Heterogenous Treatment Effects: Machine Learning Methods

Fan Li

Department of Statistical Science

Duke University

May 31, 2018

HTE with Machine Learning

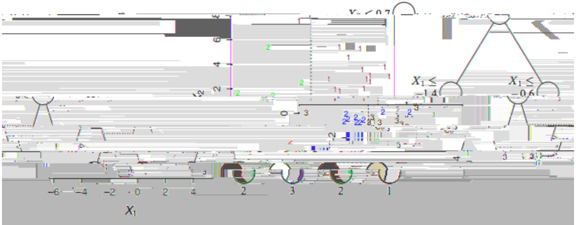
- Two buzzy words in comparative effectiveness research: Heterogenous treatment effects (HTE) and machine learning (ML)
- A very hot trend in causal inference: use ML to infer HTE, particularly in “big data” and high-dimensional cases
- The central goal is the same as traditional regression methods: accurately learn the outcome function given the covariates and treatment variable
- ML methods are usually more flexible and adaptive, but with limitations and certainly no panacea

Popular ML methods for HTE

- Penalized regression (e.g. LASSO, elastic net)
- Regression-tree based methods (e.g. CART, random forests)
-

Regression Trees

- Partition of the covariate space into “leaves” (subgroups)
- Predict responses in each leaf using the sample mean in that region
- Go through variables and leaves and decide whether and where to split leaves
- Select tree complexity using cross-validation
- Modified for HTE by Athey and Imbens (2016, PNAS), extend to “causal forest" by Wager and Athey (2017, JASA)



Regression Trees: Pros and Cons

I

Bayesian Nonparametric Methods

- Bayesian statistics: use the Bayesian theorem to combine the evidence from the previous knowledge (prior distribution) and the data
- Bayesian trees/forests: similar to regression trees but implemented under the Bayesian paradigm
- Gaussian Processes (Rasmussen and Williams, 2006): a very neat stochastic process that extend multivariate Gaussian distributions to infinite dimensional
- Gaussian Processes give much flexibility in model building

Bayesian Nonparametric Methods: Pros and Cons

- Pros: Incorporate prior knowledge, automatic uncertainty quantification, works for small samples, ELEGANT
- Cons: computational scalability, sophisticated for lay audience, choice of prior distribution, software

Applications and Software

- Much recent advance in theory in both statistics and economics, but direct application to health studies still sparse
- More translational work is needed (e.g. Powers et al. 2018)
- Software development is key, as well as effective collaboration between statisticians and clinical researcher
- One must organically fuse traditional statistical tools and ML to reach better comparative effectiveness research (Pencina, 2018)

Further Readings

S Athey and GW Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353-7360, 2016

S Wager and S Athey. (2017). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, forthcoming. `https://arxiv.org/pdf/1510.04342.pdf`

Rasmussen CE, Williams CK. Gaussian process for machine learning. MIT press; 2006.

Li, F, Morgan, LK, and Zaslavsky, AM. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390-400

S Powers, J Qian, K Jung, A Schuler, NH Shah, T Hastie, R Tibshirani. (2018). Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*. `https://doi.org/10.1002/sim.7623`

M Pencina. (2018). Dam Data: Health Systems, Machines, And Learning. *Bio.IT World*, April 9, 2018. `http://www.bio-itworld.com/2018/04/09/dam-data-health-systems-machines`