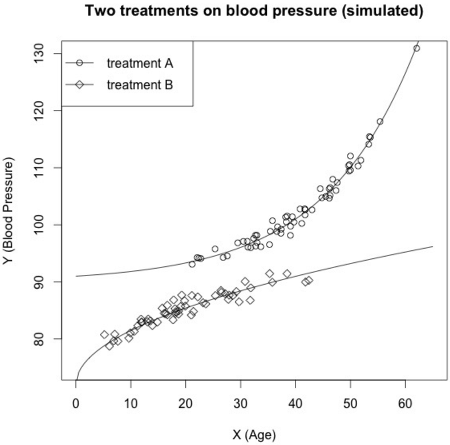
Bayesian Additive Regression Tree (BART) with application to controlled trail data analysis

Weilan Yang

wyang@stat.wisc.edu

May. 2015

Background



$$CATE_i = E(Y_i(Z_1) - Y_i(Z_0)|X_i)$$

Background

- Previous work done by Jennifer L. Hill[1] suggested that in Causal Inference, it is more convenient to user flexible non-parametric method(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 trail data. In other words, we could evaluate the prediction performance of BART in the range of non-overlapping data.

[1] Hill,J.L.(2011).Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1).

BART overview[2]

- Place CART within a Bayesian framework by specifying a prior on tree space.
- Can get multiple tree realizations by using tree-changing proposal distribution: birth/death/change/swap.
- Get multiple realizations of 1 tree, average over posterior to form predictions.

[2] http://www.stat.osu.edu/~comp_exp/jour.club/trees.pdf

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. Journal of the American Statistical Association, 93(443), 935-948.
Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics, 266-298.

BART overview, compared with CART

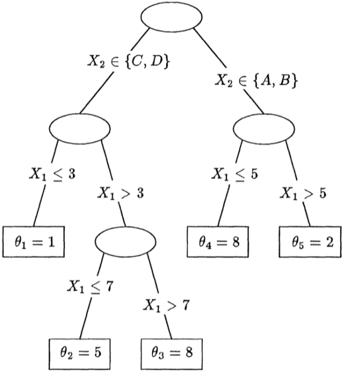


Figure 2. A Regression Tree Where $y \sim N(\theta, 2^2)$ and $\mathbf{x} = (x_1, x_2)$.

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

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] :

$$\pi(T, M, \sigma^2) = \pi(M, \sigma^2 | T)\pi(T),$$
$$M \sim N(\mu, \Sigma), \quad \sigma^2 \sim IG(v/2, v\lambda/2)$$

[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:

$P_{SPLIT}(\eta, T)$, $\quad \eta$: Further split from a leaf node?

$P_{RULE}(\rho|\eta, T)$, $\quad \rho$: Split by which variable? Which value?

complexity penalty: $P_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta}$

BART overview

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 user Metropolis-Hastings[4] to sample trees:

$$T^0, T^1, T^2, ...$$

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

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

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

[4] `http://nitro.biosci.arizona.edu/courses/EEB519A-2007/pdfs/Gibbs.pdf`

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:

$$(T_j, M_j)| T_{(j)}, M_{(j)}, \sigma^2, y$$
$$\sigma| T_1, ..., T_m, M_1, ..., M_m, y \sim IG$$

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:

$$Y|Z = \beta_{0|Z} + \sum_{\substack{i \\ i\in\mathcal{A}}} \beta_{1|Z} X_i + \sum_{\substack{i\neq j \\ (i,j)\in\mathcal{B}}} \beta_{2|Z} X_j X_k + \sum_{\substack{i\neq j\neq k \\ (i,j,k)\in\mathcal{C}}} \beta_{3|Z} X_i X_j X_k + \varepsilon$$

- 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 β_i's are set to 0, we could examine its variable selection ability.

[5] https://github.com/williamdotyang/BART_Simulation.git

Simulation study: capturing interaction

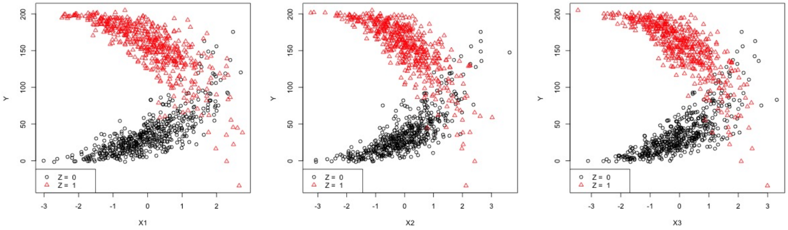
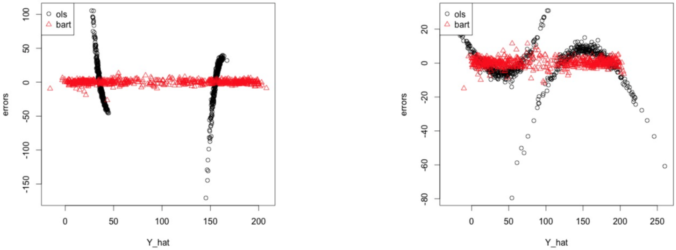


Figure: When X_i, Z have interaction)

- different trends–Z and X_i
- unequal variance – among X_i's

Simulation study: capturing interaction



(a) 整体数据拟合 (b) 分不同处理标签拟合

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: MSE of OLS and BART on test dataset

	fit on overall data	fit on separate labels
OLS	1041.30	113.39
BART	10.63	6.87

- 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

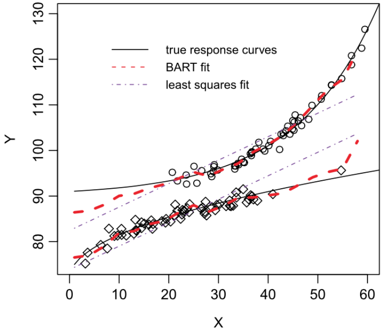
Setting: 30 candidate covariates, with 3 of them significant($|\beta_i| > 1$), 27 not significant($|\beta_i| = 0.001$), 3 2nd order interactions(significant), 1 3rd order interaction(significant).

Table: MSE of OLS and BART on test dataset

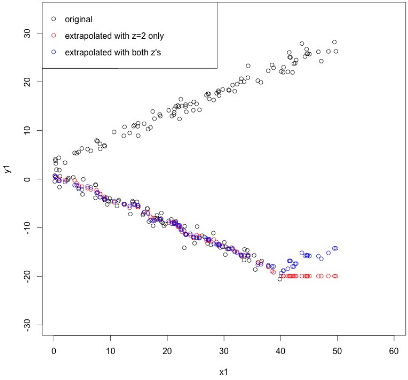
	fit on overall data	fit on separate labels
OLS	1037.11	111.42
BART	30.42	21.76

- Conclusion 3: BART has strong ability of variable selection.

Simulation study: extrapolation



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

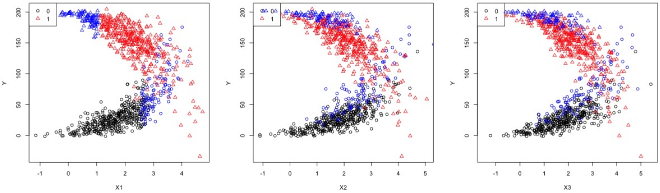


Figure: real data

Simulation study: extrapolation

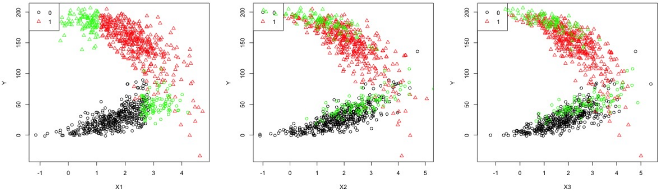


Figure: fitting overall data

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

Simulation study: extrapolation

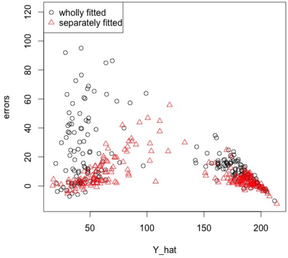


Figure: residual plot

- However, the system error is unavoidable.