Bayesian additive regression trees and the general BART model - Tan and Roy, 2019

Review for BART for binary outcomes and general BART

이재용, 2023/02/02

Contents

BART for binary outcomes

Probit model

Let $Y* = X'\beta + \epsilon$: an auxiliary variable, where $\epsilon \sim N(0,1)$. Then

$$Y = \begin{cases} 1, & Y* > 0 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & X'\beta + \epsilon > 0 \\ 0, & \text{otherwise} \end{cases}$$

Therefore,

$$P(Y=1 \mid X) = P(X'\beta + \varepsilon > 0)$$
$$= P(\varepsilon > -X'\beta)$$
$$= P(\varepsilon < X'\beta) \quad (\because \varepsilon \sim N(0,1),\ \text{Normal dstn is symmetric})$$
$$= \Phi(X'\beta)$$

cdf of standard normal dstn

For $\Phi(X'\beta) > 0.5 \Rightarrow$ prediction : 1

" $<$ " $\Rightarrow$ " 0

BART for binary outcomes

Probit model and BART

In BART's case, $Y^* = f(x) + \epsilon$: an auxiliary variable, where $\epsilon \sim N(0,1)$, $(\because \sigma = 1)$. Then

$P(Y = 1 \mid X) = \Phi(f(x))$, where $\Phi()$ is the CDF function of the standard normal dstn.

Moreover, since $\sigma = 1$, we only need priors for $(T_1, M_1), \ldots, (T_m, M_m)$ and they can be decomposed as we did in the continuous outcomes but without σ. Also, we can use similar prior specification for $\mu_{ji} \mid T_j$ and T_j as the continuous outcomes.

To estimate the posterior distribution, data augmentation is used.

We assume that $Y = I(Z > 0)$, where Z is a latent variable that is drawn as follows:

$$\begin{aligned} z &\sim N_{(-\infty,0)}[f(x), 1], \quad \text{if } \ y = 0, \\ z &\sim N_{(0,\infty)}[f(x), 1], \quad \text{if } \ y = 1, \end{aligned}$$

where $N_{(a,b)}$ is a truncated normal dstn with mean $f(x)$ and variance 1. We can treat Z as the continuous outcome for a BART model and $Z = f(x) + \epsilon$, where $\epsilon \sim N(0,1)$ and use the same MCMC procedure as continuous outcomes.

General BART model

Formal definition for continuous outcomes

Suppose we have a continuous outcome y and $x = \left\{x_1, \ldots, x_p\right\}$. Suppose we also have $w = \left\{w_1, \ldots, w_p\right\}$, such that no two columns in x and w are the same (meaning $x \neq w$?). Then we have

$$y = f(x) + h(w, \Theta) + \epsilon,$$

where $h(\cdot)$ is a function that works on w using parameters Θ, and $\epsilon \sim G(\Sigma)$ where $G(\cdot)$ can be any distribution with parameter Σ.

Assuming $\left\{(T_1, M_1), \ldots, (T_m, M_m)\right\}$, Θ, and Σ are independent, the prior dstn for y is,

$$P\left[(T_1, M_1), \ldots, (T_m, M_m)\right] P(\Theta) P(\Sigma) = \prod_{j=1}^{m} \left\{ \prod_{i=1}^{b_j} P(\mu_{ji} \mid T_j) \right\} P(T_j) P(\Theta) P(\Sigma).$$

Thus there are 4 priors ($\mu_{ji} \mid T_j$, T_j, Θ, and Σ) needed. We can also prior jointly as $P(\Theta, \Sigma)$.

General BART model

Posterior distribution

To obtain the posterior distribution of $P\left[(T_1, M_1), \ldots, (T_m, M_m), \Theta, \Sigma \,|\, y\right]$, we use gibbs sampling.

- For $P\left[(T_1, M_1), \ldots, (T_m, M_m) \,|\, \Theta, \Sigma, y\right]$

This can be seen as drawing from the following model,

$\tilde{y} = f(x) + \epsilon$, where $\tilde{y} = y - h(w, \Theta)$.

This is just a BART model with a modified outcome $\tilde{y}$. Hence, the BART algorithm we saw previously can be used.

General BART model

Posterior distribution

- For $P\left[\Theta \,|\, (T_1, M_1), \ldots, (T_m, M_m), \Sigma, y\right]$

This can be seen as drawing from the following model,

$y' = h(w, \Theta) + \epsilon$, where $y' = y - f(x)$.

This posterior draw depends on the function $h(\,\cdot\,)$ being used and the prior for Θ (the specifics are not discussed in Tan and Roy, 2019).

- For $P\left[\Sigma \,|\, \Theta, (T_1, M_1), \ldots, (T_m, M_m), y\right]$

This can be seen as drawing from the following model,

$r = y - f(x) - h(w, \Theta) = \epsilon$.

The default is usually $\epsilon \sim N(0,\sigma^2)$ $(\because \Sigma = \sigma^2)$, where $\Sigma = \sigma^2 \sim IG(\frac{v}{2}, \frac{v\lambda}{2})$

General BART model

Formal definition for binary outcomes

For binary outcomes, we use probit link as before,

$P(y = 1 \mid x) = \Phi(f(x) + h(w, \Theta))$.

Under this framework, we only need priors for $(T_1, M_1), \ldots, (T_m, M_m)$ and Θ.

We assume that $y = I(Z > 0)$, where Z is drawn as follows,

$$
\begin{aligned}
z &\sim N_{(-\infty,0)}(f(x) + h(w, \Theta), 1), \quad \text{if } y = 0, \\
z &\sim N_{(0,\infty)}(f(x) + h(w, \Theta), 1), \quad \text{if } y = 1,
\end{aligned}
$$

We can treat Z as the continuous outcome for the general BART model with

$Z = f(x) + h(w, \Theta) + \epsilon$, where $\epsilon \sim N(0,1)$.

Semiparametric BART model

Formal definition

BART is a nonparametric model and it innately loses some interpretability relative to a parametric model. However, fully parametric models rely too heavily on assumptions. Semiparametric BART combines the advantages of both of these models.

With semiparametric BART, we can model nuisance parameters nonparametrically using $f(x)$ while covariates of interest can be modeled with parametric specification using $h(w, \Theta)$ from the general BART model as follows.

$$h(w, \Theta) = \theta_0 + \theta_1 w_1 + \ldots + \theta_q w_q,$$

where $w = \left\{ w_1, \ldots, w_q \right\}, \Theta = \left\{ \theta_0, \theta_1, \ldots, \theta_q \right\}.$

As before, we use similar prior distributions for $\mu_{ji} \mid T_j$, T_j, and Σ, while $\Theta \sim MVN(\beta, \Omega)$.

Posterior estimation follows the same procedure as before using Gibbs sampling.

Future plans

Study more about bartMachine

Topics yet to study in bartMachine package:

variable importance, variable effects, partial dependence, incorporating missing data, variable selection, informed prior information on covariates, interaction effect detection, Classification

Future plans

Study more about general BART

Semiparametric BART (read Zeldow et al., 2019)

Random intercept BART for correlated outcomes

Spatially adjusted BART for a statistical matching problem

Dirichlet process mixture BART

Future plans
Soft BART

BART's shortcoming is that it is non-smooth due to BART prior being stepwise-continuous functions.

To address the lack of smoothness of BART, Linero and Yang (2018) introduced the SoftBart model, with the authors demonstrating both theoretically (through studies of posterior concentration rates) and practically (through the analysis of benchmark datasets) that leveraging smoothness often results in substantially improved prediction on real datasets.

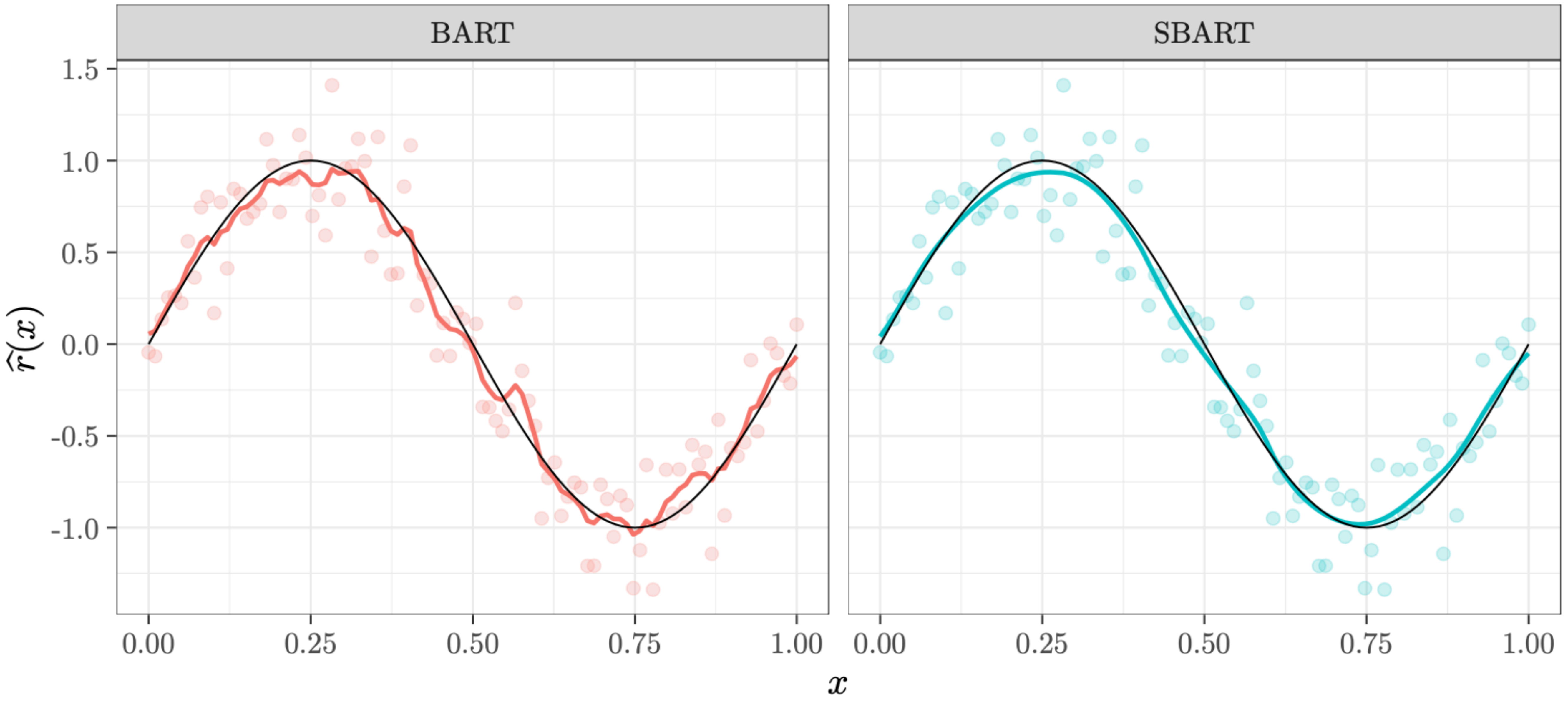

Figure 1: Comparison of the fit of the BART and `SoftBart` models to data generated from the relationship $Y_i = \sin(2\pi x) + \epsilon_i$ with $\epsilon_i \sim \text{Normal}(0, 0.1^2)$. The sine curve is overlayed in black.