Machine Learning for Effect Heterogeneity

Cyrus Samii

2019-03-27

Contents

1 Introduction

Below are notes on machine learning approaches to addressing effect heterogeneity. The methods are reviewed in light of the goals in the paper, Gechter et al. (2019, arxiv). It would be worth becoming familiar with that paper before reviewing what I have below.

2 Estimating Conditional Treatment Effects (CATES)

2.1 Overview and Goals

We first consider off-the-shel methods for estimating heterogenous treatment effects. The focus initially will be on point predictions, rather than inference on such predictions. This is because the applications that we use work with summaries of the point predictions rather than the point predictions in and of themselves. See the section below on “features of CATES”, referencing Chernozhukov et al. (2017, arxiv), for more on this point.

We are also interested in methods that work well when we entertain a high-dimensional covariate vector. I say “entertain” because it may be that, in fact, the covariates that predict heterogeneity are few, but this is something that we do not know *a priori,* and rather we have at our disposal many covariates that we want to consider as candidates for predicting effect heterogeneity. This leads us to machine learning approaches that use regularization to balance that ability to make very fine grained predictions with penalties for overfitting.

2.2 Algorithms

We consider the following algorithms:

- BART
- Generalized Random Forests
- Elastic Net on Minimum gCV Basis Expansion

3 Analyzing Features of CATES (Chernozhukov et al, 2017)

Link to the paper: arxiv

3.1 Goals

The goal is to learn about effect heterogeneity in a way that does not overfit and to provide “uniformly valid inference” on conditional average treatment effects (CATEs), or at least *features* of such CATEs. The approach is two-step: (i) build a machine learning (ML) proxy predictor of the CATE, then (ii) do inference on features of this proxy predictor. The features considered are first, best linear predictor (BLP), second, sorted group average effects (GATES), which are ATEs by heterogeneity groups, and third classification analysis (CLAN), which are the average effects of the most and least affected units.

3.2 Setting

Some notation:

- Potential outcomes, $Y(1), Y(0)$.
- Covariates Z.

- Baseline conditional average (BCA): $b_0(Z) = E[Y(0)|Z]$.
- CATE: $s_0(Z) = E[Y(1)|Z] - E[Y(0)|Z]$.
- CIA holds: $D \perp\!\!\!\perp (Y(1), Y(0))|Z$.
- Subvector of stratifying covariates, $Z_1 \subseteq Z$, such that
- propensity score is $p(Z) = P[D = 1|Z_1]$, with $0 < p_0 \leq p(Z) \leq p_1 < 1$.
- Observe $Y = DY(1) + (1 - D)Y(0)$, in which case,
- $Y = b_0(Z) + Ds_0(Z) + U$, with $E[U|Z, D] = 0$, where
- $b_0(Z) = E[Y|D = 0, Z]$ and $s_0(Z) = E[Y|D = 1, Z] - E[Y|D = 0, Z]$. CIA allows us to write b_0 and s_0 in terms of observables.
- Observe N iid draws of (Y, Z, D) with law P. Draws are indexed by $i = 1, ..., N$.

A result to keep in mind (already noted in Athey and Imbens PNAS, I believe):

Horvitz-Thompson scaling Define

$$H = H(D, Z) = \frac{D - p(Z)}{p(Z)(1 - p(Z))}.$$

Given the assumptions above, we have,

$$\begin{aligned}
E[YH|Z] &= E\left[\frac{Y(D - p(Z))}{p(Z)(1 - p(Z))}\bigg| Z\right] \\
&= \frac{1}{p(Z)(1 - p(Z))} E\left[D(D - p(Z))Y(1) + (1 - D)(D - p(Z))Y(0)|Z\right] \\
&= E[Y(1)|Z] - E[Y(0)|Z] \\
&= s_0(Z).
\end{aligned}$$

We can use YH as an "unbiased signal" about $s_0(Z)$. It is, however, a noisy signal, and so in using this, Chernozhukov et al. make adjustments (e.g., for their second BLP method). Nonetheless, it does provide a target that one can use in trying to tune methods for estimating CATEs (see below the section on choosing an ML method).

3.3 Methods

For Chernozhukov et al., directly learning about s_0 in a generic way seems impossible at the moment When P is high dimensional, ML methods will not be consistent. Moreover, inference is complicated for adaptive estimators, particularly in trying to bound biases. Finally, tuning ML models is sort of free for all at the moment, with little to guide.

For these reasons Chernozhukov et al. focus on inference for low dimensional features of s_0. First, partition the indices $\{1, ..., N\}$ into two sets, M and A for "main" and "auxiliary", respectively, to yield a main sample, Data_M, and auxiliary sample, Data_A. Suppose for the time being that the two sets are about the same size. From sample A obtain ML estimates of b_0 and s_0, to be called "proxy predictors":

$$z \mapsto B(z) = B(z; \text{Data}_A) \text{ and } z \mapsto S(z) = S(z; \text{Data}_A).$$

These proxies need not be consistent. Then construct the features of interest (BLP, GATES, CLAN) in sample M. Then do inference in a way that accounts for estimation uncertainty given sample A and splitting uncertainty given the split into M and A. Their precise method is to use many splits and then take the median of the feature estimates and then the medians of the random conditional confidence sets, adjusting them to account for splitting uncertainty. They refer to this inferential approach as "variational estimation and inference" (VEIN).

BLP method 1 involves the following:

Consider the weighted linear projection:

$$Y = \alpha_1 + \alpha_2 B(Z) + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - E[S(Z)]) + \epsilon,$$

with weights,

$$w(Z) = \frac{1}{p(Z)(1 - p(Z))}.$$

Chernozhulov et al. explain that “the interaction $(D - p(Z))(S - E[S])$ is orthogonal to $D - p(Z)$ under the weight $w(Z)$”. This is due to LIE – cf. p. 34 of the paper.

I am going to leave the BLP alone for now. I can see how it may provide for a test of heterogeneity, but not sure how useful it is for our purposes.

Sorted group ATE is quite relevant to us. We want to put units into groups based on their predicted treatment effects. E.g., suppose we want a partition,

$$G - k = \{S \in I_k\}, k = 1, ..., K,$$

where $I_k = [\ell_{k-1}, \ell_k)$ are intervals that divide the support of S. Then, we want the conditional average effects within these bins,

$$E[s_0(Z)|G_k], k = 1, ..., K.$$

given the monotonicity restriction,

$$E[s_0(Z)|G_1] \leq ... \leq E[s_0(Z)|G_K].$$

3.4 Choosing an ML method

Options that they cover: (i) “ability to predict YH using BH and S” or (ii) "ability to predict Y using B and $(D - p(Z))(S - E[S])$ given $w(Z)$.

3.5 Applications

The application that is most relevant for us is the analysis of effect heterogeneity in the Morocco microfinance study. In this case, $p(Z) = 1/2$ for everyone. They try Random Forest, Elastic Net, Boosted Tree, and Neural Network, finding that the first two do best when judged using metrics geared toward the BLP and GATES analyses.

4 Wager and Athey (2018)

Link to the paper: JASA

They start with the same HT scaling result, expressing it in a different (but equivalent) manner (keeping with the notation from above):

$$E\left[Y\left(\frac{D}{p(Z)} - \frac{1 - D}{1 - p(Z)}\right)\middle|Z\right] = s_0(Z).$$

Now I am going to move over to Aaron’s code.

5 Notes

5.1 Low dimensional features versus CATEs

In our applications, we may find ourselves doing something similar to Chernozhulov et al., in that we may have the ability to produce proxies for the target setting, but then have to take a low dimensional summary, insofar as policy makers may not be able to target on the basis of all of the covariate information available. We could adopt their approach to inference, then.

I should say that this feature of what Chernozhukov et al. are doing is very reminiscent of the “targeted learning” work of Van der Laan and coauthors.