XBART : Accelerated Bayesian Additive Regression Trees

Reviewer : Seokhun Park

Department of Statistics
Seoul National University

Contents

Outline

1. Review of Bayesian Additive Regression Trees (BART)
2. XBART : Accelerated Bayesian Additive Regression Trees

Outline

1. Review of Bayesian Additive Regression Trees (BART)
   - Preliminaries
   - Prior
   - Posterior sampling
2. XBART : Accelerated Bayesian Additive Regression Trees

Preliminaries

- $\mathbf{x} = (x_1, ..., x_p)^\top \subseteq \mathcal{X}$: p-dimensional input vector.
- Terminal node : a node with no child nodes.
- Let $\mathcal{T}$ be a binary tree structure consists of a set of interior node decision rules and a set of terminal nodes.
- For a given $\mathcal{T}$, let M be a set of height values for terminal nodes.
- Then, we define $g(\mathbf{x} : \mathcal{T}, M)$ as a decision tree for $\mathcal{T}$ and M.

Example

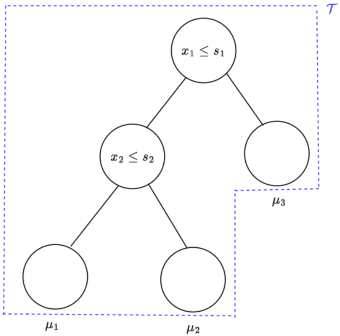

Figure 1: Example of $\mathcal{T}$ and M.

- $\mathcal{T}$ = blue dashed line.
- $M = \{\mu_1, \mu_2, \mu_3\}$.

BART model

- We consider a standard nonparametric regression model given as

$$Y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

- To approximate f, BART assumes that

$$f(\mathbf{x}) = \sum_{t=1}^{T} g(\mathbf{x} : \mathcal{T}_t, M_t). \tag{1}$$

- BART estimates $(\mathcal{T}_1, M_1), ..., (\mathcal{T}_T, M_T)$ using Bayesian approach. (Generating $\{\mathcal{T}_i, M_i\}_{i=1}^{T}$ from a posterior distribution.)

Outline

Prior

- For a given $\mathcal{T}_t$, we have

$$M_t = (\mu_{1t}, ..., \mu_{bt}), \tag{2}$$

where b_t is the number of terminal node in $\mathcal{T}_t$ for $t = 1, ..., T$

- Therefore, we need to set the prior distribution for $\mathcal{T}_t$, $\mu_{kt}|\mathcal{T}_t$ and σ, where b_t is the number of terminal node in $\mathcal{T}_t$.

Prior for $\mu_{kt}|\mathcal{T}_t$ and σ

- For $t = 1, ..., T$ and $k = 1, ..., b_t$, $\mu_{kt}|\mathcal{T}_t \sim N(0, \sigma_\mu^2)$.
- $\sigma^2 \sim IG\left(\frac{v}{2}, \frac{v\lambda}{2}\right)$, where $IG(a, b)$ is the inverse gamma distribution with the shape parameter a and scale parameter b.

Prior for tree structure $\mathcal{T}_t$

- For a given $t \in [T]$, $\pi(\mathcal{T}_t)$ consists of $\pi_{\text{split}}(b)$ and $\pi_{\text{split rule}}(b)$ for each node b in $\mathcal{T}_t$.
- $\pi_{\text{split}}(b) := \alpha(1 + \text{depth}_b)^{-\beta}$
  $\rightarrow$ Probability of splitting at node b.
- $\pi_{\text{split rule}}(b) := \pi(\text{Split variable})\pi(\text{Split value}|\text{Split variable})$
  $\rightarrow$ Probability of (uniformly) selecting rule at node b.

Outline

1. Review of Bayesian Additive Regression Trees (BART)
   - Preliminaries
   - Prior
   - Posterior sampling

2. XBART : Accelerated Bayesian Additive Regression Trees

Posterior sampling

- Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a observed data.
- We generate

$$(\mathcal{T}_1, M_1), \ldots., (\mathcal{T}_T, M_T), \sigma$$

from

$$\pi((\mathcal{T}_1, M_1), \ldots, (\mathcal{T}_T, M_T), \sigma | \mathcal{D})$$

using a Gibbs sampling and MH algorithm.

Gibbs sampling

- For $t \in [T]$, we generate $(\mathcal{T}_t, M_t)$ from

$$\pi(\mathcal{T}_t, M_t | \mathcal{T}_{(-t)}, M_{(-t)}, \sigma, \mathcal{D}), \tag{3}$$

where $\mathcal{T}_{(-t)}$ means the set of all tree structure except $\mathcal{T}_t$.

- Above sampling is equal to

$$(\mathcal{T}_t, M_t) \sim \pi(\mathcal{T}_t, M_t | \text{Resid}_t, \sigma), \tag{4}$$

where $\text{Resid}_t = \{y_i - \sum_{j \neq t} g(\mathbf{x}_i : \mathcal{T}_j, M_j), i = 1, ..., n\}$.

MH algorithm

- Since

$$\pi(\mathcal{T}_t, M_t|\text{Resid}_t, \sigma) = \pi(\mathcal{T}_t|\text{Resid}_t, \sigma)\pi(M_t|\text{Resid}_t, \sigma, \mathcal{T}_t),$$

  we first generate $\mathcal{T}_t$ from $\pi(\mathcal{T}_t|\text{Resid}_t, \sigma)$ and then, generate M_t from $\pi(M_t|\text{Resid}_t, \sigma, \mathcal{T}_t)$.

- $\mathcal{T}_t \sim \pi(\mathcal{T}_t|\text{Resid}_t, \sigma)$ using MH algorithm.
- $M_t \sim \pi(M_t|\text{Resid}_t, \sigma, \mathcal{T}_t)$ using conjugate property.

Proposal distribution for MH algorithm

MH algorithm proposes $\mathcal{T}^{\text{new}}$ using one of GROW, PRUNE, and CHANGE.

- GROW : growing tree by splitting randomly selected terminal node.
- PRUNE : pruning randomly selected node among the nodes with two terminal nodes.
- CHANGE : changing rule in the randomly selected node among the nodes with two terminal nodes.

Outline

Outline

1. Review of Bayesian Additive Regression Trees (BART)
2. XBART : Accelerated Bayesian Additive Regression Trees
   - Introduction
   - Proposed method

Introduction

- In standard BART MCMC, each regression tree is updated using local, random-walk MH proposals (e.g., birth/death moves) that make only minor adjustments to the tree. As a result, the convergence of MCMC may be slow.
- They proposed a method called stochastic hill climbing algorithm to replace the MH algorithm, and experimentally demonstrated that the efficiency of this algorithm.

Outline

Grow-from-root : Generating entirely new tree structure

- When generating $\mathcal{T}_t \sim \pi(\mathcal{T}_t|\text{Resid}_t, \sigma, \mathcal{D})$, they ignore the current tree and grow an **entirely new tree structure from scratch**.
- To generate entirely new tree structure, they sample **Rule** in each terminal node using Bayes rule.
- Let Resid_t be the residual assigned to terminal node b, then **Rule** for terminal node b are generated from

$$\pi(\textbf{Rule}|\text{Resid}_t, \sigma) = \frac{\pi(\text{Resid}_t|\textbf{Rule}, \sigma)\pi(\textbf{Rule})}{\sum_{\textbf{Rule}'} \pi(\text{Resid}_t|\textbf{Rule}', \sigma)\pi(\textbf{Rule}')}.$$

Algorithm for generating entirely new tree

Algorithm 1 GROW-FROM-ROOT(Root node b, Data $\mathcal{D}$, σ)

Input: Root node b, Data $\mathcal{D}$, variance σ

Output: Grown tree starting from root node b

```
Sample Rule using Bayes' rule
if Split == TRUE then
    Create left child node b_L and right child node b_R
    Assign D to b_L and b_R based on Rule
    GROW-FROM-ROOT(b_L, D_L, σ)
    GROW-FROM-ROOT(b_R, D_R, σ)
end if
```

Drawbacks

- The proposed algorithm called stochastic hill climbing algorithm is not a fully bayesian method.
- The method of using the GROW-FROM-ROOT to generate a new tree as the proposal distribution for MH is also mentioned, but there is no experimental proof of its effectiveness.

Thank You

References