Bayesian Additive Regression Tree (BART) with application to controlled trail data analysis

Weilan Yang

wyang@stat.wisc.edu

May. 2015
Background

\[ \text{CATE}_i = \mathbb{E}(Y_i(Z_1) - Y_i(Z_0)|X_i) \]
Background

- Previous work done by Jennifer L. Hill\(^1\) suggested that in Causal Inference, it is more convenient to use flexible non-parametric methods (like BART). The better fit can be obtained without parametric assumption. BART could also perform variable selection and confidence interval construction (from posterior samples).

- Imputation and Extrapolation are the main challenges in causal inference. If BART is able to give a reasonable estimate to the test effect, it should also be able to extrapolate the data well in a controlled trial data. In other words, we could evaluate the prediction performance of BART in the range of non-overlapping data.

BART overview

- Place CART within a Bayesian framework by specifying a prior on tree space.
- Get multiple realizations of 1 tree, average over posterior to form predictions.


BART overview, compared with CART

- Use entropy to split the nodes, which may result to producing similar trees in RF;
- Hard to obtain an interval estimation of a statistic.

![Regression Tree Diagram](image)

*Figure 2. A Regression Tree Where $y \sim N(\theta, 2^2)$ and $x = (x_1, x_2)$.]*
BART overview

A general regression form:

\[ y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]

Regression trees’ form:

\[ y = g(x; M, T) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \]

With Bayesian perspective, we need to specify the joint distribution of parameters\(^3\):

\[
\pi(T, M, \sigma^2) = \pi(M, \sigma^2 | T)\pi(T),
\]

\[ M \sim N(\mu, \Sigma), \quad \sigma^2 \sim IG(\nu/2, \nu\lambda/2) \]

---

\(^3\)Here we assume equal variance, i.e. mean shift model. Alternatively, we could have mean-variance shift model.
BART overview

\[ \pi(T) \] doesn’t have a closed form, needs a process to define:

\[ PSPLIT(\eta, T), \quad \eta : \text{Further split from a leaf node?} \]
\[ PRULE(\rho|\eta, T), \quad \rho : \text{Split by which variable? Which value?} \]

complexity penalty: \[ PSPLIT(\eta, T) = \alpha(1 + d_\eta)^{-\beta} \]
Then, integrate out $\Theta = (M, \sigma^2)$

$$p(Y|X, T) = \int p(Y|X, \Theta, T)p(\Theta|T)d\Theta$$

posterior of the tree:

$$p(T|X, Y) \propto p(Y|X, T)p(T)$$
BART overview

posterior of the tree:

\[ p(T|X, Y) \propto p(Y|X, T)p(T) \]

Cannot enumerate all possible \( p(T) \)'s, so we use Metropolis-Hastings\(^4\) to sample trees:

\[ T^0, T^1, T^2, ... \]

probability of \( T^i \) to \( T^*(T^{i+1} = T^*) \):

\[
\alpha(T^i, T^*) = \min \left\{ \frac{q(T^*, T^i) p(Y|X, T^*)p(T^*)}{q(T^i, T^*) p(Y|X, T^i)p(T^i)}, 1 \right\}
\]

\( q(T^i, T^{i+1}) \equiv p(T^i \rightarrow T^{i+1}) \) produced by four situations: GROW, PRUNE, CHANGE, SWAP

BART overview

Aggregate $m$ trees:

$$y = \sum_{j=1}^{m} g(x; M_j, T_j) + \epsilon$$

Posterior of $p((T_1, M_1), (T_2, M_2), ..., (T_m, M_m), \sigma|y)$ is produced by Gibbs Sampler:

$$\begin{align*}
(T_j, M_j)|T(j), M(j), \sigma^2, y \\
\sigma|T_1, ..., T_m, M_1, ..., M_m, y \sim IG
\end{align*}$$

The former can be simplified as:

$$(T_j, M_j)|R_j, \sigma \quad R_j \equiv y - \sum_{k \neq j} g(x; T_k, M_k)$$
Simulation study\(^5\)

Data generating mechanism:

\[
Y|Z = \beta_0|Z + \sum_{i \in A} \beta_1|Z X_i + \sum_{i \neq j; (i,j) \in B} \beta_2|Z X_j X_k + \sum_{i \neq j \neq k; (i,j,k) \in C} \beta_3|Z X_i X_j X_k + \varepsilon
\]

- Previous work (Hill, 2011) indicates that BART captures non-linear trend;
- When \( \beta_i|Z=0 \neq \beta_i|Z=1 \), \( Z \) and \( X_i \) have interaction, parameters in data generating model doubles, more complex;
- When some \( \beta_i \)'s are set to 0, we could examine its variable selection ability.

\(^5\)https://github.com/williamdottyang/BART_Simulation.git
Simulation study: capturing interaction

Figure: When $X_i$, $Z$ have interaction

- different trends—$Z$ and $X_i$
- unequal variance – among $X_i$’s
Simulation study: capturing interaction

Figure: residual plot

- Conclusion 1: BART captures interaction between $X_i, Z$
Simulation study: capturing interaction

Setting: 3 covariates, with all interactions significant, $\beta_{i|Z=0} \neq \beta_{i|Z=1}$

<table>
<thead>
<tr>
<th></th>
<th>fit on overall data</th>
<th>fit on separate labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1041.30</td>
<td>113.39</td>
</tr>
<tr>
<td>BART</td>
<td>10.63</td>
<td>6.87</td>
</tr>
</tbody>
</table>

Conclusion 2: When $\beta_{i|Z=0} \neq \beta_{i|Z=1}$ (more parameters, complex model), fitting BART on separate labels gives better result.
Simulation study: variable selection

Setting: 30 candidate covariates, with 3 of them significant (|β_i| > 1), 27 not significant (|β_i| = 0.001), 3 2nd order interactions (significant), 1 3rd order interaction (significant).

Table: MSE of OLS and BART on test dataset

<table>
<thead>
<tr>
<th></th>
<th>fit on overall data</th>
<th>fit on separate labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>1037.11</td>
<td>111.42</td>
</tr>
<tr>
<td>BART</td>
<td>30.42</td>
<td>21.76</td>
</tr>
</tbody>
</table>

- **Conclusion 3:** BART has strong ability of variable selection.
Simulation study: extrapolation

![Graph showing true response curves, BART fit, and least squares fit.](image)
Simulation study: extrapolation

Conclusion 4: With only one covariate, BART cannot extrapolate well, trend will follow the existing ones if fitted together, and constant prediction is made if fitted separately.
Simulation study: extrapolation

Figure: real data
Simulation study: extrapolation

Figure: fitting overall data

- Conclusion 5: Multiple covariates with interaction will help BART in extrapolation.
Simulation study: extrapolation

Figure: residual plot

- However, the system error is unavoidable.