Economics 2150: The Econometrics of Machine Learning (and other 'Big Data' Techniques)

January 2018

1 Course Info

Spring 2018, Mondays 1-4, Emerson 108

Sendhil Mullainathan
mullain@fas.harvard.edu
Office Hours: Mondays 4-6pm.

TF: Jonathan Roth
jonathanroth@g.harvard.edu
Office Hours: W 12:30-1:30, Littauer Basement

2 Course Description

Innovations in machine learning ('big data') have created many engineering breakthroughs from real time voice recognition to automatic categorization (and in some cases production) of news stories. Since these techniques are at their essence novel ways to work with data, they should also have implications for social science.

This course explores the intersection of machine learning and social science and aims to answer a few questions about these new techniques:

1. How does machine learning work? There are textbooks to teach you how to implement machine learning. In fact, existing statistical packages make it trivial to do this in practice. But what makes them work? What statistical guarantees do they provide? In a way, machine learning is too easy to implement. By gaining an understanding of the mathematical basis and econometric underpinnings, it can be used more accurately.

2. What can machine learning tools do that our current toolbox cannot? Or put more positively, where does it fit in the toolbox? This class will give

a sense of how it relates to the other existing tools, from causal inference to basic regression.

3. Where can machine learning be used to generate new research output? As a PhD class, one of the primary goals is to bring you to the research frontier and to give you a sense of profitable directions.

As a result, while we will cover standard machine learning techniques such as supervised and unsupervised learning, statistical learning theory and nonparametric and Bayesian approaches, this will *not* be a detailed tutorial on those techniques. Instead the goal is to create a conceptual understanding of when and how they can be profitably applied.

The key objective of this class is to help you write a chapter or chapter(s) in your thesis that uses these tools; and hopefully they profit your research career going forward. The class is structured with this specific objective in mind. Students will be required to apply some of these techniques themselves, but we will not cover the computational aspects of the underlying methods. The course is aimed at PhD students with a solid background in statistical techniques, such as comes from the equivalent of a first year economics PhD econometrics sequence.

3 Relation to Other Machine Learning Courses

There are some excellent more standard machine learning classes at Harvard (CS 181, CS 281). Those classes provide a computational introduction to machine learning. The goal of this course instead is to integrate machine learning into a standard econometrics framework.

We have worked to minimize overlap between this course and other, more standard ML courses. Some of the material that will be covered may appear to overlap (e.g. regularization, models such as decision trees and random forests) but the conceptual focus even on this material is quite distinct: put this material into an applied econometrics framework. Given this focus, it will not cover:

- The computational aspects of the underlying methods. There are some important innovations that have made these techniques computationally feasible. We will not discuss these, as there are computer science courses better equipped to cover them.

- The nitty-gritty of how to use these tools. The mechanics of implementation, whether it be programming languages or learning to use APIs, will not be covered. Students will be expected to learn this material on their own.

This is not a good course for people simply looking to learn the mechanics of using machine learning tools. You will learn this implicitly, on your own, through the project work in the class.

4 Prerequisites and Target Audience

The course is aimed at PhD students interested looking to deepen and expand their research toolset. Students should have taken the first year economics PhD econometrics sequence (Econ 2120 and ideally 2140) or the equivalent. For economics PhD students, this class is approved as a distribution course and as an econometrics field course.

5 Assignments and Grades

There will be one empirical exercise during the first 3 weeks of the course ($\frac{1}{3}$ of grade) and two problem sets ($\frac{1}{3}$ of grade). The problem sets will involve a combination of working with real data, solving theoretical exercises, and answering conceptual questions. There will not be a final exam. A final project proposal ($\frac{1}{3}$ of grade, see below) will constitute the remaining $\frac{1}{3}$ of the grade. The proposal is meant to be a fully fleshed out description of a research project – it is meant to be a proposal for what the project is, why it is interesting and how it would be conducted (examples will be handed out in class). It is due in lieu of exam. A preliminary draft is due on April 2nd - this is mandatory but will not be graded but feedback will be given. Based on these proposals, some students will be asked to make presentations on April 23rd.

6 Section

There will be a weekly section (W 1:30-2:30, Littauer M-15).

7 Readings

Since the material for this class is largely new, there is no clear book to follow. Jann Spiess and I also putting together a manuscript for a book. We will look to hand out chapters from the manuscript as available. Below, we provide a listing of relevant papers for each class; we highly recommend these for those interested in understanding the current research frontier, although they are not explicitly required for the course.

We will also recommend chapters from these three books:

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Springer, Berlin: Springer series in statistics, 2001. Available here `http://statweb.stanford.edu/~tibs/ElemStatLearn/` as a pdf download. Many courses look to this book and you may also find it helpful. In learning this material, we personally found it a very useful introduction but also one that raised more questions than it answered at times. There are a few caveats as you use this book: (i) as it is aimed towards a new reader, it is hard for those who know econometrics to understand the conceptual distinctions between models you already understand (say splines) and the new models (say LASSO); (ii) as it is not aimed at economists, there is no integration with econometrics or economic applications; and (iii) as an introductory book, it is hard to tell what is a mathematically proven claim and what is a heuristic.

- Murphy, Kevin P. *Machine learning: a probabilistic perspective.* MIT press, 2012. This is a very useful reference book as it is relatively comprehensive. It is a great place to go to read about a large number of procedures to see them presented in a probabilistic framework.

- Bishop, Christopher M. *Pattern recognition and machine learning.* Springer, 2006. This is another standard books used in computer science classes. It is again a useful place to read about various methods, including many we will not cover due to lack of time.

8 Schedule

January 22 - An Applied Perspective

An Applied Perspective on Machine Learning

How does ML differ from existing empirical tools? Where does it fit in the toolbox? What are its strengths and weaknesses? Any meaningful application of machine learning requires crisp answers to these questions. They dictate which topics will be profitable and which will not. They dictate the complexity and nuances that must be resolved in any research project.

By way of answer, this lecture first provides a basic organizing framework of "coefficients" versus "predictions". We illustrate how existing estimators provide good coefficients $\hat{\beta}$, whereas machine learning provides good predictions $\hat{y}$. We have found this distinction eye-opening to people (and us), including the exposition of why good coefficient estimators are not automatically good predictors.

Second it explains what makes prediction in high dimensions possible: *regularization and adaptive tuning.* We walk through in depth how these work and why they are, in their own way, amazing. We aim to develop a strong intuition for these two ingredients because that in turn will help you see why machine learning makes such powerful predictions, but also why as a coefficient estimator it can be somewhat suspect.

Readings:

- Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87–106

- Athey and Imbens - NBER 2015 SI methods lectures, Lectures 1 and 2 (`http://www.nber.org/econometrics_minicourse_2015/`)

Assignments: Long Empirical Project handed out: American Housing Survey from Census.

Exercise due next week (Jan 29): Load data and form table of summary statistics

Exercise due Feb 8: Prediction "challenge". Come up with best predictor of house price.

Exercise due Feb 11: Prediction challenge follow-up

January 29: Mechanics of Application

A Cookbook

Having laid out at an abstract level the essential ingredients, we become very concrete in this lecture. We go through each of the steps needed to implement it as a machine learning application: predict wages using a large set of variables. The goal is that with the concepts in mind, it is useful to see, mechanistically, all the steps.

A Menu of Models

We describe here the diverse function classes (typically thought of as different algorithms) that machine learning can draw on. We describe how these procedures control complexity, and how their performance relates to features of the data. In particular, we will cover penalized linear regression (lasso and ridge), decision and regression trees, random forests, kernel regression. We will also provide an introduction to "ensemble techniques" as well as boosting: these methods all involve combining several predictors to improve prediction quality. We will address how they work as well as why they might work.

February 5: Nitty Gritties

Statistical Guarantees; Prediction and Parameter Estimation

How can we assess a predictor's performance? How can we put standard errors on this assessment? Starting with the observation that prediction performance is observable given a new sample, we discuss how we can obtain statistical guarantees on out-of-sample prediction quality.

Next we will describe which guarantees we have – and importantly do not have – about the prediction function. This lecture connects the predictions $\hat{y}$ given by machine learning algorithms back to the estimates $\hat{\beta}$ we are used to looking at. Under certain conditions the coefficients are interpretable; but we also highlight the numerous times when they are not, and lay out the dangers and pitfalls in interpreting the representation of prediction functions themselves.

As an illustration of the limitations, we discuss biases arising from regularization. As an example for the positive results that are available in this literature, we run through the standard estimation consistency results for the lasso (Zhao and Yu, 2006; Belloni et al., 2011a), pointing out the strength of their conclusions as well as their assumptions.

Nitty Gritties

Finally we will describe nitty gritties:

Standard errors

- How would we compute a standard error on prediction quality?
- What happens if I form them using cross-validation?
- What hypotheses can be tested?
- How do I handle clustering and stratification?

Understanding the predictor

- AUC, and different measures of loss
- What do calibration curves teach us?
- What can I say about variable importance?
- Downstream uses, e.g. data or variable creation
- If I use the predictor in some other estimator how should I handle that?

Feature engineering

- Will the algorithm "discover" features?
- Can I just throw it all in?
- How to bake in prior information while still maintaining the algorithm's flexibility?

Sample size

- K-fold: what is k?
- How big is test and train?
- Regularization and sample size?

Readings:

1. Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563
2. Belloni, A., Chernozhukov, V., and Hansen, C. (2011a). Inference for High-Dimensional Sparse Econometric Models. *arXiv:1201.0220 [econ, stat]*
3. Wager, S. (2014). Asymptotic Theory for Random Forests. *arXiv:1405.0352 [math, stat]*
4. Friedman et al, Ch. 3.8.5; Murphy 13.3.5
5. Friedman et al, Chs. 7.10-7.12

February 12th: Debrief on Prediction Challenge, and Prediction Policy Problems

Debrief on Prediction Challenge

This lecture will be partly discussion of the exercise–answering questions that have arisen.

Prediction Policy Problems

Here we describe a set of economic problems where the predictions $\hat{y}$ is directly the policy issue of interest. We show how a focus on prediction – rather than estimation – together with the availability of strong prediction tools leads us to asking new research questions, with practical applications in policy (Kleinberg et al., 2015). We describe in particular three issues that arise in applying ML to these problems:

- Selective labels
- Omitted payoff biases
- The pernicious issue of measurement error

Readings:

- Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *The American economic review*, 105(5):491–495
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794
- Mullainathan, S. and Obermeyer, Z. (2017). Does Machine Learning Automate Moral Hazard and Error? *American Economic Review*, 107(5):476–480
- Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., and Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, 359(6373):325–329

Assignments: Exercise for Feb 26th: Describe in detail an innovative prediction policy problem that could feasibly be done.

February 19 - Holiday

February 26th - Frontiers in Prediction Policy and Economically Meaningful Uses of Prediction

Here we will discuss some of the frontier issues in using prediction:

- Fairness
- Interpretability
- People's use of algorithms
- Using prediction to test theories

We are often used to thinking about causal inference as a way to test theories. Here we show how prediction provides a different – and complementary – mode to test theories. We lay out how relative prediction quality using restricted sets of regressors can help us to test a given theory while controlling for other, potentially unknown theories. Using the disposition effect (Odean, 1998) as an example, we show how the prediction approach to theory testing complements the applied researcher's toolkit.
Readings:

- Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv:1701.08230 [cs, stat]*
- Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, 144(1):114–126
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*
- Einav, L., Finkelstein, A., Mullainathan, S., and Obermeyer, Z. (2017). Does high healthcare spending at end of life imply waste? Predictive modeling suggests not necessarily

- Kleinberg, J., Liang, A., and Mullainathan, S. (2017). The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness. *arXiv:1706.06974 [cs, stat]*
- Fudenberg, D. and Liang, A. (2018). Predicting and Understanding Initial Play. SSRN Scholarly Paper ID 3076682, Social Science Research Network, Rochester, NY
- Peysakhovich, A. and Naecker, J. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior & Organization*, 133:373–384

March 5 - Causal inference

We illustrate how a variety of traditional causal inference techniques – from experiments to propensity scores and controlling – can benefit from the ML tools we have laid out here.

We first show how machine learning techniques can enhance typical tasks in experimental data analysis, specifically balance/randomization checks and testing for effects on the distribution or whole groups of outcome variables.

We then review the emerging literature on estimating heterogeneous treatment effects in high-dimensional data that adapts, and significantly extends on, techniques from machine learning.

Absent an experiment - We show how controlling for observable covariates can be seen as a prediction task, and which challenges arise in its implementation; specifically, we use this example to highlight how a naive application of machine learning tools may fail, while a careful adaption yields valid inference with provable guarantees.

As a further example, we show how instrumental variable estimation can be interpreted as a prediction problem, which allows for the use of prediction techniques from machine learning in the first stage.

Readings:

- Athey, S. and Imbens, G. (2015). Recursive Partitioning for Heterogeneous Causal Effects. *arXiv:1504.01132 [econ, stat]*
- Wager, S. and Athey, S. (2015). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv:1510.04342 [math, stat]*
- Imai, K., Ratkovic, M., and others (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470
- Grimmer, J., Messing, S., and Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/Debiased Machine Learning for Treatment and Causal Parameters. *arXiv:1608.00060 [econ, stat]*
- Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain. *Econometrica*, 80(6):2369–2429
- Belloni, A., Chernozhukov, V., and Hansen, C. (2011b). Inference on Treatment Effects After Selection Amongst High-Dimensional Controls. *arXiv:1201.0224 [econ, stat]*
- Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013). Program Evaluation and Causal Inference with High-Dimensional Data. *arXiv:1311.2645 [econ, math, stat]*
- Spiess, J. (2017a). Bias Reduction in Instrumental Variable Estimation through First-Stage Shrinkage. *arXiv:1708.06443 [econ, math, stat]*
- Ludwig, J., Mullainathan, S., and Spiess, J. (2017). Machine Learning Tests for Effects on Multiple Outcomes. *arXiv:1707.01473 [econ, stat]*
- Spiess, J. (2017b). Optimal Estimation when Researcher and Social Preferences are Misaligned. *Job Market Paper*
- Dubé, J.-P. and Misra, S. (2017). Scalable Price Targeting. Working Paper 23775, National Bureau of Economic Research
- Athey and Imbens - NBER 2015 SI methods lectures, Lectures 3 and 4 (`http://www.nber.org/econometrics_minicourse_2015/`)

March 12 - No class, Spring Break

March 19 - Unsupervised learning

Here we cover briefly the following unsupervised techniques:

- Clustering
- Latent factor and hidden markov models
- LDA and related models.

In particular, we discuss clustering when cluster identities are the object of interest itself (Bonhomme and Manresa, 2015) or as a preprocessing step for further economic analysis (Bonhomme et al., 2015).

Readings:

- Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184
- Bonhomme, S., Lamadon, T., and Manresa, E. (2015). Approximate Clustering
- Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic Research Evolves: Fields and Styles. *American Economic Review*, 107(5):293–297

March 26 - Language

Machine learning tools now allow us to take qualitative data – at large scale – and treat them as quantitative. We review standard methods in natural language processing, including bag-of-words techniques and topic modelling.

Readings:

- Enke, B. (2017). Moral Values and Voting: Trump and Beyond. SSRN Scholarly Paper ID 2979591, Social Science Research Network, Rochester, NY
- Wu, A. (2017). Gender Stereotyping in Academia: Evidence from Economics Job Market Rumors Forum. SSRN Scholarly Paper ID 3051462, Social Science Research Network, Rochester, NY
- Gentzkow, M., Shapiro, J. M., and Taddy, M. (2016). Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech. Working Paper 22423, National Bureau of Economic Research
- Fu, L., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Tie-breaker: Using language models to quantify gender bias in sports journalism. *arXiv:1607.03895 [physics]*

April 2 - Computer Vision, Neural Nets and Deep Learning

Readings:

- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113
- Igami, M. (2017). Artificial Intelligence as Structural Estimation: Economic Interpretations of Deep Blue, Bonanza, and AlphaGo. *arXiv:1710.10967 [cs, econ]*
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. In *PMLR*, pages 1414–1423

- Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *arXiv:1706.05806 [cs, stat]*

Assignments: Draft of project proposal due

April 9 - No class

April 16 - New Data and Economic Effects

New Data

We describe here how search, Twitter and Facebook data have all been used profitably in social science. For example, Facebook data has been used to study the spread of information in social networks (Friggeri et al., 2014), and restaurant review data can be used to improve the targeting of hygiene inspections (Kang et al., 2013). Importantly visual data is proving useful in satellite analysis. Here we describe some nitty-gritty uses of prediction in the construction and cleaning of datasets. We point to a growing set of applications that use machine learning as a prepossessing tool, from combining experimental to survey data (Bernheim et al., 2015) to census record linking in describing intergenerational mobility (Feigenbaum, 2015a,b).

Economic and Social Consequences

What are the economic and social consequences of these tools? Once firms can screen credit risk or health risks differently, how will credit and insurance markets be affected? How are market equilibria affectded by who “owns” the data? Should regulations be enacted?
Readings (new data):

- Friggeri, A., Adamic, L. A., Eckles, D., and Cheng, J. (2014). Rumor cascades. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*
- Kang, J. S., Kuznetsova, P., Choi, Y., Luca, M., and others (2013). Using Text Analysis to Target Government Inspections: Evidence from Restaurant Hygiene Inspections and Online Reviews. Technical report.
- Bernheim, B. D., Bjorkegren, D., Naecker, J., and Rangel, A. (2015). Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions
- Feigenbaum, J. J. (2015a). Automated Census Record Linking: A Machine Learning Approach
- Feigenbaum, J. J. (2015b). Intergenerational Mobility during the Great Depression.

- Donaldson, D. and Storeygard, A. (2016). The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives*, 30(4):171–198
- Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794
- Stephens-Davidowitz, S. (2013a). The Cost of Racial Animus on a Black Presidential Candidate: Using Google Search Data to Find What Surveys Miss. SSRN Scholarly Paper ID 2238851, Social Science Research Network, Rochester, NY
- Stephens-Davidowitz, S., Varian, H., and Smith, M. D. (2017). Super returns to Super Bowl ads? *Quantitative Marketing and Economics*, 15(1):1–28
- Stephens-Davidowitz, S. (2013b). Unreported Victims of an Economic Downturn
- Glaeser, E. L., Kominers, S. D., Luca, M., and Naik, N. (2015). Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. Working Paper 21778, National Bureau of Economic Research
- Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113
- Chen, M. K. and Rohla, R. (2017). Politics Gets Personal: Effects of Political Partisanship and Advertising on Family Ties. *arXiv:1711.10602 [econ]*

Readings (economics and social consequences):

- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2017). Predictably Unequal? The Effects of Machine Learning on Credit Markets. SSRN Scholarly Paper ID 3072038, Social Science Research Network, Rochester, NY
- Liberman, A., Neilson, C., Opazo, L., and Zimmerman, S. (2017). The Equilibrium Effects of Asymmetric Information: Evidence from Consumer Credit Markets
- Dubé, J.-P. and Misra, S. (2017). Scalable Price Targeting. Working Paper 23775, National Bureau of Economic Research

- Reimers, I. and Shiller, B. R. (2018). Proprietary Data, Competition, and Consumer Effort: An Application to Telematics in Auto Insurance. Technical Report 119, Brandeis University, Department of Economics and International Businesss School

- Shiller, B. (2013). First Degree Price Discrimination Using Big Data. SSRN Scholarly Paper ID 2314997, Social Science Research Network, Rochester, NY

- Arrieta Ibarra, I., Goff, L., Jiménez Hernández, D., Lanier, J., and Weyl, E. G. (2017). Should We Treat Data as Labor? Moving Beyond 'Free'. SSRN Scholarly Paper ID 3093683, Social Science Research Network, Rochester, NY

- Lanier, J. (2013). *Who Owns the Future?* Simon & Schuster

April 23 - Student Presentations