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

Jann Spiess*
Assistant Professor of Operations, Information and Technology, Stanford GSB
jspiess@stanford.edu

Preliminary Syllabus
July 10, 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.

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

*Based on work with Sendhil Mullainathan.

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:

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The Elements of Statistical Learning. Springer, Berlin: Springer Series in Statistics, 2001. Available online as a pdf download.
- Murphy, Kevin P. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
- Bishop, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.

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?
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?
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?
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
   (b) Prediction in the service of estimation: where prediction techniques can enhance causal inference
   (c) Predictability as the question of interest: where properties of a prediction solution provide tests for theories
   (d) New data: where prediction solutions allow us to use new data sources to construct variables for further analysis

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?

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

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

   (b) Images: neural nets

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

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?

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?

References

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