Bayesian Additive Regression Trees (BART)

Yaoyuan Vincent Tan and Jason Roy

January 23, 2019

Introduction

Bayesian Additive Regression Trees (BART) is a flexible prediction model/machine learning approach.

- Popularity in various applications: biomarker discovery, causal effect estimation, genomic studies, etc.
- Extensions include survival outcomes, multinomial outcomes, semi-continuous outcomes, and more.

BART Model Overview

BART combines multiple regression trees to model complex, non-linear relationships.

$$Y_i = \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i \tag{1}$$

- Y_i: Outcome variable
- X_i: Predictor variables
- T_j, M_j: Structure and parameters of tree j
- ϵ_i: Error term

Single Regression Tree

A single regression tree partitions the predictor space and fits simple models within each partition.

Figure: Example of a single regression tree.

Sum of Regression Trees

BART models use the sum of regression trees to capture complex interactions.

$$Y_i = \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i \tag{2}$$

- Each tree captures different aspects of the data.
- Combined result is a flexible model that can approximate non-linear functions.

BART Algorithm

The BART algorithm iteratively updates the trees using Markov Chain Monte Carlo (MCMC).

1. Initialize trees to root nodes.
2. Iteratively update tree structures and parameters.
3. Ensure convergence to posterior distribution.

Posterior Performance

BART provides posterior distributions for predictions, allowing for uncertainty quantification.

- Example: Posterior performance evaluated using synthetic data.
- Real-world application: Predicting Standardized Hospitalization Ratio.

Extensions of BART

- Semiparametric BART: Combines parametric and nonparametric components.
- Random intercept BART: Models correlated outcomes.
- Spatial BART: Addresses statistical matching problems.
- Dirichlet Process Mixture BART: Enhances robustness by modeling error terms with a Dirichlet process.

Example: Semiparametric BART

$$Y_i = X_i\beta + \sum_{j=1}^{m} g(X_i; T_j, M_j) + \epsilon_i \tag{3}$$

- Combines linear predictors with nonparametric regression trees.
- Useful for models with both fixed and random effects.

Example: Dirichlet Process Mixture BART

$$\epsilon_i \sim N(a_i, \sigma_i^2), \quad (a_i, \sigma_i^2) \sim D, \quad D \sim DP(D_0, \alpha) \tag{4}$$

- Models the error term with a Dirichlet process.
- Allows for flexible error distributions.

Discussion

BART is a powerful and flexible tool for regression and classification.

- Handles complex, non-linear relationships without explicit specification.
- Provides a unified framework for various extensions and applications.
- Further research can expand BART's applicability and efficiency.