# Machine Learning Report

## BART: Bayesian Additive Regression Trees

Dadi Guo Chiyi Wang Jiayi Huang Jiayuan Wu

Peking University

December 24, 2021

# Contribution

| | |
|---|---|
| Dadi Guo[1] | Introduction |
| Chiyi Wang[1] | The BART model, Model training |
| Jiayi Huang[1] | Experiments |

[1] Making PPT

1. Introduction
2. The BART model
3. Model training
4. Experiments

# Introduction

**Recall**

- Bagging and random forests make predictions from an average of regression trees.And each tree is built separately from others.
- Boosting method uses a weighted sum of trees, each of which is constucted by fitting a tree to the residual of the current fit.

# Introduction

BART(Bayesian Additive Regression Trees) is related to both approaches: each tree is constructed in a random manner as in bagging and random forests, and each tree tries to capture signal not yet accounted for by the current model, as in boosting.

## Introduction

### Notations

- $K$: The number of regression trees.
- $B$: The number of iterations .
- $n$: The size of dataset.
- $L$: The number of burn-in iterations.

# Algorithm

1. Let $\hat{f}_1^1(x) = \hat{f}_2^1(x) = \cdots = \hat{f}_K^1(x) = \frac{1}{nK}\sum_{i=1}^n y_i$.
2. Compute $\hat{f}^1(x) = \sum_{k=1}^K \hat{f}_k^1(x) = \frac{1}{n}\sum_{i=1}^n y_i$.
3. For $b = 2, \ldots, B$:
   (a) For $k = 1, 2, \ldots, K$:
      i. For $i = 1, \ldots, n$, compute the current partial residual
      $$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^b(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i).$$
      ii. Fit a new tree, $\hat{f}_k^b(x)$, to $r_i$, by randomly perturbing the $k$th tree from the previous iteration, $\hat{f}_k^{b-1}(x)$. Perturbations that improve the fit are favored.
   (b) Compute $\hat{f}^b(x) = \sum_{k=1}^K \hat{f}_k^b(x)$.
4. Compute the mean after $L$ burn-in samples,
$$\hat{f}(x) = \frac{1}{B-L}\sum_{b=L+1}^{B} \hat{f}^b(x).$$

# Introduction

## Algorithm

1. Let $\hat{f}_1^1(x) = \hat{f}_2^1(x) = \cdots = \hat{f}_K^1(x) = \frac{1}{nK} \sum_{i=1}^n y_i$.
2. Compute $\hat{f}^1(x) = \sum_{k=1}^K \hat{f}_k^1(x) = \frac{1}{n} \sum_{i=1}^n y_i$.
3. For $b = 2, \ldots, B$ :

   (a) For $k = 1, 2, \ldots, K$ :

   i. For $i = 1, \ldots, n$, compute the current partial residual

$$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^b(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i).$$

   ii. Fit a new tree, $\hat{f}_k^b(x)$, to $r_i$, by randomly perturbing the $k$ th tree from the previous iteration, $\hat{f}_k^{b-1}(x)$. Perturbations that improve the fit are favored.

   (b) Compute $\hat{f}^b(x) = \sum_{k=1}^K \hat{f}_k^b(x)$.

## Introduction

### Algorithm

- Compute the mean after $L$ burn-in samples,

$$\hat{f}(x) = \frac{1}{B - L} \sum_{b=L+1}^{B} \hat{f}^b(x).$$

## A sum-of-trees model

BART can be considered a sum-of-trees ensemble, with a novel estimation approach relying on a fully Bayesian probability model.Specifically, the BART model can be expressed as

$$\boldsymbol{y} = f(\boldsymbol{X}) + \mathcal{E} \approx \sum_{t=1}^{m} \mathcal{T}_t^{\mathcal{M}_t}(\boldsymbol{X}) + \mathcal{E}, \quad \mathcal{E} \sim \mathcal{N}_n\left(0, \sigma^2 \boldsymbol{I}_n\right) \tag{1}$$

## A sum-of-trees model

The structure of a given tree $\mathcal{T}_t$ includes information on how any observation recurses down the tree. For each nonterminal (internal) node of the tree, there is a **splitting rule** taking the form $\boldsymbol{x}_j < c$ consisting of the **splitting variable** $\boldsymbol{x}_j$ and the **splitting value** $c$. Given the tree structure, any observation can receive the leaf value of the terminal node. The sum of the $m$ leaf values becomes its predicted value. We denote the set of tree's leaf parameters as $\mathcal{M}_t = \left\{\mu_{t,1}, \mu_{t,2}, \ldots, \mu_{t_{b_t}}\right\}$ where $b_t$ is the number of terminal nodes for a given tree.

# A sum-of-trees model

BART can be distinguished from other ensemble-of-trees models due to its underlying probability model. As a Bayesian model, BART consists of a set of priors for the structure and the leaf parameters and a likelihood for data in the terminal nodes. The aim of the priors is to provide regularization, preventing any single regression tree from dominating the total fit.

# A regularization prior

## Prior independence

$$
\begin{aligned}
&\mathbb{P}\left(\mathcal{T}_1, \mathcal{M}_1, \ldots, \mathcal{T}_m, \mathcal{M}_m, \sigma^2\right) \\
&= \left[\prod_t \mathbb{P}\left(\mathcal{T}_t, \mathcal{M}_t\right)\right] \mathbb{P}\left(\sigma^2\right) \\
&= \left[\prod_t \mathbb{P}\left(\mathcal{M}_t \mid \mathcal{T}_t\right) \mathbb{P}\left(\mathcal{T}_t\right)\right] \mathbb{P}\left(\sigma^2\right) \\
&= \left[\prod_t \prod_\ell \mathbb{P}\left(\mu_{t,\ell} \mid \mathcal{T}_t\right) \mathbb{P}\left(\mathcal{T}_t\right)\right] \mathbb{P}\left(\sigma^2\right)
\end{aligned}
$$

- the tree structure $\mathcal{T}_t$
- the leaf parameters $\mathcal{M}_t$ given $\mathcal{T}_t$
- the error variance $\sigma^2$ which is independent of $\mathcal{T}_t$ and $\mathcal{M}_t$

# A regularization prior

## The $\mathbb{P}(\mathcal{T}_t)$ prior

- The probability that node $\eta$ at depth $d_\eta (= 0, 1, 2, \dots)$ is nonterminal, given by

$$p_{SPLIT}(\eta, \mathcal{T}) = \alpha (1 + d_\eta)^{-\beta} \quad \alpha \in (0, 1) \quad \beta \in [0, +\infty]$$

- The distribution on the **splitting variable** assignments at each interior node. By default, we use the uniform prior.
- The distribution on the **splitting value** assignment in each interior node, conditional on the splitting variable. Again, we use the uniform prior.

# A regularization prior

$$\mathbb{P}(\mathcal{T}) = \prod_{\eta \in H_{\textbf{terminals}}} \left(1 - \mathbb{P}_{\text{SPLIT}}(\eta)\right) \prod_{\eta \in H_{\textbf{internals}}} \mathbb{P}_{\text{SPLIT}}(\eta) \prod_{\eta \in H_{\textbf{internals}}} \mathbb{P}_{\text{RULE}}(\eta) \tag{2}$$

where $H_{\text{terminals}}$ denotes the set of terminal nodes and $H_{\text{internals}}$ denotes the internal nodes. Recall that

$$\mathbb{P}_{\text{SPLIT}}(\eta) = \alpha / (1 + d_\eta)^\beta$$
$$\mathbb{P}_{\text{RULE}}(\eta) = 1/p_{\text{adj}}(\eta) \times 1/n_{j\cdot\ \text{adj}}(\eta).$$

Default values are $\alpha = 0.95$ and $\beta = 2$. With this choice, trees with 1, 2, 3, 4 and $\cancel{\leq}\ \geq$ 5 terminal nodes receive prior probability of 0.05, 0.55, 0.28, 0.09 and 0.03, respectively.

# A regularization prior

## The $\mathbb{P}(\mathcal{M}_t \mid \mathcal{T}_t)$ prior

This parameter is the fitted value assigned to any observation that lands in that node. Note that $\forall t \in \{1, 2, \ldots, m\}$ and $\ell \in \{1, 2, \ldots, b_t\}$, given other parameters, the likelihood of each of the leaf parameters $\mu_{t,\ell}$ is normal distribution with known variance, so we use the conjugate normal distribution as follows:

$$\mu_{t,\ell} \mid \mathcal{T}_t \overset{iid}{\sim} \mathcal{N}\left(\mu_\mu, \sigma_\mu^2\right).$$

## A regularization prior

Since $\mathbb{E}(y \mid x) \sim \mathcal{N}\left(m\mu_\mu, m\sigma_\mu^2\right)$, and the high probability that $\mathbb{E}(y \mid x) \in [y_{min}, y_{max}]$, we can choose $\mu_\mu$ and $\sigma_\mu^2$ so that

$$\begin{cases} m\mu_\mu - 2\sqrt{m}\sigma_\mu = y_{\min} \\ m\mu_\mu + 2\sqrt{m}\sigma_\mu = y_{\max}. \end{cases}$$

To eliminate the influence of outliers, we shift and rescale $\boldsymbol{y}$ so that

$$y_{min} \to -0.5 \quad \text{and} \quad y_{max} \to 0.5,$$

which leads to

$$\mu_{t,\ell} \mid \mathcal{T}_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right), \quad \text{where} \quad \sigma_\mu = \frac{1}{4\sqrt{m}}.$$

# A regularization prior

## The $\sigma^2$ prior

Given other parameters, the full model residuals can be written as

$$\mathcal{E} = \boldsymbol{y} - \sum_{t=1}^{m} \mathcal{T}_t^{\mathcal{M}_t}(\boldsymbol{X}).$$

The likelihood of the residuals is

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{\mathcal{E}^\top\mathcal{E}}{2\sigma^2}\right\}.$$

We use a conjugate prior, the inverse-gamma distribution

$$\sigma^2 \sim \mathrm{InvGamma}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right).$$

# A regularization prior

## The $\sigma^2$ prior

We pick a value of $v = 3$(by default) to get an appropriate shape, and the value of $\lambda$ is determined from the data so that there is a q = 90%(by default) priori chance that the BART model will improve upon the RMSE from an ordinary least squares regression:

$$\mathbb{P}\left(\sigma < \hat{\sigma}_{OLS}\right) = q$$

1. Introduction
2. The BART model
3. Model training
4. Experiments

## Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm can draw samples from any probability distribution with probability density $\pi(x)$, provided that we know a function $f(x) \propto \pi(x)$.

1. initialize $x^{(0)}$
2. **for** $\mathrm{i} = 0$ **to** $N - 1$:

$$
\begin{aligned}
& u \sim U(0,1) \\
& x^* \sim q\left(x^* \mid x^{(i)}\right) \\
& r = \min\left(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)}\right) \\
& \textbf{if } u < r: \\
& \qquad x^{(i+1)} = x^* \\
& \textbf{else}: \\
& \qquad x^{(i+1)} = x^{(i)}
\end{aligned}
$$

- The take-home message here, is that it does not "disgard" samples like rejection sampling. It simply "repeats" samples.

## Gibbs sampling algorithm

- given a starting sample $(x_0, y_0, z_0)^\top$
- you want to sample

$$\left\{(x_1, y_1, z_1)^\top, (x_2, y_2, z_2)^\top, \ldots, (x_N, y_N, z_N)^\top\right\} \sim P(x, y, z)$$

- Then the algorithm goes

$$\begin{aligned}
x_2 &\sim P\left(x \mid y_1, z_1\right) \\
y_2 &\sim P\left(y \mid x_2, z_1\right) \\
z_2 &\sim P\left(z \mid x_2, y_2\right) \\
\\
x_3 &\sim P\left(x \mid y_2, z_2\right) \\
y_3 &\sim P\left(y \mid x_3, z_2\right) \\
z_3 &\sim P\left(z \mid x_3, y_3\right) \\
&\cdots
\end{aligned}$$

## The Gibbs sampler for BART

A Gibbs sampler is employed to generate draws from the posterior distribution of

$$\mathbb{P}\left(\mathcal{T}_1, \mathcal{M}_1, \mathcal{T}_2, \mathcal{M}_2, \ldots, \mathcal{T}_m, \mathcal{M}_m, \sigma^2 \mid \boldsymbol{y}\right)$$

A key feature of the Gibbs sampler for BART is to employ a form of "Bayesian backfitting", where the $j^{th}$ tree is fit iteratively, holding all other $m-1$ trees constant by exposing only the residual response that remains unfitted:

$$\boldsymbol{R}_{-j} := \boldsymbol{y} - \sum_{\substack{1 \le t \le m \\ t \ne j}} \mathcal{T}_t^{\mathcal{M}_t}(\boldsymbol{X})$$

# The Gibbs sampler for BART

The Gibbs sampler works through the following $2m+1$ steps:

$$
\begin{aligned}
1:&\quad \mathcal{T}_1 \mid \boldsymbol{R}_{-1}, \sigma^2\\
2:&\quad \mathcal{M}_1 \mid \mathcal{T}_1, \boldsymbol{R}_{-1}, \sigma^2\\
3:&\quad \mathcal{T}_2 \mid \boldsymbol{R}_{-2}, \sigma^2\\
4:&\quad \mathcal{M}_2 \mid \mathcal{T}_2, \boldsymbol{R}_{-2}, \sigma^2\\
&\quad \vdots\\
2m-1:&\quad \mathcal{T}_m \mid \boldsymbol{R}_{-m}, \sigma^2\\
2m:&\quad \mathcal{M}_m \mid \mathcal{T}_m, \boldsymbol{R}_{-m}, \sigma^2\\
2m+1:&\quad \sigma^2 \mid \mathcal{T}_1, \mathcal{M}_1, \ldots, \mathcal{T}_m, \mathcal{M}_m, \mathcal{E}
\end{aligned}
$$

# The Gibbs sampler for BART

step $1, 3, \ldots, 2m-1$

Sampling from the posterior of the tree structure **does not** depend on the leaf parameters, as they can be analytically integrated out of the computation.These steps rely on Metropolis-Hastings draws from the posterior of the tree distributions. These involve introducing small perturbations to the tree structure:

- **GROW**: growing a terminal node by adding two child nodes
- **PRUNE**: pruning two child nodes (rendering their parent node terminal)
- **CHANGE**: changing a split rule (variable & value)

Probabilities of the GROW / PRUNE / CHANGE steps is 28% / 28% /44% by default.

## The Gibbs sampler for BART

### step $1, 3, \ldots, 2m-1$

To calculate the Metropolis ratio, where the parameter sampled is the tree and the data is the responses unexplained by other trees denoted by R,

$$r = \frac{\mathbb{P}\left(\mathcal{T} \mid \mathcal{T}_*\right)}{\mathbb{P}\left(\mathcal{T}_* \mid \mathcal{T}\right)} \frac{\mathbb{P}\left(\mathcal{T}_* \mid \boldsymbol{R}, \sigma^2\right)}{\mathbb{P}\left(\mathcal{T} \mid \boldsymbol{R}, \sigma^2\right)}, \tag{3}$$

we employ Bayes' Rule,

$$\mathbb{P}\left(\mathcal{T} \mid \boldsymbol{R}, \sigma^2\right) = \frac{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}, \sigma^2\right) \mathbb{P}\left(\mathcal{T} \mid \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \mid \sigma^2\right)}. \tag{4}$$

## The Gibbs sampler for BART

### step $1, 3, \ldots, 2m-1$

We plug 4 into Equation 3 to obtain:

$$r = \underbrace{\frac{\mathbb{P}\left(\mathcal{T}_* \to \mathcal{T}\right)}{\mathbb{P}\left(\mathcal{T} \to \mathcal{T}_*\right)}}_{\text{transition ratio}} \times \underbrace{\frac{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}_*, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}, \sigma^2\right)}}_{\text{likelihood ratio}} \times \underbrace{\frac{\mathbb{P}\left(\mathcal{T}_*\right)}{\mathbb{P}(\mathcal{T})}}_{\text{tree structure ratio}} . \tag{5}$$

Note that the probability of the tree structure is independent of $\sigma^2$.Now we are going to explicitly calculate r for all possible tree proposals — **GROW**, **PRUNE** and **CHANGE**.

## The Gibbs sampler for BART

step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

### Case 1: GROW proposal

**Transition ratio:** Transitioning from the original tree to a new tree involves growing two child nodes from a current terminal node:

$$
\begin{aligned}
\mathbb{P}\left(\mathcal{T} \to \mathcal{T}_*\right) =& \mathbb{P}(\text{GROW})\mathbb{P}(\text{selecting } \eta \text{ to grow from})\times \\
& \mathbb{P}(\text{selecting the } j^{th} \text{ attribute to split on})\times \\
& \mathbb{P}(\text{selecting the } i^{th} \text{ value to split on}) \\
=& \mathbb{P}(\text{GROW})\frac{1}{b}\frac{1}{p_{\text{adj}}(\eta)}\frac{1}{n_{j\cdot\text{adj}}(\eta)}.
\end{aligned} \tag{6}
$$

$p_{\text{adj}}(\eta)$ denotes the number of predictors left available to split on. $n_{j\cdot\text{adj}}(\eta)$ denotes the number of unique values left in the $j^{th}$ attribute after adjusting for parents' splits.

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

(Cont'd)Transitioning from the new tree back to the original tree involves pruning that node:

$$\begin{aligned}\mathbb{P}(\mathcal{T}_* \to \mathcal{T}) &= \mathbb{P}(\mathrm{PRUNE})\mathbb{P}(\text{selecting } \eta \text{ to prune from}) \\ &= \mathbb{P}(\mathrm{PRUNE})\frac{1}{w_2^*}\end{aligned} \tag{7}$$

where $w_2^*$ denotes the number of second generation internal nodes (nodes with two terminal child nodes) in the new tree. Thus, the full **transition ratio** is:

$$\frac{\mathbb{P}(\mathcal{T}_* \to \mathcal{T})}{\mathbb{P}(\mathcal{T} \to \mathcal{T}_*)} = \frac{\mathbb{P}(\mathrm{PRUNE})}{\mathbb{P}(\mathrm{GROW})} \frac{b \cdot p_{\mathrm{adj}}(\eta) \cdot n_{j \cdot \mathrm{adj}}(\eta)}{w_2^*} \tag{8}$$

Note that when $p_{\mathrm{adj}}(\eta) = 0$, the step will be automatically rejected.

## The Gibbs sampler for BART

### step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

**Likelihood ratio:** To calculate the likelihood, the tree structure determines which responses fall into which of the $b$ terminal nodes:

$$\mathbb{P}\left(R_1, \ldots, R_n \mid \mathcal{T}, \sigma^2\right) = \prod_{\ell=1}^{b} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \sigma^2\right), \qquad (9)$$

where each term on the right hand side is the probability of responses in one of the $b$ terminal nodes, which are independent by assumption. The $R_\ell$ 's denote the data in the $\ell$ th terminal node and where $n_\ell$ denotes how many observations are in each terminal node and $n = \sum_{\ell=1}^{b} n_\ell$. Remember, if the mean in each terminal node, which we denote $\mu_\ell$, was known, then we would have

$$R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \mu_\ell, \sigma^2 \overset{iid}{\sim} \mathcal{N}\left(\mu_\ell, \sigma^2\right). \qquad (10)$$

## The Gibbs sampler for BART

### step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

Recall that one of the BART model assumptions is a prior on the average value of $\mu_\ell \sim \mathcal{N}\left(0, \sigma_\mu^2\right)$ and thus,

$$\mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \sigma^2\right) = \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \mu_\ell, \sigma^2\right) \mathbb{P}\left(\mu_\ell; \sigma_\mu^2\right) d\mu_\ell \tag{11}$$

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

$$
\begin{aligned}
&\mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \sigma^2\right) \\
=& \int_{\mathbb{R}} \mathbb{P}\left(R_{\ell_1}, \ldots, R_{\ell_{n_\ell}} \mid \mu_\ell, \sigma^2\right) \mathbb{P}\left(\mu_\ell; \sigma_\mu^2\right) d\mu_\ell \\
=& \int_{\mathbb{R}} \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_\ell}{2}} \left(\frac{1}{2\pi\sigma_\mu^2}\right)^{\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\ell}\left(R_{\ell_i}-\mu_\ell\right)^2 - \frac{1}{2\sigma_\mu^2}\mu_\ell^2\right\} d\mu_\ell \\
=& \frac{1}{(2\pi\sigma^2)^{n_\ell/2}} \sqrt{\frac{\sigma^2}{\sigma^2 + n_\ell\sigma_\mu^2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_\ell} R_{\ell_i}^2 - \frac{\bar{R}_\ell^2 n_\ell^2}{n_\ell + \frac{\sigma^2}{\sigma_\mu^2}}\right)\right\}
\end{aligned}
\tag{12}
$$

where $\bar{R}_\ell = \sum_{i=1}^{n_\ell} R_{\ell_i}/n_\ell$.

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

Note that the likelihoods are solely determined by the terminal nodes.

- select leaf node $\ell$ from tree $\mathcal{T}$
- $\ell \overset{SPLIT}{\rightarrow} \{\ell_L, \ell_R\}$

$$\frac{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}_*, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}, \sigma^2\right)} = \frac{\mathbb{P}\left(R_{\ell_{L,\mathbf{1}}}, \ldots, R_{\ell_{L,n_{\ell,L}}} \mid \sigma^2\right) \mathbb{P}\left(R_{\ell_{R,\mathbf{1}}}, \ldots, R_{R,\ell_{n_{\ell,R}}} \mid \sigma^2\right)}{\mathbb{P}\left(R_{\ell_{\mathbf{1}}}, \ldots, R_{\ell_{n_\ell}} \mid \sigma^2\right)} \tag{13}$$

Plugging Equation 12 into Equation 13 three times yields:

$$\sqrt{\frac{\sigma^2\left(\sigma^2 + n_\ell \sigma_\mu^2\right)}{\left(\sigma^2 + n_{\ell_L}\sigma_\mu^2\right)\left(\sigma^2 + n_{\ell_R}\sigma_\mu^2\right)}} \times \exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{\left(\sum_{i=1}^{n_{\ell_L}} R_{\ell_L,i}\right)^2}{\sigma^2 + n_{\ell_L}\sigma_\mu^2} + \frac{\left(\sum_{i=1}^{n_{\ell_R}} R_{\ell_R,i}\right)^2}{\sigma^2 + n_{\ell_R}\sigma_\mu^2} - \frac{\left(\sum_{i=1}^{n_\ell} R_{\ell,i}\right)^2}{\sigma^2 + n_\ell \sigma_\mu^2}\right)\right) \tag{14}$$

where $n_{\ell_L}$ and $n_{\ell_R}$ denote the number of data points in $\ell_L$ and $\ell_R$.

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

**Tree structure ratio:** For the entire tree,

$$\mathbb{P}(\mathcal{T}) = \prod_{\eta \in H_{\textbf{terminals}}} \left(1 - \mathbb{P}_{\text{SPLIT}}(\eta)\right) \prod_{\eta \in H_{\textbf{internals}}} \mathbb{P}_{\text{SPLIT}}(\eta) \prod_{\eta \in H_{\textbf{internals}}} \mathbb{P}_{\text{RULE}}(\eta) \tag{15}$$

where $H_{\text{terminals}}$ denotes the set of terminal nodes and $H_{\text{internals}}$ denotes the internal nodes. Recall that

$$\mathbb{P}_{\text{SPLIT}}(\eta) = \alpha / \left(1 + d_{\eta}\right)^{\beta}$$
$$\mathbb{P}_{\text{RULE}}(\eta) = 1/p_{\text{adj}}(\eta) \times 1/n_{j\cdot\ \text{adj}}(\eta).$$

## The Gibbs sampler for BART

### step $1, 3, \ldots, 2m-1$ Case 1: GROW proposal

We can now form the ratio:

$$\frac{\mathbb{P}(\mathcal{T}_*)}{\mathbb{P}(\mathcal{T})} = \frac{(1-\mathbb{P}_{\text{SPLIT}}\ (\eta_L))(1-\mathbb{P}_{\text{SPLIT}}\ (\eta_R))\,\mathbb{P}_{\text{SPLIT}}\ (\eta)\mathbb{P}_{\text{RULE}}\ (\eta)}{(1-\mathbb{P}_{\text{SPLIT}}\ (\eta))}$$

$$= \frac{\left(1-\frac{\alpha}{\left(1+d_{\eta_L}\right)^\beta}\right)\left(1-\frac{\alpha}{\left(1+d_{\eta_R}\right)^\beta}\right)\frac{\alpha}{(1+d_\eta)^\beta}\frac{1}{p_{\text{adj}}\ (\eta)}\frac{1}{n_{j\cdot\ \text{adj}}\ (\eta)}}{1-\frac{\alpha}{(1+d_\eta)^\beta}}$$

$$= \alpha\frac{\left(1-\frac{\alpha}{(2+d_\eta)^\beta}\right)^2}{\left((1+d_\eta)^\beta-\alpha\right)p_{\text{adj}}\ (\eta)n_{j\cdot\text{adj}}(\eta)}$$

note the fact that $d_{\eta_L} = d_{\eta_R} = d_\eta + 1$.

## The Gibbs sampler for BART

step $1, 3, \ldots, 2m-1$ Case 2: Prune proposal

### Case 2: Prune proposal

**Transition ratio:** A prune proposal is the **opposite** of a grow proposal. Prune selects a second generation internal node and removes both of its children.

$$\frac{\mathbb{P}\left(\mathcal{T}_* \to \mathcal{T}\right)}{\mathbb{P}\left(\mathcal{T} \to \mathcal{T}_*\right)} = \frac{\mathbb{P}(\mathrm{GROW}) \frac{1}{b-1} \frac{1}{p_{\mathrm{adj}}(\eta^*)} \frac{1}{n_{j^* \cdot \mathrm{adj}}(\eta^*)}}{\mathbb{P}(\mathrm{PRUNE}) \frac{1}{w_2}}$$

$$= \frac{\mathbb{P}(\mathrm{GROW})}{\mathbb{P}(\mathrm{PRUNE})} \frac{w_2}{(b-1) p_{\mathrm{adj}}\left(\eta^*\right) n_{j^* \cdot \mathrm{adj}}\left(\eta^*\right)}$$

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 2: Prune proposal

**Likelihood ratio:**

$$\frac{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}_*, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}, \sigma^2\right)} = \sqrt{\frac{\left(\sigma^2 + n_{\ell_L}\sigma_\mu^2\right)\left(\sigma^2 + n_{\ell_R}\sigma_\mu^2\right)}{\sigma^2\left(\sigma^2 + n_\ell \sigma_\mu^2\right)}} \times$$

$$\exp\left(\frac{\sigma_\mu^2}{2\sigma^2}\left(\frac{\left(\sum_{i=1}^{n_\ell} R_{\ell,i}\right)^2}{\sigma^2 + n_\ell \sigma_\mu^2} - \frac{\left(\sum_{i=1}^{n_{\ell_L}} R_{\ell_L,i}\right)^2}{\sigma^2 + n_{\ell_L}\sigma_\mu^2} - \frac{\left(\sum_{i=1}^{n_{\ell_R}} R_{\ell_R,i}\right)^2}{\sigma^2 + n_{\ell_R}\sigma_\mu^2}\right)\right)$$

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 2: Prune proposal

**Tree structure ratio:**

$$\frac{\mathbb{P}\left(\mathcal{T}_*\right)}{\mathbb{P}(\mathcal{T})} = \frac{\left(\left(1+d_\eta\right)^\beta - \alpha\right) p_{\text{adj}}\ \left(\eta^*\right) n_{j^* \cdot \text{adj}}\left(\eta^*\right)}{\alpha\left(1-\frac{\alpha}{\left(2+d_\eta\right)^\beta}\right)^2}$$

## The Gibbs sampler for BART

step $1, 3, \ldots, 2m-1$ Case 3: Change proposal

### Case 3: Change proposal

**Transition ratio:**

$$\begin{aligned}\mathbb{P}\left(\mathcal{T} \to \mathcal{T}_*\right) = & \mathbb{P}(\text{ CHANGE })\mathbb{P}(\text{ selecting node } \eta \text{ to change })\times \\ & \mathbb{P}(\text{ selecting the new attribute to split on })\times \\ & \mathbb{P}(\text{ selecting the new value to split on })\end{aligned}$$

$$\Rightarrow \frac{\mathbb{P}\left(\mathcal{T}_* \to \mathcal{T}\right)}{\mathbb{P}\left(\mathcal{T} \to \mathcal{T}_*\right)} = \frac{n_{j^* \cdot \mathrm{adj}}\left(\eta^*\right)}{n_{j \cdot \mathrm{adj}}(\eta)}$$

# The Gibbs sampler for BART

## step $1, 3, \ldots, 2m-1$ Case 3: Change proposal

### Likelihood ratio:

$$
\begin{aligned}
&\frac{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}_*, \sigma^2\right)}{\mathbb{P}\left(\boldsymbol{R} \mid \mathcal{T}, \sigma^2\right)} \\
=&\frac{\mathbb{P}\left(R_{1^*,1}, \ldots, R_{1^*,n_{1^*}} \mid \sigma^2\right) \mathbb{P}\left(R_{2^*,1}, \ldots, R_{2^*,n_{2^*}} \mid \sigma^2\right)}{\mathbb{P}\left(R_{1,1}, \ldots, R_{1,n_1} \mid \sigma^2\right) \mathbb{P}\left(R_{2,1}, \ldots, R_{2,n_2} \mid \sigma^2\right)} \\
=&\sqrt{\frac{\left(\frac{\sigma^2}{\sigma_\mu^2}+n_1\right)\left(\frac{\sigma^2}{\sigma_\mu^2}+n_2\right)}{\left(\frac{\sigma^2}{\sigma_\mu^2}+n_1^*\right)\left(\frac{\sigma^2}{\sigma_\mu^2}+n_2^*\right)}} \times \\
&\exp\left(\frac{1}{2\sigma^2}\left(\frac{\left(\sum_{i=1}^{n_1*} R_{1^*,i}\right)^2}{n_{1^*}+\frac{\sigma^2}{\sigma_\mu^2}}+\frac{\left(\sum_{i=1}^{n_2*} R_{2^*,i}\right)^2}{n_{2^*}+\frac{\sigma^2}{\sigma_\mu^2}}-\frac{\left(\sum_{i=1}^{n_1} R_{1,i}\right)^2}{n_1+\frac{\sigma^2}{\sigma_\mu^2}}-\frac{\left(\sum_{i=1}^{n_2} R_{2,i}\right)^2}{n_2+\frac{\sigma^2}{\sigma_\mu^2}}\right)\right)
\end{aligned}
$$

## The Gibbs sampler for BART

step $1, 3, \ldots, 2m-1$ Case 3: Change proposal

**Tree structure ratio:**

$$\frac{\mathbb{P}\left(\mathcal{T}_*\right)}{\mathbb{P}(\mathcal{T})} = \frac{\left(1-\mathbb{P}_{\text{SPLIT}}\left(\eta_{1^*}\right)\right)\left(1-\mathbb{P}_{\text{SPLIT}}\left(\eta_{2^*}\right)\right)\mathbb{P}_{\text{SPLIT}}\left(\eta_*\right)\mathbb{P}_{\text{RULE}}\left(\eta_*\right)}{\left(1-\mathbb{P}_{\text{SPLIT}}\left(\eta_1\right)\left(1-\mathbb{P}_{\text{SPLIT}}\left(\eta_2\right)\right)\right)\mathbb{P}_{\text{SPLIT}}(\eta)\mathbb{P}_{\text{RULE}}\left(\eta\right)}$$

$$= \frac{\mathbb{P}_{\text{RULE}}\left(\eta_*\right)}{\mathbb{P}_{\text{RULE}}\left(\eta\right)} = \frac{n_{j\cdot\text{adj}}(\eta)}{n_{j^*\cdot\text{adj}}\left(\eta^*\right)}$$

## The Gibbs sampler for BART

### step $2, 4, \ldots, 2m$

Now comes step $2, 4, \cdots 2m$. Within a given terminal node, since both the prior and likelihood are normally distributed, the posterior of each of the leaf parameters in $\mathcal{M}$ is conjugate normal with its mean being a weighted combination of the likelihood and prior parameters.

# The Gibbs sampler for BART

step $2, 4, \ldots, 2m$

Let $\boldsymbol{R}_{-j,\ell} = (R_{-j,\ell,1}, \ldots, R_{-j,\ell,n_\ell})^\top$ be a subset from $\boldsymbol{R}_{-j}$ where $n_\ell$ is the number of $R_{-j,\ell,h}$ 's allocated to the $\ell^{th}$ leaf node of $j^{th}$ tree with parameter $\mu_{j,\ell}$ and $h$ indexes the subjects allocated to the terminal node with parameter $\mu_{j,\ell}$. We note that

$$R_{-j,\ell,h} \mid \mathcal{T}_j, \mu_{j,\ell}, \sigma^2 \overset{iid}{\sim} \mathcal{N}\left(\mu_{j,\ell}, \sigma^2\right)$$

and

$$\mu_{j,\ell} \mid \mathcal{T}_j \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\mu^2\right).$$

# The Gibbs sampler for BART

## step $2, 4, \ldots, 2m$

Then, the posterior distribution of $\mu_{j,\ell}$ is given by

$$
\begin{aligned}
\mathbb{P}\left(\mu_{j,\ell} \mid \mathcal{T}_j, \sigma^2, \boldsymbol{R}_{-j}\right) &\propto \mathbb{P}\left(\boldsymbol{R}_{-j,\ell} \mid \mathcal{T}_j, \mu_{j,\ell}, \sigma^2\right) \mathbb{P}\left(\mu_{j,\ell} \mid \mathcal{T}_j\right) \\
&\propto \exp\left[-\frac{\sum_h \left(R_{-j,\ell,h} - \mu_{j,\ell}\right)^2}{2\sigma^2}\right] \exp\left[-\frac{\mu_{j,\ell}^2}{2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{\left(n_\ell \sigma_\mu^2 + \sigma^2\right) \mu_{j,\ell}^2 - 2\sigma_\mu^2 \sum_h R_{-j,\ell,h} \cdot \mu_{j,\ell}}{2\sigma^2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{\left(\mu_{j,\ell} - \frac{\sigma_\mu^2 \sum_h R_{-j,\ell,h}}{n_\ell \sigma_\mu^2 + \sigma^2}\right)^2}{2\frac{\sigma^2\sigma_\mu^2}{n_\ell \sigma_\mu^2 + \sigma^2}}\right]
\end{aligned}
$$

# The Gibbs sampler for BART

## step $2m + 1$

Finally, due to the normal-inverse-gamma conjugacy, the posterior of $\sigma^2$ is inverse gamma as well.

$$
\begin{aligned}
& \mathbb{P}\left(\sigma^2 \mid (\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \boldsymbol{y}\right) \\
\propto & \mathbb{P}\left(\boldsymbol{y} \mid (\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_m, \mathcal{M}_m), \sigma^2\right) \mathbb{P}\left(\sigma^2\right) \\
\propto & \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{\mathcal{E}^\top \mathcal{E}}{2\sigma^2}\right\} \cdot \left(\sigma^2\right)^{-\left(\frac{v}{2}+1\right)} \exp\left(-\frac{v\lambda}{2\sigma^2}\right) \\
\propto & \left(\sigma^2\right)^{-\left(\frac{v+n}{2}+1\right)} \exp\left[-\frac{v\lambda + \mathcal{E}^\top \mathcal{E}}{2\sigma^2}\right].
\end{aligned}
$$

## Boston Housing Data Set

- from the 1970 US Census
- 506 observations, 12 covariates, 1 response variable
- each observation represents a Census tract in Boston
- predict the median value of owner-occupied homes `y = mdev` from other 12 covariates

## Experimental Result

In the experiment, we fit the following BART model for continuous outcomes:

$$\begin{aligned} y_i = \mu_0 + f(x_i) + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma^2) \\ \mathbb{P}_{\text{prior}}(f, \sigma^2) \sim \text{BART} \end{aligned} \quad (16)$$

with $i$ indexing subjects; $i = 1, \ldots, N$. We use Markov chain Monte Carlo (MCMC) to get draws from the posterior distribution of the parameter $(f, \sigma^2)$.

# Assessing Convergence of BART

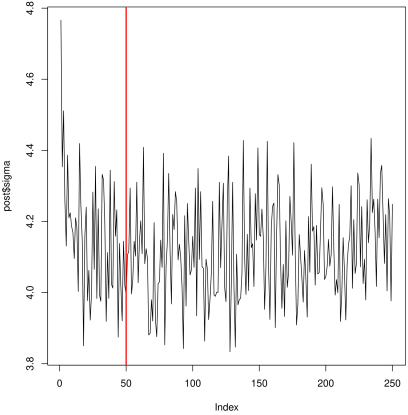


Figure 1: Trace plot of the error variance, $\sigma$, which demonstrates convergence for BART rather quickly, i.e., by 50 iterations or earlier.

# Comparison with Linear Regression

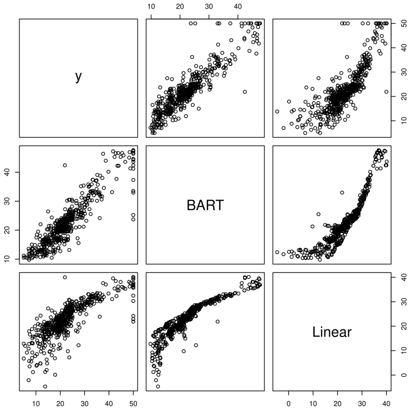


Figure 2: Scatter plots comparing `y = mdev`, the BART fit ("BART") and multiple linear regression ("Linear").

# Prediction Result

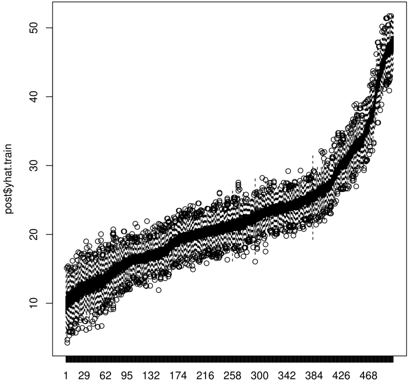


Figure 3: Boxplots of the posterior samples of predictions (on the $y$-axis) ordered by the average predicted home value per tract (on the $x$-axis).

## Comparison with Other Tree-based Methods

| | Random | Adaptive | Number of Trees | MSE |
|---|---|---|---|---|
| Linear | | | - | 27.65 |
| Bagging | $\surd$ | | 500 | 22.98 |
| Random Forest | $\surd$ | | 500 | 19.87 |
| Boosting | | $\surd$ | 5000 | 18.96 |
| BART | $\surd$ | $\surd$ | 200 | **15.29** |

Table 1: Comparison of BART with linear model and other tree-based methods, namely Bagging, Random Forest and Boosting.

# References I

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch.
BART: Bayesian additive regression trees.
*The Annals of Applied Statistics*, 4(1), mar 2010.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
*An Introduction to Statistical Learning*.
Springer US, 2021.

Adam Kapelner and Justin Bleich.
bartMachine: Machine learning with bayesian additive regression trees.
*arXiv preprint arXiv:1312.2171*, 2013.

# References II

Rodney Sparapani, Charles Spanbauer, and Robert McCulloch.
Nonparametric machine learning and efficient computation with bayesian additive regression trees: The BART R package.
*Journal of Statistical Software*, 97(1), 2021.