Big Data, Machine Learning, and Artificial Intelligence: Methods, Lectures and Applications

Susan Athey, Professor, Stanford GSB

Introduction to Machine Learning Methods

Thursday, August 26, 2021
Selected Overview Articles: Econometrics and ML


Tutorials: [https://bookdown.org/stanfordgsbsilab/tutorial/](https://bookdown.org/stanfordgsbsilab/tutorial/)

Survey

Prediction v. Estimation

Prediction policy

Prediction v. Causal Inference
Introduction and Themes
Two Types of Machine Learning

**SUPERVISED**

Independent observations
Stable environment
Regression/prediction:
  - $E[Y|X=x]$
Classification
  - $Pr(Y=y|X=x)$

**UNSUPERVISED**

Collections of units characterized by features
  - Images
  - Documents
  - Individual internet activity history

Find groups of similar items

Also of interest for causal inference:
Bandits/Reinforcement Learning; Generative Adversarial Networks
Classification

Advances in ML dramatically improve quality of image classification
Classification

Neural nets figure out what features of image are important

Features can be used to classify images

Relies on stability

Given \( X_i \) is this a cat?

\[
\begin{align*}
\Pr(Y_i = CAT | X_i) &= .95 \\
\Pr(Y_i = DOG | X_i) &= .05
\end{align*}
\]
What’s New About ML?

Flexible, rich, data-driven models

Increase in personalization and precision

Methods to avoid over-fitting
Prediction in a Stable Environment

Goal: estimate $\mu(x) = E[Y|X = x]$ and minimize MSE in a new dataset where only $X$ is observed

- $MSE = \frac{1}{T} \sum_i (Y_i - \hat{\mu}(X_i))^2$
- No matter how complex the model, the output, the prediction, is a single number
- Can hold out a test set and evaluate the performance of a model
- Ground truth is observed in a test set
- Only assumptions required: independent observations, and joint distribution of $(Y,X)$ same in test set as in training set

Note: minimizing MSE entails bias-variance tradeoff, and always accept some bias

- Idea: if estimator too sensitive to current dataset, then procedure will be variable across datasets
- Models are very rich, and overfitting is a real concern, so approaches to control overfit necessary

For work on stability:
- See series of papers with Athey, Kuang, Cui, Li, et al
Prediction in a Stable Environment

Idea of ML algorithms

- Consider a family of models
- Use the data to select among the models or choose tuning parameters
- Common approach: cross-validation
  - Break data into 10 folds
  - Estimate on 9/10 of data, estimate MSE on last tenth, for each of a grid of tuning parameters
  - Choose the parameters that minimize MSE

ML works well because you can accurately evaluate performance without add’l assumptions

- Your robotic research assistant then tests many models to see what performs best
Economists have focused on the case with substantially more observations than covariates (N>>P)

- In-sample MSE is a good approximation to out-of-sample MSE
- OLS is BLUE, and if overfitting is not a problem, then no need to incur bias
- OLS uses all the data and minimizes in-sample MSE

OLS obviously fails due to overfitting when P~N and fails entirely when P>N
- ML methods generally work when P>N

Economists worry about estimating causal effects and identification
- Causal effects
- Counterfactual predictions
- Separating correlation from causality
- Standard errors
- Structural models incorporating behavioral assumptions

Identification problems can not be evaluated using a hold-out set
- If joint dist’n of observable same in training and test, will get the same results in both

Causal methods sacrifice goodness-of-fit to focus only on variation in data that identifies parameters of interest
What We Say v. What We Do (Econometrics)

**What We Say**
- Causal inference and counterfactuals
- God gave us the model
- We report estimated causal effects and appropriate standard errors
- Plus a few additional specifications for robustness

**What we do**
- Run OLS or IV regressions
  - Try a lot of functional forms
  - Report standard errors as if we ran only one model
  - Have research assistants run hundreds of regressions and pick a few “representative” ones
- Use complex structural models
  - Make a lot of assumptions without a great way to test them
Key Lessons for Econometrics

Many problems can be decomposed into predictive and causal parts
- Can use off-the-shelf ML for predictive parts

Data-driven model selection
- Tailored to econometric goals
  - Focus on parameters of interest
  - Define correct criterion for model
  - Use data-driven model selection where performance can be evaluated
- While retaining ability to do inference

ML-Inspired Approaches for Robustness

Validation
- ML always has a test set
- Econometrics can consider alternatives
  - Ruiz, Athey and Blei (2017) evaluate on days with unusual prices
  - Donnelly, Ruiz, Blei and Athey (2019) evaluate change in purchases before and after price changes
  - Tech firm applications have many A/B tests and algorithm changes

Other computational approaches for structural models
- Stochastic gradient descent
- Variational Inference (Bayesian models)

See Sendhil Mullainathan et al (JEP, AER) for key lessons about prediction in economics
- See also Athey (Science, 2017)
Empirical Economics in Five Years: My Predictions

<table>
<thead>
<tr>
<th>Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regularization/data-driven model selection will be the standard for economic models</td>
</tr>
<tr>
<td>Principled ways to interpret and describe results of resulting models</td>
</tr>
<tr>
<td>Prediction problems better appreciated</td>
</tr>
<tr>
<td>Measurement using ML techniques an important subfield</td>
</tr>
<tr>
<td>Textual analysis standard (already many examples)</td>
</tr>
<tr>
<td>Papers will bring in much more auxiliary and unstructured data</td>
</tr>
<tr>
<td>Models will explicitly distinguish causal parts and predictive parts</td>
</tr>
<tr>
<td>Reduced emphasis on sampling variation</td>
</tr>
<tr>
<td>Model robustness emphasized on equal footing with standard errors</td>
</tr>
<tr>
<td>Models with lots of latent variables</td>
</tr>
</tbody>
</table>
How can economics and statistics improve AI/ML?
Challenges for Management/Regulation of ML

Credit Scoring Example
- **Instability** of joint distribution of outcomes, novel features
- **Poor performance** when extrapolating
- **Manipulation** of novel features
- **Discrimination** and **Fairness**
- Ever-changing **adverse selection** problem as competing firms change models, marketing strategies
- When are results more or less **reliable**?

**Equilibrium** effects
- Agents using ML interact
- Collusion (airline prices)
- Instability (financial market crashes, correlated mistakes across firms)
- Google maps examples

Need models of individual behavior and eqm selection to study eqm changes
- Why existing AI/ML is a long way from solving “harder” problems

Algorithms have demonstrable errors
Engineers build black-box algorithms, but are not trained to evaluate
Need “best practices” to analyze the black box
Artificial Intelligence/Machine Learning

Desired Properties for Applications

**DESIRED PROPERTIES**

- Interpretability
- Stability/Robustness
- Transferability
- Fairness/Non-discrimination
- “Human-like” decision-making
  - Reasonable decisions in never-experienced situations

**CAUSAL INFEREN GE FRAMEWORK**

Goal: learn model of how the world works
- Impact of interventions can be context-specific
- Model maps contexts and interventions to outcomes
- Formal language to separate out correlates and causes

Ideal causal model is by definition stable, interpretable

Transferability: straightforward for new context dist’n
- If you estimate treatment effect heterogeneity

Fairness: Many aspects of algorithmic discrimination relate to correlation v. causation
- Gender and race may be correlated with factors that shift distributions of characteristics like test scores or credit scores, relatively limited direct causal effects
Machine Learning and Causal Inference
Goal: estimate the causal impact of interventions or treatment assignment policies
  ◦ Low dimensional intervention
  ◦ Desire confidence intervals

Estimands
  ◦ Average effect
  ◦ Heterogeneous effects
  ◦ Optimal policy

Designs that enable identification and estimation of these effects
  ◦ Randomized experiments
  ◦ Unconfoundedness
  ◦ “Natural” experiments (IV)
  ◦ Regression discontinuity
  ◦ Difference-in-difference
  ◦ Longitudinal data
  ◦ Randomized and natural experiments in social network/settings w/ interference
1. Consider **identification**, then **estimation**
   - Could you solve problem with infinite data?
   - Design-based approach determines objective function
   - What parts of problem need flexible functional form?

2. **Semi-parametric efficiency theory** brings insights not commonly exploited in traditional ML
   - *ML methods replace kernels/sieves in semi-parametric estimation*
   - Orthogonal moments/double robustness
   - Important since hard to estimate flexible functions well

3. Regularization induces **omitted variable bias**
   - Optimizing for goodness of fit (MSE) prioritizes predictive covariates
   - Choose objective fn carefully, e.g. prioritize control for confounders
   - Omitted variables challenge causal inference, interpretability, fairness

4. Sample splitting, cross-fitting restore **statistical guarantees**
   - Different data for model selection and model estimation
   - Cross-fitting/out of bag estimation of nuisance parameters

5. Exploit **structure of problem** carefully for better counterfactual predictions
   - Black-box algorithms reserved for nuisance parameters, parameter heterogeneity

6. Carefully crafted approaches to **validate** and **tune** ML models to optimize for structural/causal parameters