The Art of BART: On Flexibility of Bayesian Forests

Seonghyun Jeong

(Joint work with Veronika Ročková)

Department of Applied Statistics

Yonsei University

Nonparametric Regression

- Nonparametric regression: $f : [0, 1]^p \mapsto \mathbb{R}$,

$$y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

- How to model f? (Spline, Kernel, GP, DNN, ...)
- Bayesian Additive Regression Trees (BART).

Success of BART

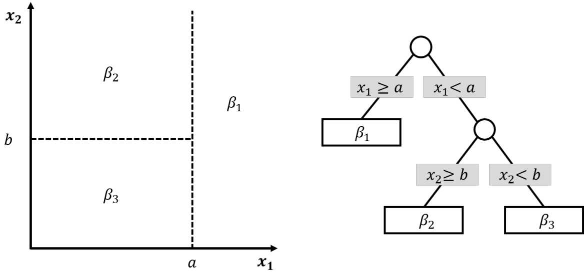
BART: Bayesian additive regression trees

HA Chipman, EI George... - The Annals of Applied ..., 2010 - projecteuclid.org

We develop a Bayesian "sum-of-trees" model where each tree is constrained by a regularization prior to be a weak learner, and fitting and inference are accomplished via an iterative Bayesian backfitting MCMC algorithm that generates samples from a posterior. Effectively, BART is a nonparametric Bayesian regression approach which uses dimensionally adaptive random basis elements. Motivated by ensemble methods in general, and boosting algorithms in particular, BART is defined by a statistical model: a prior and a ...

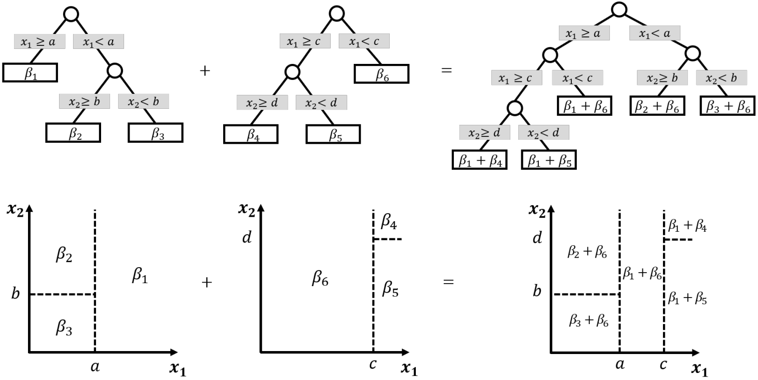
☆ 99 Cited by 1211 Related articles All 15 versions

Bayesian CART



- CART: Trees partition $[0, 1]^p$ through axis-parallel splits.
- Piecewise constant functions defined on a tree-based partition: $f_{\mathcal{T},\beta}(x) = \sum_k \beta_k \mathbb{1}_{\Omega_k}(x)$.

Bayesian Additive Regression Trees



- BART overlaps trees for more flexible partitions.
- Piecewise constant functions on an ensemble-based partition: $f_{\mathcal{E},B} = \sum_{t=1}^{T} \sum_k \beta_k^t \mathbb{1}_{\Omega_k^t}(x)$.

Theoretical Developments of BART

- L_2 posterior contraction (for isotropic Hölder classes):
  - Ročkovà and van der Pas (2020)
  - Ročkovà and Saha (2019)
  - Linero and Yang (2018)
- Sup-norm posterior contraction (under some restricted setup):
  - Castillo and Ročkovà (2020)
- Bernstein-von Mises theorems, uncertainty quantification:
  - Ročkovà (2020)
  - Wang and Ročkovà (2020)
  - Castillo and Ročkovà (2020)

Our Contribution

- Enhanced understanding of posterior contraction in BART:
  - for function spaces beyond the ordinary Hölder classes
  - under relaxed conditions
- Minimax optimality
- Beyond regression problems (only in the paper)

Bayesian Framework

- Data generation: $y = f_0(x) + \varepsilon$, $\varepsilon \sim N(0, \sigma_0^2)$, $x \in [0,1]^p$
- High-dimensional setup: $p > n$
- Sparsity: d-sparse functions
- Priors:

$$\Pi((\mathcal{T}^1, \beta^1), \dots, (\mathcal{T}^T, \beta^T), \sigma^2, S) = \Pi(S)\Pi(\sigma^2) \prod_{t=1}^{T} \Pi(\beta^t | \mathcal{T}^t, S)\Pi(\mathcal{T}^t | S)$$

- Gaussian priors on β^t and an inverse gamma prior on σ^2
- Marginal likelihood $m(\mathcal{T}^1, \dots, \mathcal{T}^T, S)$ is available.

Prior Specification: Spike-and-Slab Tree Priors

- Spike-and-slab prior $\Pi(S)$
- Tree Prior $\Pi(\mathcal{T}|S)$
  - Denison et al. (1998):
    - Exponential-tailed prior on the tree size K: $\log \pi(K = k) \asymp -k \log k$.
    - Uniform prior over trees:

$$\pi(\mathcal{T}|S, K) = \frac{1}{\#\mathbb{T}_{S,K}} \mathbb{1}(\mathcal{T} \in \mathbb{T}_{S,K}),$$

  - Chipman et al. (1998):
    - Splitting probability: Each node at depth ℓ is split with probability ν^ℓ, $\nu \in (\nu_0, 1/2)$ for some constant $\nu_0 > 0$.
    - Each splitting variable is uniformly chosen; each splitting point is uniformly chosen.

Posterior Contraction Rates

- A sequence ϵ_n such that for every $M_n \to \infty$,

$$\Pi(\theta : d(\theta, \theta_0) \geq M_n \epsilon_n | Y^{(n)}) = o_P(1).$$

  - Measures how fast the posterior distribution contracts to the true parameter
  - Comparable to the frequentist convergence rates
  - $1/\sqrt{n}$ for regular i.i.d. parametric models of fixed dimension

- For example, if we take the empirical L_2-norm for d in our nonparametric setup, we are interested in the smallest ϵ_n such that

$$\Pi(f : \|f - f_0\|_n \geq M_n \epsilon_n | Y^{(n)}) = o_P(1).$$

Posterior Contraction Theory

- Ghosal, Ghosh, and van der Vaart (2000); Ghosal and van der Vaart (2007)
- A test function ϕ_n: for every $\epsilon > 0$ and every $\theta_1 \in \Theta_n$ with $d_n(\theta_0, \theta_1) > \epsilon$, for some $K, \xi > 0$,

$$P_0^{(n)} \phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{\theta \in \Theta_n : d_n(\theta, \theta_1)} P_\theta^{(n)} (1 - \phi_n) \leq e^{-Kn\epsilon^2}.$$

- If, for $\epsilon_n \geq \bar{\epsilon}_n$ with $n\bar{\epsilon}_n^2 \to \infty$, there exists $\Theta_n \subset \Theta$ such that
  - $\Pi(\sum_{i=1}^n \int \log \frac{p_{0i}}{p_i} dP_{0i} < n\bar{\epsilon}_n^2, \int (\log \frac{p_{0i}}{p_i} - \int \log \frac{p_{0i}}{p_i} dP_{0i})^2 dP_{0i} < n\bar{\epsilon}_n^2) \geq e^{-cn\bar{\epsilon}_n^2}$;
  - $\log N(\xi\epsilon_n, \Theta_n, d_n) \leq n\epsilon_n^2$;
  - $\Pi(\Theta \setminus \Theta_n) \leq e^{-(c+4)n\bar{\epsilon}_n^2}$,

  then, $\Pi(\theta : d_n(\theta, \theta_0) \geq M_n\epsilon_n | Y^{(n)}) = o_{P_0}(1)$ for every $M_n \to \infty$.

Posterior Contraction Theory

- The entropy condition and the prior complement mass condition are satisfied with the BART priors.
- What kind of functions classes can be used to satisfy the prior concentration condition?
  $\rightarrow$ Approximation theory!

Piecewise Heterogeneous Anisotropic Hölder Space

- Piecewise heterogeneous anisotropic Hölder space:

$$\mathcal{H}_\lambda^{A,d}(\mathfrak{X};\alpha_\star) = \left\{ h : [0,1]^d \mapsto \mathbb{R};\ h|_{\Xi_r} \in \mathcal{H}_\lambda^{\alpha_r,d}(\Xi_r),\ r = 1,\ldots,R \right\}.$$

  - $\mathfrak{X} = (\Xi_1,\ldots,\Xi_R)$: a box partition of $[0,1]^d$;
  - $A = (\alpha_r)_{r=1}^R$: smoothness parameters with $\alpha_r = (\alpha_{r1},\ldots,\alpha_{rd})' \in (0,1]^d$ such that $\alpha_\star = (\frac{1}{d}\sum_{j=1}^d \frac{1}{\alpha_{rj}})^{-1} \in (0,1]$, $r = 1,\ldots,R$.

- Anisotropic Hölder space:

$$\mathcal{H}_\lambda^{\alpha,d}(\Xi) = \left\{ h : \Xi \mapsto \mathbb{R};\ |h(x) - h(y)| \le \lambda \sum_{j=1}^d |x_j - y_j|^{\alpha_j},\ x, y \in \Xi \right\}.$$

- $h \in \mathcal{H}_\lambda^{A,d}(\mathfrak{X};\alpha_\star)$ can be discontinuous!

Piecewise Heterogeneous Anisotropic Hölder Space

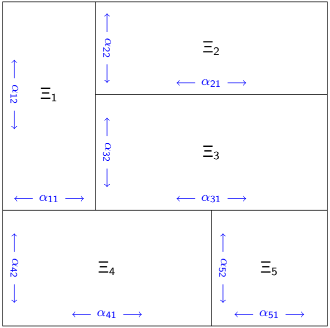
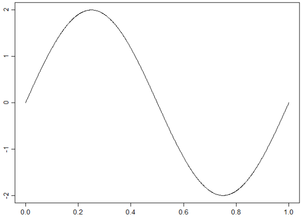


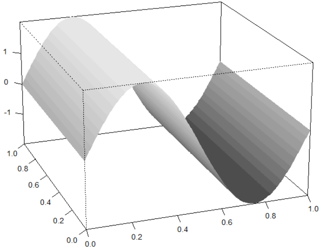
Figure: Piecewise heterogeneous anisotropic Hölder space

Sparse Function Spaces

- $W_S^p : \mathcal{C}(\mathbb{R}^{|S|}) \mapsto \mathcal{C}(\mathbb{R}^p)$ is the map that transmits $h \in \mathcal{C}(\mathbb{R}^{|S|})$ onto $W_S^p h : x \mapsto h(x_S)$, $S \subseteq \{1, \ldots, p\}$.
- *d*-sparse piecewise heterogeneous anisotropic Hölder space:

$$\Gamma_\lambda^{A,d,p}(\mathfrak{X};\alpha_\star) = \bigcup_{S:|S|=d} W_S^p\big(\mathcal{H}_\lambda^{A,d}(\mathfrak{X};\alpha_\star)\big).$$

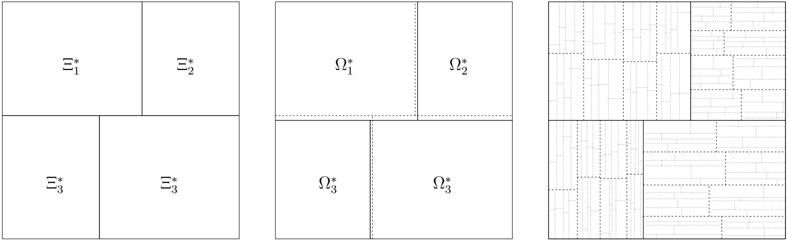




Approximating the True Functions

- The ability of approximating the true function (with a suitable error bound) is very important for the optimal posterior contraction.
- In our setup, this ability is directly dependent on the locations for possible splits.
- We call the possible locations for splits a split-net.
  - Too many splitting locations: good approximability, complexity issue.
  - Insufficient splitting locations: controlled complexity, bad approximability.
- The cardinality of a split-net should be balanced well!

Global-Local Approximability



- A split-net should be chosen such that:
  - The boundaries of the box partition $\mathfrak{X}$ should be detected;
  - The local feature of the function should be detected on each box.

Main Results: Posterior Contraction of BART

- $y = f_0(x) + \varepsilon$, $\varepsilon \sim N(0, \sigma_0^2)$.
- $f_0 \in \Gamma_\lambda^{A,d,p}(\mathfrak{X}; \alpha_\star)$
- BART prior for f; inverse gamma prior for σ^2
- Posterior contraction: There exists a constant $M > 0$ such that

$$\mathbb{E}_0 \Pi \Big\{ (f, \sigma^2) : \|f - f_0\|_n + |\sigma^2 - \sigma_0^2| > M\epsilon_n \,\Big|\, Y_1, \dots, Y_n \Big\} \to 0,$$

where

$$\epsilon_n = \sqrt{\frac{d \log p}{n}} + (\lambda d)^{d/(2\alpha_\star + d)} \left(\frac{R \log n}{n} \right)^{\alpha_\star/(2\alpha_\star + d)}.$$

Main Results: Minimax Optimality

- Under some restricted design, the minimax rate can be found.
- Derivation is similar to Yang and Tokdar (2015).
- Minimax L_2-risk:

$$\gamma_n = \sqrt{\frac{\log \binom{p}{d}}{n}} + \left(\frac{\lambda^d}{n^{\alpha_\star}}\right)^{1/(2\alpha_\star + d)}.$$

Conclusion

- BART is adaptive to heterogeneous anisotropic Hölder space with near-minimaxity.
- No major modification is required for the BART prior.

Thank you!