Machine Learning and Econometrics (Optional Read)

Page
Settings
More

This is a discussion I had with some students in my Econ 414 (Urban Economics) course some years ago on this topic. Thought I would share with you (again, neither required nor recommended -- strictly for your continued education/interest):

Following up on our brief chat during Econ 414 class today concerning Machine Learning (ML) and Econometrics (EM).

(I am cc:ing others who might also be interested. Entirely FYI, won't be tested in Econ 361 or 414 ;-)

There is a recent issue (Spring 2017) of the Journal of Economic Perspectives that surveyed "recent" ideas in Econometrics, including the use of popular machine learning approaches:

https://www.aeaweb.org/issues/453

The Journal has free access and the articles are written with an educated but general audience in mind. (Don't need to know too much ML or EM to be able to read the articles)

The articles of particular interest are

1. "Machine Learning: An Applied Econometric Approach" by Sendhil Mullainathan and Jann Spiess
2. "The State of Applied Econometrics: Causality and Policy Evaluation" by Susan Athey an Guido Imbens

FYI — Susan Athey used to be chief economist at Microsoft; micro theorist who does research in mathematical economics and IO. Sendhil Mullainathan is a friend/former classmate of Jessica Reyes from her years as a graduate student at Harvard

The key takeaway is as follows:

Machine Learning is good at addressing the class of problems it was designed to tackle: namely, predicting Y given some X in a largely data-driven manner.

[For those of you who took Econ 361 with me: this is part of the reason why I spend much of the early part of the course focused on Best Predictor and Best Linear Predictor of Y given X. It provides the connection between modern economics and modern statistics]

To the extent that the analyst does not have useful information about the data generating process (joint distribution of Y and X) that is not contained in the observed data, it is hard to come up with statistical methods that lead to better predictive outcomes than modern ML (e.g. LASSO, Tree, neural nets)

For those of you not familiar with ML, the easiest way to see this is to think about Mean Squared Error (MSE), which is implicitly the loss/risk function rationalizing most econometrics (and even statistical) analysis. MSE is the sum of the variance of the predictor and the bias^2 of the predictor. In econometrics, there is an obsession with predictors (estimators) that are unbiased (consistent) — this is mainly for hypothesis testing reason. But this implies that the prediction/estimation methods being used in econometrics may not have the best MSE properties as they forgo potentially valuable trade-offs between variance and bias. This is basically the intuition underlying the early ML methods like Ridge and LASSO (which are basically constrained regressions where the constraints on the possible coefficient values introduce some bias but also significant reduction in variance)

So, from a pure prediction perspective (using MSE as the loss/risk function), standard econometric methods which emphasize unbiasedness/consistency can be suboptimal.

However, a drawback to ML methods is that hypothesis testing of the underlying parameter estimates is difficult. This is why most analysis involving ML methods eschew hypothesis testing and, instead, explore the “precision” of their estimate/prediction using interval estimation (e.g. confidence intervals).

But if the objective is less pure data-driven prediction and more precise estimates of parameters underlying the presumed data generating process, then ML methods can suffer. When would this be the objective? One situation: when you have information about the data generating process that is not contained in the data — e.g. economic theory/models that may be informative about the data generating process. So, instead of focusing on so-called “supervised ML” methods where the supervision is largely/entirely data-driven, econometrics can do better if economic theory/models provide supervision superior to simply data-driven supervision.

This leads to a popular characterization of the ML vs. EM debate: the “superior” choice depends on how informative economic theory/models are for the data generating process being studied/explored.

For a better discussion of this point, see
http://science.sciencemag.org/content/346/6210/1243089
("Economics in the Age of Big Data" by L. Einav and J. Levin, Science Nov 2014 — Einav and Levin are leading IO economists)

Of course, this dichotomy between ML and Econometrics is a bit false. One could do supervised ML where some of the supervision is data-driven and some from relevant economic theory/models. But such hybrid methods are tricky as the exact implementation is difficult to standardize — depends on the situation. This is why some major data science teams at top tech firms are led by economists and computer scientists — economists to help adapt the supervision process and CS types to help write the efficient/effective code implementing the supervision scheme (a non-trivial problem when you are dealing with "Big Data")

By the way, the Journal issue also has articles discussing structural models. Structural models are econometric models that explicitly use economic theory/models. So called "structuralists" are in the minority in empirical economics; most empirical economist follow the footstep of Angrist and Pischke (there is an article by them in the issue) too, which uses economic theory/models more casually. Within economics, the major debate is less ML vs EM and more structural vs. so-called "reduced form" (i.e. how does/should one use economic theory/models)

Disclosure: I am mostly in the structural camp, as are most IO economists. The other empirical economists on the Amherst faculty are mostly in the reduced form camp. My personal view is that if you are doing reduced form (so not using economic theory/model as much), you might as well do ML.

Jun Ishii
Associate Professor of Economics