Machine Learning: An Applied Econometric Approach
BSE Short Course, September 2–4, 2019

Jann Spiess*
Assistant Professor of Operations, Information & Technology
Stanford Graduate School of Business
jspiess@stanford.edu

Preliminary Syllabus
July 12, 2019

1 Overview of class purpose and content

Machine learning has created many engineering breakthroughs from real-time voice recognition to automatic categorization (and in some cases production) of news stories. What is particularly tantalizing though is that machine learning is, at its heart, an empirical tool. It takes as input large data sets and produces outputs such as a functions that relate one variable to others. The language is different: estimators are called algorithms; outcome variables are called labels; and so on. But at their core they are econometric tools: they find empirical relationships in data. Given the similarity to tools we know, it is tempting to ask whether it is merely old (econometric) wine in a new (machine learning) bottle.

In this course, we will argue that it is not. Far from it, we will discuss how these tools can powerfully improve and expand on the kind of empirical work we tend to do. At the same time, we will discuss their limitations and how they fit into the "econometric toolbox". At a high level, this class will address these three questions:

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.

*Based on work with Sendhil Mullainathan.

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, specifically causal inference and basic regression.

3. Where can machine learning be used to generate new research output? New computational tricks, statistical advances, and novel data sources allow us to improve answers to old questions as well as ask new questions.

We will cover standard machine learning techniques with a focus on supervised learning (such as regularized regression and methods based on decision trees). Towards the end of the class, we will also briefly discuss some unsupervised learning techniques (e.g. clustering).

Relative to the BeNA "Applications in Empirical Microeconomics" course, this class will focus more on econometric underpinnings and spend less time on reviewing applications in the empirical literature. The two classes overlap significantly and are not designed to be taken together.

1.1 Target audience and prerequisites

The course is aimed at graduate students looking to deepen and expand their research toolset, both those interested in empirical research using machine learning and those interested in developing methods and econometric theory themselves. Students should have some basic graduate training in econometrics.

There will be examples and small exercises in the statistical programming language `R`. While we will not be able to teach `R`, knowledge of the specific language is not essential, but some familiarity with statistical programming (such as in Stata, Matlab, or Python) is helpful.

1.2 Limitations

Given time limitations and the availability of numerous resources on machine learning, we 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 technical mechanics of implementation, whether it be programming languages or learning to use APIs, will not be covered in detail. We will instead focus on the conceptual aspects of applying available implementations in economics.

Given the time constraints of the course, even for many other topics that are covered in the class we will only be able to give a high-level conceptual understanding as well as pointers to more detailed material.

1.3 General references

There is no required reading or unified textbook for the course. Some references will be given for specific topics. The class is largely following the framing and structure from Mullainathan and Spiess (2017). Helpful textbooks with background on machine learning are:

- Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer. Available online as a pdf download.
- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

2 Preliminary structure of the class

1. **The rise of machine learning.** Where do recent breakthroughs in machine intelligence come from? How is machine learning (ML) different from classical artificial intelligence? What is the relationship to statistics?
   - Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.
   - Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An agenda*. University of Chicago Press.
2. **The secret sauce of ML.** What allows machine learning to predict well in very high-dimensional data? What are common features of supervised machine-learning algorithms? How are they implemented and how can we choose the right algorithms and parameters for the prediction task at hand?
   - James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.
   - Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.
3. **Prediction ($\hat{y}$) vs estimation ($\hat{\beta}$).** How does ML relate to standard regression analysis in econometrics? Which guarantees do we need and obtain? What are the limits of interpreting parameters coming out of ML?
   - Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

– Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

4. **Applications of ML in empirical work.** Given that machine learning provides high-quality predictions but we often care about (causal) estimation in applied econometric work, how can we adapt techniques and insights from machine learning in a program-evaluation context?

   (a) Prediction policy problems: where prediction solutions directly solve the problem
   - Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.
   - Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017a). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

   (b) Prediction in the service of estimation: where prediction techniques can enhance causal inference
   - Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
   - Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.
   - 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(1):C1–C68.
   - 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.

   (c) Predictability as the question of interest: where properties of a prediction solution provide tests for theories
   - Kleinberg, J., Liang, A., and Mullainathan, S. (2017b). The theory is predictive, but is it complete? An application to human perception of randomness. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 125–126. ACM.
   - Gagnon-Bartsch, J. and Shem-Tov, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance. *Annals of Applied Statistics*.

   (d) New data: where prediction solutions allow us to use new data sources to construct variables for further analysis
   - Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–98.

- Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

5. **Beyond supervised learning.** Which techniques from machine learning can we use beyond prediction algorithms?

   (a) Unsupervised learning: What ML techniques are available that find structure in data without a designated outcome variable? How can they be used in empirical applications?

   - Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

   (b) Reinforcement learning (RL): What is RL and how does it relate to techniques in economics?

   - Igami, M. (2017). Artificial intelligence as structural estimation: Economic interpretations of Deep Blue, Bonanza, and AlphaGo. *arXiv preprint arXiv:1710.10967*.

6. **Working with new data.** Which new data sources does machine learning make available to economists, and how do these techniques work on a high level?

   (a) Text: bag-or-words techniques, topic modelling, sentiment analysis

   - Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.
   - Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
   - Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–97.

   (b) Images: neural nets

   - Jean, N., Burke, M., Xie, M., Davis, M., Lobell, D., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

   (c) Medical and administrative records: challenges and best practices

   - Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13):1216.

7. **Implications of the availability of machine learning.** How would we expect markets to change when these techniques become more and more available? Who wins, who loses?

   - Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

8. **Transparency and fairness.** Which technical, ethical, and legal challenges come when machine learning methods distinguish between people in policy and commercial applications? How can they be addressed?

   - Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM.
   - Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.
   - Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2017). Predictably unequal? The effects of machine learning on credit markets. Technical report, CEPR Discussion Papers.
   - Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, pages 22–27. American Economic Association.