Prospects for Economics in the Machine Learning and Big Data Era

SEPT 29, 2018

WILLIAM CHOW

Introduction

- Is Big Data and the associated treatment of it a fad?
- No.
  - Because the economics profession has been engaging in related datamining techniques for quite some time.
  - Artificial Intelligence – Genetic Algorithm and Neural Networks.
  - Machine Learning – Regression Trees, Ridge Regression and Principal Component Analysis.

Introduction

- Yes (maybe).
  - Because we may not want to let go conventional analytics structured to data of much smaller dimension and with rather different theoretical virtues.
  - Causal inference instead of mere prediction.
  - Macroeconomics seems more sceptical.

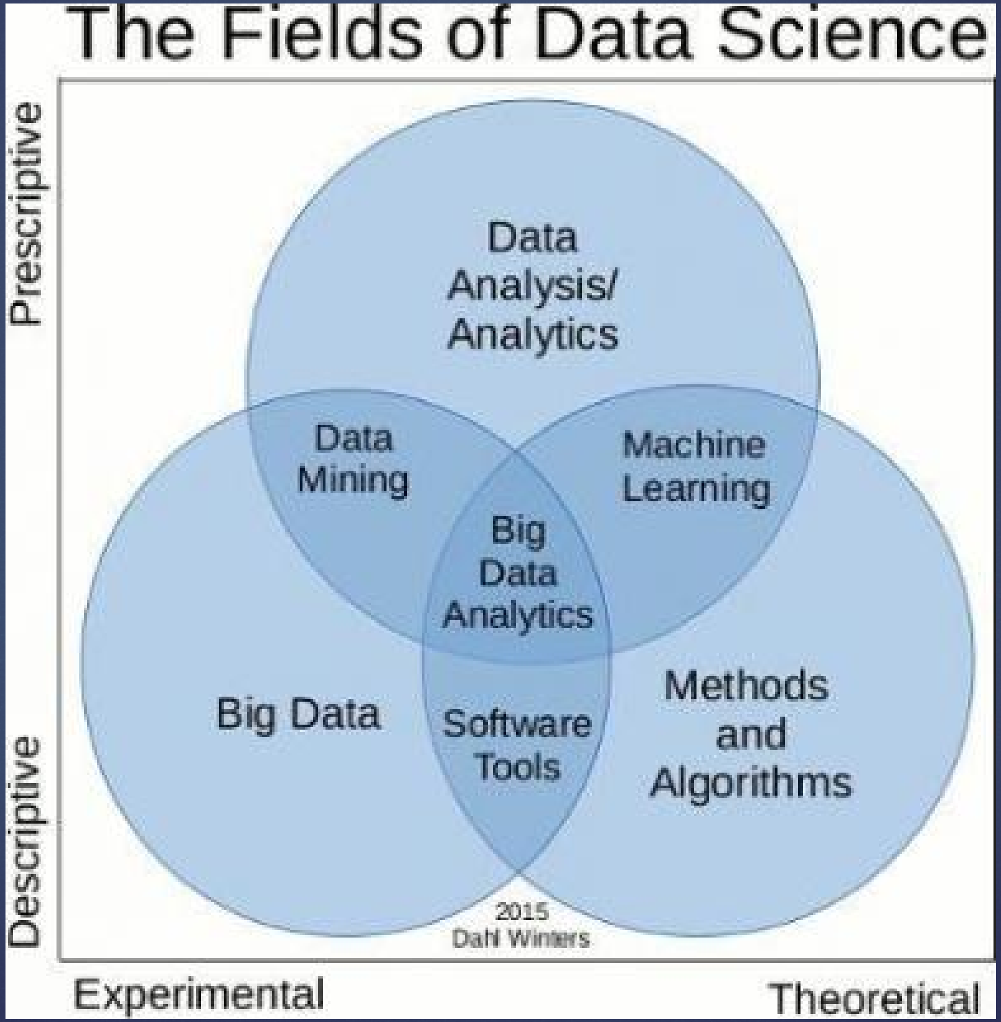
Introduction

- The proliferation of e-communication and e-commerce offers a wealth of data.
  - Google searches reach 5.2 billion a day in 2017.
- The “volume”, “variety” and “velocity” of data are much different from what economists are accustomed to.
- We need new methods to analyse these less structured data e.g. weblogs, scanner data and photos.

Current State of Play

- Collaboration of Economics profession and industry.
  - Susan Athey with Microsoft.
  - Hal Varian with Google.
- A non-exhaustive list of Machine Learning works:
  - Athey, S., Imbens, G. (2016): Treatment Effects.
  - Bajari et al. (2015): Demand estimation.
  - Glaeser et al. (2018): Urban economics.
  - Kleinberg (2017): Human decision vs Machine prediction.
  - Peysakhovich, A., Naecker, J. (2017): Behavioral Economics.
- Software: R and Python

What Machine Learning is About



What Machine Learning is About

- Using computer algorithms to perform predictions and classification.
    1. Simulation methods are common in economics, not always.
- Simply put, the task is to predict future outcome Y given observed characteristics X and past Y.
    1. This is what economists care as well.
    2. But identifying structural parameters is relatively downplayed – Lucas Critique.
    3. Also emphasize hypothesis testing.

What Machine Learning is About

- ML methods are particularly suited for Big Data which can come in a form $N \ll K$, or no. of observations smaller than that of covariates.
  1. Standard econometrics usually cannot handle that.
  2. Exceptions are Bayesian methods, shrinkage regressions and stepwise regressions.
- Regularization (Dimensionality-reduction) – to avoid over-fitting and excessive complexity.
  1. In-sample fit $\nRightarrow$ Out-of-sample accuracy.
  2. Economists care about this but also consider theories and signal extraction.

What Machine Learning is About

- Regularization requires tuning the model and is manifested as model and variable selection.
  1. Algorithms pick the estimates that minimize the predictive loss function subject to a penalty for complexity.
  2. Model selection less common in economics as there is usually a designated model to estimate.
  3. Cross-validation as a means to tune the model.
  4. Such validation methods are less regularly practised in economics.

What Machine Learning is About

- The prediction task is manifested in sample-splitting, in-sample data training and tuning, and out-of-sample prediction.
  1. Training model with data is usually ignored in economics due to small data set.
  2. More common in forecasting exercises.
- Sometimes accompanied by model averaging.
  1. Idea is the combination of predictions from individual models can produce better predictive performance.
  2. Common in forecasting and Bayesian exercises.

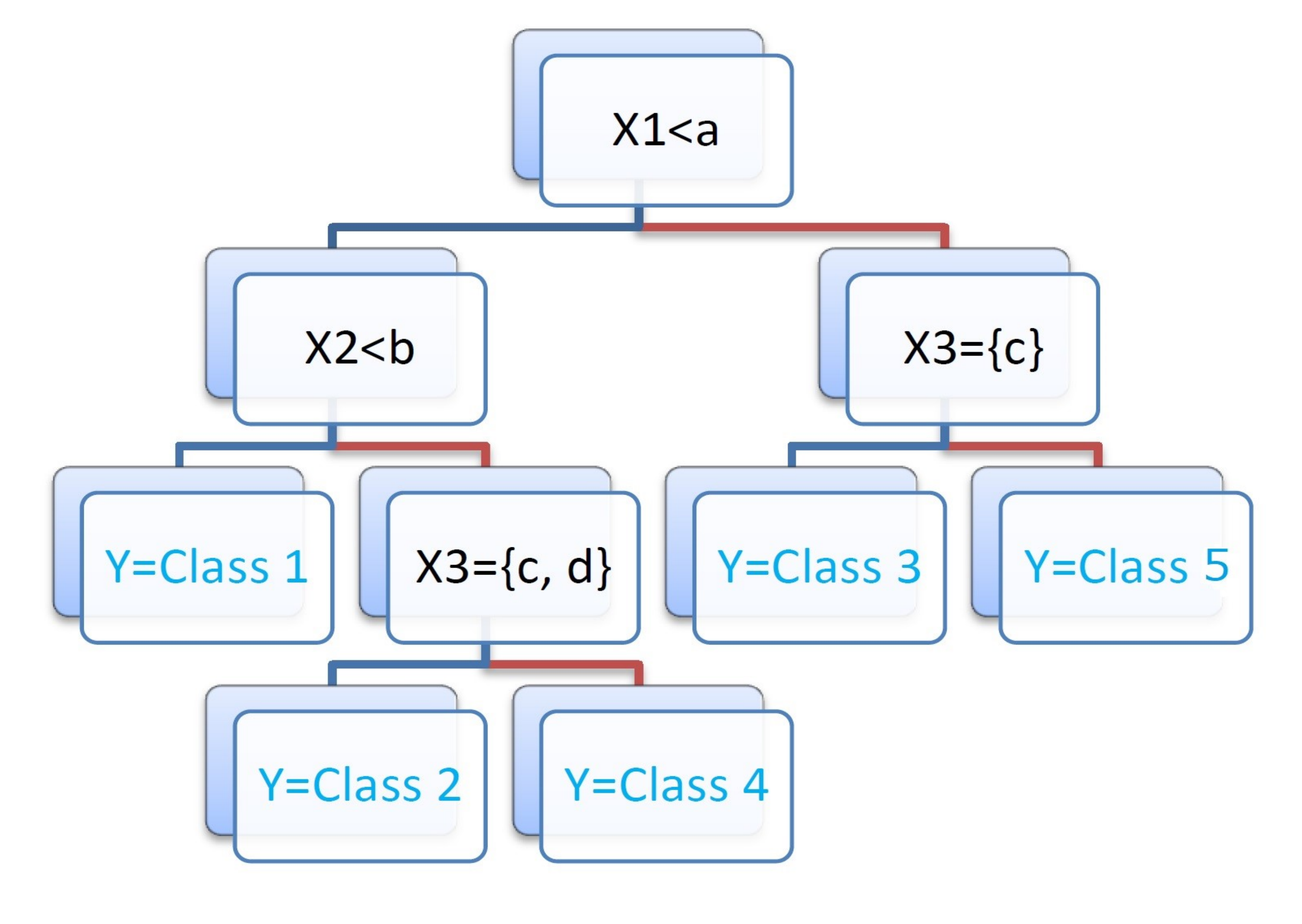
Common Machine Learning Methods

- Classification and Regression Trees (CART).
- Ensemble Methods:
  - Random Forests.
  - Boosting.
- Penalized Regressions:
  - Least Absolute Shrinkage and Selection Operators (LASSO).
  - Support Vector Machines (SVM).

Example 1: CART

- Continuous Y – Regression Trees
- Categorical Y – Classification Trees
- Identify nodes for predictors Xs so that the predicted Y under the node (average within node) has the smallest sum of squared errors.
- Do this recursively until no further splits possible.
- Trees tend to overfit with nodes and leaves.

Example 1: CART



Example 1: CART

- Prune the tree using cross-validation method suggested.
  1. Split data into training and testing set.
  2. Grow the tree with training data set.
  3. Check if error is smaller for the tree using testing data set if we trim a certain node.
- Make prediction, e.g. using empirical frequencies implied by the tree

$$P(Y=\text{fail} \mid X_1<0.4, X_2>5) = 33/44$$

Example 2: LASSO

- Penalized regression of the form:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} (Y_i - X\beta)^2 + \lambda \sum_{j=1}^{K} |\beta_j| \gamma_j \right]$$

- $\lambda = 0$ gives the usual OLS estimator.
- $\lambda > 0$ is the penalty level chosen by cross-validation.
- A non-zero λ puts heavier weight on the 2nd term in the square bracket.
- This forces the β_j to be smaller.
- More capable of delivering sparse system than ridge regressions.

Example 2: LASSO

- γ_j (optional) is the penalty loadings that serve to rescale the Xs.
- For orthonormal regressors X'X=I,

$$\hat{\beta}_j = 0 \ \ if \ \left|[X'Y]_j\right| < \lambda/2$$

- No closed form solution. Solve by convex optimization.
- Solution with non-zero coefficients tend to be biased to zero.

Prospects for Economics

- Econometrics:
  - Model checking and Hypothesis testing.
  - Bayesian methods may have an edge, e.g. Dirichlet Process.
  - Forecasting and nowcasting.
- Microeconomics:
  - Treatment effects and policy analysis e.g. Healthcare.
  - Mechanism design.
  - Network analysis e.g. Trade flows, Market bubbles.

Prospects for Economics

- Macroeconomics:
  - Little done so far.
  - General resistance to paradigm shifts e.g. long transition from representative agent to heterogeneous agents.
  - Structural parameters remain important.
  - Big data allows evaluation of Rationality principle.
  - Agent Based Modeling e.g. Market evolution and Market anomalies.
  - Model and data Validation needed.

Thank You

References

- Arifovic, J. (1994). Genetic Algorithm Learning and the Cobweb Model. Journal of Economic Dynamics and Control, 18(1), 3-28.
- Athey, S., Imbens, G. (2016). Recursive Partitioning for Heterogeneous Causal Effects. Proceedings of the National Academy of Sciences, 113(27), 7353-7360.
- Athey, S. (2018). The Impact of Machine Learning on Economics. In Agrawal, A.K., Gans, J. and Goldfarb, A. (eds.) Economics of Artificial Intelligence: An Agenda. NBER Conference Series, University of Chicago Press.
- Bajari, P., Nekipelov, D., Ryan, S.P., Yang, M. (2015). Demand Estimation with Machine Learning and Model Combination. National Bureau of Economic Research, Working Paper 20955.
- Blake, A.P. (1999). An Artificial Neural Network System of Leading Indicators. National Institute of Economic and Social Research, Discussion Paper.

References

- Einav, L., Levin, J. (2014). Economics in the Age of Big Data. Science, Vol. 326, Issue 6210.
- Glaeser, E.L., Kominers, S.D., Luca, M., Naik, N. (2018). Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life. Economic Inquiry, 56(1), 114-137.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2017). Human Decisions and Machine Predictions. National Bureau of Economic Research, Working Paper 23180.

References

- Mullainathan, S., Mol, C. D., Gautier, E., Giannone, D., Reichlin, L., van Dijk, H., & Wooldridge, J. (2017). Big Data in Economics: Evolution or Revolution? In Matyas, L., Blundell, R., Cantillon, E., Chizzolini, B., Ivaldi, M., Leininger, W., and Mariman, R. (eds) Economics without Borders: Economic Research for European Policy Challenges. Cambridge University Press.
- Peysakhovich, A., Naecker, J. (2017). Using Methods from Machine Learning to Evaluate Models of Human Choice under Risk and Ambiguity.Journal of Economic Behavior & Organization, 133, 373-384.
- Varian, H. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3-28.