistical Inference: Algorithms, Evidence and Data Science*, Cambridge University Press.

Achim Ahrens, Christian Hansen, Mark Schaffer (2019), "lassopack: Model selection and prediction with regularized regression in Stata," arXiv:1901.05397

Susan Athey (2018), "The Impact of Machine Learning on Economics".
<http://www.nber.org/chapters/c14009.pdf>

Susan Athey and Guido Imbens (2019), "Machine Learning Methods Economists Should Know About."

Alex Belloni, Victor Chernozhukov and Christian Hansen (2011), "Inference Methods for High-Dimensional Sparse Econometric Models," *Advances in Economics and Econometrics*, ES World Congress 2010, ArXiv 2011.

Alex Belloni, D. Chen, Victor Chernozhukov and Christian Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain", *Econometrica*, Vol. 80, 2369-2429.

Alex Belloni, Victor Chernozhukov, Ivan Fernandez-Val and Christian Hansen (2017), "Program Evaluation and Causal Inference with High-Dimensional Data," *Econometrica*, 233-299.

Max Farrell (2015), "Robust Estimation of Average Treatment Effect with Possibly more Covariates than Observations", *Journal of Econometrics*, 189, 1-23.

Max Farrell, Tengyuan Liang and Sanjog Misra (2018), "Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands," arXiv:1809.09953v2.

Jon Kleinberg, H. Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan (2018), "Human decisions and Machine Predictions", *Quarterly Journal of Economics*, 237-293.

Hal Varian (2014), "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives*, Spring, 3-28.

Stefan Wager and Susan Athey (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *JASA*, 1228-1242.