Machine Learning: An Applied Econometric Approach

Mullainathan, S. & Spiesss, J.
JEP (2017)

Is it econometrics?

- It solves a different problem using different set of tools.
- It generates *predictions*, $\hat{y}$. E.g., face recognition, translation of websites, understand voices, etc.
- It does that by: (i) discovering complex structures (w/o specification in advance) and (ii) finding functions that work well OOS.
- ML in economics, requires a relevant $\hat{y}$ task:
  1. New kind of data for traditional questions.
  2. The inference procedure ($\hat{\beta}$) can contain a prediction task.
  3. Direct policy applications.
- Easy to use, but...

How ML works

- Once again, it seeks functions that predict well OOS.
- In performance, it does better than OLS (some algorithms).

	Prediction performance (R^2)	
Method	*Training sample*	*Hold-out sample*
Ordinary least squares	47.3%	41.7% [39.7%, 43.7%]
Regression tree tuned by depth	39.6%	34.5% [32.6%, 36.5%]
LASSO	46.0%	43.3% [41.5%, 45.2%]
Random forest	85.1%	45.5% [43.6%, 47.5%]
Ensemble	80.4%	45.9% [44.0%, 47.9%]

- ML searches for the interactions needed automatically. But...?
- Curse of dimensionality: more flexible f forms, better fit, worse OOS prediction.
- Nothing is lost, solution through some structure (i) regularization and (ii) empirical tuning.

How ML works (cont'd)

- This structure helps us organize the variety of prediction algorithms.

Some Machine Learning Algorithms

Function class $\mathcal{F}$ (and its parametrization)	*Regularizer $R(f)$*
Global/parametric predictors	
Linear $\beta' x$ (and generalizations)	Subset selection $\|\|\beta\|\|_0 = \sum_{j=1}^{k} 1_{\beta_j \neq 0}$
	LASSO $\|\|\beta\|\|_1 = \sum_{j=1}^{k} \|\beta_j\|$
	Ridge $\|\|\beta\|\|_2^2 = \sum_{j=1}^{k} \beta_j^2$
	Elastic net $\alpha \|\|\beta\|\|_1 + (1-\alpha) \|\|\beta\|\|_2^2$
Local/nonparametric predictors	
Decision/regression trees	Depth, number of nodes/leaves, minimal leaf size, information gain at splits
Random forest (linear combination of trees)	Number of trees, number of variables used in each tree, size of bootstrap sample, complexity of trees (see above)
Nearest neighbors	Number of neighbors
Kernel regression	Kernel bandwidth
Mixed predictors	
Deep learning, neural nets, convolutional neural networks	Number of levels, number of neurons per level, connectivity between neurons
Splines	Number of knots, order
Combined predictors	
Bagging: unweighted average of predictors from bootstrap draws	Number of draws, size of bootstrap samples (and individual regularization parameters)
Boosting: linear combination of predictions of residual	Learning rate, number of iterations (and individual regularization parameters)
Ensemble: weighted combination of different predictors	Ensemble weights (and individual regularization parameters)

How ML works (cont'd)

We can help us with econometrics to answer the following questions:

- How do we choose the function we fit?
- How do we regularize them?
- How to encode and transform the underlying variables?
- Should OOS performance be measure using a CV or correction for overfitting?
- How many fold should we used when CV?
- How should the final tuning parameter be chosen?

The answer relies on economic theory and content expertise. **There is no a definitive answer to this.**

Drawbacks

- It cannot be used to learn about the underlying model.
- Lack of standard errors on the coefficients in order to make inference. Also have to take into account model selection.
- A variable used in one partition may be unused in another. The algorithm can return very unstable patterns (this are not reflected in R^2). $\rightarrow$ Variables are correlated with each other.
- Regularization is a problem itself: (i) choice less complex models, but these might be the wrong models. (ii) It cause omitted variable error.

Applicability

- New data: deals with unconventional data - high-dimensional for standard methods (e.g., Google Street View to measure block-level income in NYC and BOS).
- Prediction in the service of estimation: tasks that we approach as estimation problems (e.g., IV 1st stage).
- Prediction in policy: prediction is really related to questions we already seek to answer (e.g., impact of an extra teacher depends on how she is chosen).
- Testing theories: inherently about predictability (e.g., efficient markets theory).