Bayesian causal machine learning using BART

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

Leuven, September 2024

DEPARTMENT OF
STATISTICS

Introduction

BART has emerged as one of the most popular Bayesian nonparametric (predictive) methods – around 2500 citations – and is particularly well suited to causal inference

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." (2010), Annals of Applied Statistics 266-298.

Acknowledgement: Thanks to Ed George for kindly providing his overview slides on BART

Part I. BART (Bayesian Additive Regression Trees)

Data: n observations of y and $x = (x_1, ..., x_p)$

Suppose: $Y = f(x) + \varepsilon, \quad \varepsilon$ symmetric with mean 0

Bayesian Ensemble Idea: Approximate unknown $f(x)$ by a form

$$f(x) = g(x; \theta_1) + g(x; \theta_2) + ... + g(x; \theta_m)$$

$$\theta_1, \theta_2, \ldots, \theta_m \quad \text{iid} \sim \pi(\theta)$$

and use the posterior of f given y for inference.

BART is obtained when each $g(x; \theta_j)$ is a regression tree.

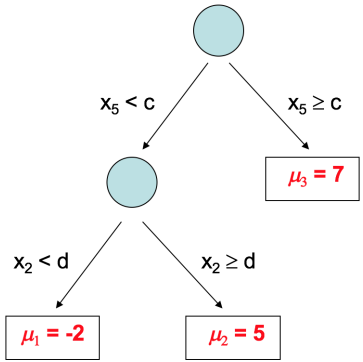
Key calibration: Using y, set $\pi(\theta)$ so that $\text{Var}(f) \approx \text{Var}(y)$.

Beginning with a Single Regression Tree Model

Let T denote the tree structure including the decision rules

Let $M = \{\mu_1, \mu_2, \ldots \mu_b\}$ denote the set of bottom node μ's.

Let $g(x; T, M)$ be a regression tree function that assigns a μ value to x



A single tree model:

$$Y = g(x; T, M) + \sigma z, \quad z \sim N(0,1)$$

Bayesian CART: Just add a prior $\pi(M, T)$

Bayesian CART Model Search
(Chipman, George, McCulloch 1998)

$$\pi(M, T) = \pi(M \mid T)\pi(T)$$

$\pi(M \mid T) : (\mu_1, \mu_2, \ldots, \mu_b)' \sim N_b(0, \tau^2 I)$

$\pi(T)$: Stochastic process to generate tree skeleton plus uniform prior on splitting variables and splitting rules.

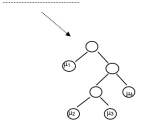
Closed form for $\pi(T \mid y)$ facilitates MCMC stochastic search for promising trees.

Moving on to BART

Bayesian Additive Regression Trees
(Chipman, George, McCulloch 2010)

The BART ensemble model

$$Y = g(x; T_1, M_1) + g(x; T_2, M_2) + \ldots + g(x; T_m, M_m) + \sigma z, \quad z \sim N(0,1)$$



Each (T_i, M_i) identifies a single tree.

$E(Y \mid x, T_1, M_1, \ldots, T_m, M_m)$ is the sum of m bottom node μ's, one from each tree.

Number of trees m can be much larger than sample size n.
$g(x; T_1, M_1), g(x; T_2, M_2), ..., g(x; T_m, M_m)$ is a highly redundant "over-complete basis"

Complete the Model with a Regularization Prior

$$\pi((T_1, M_1), (T_2, M_2), \ldots, (T_m, M_m), \sigma)$$

π applies the Bayesian CART prior to each (T_j, M_j) independently so that:

- Each T small.
- Each μ small.

The observed variation of y is used to calibrate the choice of the hyperparameters for the μ and σ priors.

π is a "regularization prior" as it keeps the contribution of each $g(x; T_i, M_i)$ small, explaining only a small portion of the fit.

Connections to Other Modeling Ideas

$$Y = g(x;T_1,M_1) + ... + g(x;T_m,M_m) + \sigma z$$
plus
$$\pi((T_1,M_1),....(T_m,M_m),\sigma)$$

Bayesian Nonparametrics:

- Lots of parameters (to make model flexible)
- A strong prior to shrink towards simple structure (regularization)
- BART shrinks towards additive models with some interaction

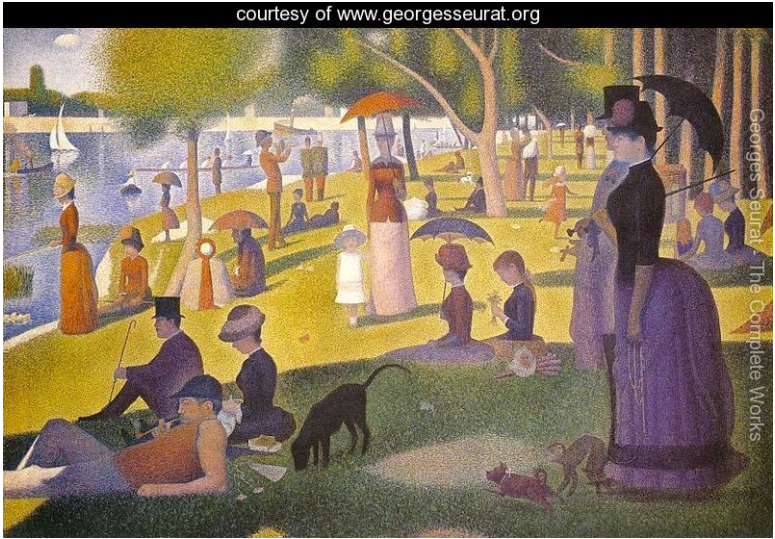
Dynamic Random Basis Elements:

- $g(x; T_1, M_1), ..., g(x; T_m, M_m)$ are each dimensionally adaptive

Gradient Boosting:

- Fit becomes the cumulative effort of many *weak learners*

Build up the fit, by adding up tiny bits of fit ..



A Sketch of the BART MCMC Algorithm

$$Y = g(x;T_1,M_1) + \ldots + g(x;T_m,M_m) + \sigma z$$

plus

$$\pi((T_1,M_1),\ldots(T_m,M_m),\sigma)$$

Outer Loop is a "simple" Gibbs sampler: 0.0in

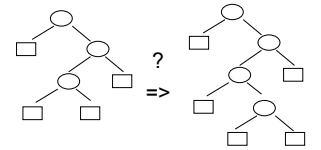
- Iteratively sample each (T_i, M_i) given Y, σ and all other (T_j, M_j)'s (Bayesian Backfitting)
- Sample σ given Y and $(T_1, M_1, \ldots, \ldots, T_m, M_m)$

To sample (T_i, M_i) above,

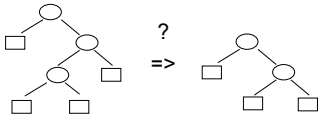
1. Subtract the contributions of all the other trees from both sides to get a simple one-tree model update.
2. Integrate out M to sample T and then sample $M \mid T$.

For the draw of T we use a Metropolis-Hastings within Gibbs step.

Our proposal moves around tree space by proposing local modifications such as the "birth-death" step:



propose a more complex tree



propose a simpler tree

... as the MCMC runs, each tree in the sum will grow and shrink, swapping fit amongst

Using the MCMC Output to Draw Inference

After convergence (which happens surprisingly quickly), each iteration d results in a draw from the posterior of f

$$\hat{f}_d(\cdot) = g(\cdot; T_1, M_1) + \cdots + g(\cdot; T_m, M_m)$$

To estimate $f(x)$ we simply average the $\hat{f}_d(\cdot)$ draws at x

Posterior uncertainty is captured by variation of the $\hat{f}_d(x)$
eg, 95% HPD region estimated by middle 95% of values

Can do the same with functionals of f.

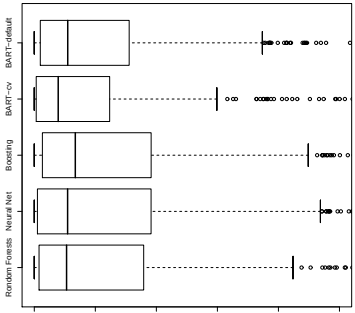
Out of Sample Prediction

Predictive comparisons on 42 data sets.

Data from Kim, Loh, Shih and Chaudhuri (2006) (thanks Wei-Yin Loh!)

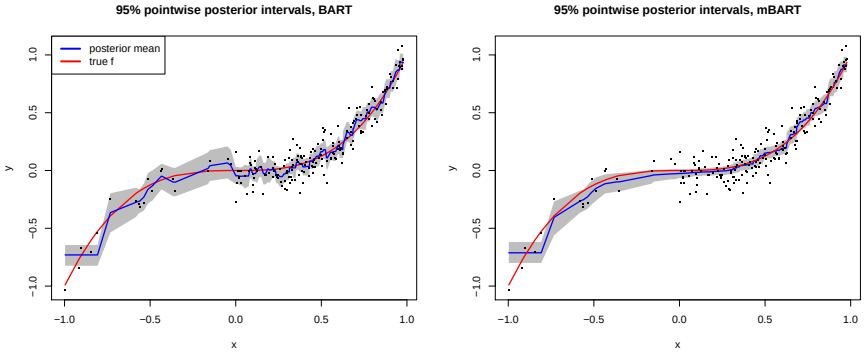
- $p = 3$ to 65, $n = 100$ to 7,000.
- for each data set 20 random splits into 5/6 train and 1/6 test
- use 5-fold cross-validation on train to pick hyperparameters (except BART-default!)
- gives 20*42 = 840 **out-of-sample predictions**, for each prediction, divide rmse of different methods by the smallest

+ each boxplots represents 840 predictions for a method
+ 1.2 means you are 20% worse than the best
+ BART-cv best
+ BART-default (use default prior) does amazingly well!!



Automatic Uncertainty Quantification

A simple simulated 1-dimensional example



Note: MBART on the right plot still to be discussed

Example: Friedman's Simulated Data

$$Y = f(x) + \varepsilon, \quad \varepsilon \sim N(0,1)$$

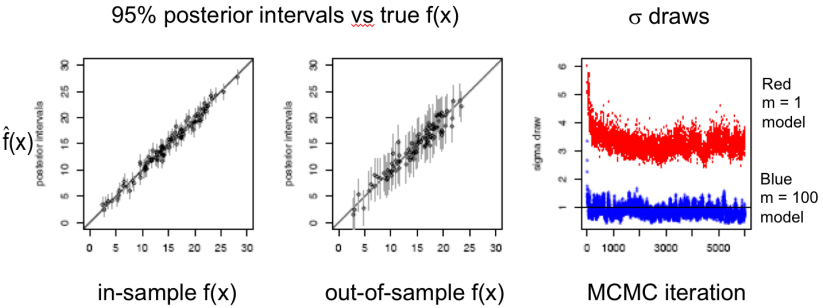
where

$$f(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - .5)2 + 10x_4 + 5x_5 + 0x_6 + \cdots + 0x_{10}$$

- x_i's iid $\sim$ *Uniform*$(0,1)$
- Only the first 5 x_i's matter!
- Friedman (1991) used $n = 100$ observations from this model to illustrate the potential of MARS
- BART handily outperforms competitors including random forests, neural nets and gradient boosting on this example.

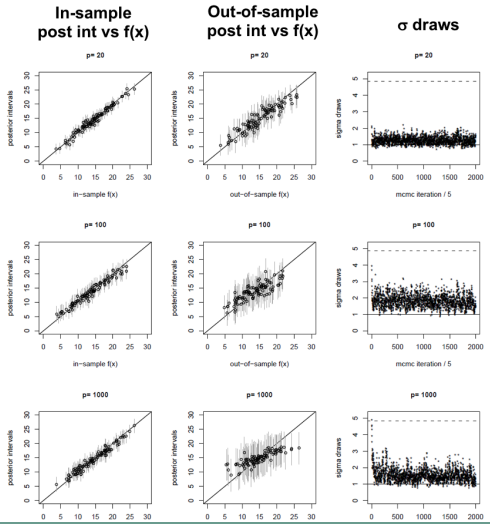
Applying BART to the Friedman Data

With $n = 100$ observations and $m = 100$ trees



Added many useless x's to Friedman's example

With only 100 observations on y and *1000 x's, BART yielded "reasonable" results !!!!*



20 x's

100 x's

1000 x's

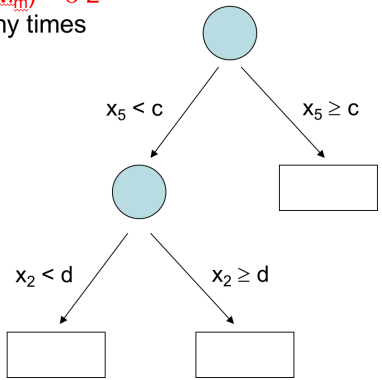
Measuring Variable Importance in BART

When $Y = g(x;T_1,M_1) + \ldots + g(x;T_m,M_m) + \sigma z$ is fit to data, we can count how many times a predictor is used in the trees.

For example, in the tree here, x_2 and x_5 are each used once.

The importance of each x_k can thus be measured by its overall usage frequency.

This approach is most effective when the number of trees m is small.



Creating a Competitive Bottleneck

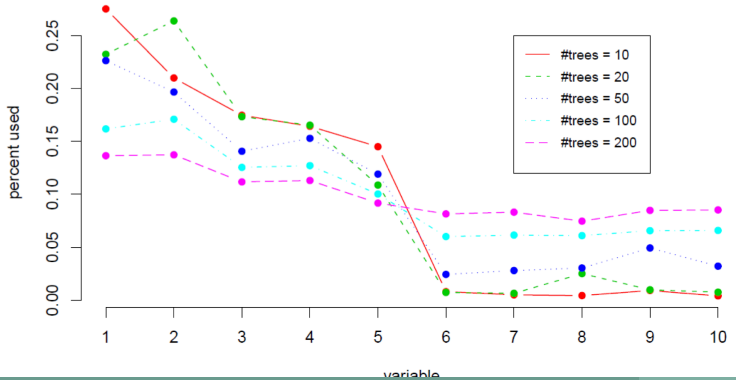
BART is an automatic attractor of x's to explain Y

Those x's which most explain Y are attracted most

Small number of trees m creates a bottleneck which excludes

Variable Selection via BART

Variable usage frequencies as the number of trees m is reduced



Extensions to causal inference

Hill (2011) applied BART to problems in causal inference, noting that BART's response surface estimation accuracy made it well suited to prediction of potential outcomes.

BART is particularly well suited to detecting treatment heterogeneity due to it's nonlinear nature

In Hill (2011) the treatment indicator was viewed as simply another covariate and BART was used to estimate

$$E[Y \mid x, T = 1] - E[Y \mid T = 0]$$

Regularization-induced confounding

However, in the presence of strong confounding (association between X and T), even when all confounders have been included, the use of standard BART priors may introduce bias into the estimation of treatment effects

Picture the likelihood function for two predictors (covariates) in a linear regression function as the correlation increases. A prior that "shrinks towards zero" will favour spreading the combined effect across the two parameters

Hahn *et. al.* (2020) noted standard priors can unwittingly lead to extreme bias in **estimating the target parameter**

Bayesian Causal Forests

The solution reported in Hahn *et. al.* (2020) is to re-parameterise the model to separate out the prior on the treatment effect into two BART models

$$f(x_i, t_i) = \mu(x_i, \hat{\pi}_i) + \tau(x_i)t_i$$

At first sight this may look odd, but note that under this representation we have

$$E[Y \mid x, T = 1] - E[T \mid x, T = 0] = \{\mu(x, \pi) + \tau(x)\} - \mu(x, \pi)\} = \tau(x)$$

which isolates the heterogeneous treatment effect model in $\tau(\cdot)$

This allows for a special prior – little shrinkage – to be placed on the treatment effect model

Concluding Remarks

- Despite its many compelling successes in practice, theoretical frequentist support for BART is only now just beginning to appear.
- In particular, Rockova and van der Pas (2017) *Posterior Concentration for Bayesian Regression Trees and Their Ensembles* recently obtained the first theoretical results for Bayesian CART and BART, showing near-minimax posterior concentration when $p > n$ for classes of Holder continuous functions.
- BART can be extended to causal inference settings for heterogeneous treatment effects through careful prior specification and use of two BART models
- Software for BART is on CRAN, for MBART at https://bitbucket.org/remcc/mbart (use git clone to get everything), and for HBART coming soon as LISA-BART on CRAN.

Thank You!