The Impact of Machine Learning on Economics

by: Susan Athey

Discussion by:

Mara Lederman

Rotman School of Management, University of Toronto

NBER Conference on the Economics of AI

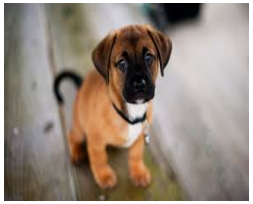
September 2017

Unsupervised ML

Finding clusters of "observations" that are similar in their "covariates"



The items in a group are similar to each other and different from items in the other group


Supervised ML

Using a set of covariates ("x"s) to predict an outcome ("y")

Employee Turnover

Fraudulent Transactions

Engine Maintenance

"Typical" covariates/ structured data

Unstructured data

Image Recognition

Document relevance

Customer Service ("Bots")

Cancer Diagnosis

The output is a variable we can work with

Supervised ML	Standard Econometrics
Predicting $\hat{y}$	Estimating $\hat{\beta}$
Model selected by the machine	Model specified in advance
Correlation	Causation
Maximizing goodness of fit	Sacrificing goodness of fit for identification

Impact on (Research in) Economics

1. Improving prediction in research problems that are fundamentally prediction exercises
   - See Varian (JEP 2014)

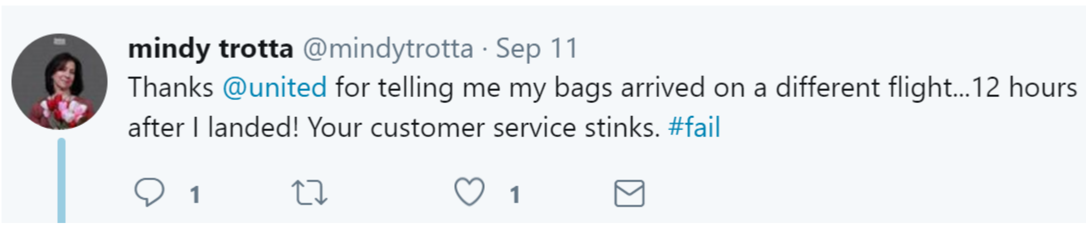
2. Using ML techniques for model/variable selection; robustness
   - Varian (JEP 2014)
   - Athey (today's paper)

3. Prediction policy problems
   - Mullainathan and Spiess (JEP 2017)

4. New data/variables to incorporate into traditional econometric analyses

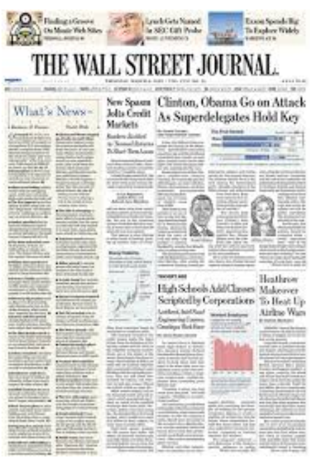
5. ML as the "data-generating process"

Machine learning can turn stuff that doesn't look like data into data



Y=1 if negative
Y=0 otherwise

Gans, Goldfarb and Lederman (2017)


THE WALL STREET JOURNAL.

What's News

Clinton, Obama Go on Attack As Superdelegates Hold Key

Y="slant"

Gentzkow and Shapiro (2010)

Many Other Possibilities?

- Financial filings?
- Job descriptions?
- Contracts?
- Patent applications?
- Online reviews?
- Performance evaluations?
- Facial expressions/eye contact/body movement?
- Court transcripts?
- Email text?
- Social media connections?
- Interview transcripts and videos?
- Health trackers/wearables?
- Calendars?
- Etc…

Substituting Predictions for Missing Data

- ML allows us to create predicted values of a variable IF we have:
  - Information on the variable in another dataset (the training data)
  - Things that correlate with the variable
  - Information on the covariates in the main dataset

- Could be used in several ways:
  1. **Impute missing values (or all values) of a control variable**
     - Eg: no info on income but can predict it from other characteristics if have income and those characteristics in some training data; health conditions?
     - Privacy concerns?
  2. **Predicting LHS variables at a higher frequency than they are measured**
     - Glaeser et al (2016) predict hygiene scores from Yelp reviews. Could use predicted score as dependent variable when investigating impact of policies on hygiene performance (Glaeser et al, 2015)
  3. **Predict a variable that may be easy to gather in one setting but hard to gather in others**
     - Glaeser et al (2016) predict from Google street view images in NYC. Can be used to predict income in places where measurement of income is difficult

ML as the Data-Generating Process

- The data we analyze in empirical work comes from the “real world”

- We care about the **data-generating process**: the sets of behaviors, decisions, actions, random events that generate the data we observe

    - Observed prices are the result of the interaction of supply and demand
    - Educational outcomes are the result from investments in human capital
    - Firm boundaries are the result of economizing on transaction costs

(How) is our empirical work affected when the data we observe results from decisions made by AIs and not the agents we typically model?

Do We Need to Adjust Our Models?

Think about structural estimation where we take the models written down seriously and use them to estimate structural parameters

- For example, in IO, common to estimate structural models of demand and supply (eg: to consider impact of mergers or estimate elasticities)

- Typically specify a some sort of pricing game on the supply-side

If firms are using AI to set prices, do we need to model how the AI makes decisions? Do we need to model a firm's decision to delegate pricing to AI?

Does the AI get closer to our model of decision-making or further?

Does Adoption of ML Make Casual Inference Even Harder?

- More often than not, the data-generating process is not random assignment

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{\text{Observed difference in average health}} = \underbrace{E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]}_{\text{average treatment effect on the treated}} + \underbrace{E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{\text{selection bias}}$$

Angrist and Pischke (2009)

- We model the data-generating process to understand the selection bias:
  - Higher ability students apply obtain MBAs
  - Firms advertise when and where they expect to find their customers

- Now, decisions are made by AIs which are targeting the treatment at the individuals/groups that are believed to have the largest treatment effects
  - MBA programs admit students predicted to have high salaries upon graduation
  - Firms show ads to customers who have already viewed a product

If this targeting makes the two groups even more different in terms of pre-treatment outcomes, it will make the data-generated even more problematic for causal inference