Mini-course Machine Learning in Empirical Economic Research

Lecture 1: Introduction

Andreas Dzemski[1]

[1]University of Gothenburg

May 28, 2019

What is econometrics

- structural function

$$g(x_1, x_2)$$

- used to express an economic idea

$$\text{wage gap} = \int (g(1, x_2) - g(0, x_2))\ dF(x_2)$$

- often a *counterfactual*
- structural empirical model

$$y = g(x_1, x_2) + \epsilon$$

- links structural function to data

What is econometrics?

- identifying assumptions

$$y = \underbrace{g(x_1, x_2)}_{\text{rank assumption}} \overbrace{+}^{\text{separability}} \underbrace{\epsilon}_{\text{exogeneity}}$$

- estimation: functional form assumptions about g
- an incorrect model is properly not very useful

What is machine learning?

- supervised learning
- unsupervised learning

Ways to think about machine learning

1. statistical tools for prediction
2. methods for "high-dimensional" data
   - everything is data!
3. methods to fit flexible, non-smooth functional forms
   - regression trees, random trees/forests, neural networks
4. data-driven model selection
5. often targeted toward data-rich environments (?)

Focus on prediction

"[. . .] few assumptions are required for off-the-shelf prediction techniques to work: The environment must be stable, and the units whose behavior is being studied should not interact or "interfere" with one another. In many applications, SML [supervised machine learning] techniques can be successfully applied by data scientists with little knowledge of the problem domain." (Athey 2017)

Focus on prediction

- "wrong" models may be very useful!
- for prediction we check the predictive properties of our model on "new" data
- more generally: supervised learning works well if we can formulate a loss function
  - criterion that tells us, using only observable data, how good our model is
- loss function for causal relationship?

But: Causal inference requires predictive tasks

non-parametric IV model (Newey and Powell 2003)

$$y = g(x) + \epsilon$$

instrumental variable z

$$\mathbb{E}[\epsilon \mid z] = 0$$

Identification of g in IV model

identification from *Fredholm integral equation*

$$\underbrace{\mathbb{E}[y \mid z]}_{\text{identified}} = \int g(x) \underbrace{dF(x \mid z)}_{\text{identified}}$$

- estimation of $\mathbb{E}[y \mid z]$ and $F(x \mid z)$ are prediction problems (why?)
- then, recovering g is a deconvolution problem

"DeepIV" approach by Hartford et al. (2017)

- first-stage estimator $\hat{F}(\cdot \mid \cdot)$
- formulate deconvolution as a supervised learning problem
- loss function

$$L(g) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i - \int g(x)\, dF(x \mid z_i) \right\}^2$$

Other prediction tasks in empirical economics

- prediction may be a way to address the research question
  - predictability of bankruptcy (Becerra, Galvão, and Abou-Seada 2005)
  - predictability of volatility from public statements (Kogan et al. 2009)
- imputation of variables

  The estimation of the treatment effect on the treated requires only predicting the control outcome of the treated. This task is a supervised learning problem.

- Do you agree with this statement?

High-dimensional data

- many "features" are considered for prediction
- consider many economic variables
  - examples: detailed census data, social network activity
- uncertainty about functional form
  - consider many transformations/interactions of small set of economic variables

Everything is data

- images
- texts
- social media usage
- pre-processing step converts these into *high-dimensional* datasets
- *unsupervised learning* in pre-processing

Estimating poverty from satellite images

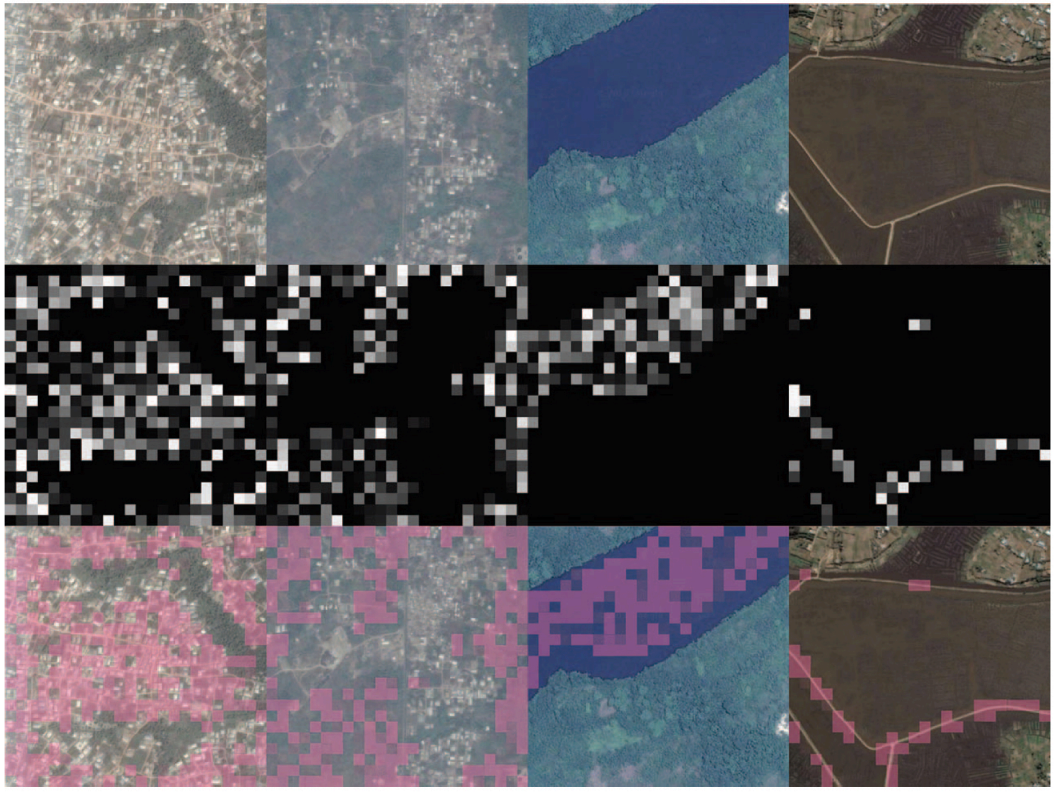

Fig. 2. Visualization of features. By column: Four different convolutional filters (which identify, from left to right, features corresponding to urban areas, nonurban areas, water, and roads) in the convolutional neural network model used for extracting features. Each filter "highlights" the parts of the image that activate it, shown in pink. By row: Original daytime satellite images from Google Static Maps, filter activation maps, and overlay of activation maps onto original images

Figure: Source: Jean et al. (2016)

Estimating partisan slant in newspapers

TABLE I

MOST PARTISAN PHRASES FROM THE 2005 *CONGRESSIONAL RECORD*[a]

Panel A: Phrases Used More Often by Democrats

Two-Word Phrases		
private accounts	Rosa Parks	workers rights
trade agreement	President budget	poor people
American people	Republican party	Republican leader
tax breaks	change the rules	Arctic refuge
trade deficit	minimum wage	cut funding
oil companies	budget deficit	American workers
credit card	Republican senators	living in poverty
nuclear option	privatization plan	Senate Republicans
war in Iraq	wildlife refuge	fuel efficiency
middle class	card companies	national wildlife
Three-Word Phrases		
veterans health care	corporation for public broadcasting	cut health care
congressional black caucus	additional tax cuts	civil rights movement
VA health care	pay for tax cuts	cuts to child support
billion in tax cuts	tax cuts for people	drilling in the Arctic National
credit card companies	oil and gas companies	victims of gun violence
security trust fund	prescription drug bill	solvency of social security
social security trust	caliber sniper rifles	Voting Rights Act
privatize social security	increase in the minimum wage	war in Iraq and Afghanistan
American free trade	system of checks and balances	civil rights protections
central American free	middle class families	credit card debt

(*Continues*)

Figure: Source: Gentzkow and Shapiro (2010)

Predicting wealth from mobile phone logs

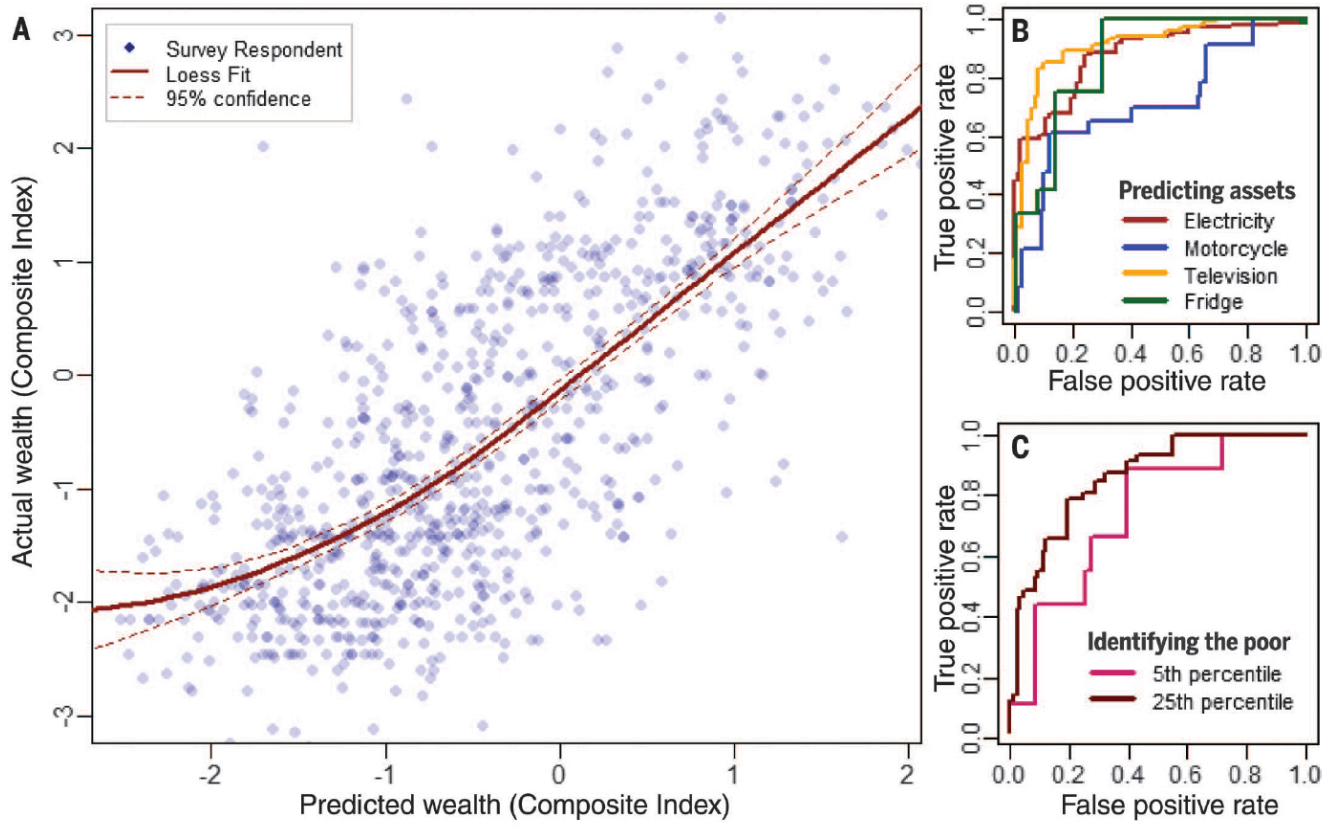

Figure: Source: Blumenstock, Cadamuro, and On (2015)

Impressive results in data-rich environments

- predicting related products in a large online shop (customers are continuously creating new data through "clicks")
- predicting viewing behavior on Netflix
- Prediction tasks in AI: teaching a robot how to walk, teach an AI to play a video game

"The total playing time between humans and the AI equated to 10.7 years, which is impressive until you learn that OpenAI Five generates this much data every 12 minutes of training by playing against itself" (The Verge, 2019-04-23)

Goal of this course: Have this makes sense to you

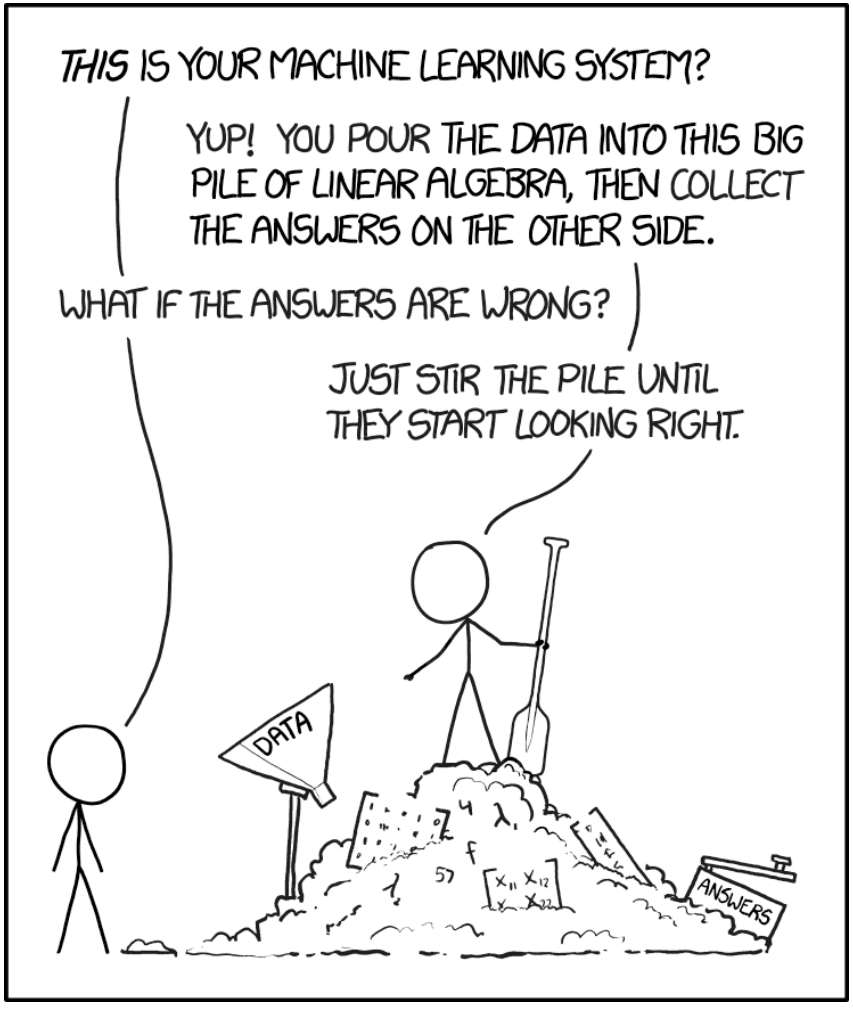

Figure: Source: https://xkcd.com/1838/

Tools for machine learning

- implement machine learning methods (train and evaluate)
- data scraping and handling
- distributed computing

Programming languages for ML

- a lot of "standard" methods are implemented in Stata
  - the usual caveat applies
- Python or R if you want to get real
  - don't share Stata's caveat
  - extensive library support (e.g. caret for R, scikit-learn for Python)
  - free and pre-installed on many clusters
  - easy file system and string manipulation
- libraries with bindings to different languages
  - Google's tensorflow (including R and Python)

Don't underestimate the power of system admin tools

Figure: Source: https://xkcd.com/208/

Computational resources

- we can use the Swedish National Infrastructure for Computing (SNIC)
- Hebbe cluster at Chalmers
- you have to apply with a project
- batch jobs
- drawback/advantage: have to know basics of
  - using Linux from the command line
  - Bash
- for sensitive data BIANCA