Causal Random Forests

Gregory M Duncan

Department of Economics

University of Washington-Seattle

Outline

- Regression Trees
- Random Forests
- Causal Random Forests
  - Identification in non-parametric control function models
    - Blundell and Powell (2002)
    - Blundell and Matzkin (2010)

Trees

- $(y, x) \in R^{1+p}$ estimate $E(y|x)$
- In ML
  - x are called features and can be complicated functions of what we usually call independent variables
  - y is called the target
    - when target is continuous
      - they say regression problem
    - when target is a label (0/1)
      - they say classification problem

Trees

- Partition the feature space into a set of rectangles, R_m,
- Fit a simple model in each
  - Average for regressions
- In essence, a piecewise constant approximation

$$E(y|x) \approx \sum_{m=1}^{M} c_m I(x \in R_m)$$

  - $I(x \in R_m)$ is the indicator function
- Choose M, c_m and R_m to minimize

$$\sum_{n=1}^{N} \left(y_n - \sum_{m=1}^{M} c_m I(x_n \in R_m) \right)^2$$

- If we knew the regions R_m, this would be a straightforward dummy variable regression

Trees

- But we don't know the regions.
  - So find them by searching
- As stated, this problem is too hard to solve
- So solve a simpler one using simple regions defined recursively
- Brings us to Brieman, et al. 1984

CART

(Classification And Regression Trees)

- Successively choose each variable
  - Split the range of that variable into two regions,
  - Calculate mean and sum of squares of Y in each region.
  - Choose the split point as the one that minimizes the sums of squares of the two regions
- Choose the variable to split on as the one with the lowest SSR
- Continue process
  - both regions are split into two more regions,
    - » The split points are called nodes
  - until some stopping rule is applied.
    - The end points are called leaves
- Stop splitting a leaf if
  - all the values in the leaf are the same
  - the number of branches exceeds some tuning parameter
  - the tree gets too complex by some criteria
- Quit when no leaves can be split.

Problems

- CART badly overfits out-of-sample
- Nonetheless used a lot in pharma, medicine, fraud detection
- Brings us to

Brieman (2001)

Random Forests

- If simple trees were independent and unbiased
  - average would be unbiased and have a small variance.
- Such averaging is called ensemble learning
  - averaging over many models tends to give better out-of-sample prediction than choosing a single complicated model.
- Draw B bootstrap samples
- Grow B trees
  - Predictions from each tree turn out to be nearly uncorrelated

Pseudo Code

- For $b = 1 \; to \; B$
  - Draw a bootstrap sample
  - Grow a single tree $T(b)$ from the bootstrapped data
    - Select m variables at random from the p variables.
    - Pick the best variable/split-point among the m.
    - Split the node into two daughter nodes.
    - repeat for each terminal node of the tree, until the some minimum node size is reached.
  - The output of each tree is the average of the y in each leaf

Pseudo Code

- Calculate average of the y in leaves, $\bar{y}_b(x)$
- Output the trees $T(b)\ b = 1, \dots, B$
- To make a prediction at a new point x_0:

Average the averages

$$f(x_0) = \sum_{b=1}^{B} \frac{\bar{y}_b(x_0)}{B}$$

Problem

- Random Forests are hard to interpret
  - Give astoundingly good predictions
  - BUT yield little insight into data generation mechanism
- However, astoundingly good predictions suggests a solution to a classic econometrics problem.
- One problem with instrumental variables is the poor quality of the instruments without overfitting
- Idea:
  - use RF for predicting endogeneous from instruments.

Causal Models

- Economists beginning with Frisch handled causal models.
- These methods had fallen out of favor in economics
- Only to be picked up by Computer Science
  - Pearl
- Recently economists have begun resurrecting these methods
  - Matzkin, Chesher, Blundell and Powell, Heckman
- James Heckman has forcefully pointed out that the methods being developed in Computer Science by Pearl, and others, were well known in economics, sociology and agronomics generations ago

Causal Models

- Problem was they did not predict well
  - Because of reliance on truly simplistic linear and economic models
- Now may be possible to translate the insights of economists in these simplistic situations to utilize modeling methods such as Random Forests
- Recast RF to accommodate causal modeling

Adaptive Nearest Neighbors

- RF recast as an adaptive nearest neighbor estimator
  - (Lin and Jeon (2001))
- Draw M bootstrap samples
- For sample m

$$\hat{Y}_m(X) = \sum_{i=1}^{N} W_{im}(X) Y_{im}$$

- $W_{im}(X) > 0$ and $\sum_{i=1}^{N} W_{im}(X) = 1$.
- When $W_{im}(X_j) > 0$ X_j is called a neighbor of X.
  - In the k-NN algorithm $W_{im}(X) = \frac{1}{k_m}$ if X_j is among the k closest points to X in the m-bootstrap sample and 0 otherwise.
    - Bootstrap data called “inbag”

Adaptive Random Forest

- k_m: number of inbag observations which fall in the same leaf as X in the m-th tree.
- The neighbors of X are X_i which fall in the same leaf as X in at least one tree of the forest.
- The prediction of the whole forest is the weighted average

$$\hat{Y}(X) = \frac{1}{M}\sum_{i=1}^{n} \hat{Y}_m(X) = \sum_{i=1}^{n} \left(\frac{1}{M}\sum_{m=1}^{M} W_{im}(X)\right) Y_i$$

$$\hat{Y}(X) = \sum_{i=1}^{n} W_i(X)\, Y_i$$

- A weighted average with weights

$$W_i(X) = \frac{1}{M}\sum_{m=1}^{M} W_{im}(X)$$

- This puts the model into the framework of Stone (1977)

A Causal Random Forest

- Endogenous variables $y, Y_s, s = 1, \dots, S$
- Exogenous variables X, W, Z
- Reduced form equations

$$Y_s = h_s(W, Z) + e_s$$

- Estimate each using a random forest

$$\hat{Y}_s(W, Z) = \frac{1}{M}\sum_{i=1}^{n} \hat{Y}_{sm}(W, Z)$$

$$= \frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{n} W_{sim}(W, Z) Y_{si}$$

$$= \sum_{i=1}^{n}\left(\frac{1}{M}\sum_{m=1}^{M} W_{sim}(W, Z)\right) Y_{si}$$

A Causal Random Forest

- Define the reduced form residuals as $\hat{e}_s(W_i, Z_i) = Y_{si} - \hat{Y}_s\,(W_i, Z_i)\;\; s = 1, \ldots, S.$
- Structural Equation

$$y = f(Y, X, Z, U)$$

- Control function approach

$$y = f(h(W, Z) + e, X, Z, U)$$
$$y = g(h, e, X, Z\,) + v$$

- suggests random forest prediction of y on Y_0, $\hat{e}_m$, X and Z,

A Causal Random Forest

- obtain $\hat{y}_m = rf(Y_0, \hat{e}_m, X, Z),\ m = 1, \dots, N$.
  - Holding Y_0, X and Z constant
  - Varying $\hat{e}_m$
- Average $\hat{y}_m$ for an average structural function (Blundell and Powell (2002))

$$\hat{y} = \sum_{m=1}^{N} rf(Y_0, \hat{e}_m, X, Z)/N$$

$$= \sum_{m=1}^{N} \hat{y}_m/N$$

Proofs and Simulations

To come later

The End

References

Angrist ,J. and Alan B. Krueger, (2001) Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments, *Journal of Economic Perspectives* –Vol. 15, 69-85

Angrist, J., Imbens, G., and Rubin, D.,(1996) Identification of Causal effects Using Instrumental Variables, *Journal of the American Statistical Association*.

Angrist, Joshua D. and Jörn-Steffen Pischke, (2008) *Mostly Harmless Econometrics*, Princeton University Press

Bates, J., and C. Granger (1969), The Combination of Forecasts, *Operations Research Quarterly*, 20, 451-468.

Berk, Richard A., (2009) *Statistical Learning from a Regression Perspective*, Springer

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984) *Classication and Regression Trees*. Wadsworth and Brooks/Cole, Monterey.

Blundell, Richard; Matzkin, Rosa L. (2010) Conditions for the existence of control functions in nonseparable simultaneous equations models, cemmap working paper, No CWP28/10, http://dx.doi.org/10.1920/wp.cem.2010.2810

Blundell R.W. & J.L. Powell (2003) Endogeneity in Nonparametric and Semiparametric Regression Models. In Dewatripont, M., L.P. Hansen, and S.J. Turnovsky, eds., *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II. Cambridge: Cambridge University Press.

Blundell, R.W. & J.L. Powell (2004) Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* 71, 655-679.

Chesher, A.D. (2005) Nonparametric Identification under Discrete Variation *Econometrica* 73, 1525-1550.

References

Chesher, A.D. (2007) Identification of Nonadditive Structural Functions. In R. Blundell, T. Persson and W. Newey, eds., *Advances in Economics and Econometrics, Theory and Applications*, 9th World Congress, Vol III. Cambridge: Cambridge University Press.

Chesher, A.D. (2010) Instrumental Variable Models for Discrete Outcomes. *Econometrica* 78, 575-601.

Duncan, Gregory M. (1980) Approximate Maximum Likelihood with Datasets That Exceed Computer Limits, *Journal of Econometrics* 14 257-264.

Einav, Liran and Jonathan Levin. The data revolution and economic analysis (2013) NBER Innovation Policy and the Economy Conference,2013.

Friedman, Jerome and Bogdan E. Popescu (2005) Predictive learning via rule ensembles. Technical report, Stanford University http://www-stat.stanford.edu/~jhf/R-RuleFit.html

Friedman, Jerome and Peter Hall (2005) On bagging and nonlinear estimation. Technical report, Stanford University, http://www-stat.stanford.edu/~jhf/ftp/bag.pdf

Friedman, Jerome (1999) Stochastic gradient boosting. Technical report, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/stobst.pdf

Granger, C., and R. Ramanathan (1984), Improved methods of combining forecasts, *Journal of Forecasting*, 3, 197-204.

Haavelmo, T. (1943). "The Statistical Implications of a System of Simultaneous Equations". *Econometrica*, Vol. 11, 1–12.

Haavelmo, T. (1944). "The Probability Approach in Econometrics" *Econometrica*, Vol. 12, Supplement, iii-115

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2 ed http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html

Heckman James J. (2010) Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy, *Journal of Economic Literature*, Vol. 48, No. 2

Heckman, J.J. (2008) Econometric Causality. *International Statistical Review*, Vol. 76, 1--27.

Heckman, J.J., (2005) The scientific model of causality, *Sociological Methodology*, Vol. 35, 1-97.

Heckman, James J. and Rodrigo Pinto (2012) Causal Analysis After Haavelmo: Definitions and a Unified Analysis of Identification of Recursive Causal Models, *Causal Inference in the Social Sciences*, University of Michigan

Hendry, David F. and Hans-Martin Krolzig(2004) We ran one regression. *Oxford Bulletin of Economics and Statistics*, 66(5):799-810

Holland, Paul W., (1986) Statistics and Causal Inference, *Journal of the American Statistical Association*, Vol. 81, No.396., pp. 945–960

Hurwicz, L. (1950) Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Cowles Commission, Monograph 10, Wiley, New York, 245–257.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013) *An Introduction to Statistical Learning with Applications in R*. Springer, NewYork.

Lin, Y. and Y. Jeon. (2006) Random Forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101 (474) pp 578–590.

Morgan, James N. and John A. Sonquist (1963) Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415-434. URL http://www.jstor.org/stable/2283276.

Morgan, Stephen L. and Christopher Winship, (2007) *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge University Press

References

Pearl, J., (2000) *Causality: Models, Reasoning, and Inference*, Cambridge University Press .

Pearl, J., (2009) Causal inference in statistics: An overview, *Statistics Surveys*, Vol. 3, 96-146.

Reiersol, Olav. (1941) Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis, *Econometrica* , Vol. 9, 1-24.

Rubin, Donald (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, 66 (5), pp. 688–701.

Rubin, Donald (1978) Bayesian Inference for Causal Effects: The Role of Randomization, *The Annals of Statistics*, 6, pp. 34–58.

Rubin, Donald (1977) Assignment to Treatment Group on the Basis of a Covariate, *Journal of Educational Statistics*, 2, pp. 1–26.

Shpitser, I. and Pearl, J.,(2006) Identification of Conditional Interventional Distributions. In R. Dechter and T.S. Richardson (Eds.), *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, Corvallis, OR: AUAI Press, 437-444.

Siroky, David S. (2009) Navigating Random Forests and related advances in algorithmic modeling. *Statistics Surveys* 3, pp. 147-163.

Stone, Charles J. (1977) Consistent Nonparametric Regression *Annals of Statistics*, 5, 595-620.

Wu, Xindong and Vipin Kumar, editors. The Top Ten Algorithms in Data Mining. CRC Press, 2009. URL http://www.cs.uvm.edu/~icdm/algorithms/index.shtml