Wiley StatsRef:
Statistics Reference Online
DOI: 10.1002/ISBN.stat00999.pub9

Bayesian Additive Regression Trees (BART)

Rodney Sparapani[a]*and Robert McCulloch[b]

Keywords: black-box models; categorical outcomes; competing risks; continuous outcomes; dichotomous outcomes; ensemble predictive modeling; nonparametric; machine learning; recurrent events; survival analysis

Abstract: This article introduces the nonparametric machine learning technique known as Bayesian Additive Regression Trees (BART) for continuous, dichotomous, categorical and time-to-event outcomes.

1. Introduction

Bayesian Additive Regression Trees (BART) arose out of earlier research on Bayesian model fitting of an outcome to a single tree (Chipman et al., 1998); for more on approaches like this, see **Classification and Regression Trees Methods** (Loh, 2014) that carry the popular acronym CART. In that era (circa 1998), the outstanding predictive performance of ensemble models was starting to become apparent (Breiman, 1996; Krogh & Solich, 1997; Freund & Schapire, 1997; Breiman, 2001; Friedman, 2001; Baldi & Brunak, 2001). Instead of making a single prediction from a complex model, ensemble models make a prediction which is the summary of the predictions from many simple models. Generally, ensemble models have desirable properties, e.g., on the spectrum from low bias with high variance (such as CART (Breiman, 2017)) to high bias with low variance (linear regression), ensemble models generally fall somewhere in the middle which accounts for their superior performance (Kuhn & Johnson, 2013). With some similarities to bagging (Breiman, 1996), boosting (Freund & Schapire, 1997; Friedman, 2001) and random forests (Breiman, 2001), BART relies on an ensemble of trees to predict the outcome.

BART is a Bayesian nonparametric, sum of trees method for continuous, dichotomous, categorical and time-to-event outcomes. Furthermore, BART is a black-box, machine learning method which fits the outcome via an arbitrary random function, f, of the covariates. So-called black-box models generate functions of the covariates which are so

[a]**Medical College of Wisconsin, Milwaukee, WI, USA**
[b]**Arizona State University, Tempe, AZ, USA**
*__Email: rsparapa@mcw.edu__

Prepared using WileySTAT.cls [Version: v1.00]

complex that interpreting the internal details of the fitted model is generally abandoned in favor of assessment via evaluations of the fitted function, f, at chosen values of the covariates (Friedman, 2001). As shown by Chipman et al. (2010), BART's out-of-sample predictive performance is generally equivalent to, or exceeds that, of alternatives like lasso with L1 regularization (Efron et al., 2004) or black-box models such as gradient boosting (Freund & Schapire, 1997; Friedman, 2001), neural nets with one hidden layer (Ripley, 2007; Venables & Ripley, 2013) and random forests (Breiman, 2001). **Overfitting (Overtraining)** is the tendency to over-do the fit of a model to an in-sample training data set's signal, and particularly its noise, resulting in poor predictive performance for unseen out-of-sample data (Cristianini, 2014; Cook & Ranstam, 2016). Typically, BART does not over-fit to the training data due to the regularization tree-branching penalty of the BART prior, i.e., generally, each tree in the ensemble has few branches and plays a small part in the overall fit (see formula (1) below). Essentially, BART is a Bayesian nonlinear model with all the advantages of the Bayesian paradigm such as posterior inference including point and interval estimation. Conveniently, BART naturally scales to large numbers of covariates and facilitates variable selection; it does not require the covariates to be rescaled; neither does it require the covariate functional relationship, nor the interactions considered, to be pre-specified.

In this article, we will discuss the BART prior in the context of continuous outcomes in Section 2 along with posterior computations in Section 3. In Section 4, we demonstrate BART with a classic example. Next, we will briefly discuss BART extensions for dichotomous, categorical and survival outcomes in Section 5. We will then conclude with some recent developments and software implementations (as of this writing) in Section 6.

2. Binary trees and the BART prior

Here, we briefly describe binary trees and their relationship to BART; for a more detailed discussion of trees and tree-based methods, see **Tree-structured Statistical Methods** (Zhang et al., 2014). BART relies on an ensemble of H binary trees which are a type of a directed acyclic graph. For illustration, we fully exploit the wooden tree metaphor with binary trees. Each of these trees grows from the ground up starting out as a root node. The root node is generally a branch decision rule, but it doesn't have to be; occasionally there are trees in the ensemble which are only a root terminal node consisting of a single leaf output value. If the root is a branch decision rule, then it spawns a left and a right node which each can be either a branch decision rule or a terminal leaf value and so on. In binary tree, $\mathcal{T}$, there are C nodes which are made of B branches and L leaves: $C = B + L$. There is a further relationship between the number of branches and leaves: $B = L - 1$. The nodes are numbered in relation to the tree's tier level, $t(n) = \lfloor \log_2 n \rfloor$ as follows.

The key to discriminating between branches and leaves is via the algebraic relationship between a branch, n, at tree tier $t(n)$ leading to its left, $l = 2n$, and right, $r = 2n + 1$, nodes at tier $t(n) + 1$, i.e., for each node, besides root, you can determine from which branch it arose and those nodes that are not a branch (since they have no leaves) are necessarily leaves.

Tier								
t	2^t		...		$2^{t+1}-1$			
⋮								
2	4		5		6		7	
1		2				3		
0				1				

Underlying this methodology is the BART prior. The BART prior specifies a flexible class of unknown functions, f, from which we can gather randomly generated fits to the given data via the posterior. Let the regression tree function $g(\boldsymbol{x}; \mathcal{T}, \mathcal{M})$ assign a value based on the input $\boldsymbol{x}$; for more details, see **Regression Trees** (LeBlanc, 2014). The binary decision tree $\mathcal{T}$ is represented by a set of ordered triples, (n, j, k), representing branch decision rules: n for the node, j for covariate x_j and k for the cutpoint c_{jk}. The branch decision rules are of the form $x_j \leq c_{jk}$ which means branch left and $x_j > c_{jk}$, branch right; or terminal leaves where it stops. $\mathcal{M}$ represents leaves and is a set of ordered pairs, (n, μ_n): n for the node, μ_n for the outcome value and $n \in \mathcal{L}$ where $\mathcal{L}$ is the set of leaves. The function, $f(\boldsymbol{x})$, is a sum of H regression tree functions:

$$f(\boldsymbol{x}) = \sum_{h=1}^{H} g(\boldsymbol{x}; \mathcal{T}_h, \mathcal{M}_h)$$

where H is "large", let's say, 50, 100 or 200.

For a continuous outcome, y_i, we have the following BART regression on the vector of covariates, $\boldsymbol{x}_i$:

$$y_i = \mu_0 + f(\boldsymbol{x}_i) + \epsilon_i \text{ where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ w_i^2 \sigma^2\right)$$

with i indexing subjects $i = 1, \ldots, N$. The unknown random function, f, and the error variance, σ^2, follow the BART prior expressed notationally as

$$(f, \sigma^2) \overset{\text{prior}}{\sim} \mathrm{BART}(H, \mu_0, \tau, k, \alpha, \gamma; \nu, \lambda, q)$$

where H is the number of trees, μ_0 is a known constant which centers y and the rest of the parameters will be explained later in this section (for brevity, we often use the simpler shorthand $(f, \sigma^2) \overset{\text{prior}}{\sim} \mathrm{BART}$). The w_i are known standard deviation weight multiples (only available for continuous outcomes) where the unit weight vector is usually assumed. The centering parameter, μ_0: the default is often taken to be $\overline{y}$ for continuous outcomes.

BART is a Bayesian nonparametric prior. Using the Gelfand-Smith generic bracket notation for the specification of random variable distributions (Gelfand & Smith, 1990), we represent the BART prior in terms of the collection of all trees, $\boldsymbol{\mathcal{T}}$; collection of all leaves, $\boldsymbol{\mathcal{M}}$; and the error variance, σ^2, as the following product: $\left[\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}}, \sigma^2\right] = \left[\sigma^2\right] \left[\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}}\right] = \left[\sigma^2\right] \left[\boldsymbol{\mathcal{T}}\right] \left[\boldsymbol{\mathcal{M}} | \boldsymbol{\mathcal{T}}\right]$. Furthermore, the individual trees themselves are independent: $\left[\boldsymbol{\mathcal{T}}, \boldsymbol{\mathcal{M}}\right] = \prod_h \left[\mathcal{T}_h\right] \left[\mathcal{M}_h | \mathcal{T}_h\right]$. where $\left[\mathcal{T}_h\right]$

is the prior for the hth tree and $[\mathcal{M}_h|\mathcal{T}_h]$ is the collection of leaves for the hth tree. And, finally, the collection of leaves for the hth tree are independent: $[\mathcal{M}_h|\mathcal{T}_h] = \prod_{n\in\mathcal{L}_h}[\mu_{hn}|\mathcal{T}_h]$ where n indexes the leaf nodes.

The tree prior: $[\mathcal{T}_h]$. There are three prior components of $\mathcal{T}_h$ which govern whether the tree branches grow or are pruned. The first tree prior regularizes the probability of a branch at leaf node n in tree tier $t(n) = \lfloor \log_2 n \rfloor$ as

$$\mathrm{P}[B_n = 1] = \alpha(t(n)+1)^{-\gamma} \tag{1}$$

where $B_n = 1$ represents a branch while $B_n = 0$ is a leaf, $0 < \alpha < 1$ and $\gamma \geq 0$. The following defaults are recommended: $\alpha = 0.95$ and $\gamma = 2$; for a detailed discussion of these parameter settings, see Chipman et al. (1998). Note that this prior penalizes branch growth, i.e., in prior probability, the default number of branches will likely be 1 or 2. Next, there is a prior dictating the choice of a splitting variable j conditional on a branch event B_n which defaults to uniform probability $s_j = P^{-1}$ where P is the number of covariates (however, you can specify a Dirichlet prior which is more appropriate if the number of covariates is large (Linero, 2018)). Given a branch event, B_n, and a variable chosen, x_j, the last tree prior selects a cut point, c_{jk}, within the range of observed values for x_j; this prior is often chosen to be uniform for convenience.

The leaf prior: $[\mu_{hn}|\mathcal{T}_h]$. Given a tree, $\mathcal{T}_h$, there is a prior on its leaf values. We denote the collection of all leaves in $\mathcal{T}_h$ by $\mathcal{M}_h = \{(n, \mu_{hn}) : n \in \mathcal{L}_h\}$. Note that $y_i \in [y_{\min}, y_{\max}]$ for $i = 1, \ldots, N$ and denote $\left[\mu_{1(\mathbf{x}_i)}, \ldots, \mu_{H(\mathbf{x}_i)}\right]$ as the leaf output values from each tree corresponding to the vector of covariates, $\mathbf{x}_i$. If $\mu_{h(\mathbf{x}_i)}|\mathcal{T}_h \overset{\text{iid}}{\sim} \mathrm{N}\left(0,\ \sigma_\mu^2\right)$, then the model estimate for subject i is $\mu_i = \mathrm{E}[y_i|\mathbf{x}_i] = \mu_0 + \sum_h \mu_{h(\mathbf{x}_i)}$ where $\mu_i \sim \mathrm{N}\left(\mu_0,\ H\sigma_\mu^2\right)$. We choose a value for σ_μ which is the solution to the equations $y_{\min} = \mu_0 - k\sqrt{H}\sigma_\mu$ and $y_{\max} = \mu_0 + k\sqrt{H}\sigma_\mu$, i.e., $\sigma_\mu = \frac{y_{\max} - y_{\min}}{2k\sqrt{H}}$. Therefore, we arrive at $\mu_{hn} \overset{\text{prior}}{\sim} \mathrm{N}\left(0,\ \left[\frac{\tau}{2k\sqrt{H}}\right]^2\right)$ where $\tau = y_{\max} - y_{\min}$. So, the prior for μ_{hn} is weakly informed by the data, y, only via the extrema, $y_{\min}$ and $y_{\max}$. The parameter k calibrates this prior as follows.

$$\begin{aligned}
\mu_i &\sim \mathrm{N}\left(\mu_0,\ \left[\frac{\tau}{2k}\right]^2\right) \\
\mathrm{P}[y_{\min} \leq \mu_i \leq y_{\max}] &= \Phi(k) - \Phi(-k) \\
\text{Since } \mathrm{P}[\mu_i \leq y_{\max}] &= \mathrm{P}\left[z \leq 2k\frac{y_{\max} - \mu_0}{\tau}\right] \approx \mathrm{P}[z \leq k] = \Phi(k) \\
\text{Similarly } \mathrm{P}[\mu_i \leq y_{\min}] &= \Phi(-k)
\end{aligned}$$

The recommended default choice, $k = 2$, corresponds to μ_i falling within the extrema with approximately 0.95 probability. Values of $k \in [1, 3]$ generally yield good results. k is a potential candidate parameter for choice via cross-validation.

The error variance prior: $\left[\sigma^2\right]$. The prior for σ^2 is the conjugate scaled inverse Chi-square distribution, i.e., $\nu\lambda\mathcal{X}^{-2}(\nu)$. We suggest that the degrees of freedom is in the range $\nu \in [3, 10]$ and we recommend the default choice of 3. Now, λ is based on the estimate $\widehat{\sigma}$: generally, if $P < N$, then $y_i \sim \mathrm{N}\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}},\ \widehat{\sigma}^2\right)$; otherwise, $\widehat{\sigma} = s_y$. Solve for λ such that

$\mathrm{P}\left[\sigma^2 \leq \widehat{\sigma}^2\right] = q$. The suggested range for the quantity $q \in [0.75, 0.99]$ where the recommended default choice is 0.9. (ν, q) are potential candidate parameters for cross-validation.

Other details of the BART prior. We fix the number of trees at H: the proposed default number of trees is 200 for continuous outcomes, but, as shown by Bleich et al. (2014), 50 is also a reasonable choice: cross-validation can be considered. Although, typically, not considered an argument, the number of cutpoints is an implementation detail where the standard choice is 100.

3. Posterior computation

In order to generate samples from the posterior for f, we employ **Markov Chain Monte Carlo (MCMC, Metropolis-Hastings, Gibbs Sampling)** (Hancock, 2014) for the structure of all the trees $\mathcal{T}_h$, for $h = 1, \ldots, H$; the values of all leaves μ_{hn} for $n \in \mathcal{L}_h$ within tree h; and the error variance σ^2 for continuous outcomes.

The leaf and variance parameters are sampled from the posterior using Gibbs sampling (Geman & Geman, 1984; Gelfand & Smith, 1990). Since the priors on these parameters are conjugate, the Gibbs conditionals are specified analytically. For the leaves, each μ_{hn} is drawn from a Normal conditional density. The error variance, σ^2, is drawn from a scaled inverse Chi-square conditional.

Drawing a tree from the posterior requires a Metropolis-within-Gibbs sampling scheme (Mueller, 1991, 1993), i.e., a Metropolis-Hastings step (Metropolis et al., 1953; Hastings, 1970) within Gibbs sampling. For single-tree models, four different proposal mechanisms are defined: the complementary BIRTH and DEATH along with CHANGE and SWAP (Chipman et al., 1998, 2013) (N.B. other MCMC tree sampling strategies have been proposed: Denison et al. (1998); Wu et al. (2007); Pratola (2016)). For the purposes of this discussion, we restrict our attention to the BIRTH and DEATH proposals each with equal probability. BIRTH selects a leaf and turns it into a branch, i.e., selects a new variable and cutpoint with two leaves "born" as its descendants. DEATH selects a branch leading to two terminal leaves and "kills" the branch by replacing it with a single leaf.

For illustration, we present the acceptance probability for a BIRTH proposal. The algorithm assumes a fixed discrete set of possible split values for each x_j. Furthermore, the leaf values, μ_{hn}, are removed here (by integrating over them) so that our search in tree space is over a large, but discrete, set of possibilities. At the mth MCMC step, let $\mathcal{T}^m$ denote the current state for the hth tree and $\mathcal{T}^*$ denotes the proposed hth tree (subscript h is suppressed here for convenience). $\mathcal{T}^*$ are identical to $\mathcal{T}^m$ except that one terminal leaf of $\mathcal{T}^m$ is replaced by a branch of $\mathcal{T}^*$ with two terminal leaves. The proposed tree is accepted with the following probability:

$$\pi_{\mathrm{BIRTH}} = \min\left(1, \frac{\mathrm{P}[\mathcal{T}^*]}{\mathrm{P}[\mathcal{T}^m]} \frac{\mathrm{P}[\mathcal{T}^m|\mathcal{T}^*]}{\mathrm{P}[\mathcal{T}^*|\mathcal{T}^m]}\right)$$

where $\mathrm{P}[\mathcal{T}^m]$ and $\mathrm{P}[\mathcal{T}^*]$ are the posterior probabilities of T^m and T^* respectively, $\mathrm{P}[\mathcal{T}^m|\mathcal{T}^*]$ is the probability of proposing $\mathcal{T}^m$ given current state $\mathcal{T}^*$ (a DEATH) and $\mathrm{P}[\mathcal{T}^*|\mathcal{T}^m]$ is the probability of proposing $\mathcal{T}^*$ given current state

$\mathcal{T}^m$ (a BIRTH).

First, we describe the likelihood contribution to the posterior. Let $\boldsymbol{y}_n$ denote the partition of $\boldsymbol{y}$ corresponding to the leaf node n given the tree $\mathcal{T}$. Because the leaf values are a priori conditionally independent, we have $[\boldsymbol{y}|\mathcal{T}] = \prod_n [\boldsymbol{y}_n|\mathcal{T}]$. So, for the ratio $\frac{\mathrm{P}[\mathcal{T}^*]}{\mathrm{P}[\mathcal{T}^m]}$ after cancellation of terms in the numerator and denominator, we have the likelihood contribution:

$$\frac{\mathrm{P}[\boldsymbol{y}_\mathrm{L}, \boldsymbol{y}_\mathrm{R}|\mathcal{T}^*]}{\mathrm{P}[\boldsymbol{y}_\mathrm{LR}|\mathcal{T}^m]} = \frac{\mathrm{P}[\boldsymbol{y}_\mathrm{L}|\mathcal{T}^*]\,\mathrm{P}[\boldsymbol{y}_\mathrm{R}|\mathcal{T}^*]}{\mathrm{P}[\boldsymbol{y}_\mathrm{LR}|\mathcal{T}^m]}$$

where $\boldsymbol{y}_\mathrm{L}$ is the partition corresponding to the newborn left leaf node; $\boldsymbol{y}_\mathrm{R}$, the partition for the newborn right leaf node; and $\boldsymbol{y}_\mathrm{LR} = \begin{bmatrix} \boldsymbol{y}_\mathrm{L} \\ \boldsymbol{y}_\mathrm{R} \end{bmatrix}$. N.B. the terms in the ratio are the predictive densities of a Normal mean with a known variance and a Normal prior for the mean.

Similarly, the terms that the prior contributes to the posterior ratio often cancel since there is only one "place" where the trees differ and the prior draws components independently at different "places" of the tree. Therefore, the prior contribution to $\frac{\mathrm{P}[\mathcal{T}^*]}{\mathrm{P}[\mathcal{T}^m]}$ is

$$\frac{\mathrm{P}[B_n = 1]\,\mathrm{P}[B_l = 0]\,\mathrm{P}[B_r = 0]\,s_j}{\mathrm{P}[B_n = 0]} = \frac{\alpha(t(n)+1)^{-\gamma}\,[1-\alpha(t(n)+2)^{-\gamma}]^2\,s_j}{1-\alpha(t(n)+1)^{-\gamma}}$$

where $\mathrm{P}[B_n]$ is the branch regularity prior (formula (1)), s_j is the splitting variable selection probability, n is the chosen leaf node in tree $\mathcal{T}^m$, $l = 2n$ is the newborn left leaf node in tree $\mathcal{T}^*$ and $r = 2n+1$ is the newborn right leaf node in tree $\mathcal{T}^*$.

Finally, the ratio $\frac{\mathrm{P}[\mathcal{T}^m|\mathcal{T}^*]}{\mathrm{P}[\mathcal{T}^*|\mathcal{T}^m]}$ is

$$\frac{\mathrm{P}[\mathrm{DEATH}|\mathcal{T}^*]\,\mathrm{P}[n|\mathcal{T}^*]}{\mathrm{P}[\mathrm{BIRTH}|\mathcal{T}^m]\,\mathrm{P}[n|\mathcal{T}^m]\,s_j}$$

where $\mathrm{P}[n|\mathcal{T}]$ is the probability of choosing node n given tree $\mathcal{T}$. See that s_j appears in both the numerator and denominator of the acceptance probability π_BIRTH, therefore, cancelling which is mathematically convenient.

4. Example: Boston housing values and air pollution

Here, we demonstrate BART with the classic Boston housing example (Harrison Jr & Rubinfeld, 1978). This data is based on the 1970 US Census where each observation represents a Census tract in the Boston Standard Metropolitan Statistical Area. For each tract, there was a localized air pollution estimate, the concentration of nitrogen oxides, `nox`, based on a meteorological model that was calibrated to monitoring data. Restricted to tracts with owner-occupied homes, there are $N = 506$ observations. We'll predict the median value of owner-occupied homes (in thousands of dollars), `mdev`, by thirteen covariates including `nox` which is our primary interest.

However, BART does not directly provide a summary of the effect of a single covariate, or a subset of covariates, on the outcome. Friedman's partial dependence function (Friedman, 2001) can be employed with BART to summarize the marginal effect due to a subset of the covariates, x_S, by aggregating over the complement covariates, x_C, i.e., $x = [x_S, x_C]$. The marginal dependence function is defined by fixing x_S while aggregating over the observed settings of the complement covariates in the data set: $f(x_S) = N^{-1} \sum_{i=1}^{N} f(x_S, x_{iC})$. For example, suppose that we want to summarize `mdev` by `nox` while aggregating over the other twelve covariates in the Boston housing data. In Figure 1, we demonstrate the marginal estimate and its 95% credible interval: notice that BART has discerned a complex non-linear relationship between `mdev` and `nox` from the data. N.B. this example including data and source code can be found in the *BART* R package (McCulloch et al., 2019) as the `nox.R` demonstration program.

5. Dichotomous, categorical and survival outcomes

Dichotomous outcomes can be handled with either a probit or a logit link function as in other Bayesian regression models. For probit, typically, the Albert & Chib (1993) technique is employed while for the logit link there are a few common alternatives (Holmes & Held, 2006; Frühwirth-Schnatter & Frühwirth, 2010; Gramacy & Polson, 2012). There are several techniques for categorical outcomes as well; see (Albert & Chib, 1993; McCulloch & Rossi, 1994; McCulloch et al., 2000; Frühwirth-Schnatter & Frühwirth, 2010; Kindo et al., 2016; Murray, 2017). For survival analysis, the approach of **Discrete Survival-time Models** (Fahrmeir, 2014) can be used to turn the problem into a dichotomous outcome (with either the probit or logit link) (Sparapani et al., 2016) without resorting to precarious restrictive assumptions such as those of the **Cox Proportional Hazard Model** (Cai & Zeng, 2014) or **Accelerated Failure-time Models** (James, 2014). Similarly, the discrete survival-time approach is applicable to recurrent events (Sparapani et al., 2018) and competing risks (Sparapani et al., 2019). For survival analysis with large data sets, you can pair BART with Dirichlet Process Mixtures (see **Nonparametric Bayes**, Dunson (2017)) (Bonato et al., 2011; Henderson et al., 2017); however, you will need to make assumptions like proportional hazards or accelerated failure-time.

6. Recent developments and software implementations

BART is still evolving as variants emerge for new purposes. Zhang et al. (2007) paired BART with a conditional autoregressive (CAR) model for spatial data. Sequential BART is an adaptation to missing value imputation (Xu et al., 2016). Pratola et al. (2017) adapt BART to heteroscedastic data with two ensembles of trees: one that is BART and another that is multiplicative (rather than additive like BART). Hahn et al. (2017) have extended BART to causal inference. Linero & Yang (2018) present what they call SoftBART which creates smooth BART functions. George et al. (2019) marry BART with Dirichlet Process Mixtures to avoid the Normally distributed errors assumption. Big data sample sizes require developments which are helpful in using BART with high perfomance computing frameworks such as the Message Passing Interface (Walker & Dongarra, 1996; Gabriel et al., 2004): two such approaches are the

Consensus MCMC (Pratola et al., 2014) and the Modified Likelihood Inflating Sampling Algorithm (Entezari et al., 2018).

Since BART is essentially a computational approach, software is necessary for its use. As of this writing, several R package implementations of BART are available on the Comprehensive R Archive Network (CRAN): *BART* (McCulloch et al., 2019), *bartMachine* (Bleich et al., 2014), *BayesTree* (Chipman & McCulloch, 2016) and *dbarts* (Dorie et al., 2016).

Related Articles

Accelerated Failure-time Models; Classification and Regression Tree Methods; Cox Proportional Hazard Model; Discrete Survival-time Models; Markov Chain Monte Carlo (MCMC, Metropolis-Hastings, Gibbs Sampling); Nonparametric Bayes; Overfitting (Overtraining); Regression Trees; Tree-structured Statistical Methods.

References

Albert, J & Chib, S (1993), 'Bayesian analysis of binary and polychotomous response data,' *Journal of the American Statistical Association*, **88**, pp. 669–79.

Baldi, P & Brunak, S (2001), *Bioinformatics: the machine learning approach*, MIT Press, Cambridge, MA.

Bleich, J, Kapelner, A, George, EI & Jensen, ST (2014), 'Variable selection for BART: an application to gene regulation,' *The Annals of Applied Statistics*, **8**(3), pp. 1750–1781.

Bonato, V, Baladandayuthapani, V, Broom, BM, Sulman, EP, Aldape, KD & Do, KA (2011), 'Bayesian ensemble methods for survival prediction in gene expression data,' *Bioinformatics*, **27**(3), pp. 359–367.

Breiman, L (1996), 'Bagging predictors,' *Machine learning*, **24**(2), pp. 123–140.

Breiman, L (2001), 'Random forests,' *Machine learning*, **45**(1), pp. 5–32.

Breiman, L (2017), *Classification and regression trees*, Routledge, New York, NY.

Cai, J & Zeng, D (2014), 'Cox proportional hazard model,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat06880].

Chipman, H & McCulloch, R (2016), *BayesTree: Bayesian Additive Regression Trees*, [https://CRAN.R-project.org/package=BayesTree].

Chipman, HA, George, EI & McCulloch, RE (1998), 'Bayesian CART model search,' *Journal of the American Statistical Association*, **93**(443), pp. 935–948.

Chipman, HA, George, EI & McCulloch, RE (2010), 'BART: Bayesian Additive Regression Trees,' *Annals of Applied Statistics*, **4**, pp. 266–98.

Chipman, HA, George, EI & McCulloch, RE (2013), 'Bayesian regression structure discovery,' *Bayesian Theory and Applications,(Eds, P. Damien, P. Dellaportas, N. Polson, D. Stephens), Oxford University Press, USA*.

Cook, J & Ranstam, J (2016), 'Overfitting,' *British Journal of Surgery*, **103**(13), pp. 1814–1814.

Cristianini, N (2014), 'Overfitting (Overtraining),' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9780471650126.dob0513.pub2].

Denison, DG, Mallick, BK & Smith, AF (1998), 'A Bayesian CART algorithm,' *Biometrika*, **85**(2), pp. 363–377.

Dorie, V, Chipman, H & McCulloch, R (2016), *dbarts: Discrete Bayesian Additive Regression Trees Sampler*, [https://CRAN.R-project.org/package=dbarts].

Dunson, DB (2017), 'Nonparametric Bayes,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat07905].

Efron, B, Hastie, T, Johnstone, I & Tibshirani, R (2004), 'Least angle regression,' *The Annals of statistics*, **32**(2), pp. 407–499.

Entezari, R, Craiu, RV & Rosenthal, JS (2018), 'Likelihood inflating sampling algorithm,' *Canadian Journal of Statistics*, **46**(1), pp. 147–175.

Fahrmeir, L (2014), 'Discrete survival-time models,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat06012].

Freund, Y & Schapire, RE (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting,' *Journal of computer and system sciences*, **55**(1), pp. 119–139.

Friedman, JH (2001), 'Greedy function approximation: a gradient boosting machine,' *Annals of Statistics*, **29**, pp. 1189–1232.

Frühwirth-Schnatter, S & Frühwirth, R (2010), 'Data augmentation and MCMC for binary and multinomial logit models,' in *Statistical modelling and regression structures*, Springer, pp. 111–132.

Gabriel, E, Fagg, GE, Bosilca, G, Angskun, T, Dongarra, JJ, Squyres, JM, Sahay, V, Kambadur, P, Barrett, B, Lumsdaine, A, Castain, R, Daniel, D, Graham, R & Woodall, T (2004), 'Open MPI: Goals, concept, and design of a next generation MPI implementation,' in *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*, Springer, pp. 97–104.

Gelfand, AE & Smith, AF (1990), 'Sampling-based approaches to calculating marginal densities,' *Journal of the American Statistical Association*, **85**(410), pp. 398–409.

Geman, S & Geman, D (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, pp. 721–741.

George, E, Laud, P, Logan, B, McCulloch, R & Sparapani, R (2019), 'Fully nonparametric Bayesian Additive Regression Trees,' in Jeliazkov, I & Tobias, J (eds.), *Advances in Econometrics, Volume 40: Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling*, Emerald Group Publishing Limited, vol. (in press), [https://www.emeraldinsight.com/series/aeco].

Gramacy, RB & Polson, NG (2012), 'Simulation-based regularized logistic regression,' *Bayesian Analysis*, **7**(3), pp. 567–590.

Hahn, PR, Murray, JS & Carvalho, CM (2017), 'Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects,' *arXiv preprint arXiv:1706.09523*.

Hancock, JM (2014), 'Markov Chain Monte Carlo (MCMC, Metropolis-Hastings, Gibbs Sampling),' *Dictionary of Bioinformatics and Computational Biology*, [https://doi.org/10.1002/9780471650126.dob0973].

Harrison Jr, D & Rubinfeld, DL (1978), 'Hedonic housing prices and the demand for clean air,' *Journal of environmental economics and management*, **5**(1), pp. 81–102.

Hastings, W (1970), 'Monte Carlo sampling methods using Markov chains and their applications,' *Biometrika*, **57**, pp. 97–109.

Henderson, NC, Louis, TA, Rosner, GL & Varadhan, R (2017), 'Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models,' *arXiv preprint arXiv:1706.06611*.

Holmes, C & Held, L (2006), 'Bayesian auxiliary variable models for binary and multinomial regression,' *Bayesian Analysis*, **1**, pp. 145–68.

James, I (2014), 'Accelerated failure-time models,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat06002].

Kindo, BP, Wang, H & Peña, EA (2016), 'Multinomial probit Bayesian Additive Regression Trees,' *Stat*, **5**(1), pp. 119–131.

Krogh, A & Solich, P (1997), 'Statistical mechanics of ensemble learning,' *Physical Review E*, **55**, pp. 811–25.

Kuhn, M & Johnson, K (2013), *Applied Predictive Modeling*, Springer, New York, NY.

LeBlanc, M (2014), 'Regression trees,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat07514.pub2].

Linero, A (2018), 'Bayesian regression trees for high dimensional prediction and variable selection,' *Journal of the American Statistical Association*, **113**(522), pp. 626–36.

Linero, AR & Yang, Y (2018), 'Bayesian regression tree ensembles that adapt to smoothness and sparsity,' *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(5), pp. 1087–1110.

Loh, W (2014), 'Classification and regression tree methods,' *Wiley StatsRef: Statistics Reference Online*, [https://doi.org/10.1002/9781118445112.stat03886].

McCulloch, R & Rossi, PE (1994), 'An exact likelihood analysis of the multinomial probit model,' *Journal of Econometrics*, **64**(1-2), pp. 207–240.

McCulloch, R, Sparapani, R, Gramacy, R, Spanbauer, C & Pratola, M (2019), *BART: Bayesian Additive Regression Trees*, [https://CRAN.R-project.org/package=BART].

McCulloch, RE, Polson, NG & Rossi, PE (2000), 'A Bayesian analysis of the multinomial probit model with fully identified parameters,' *Journal of econometrics*, **99**(1), pp. 173–193.

Metropolis, N, Rosenbluth, AW, Rosenbluth, MN, Teller, AH & Teller, E (1953), 'Equation of state calculations by fast computing machines,' *The journal of chemical physics*, **21**(6), pp. 1087–1092.

Mueller, P (1991), 'A generic approach to posterior integration and Gibbs sampling,' Tech. Rep. 91-09, Purdue University, West Lafayette, Indiana, [http://www.stat.purdue.edu/research/technical_reports/pdfs/1991/tr91-09.pdf].

Mueller, P (1993), 'Alternatives to the Gibbs sampling scheme,' Tech. rep., Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.5613&rep=rep1&type=pdf].

Murray, JS (2017), 'Log-linear Bayesian additive regression trees for categorical and count responses,' *arXiv preprint arXiv:1701.01503*.

Pratola, M, Chipman, H, George, E & McCulloch, R (2017), 'Heteroscedastic BART using multiplicative regression trees,' *arXiv preprint arXiv:1709.07542*.

Pratola, MT (2016), 'Efficient Metropolis–Hastings proposal mechanisms for Bayesian regression tree models,' *Bayesian analysis*, **11**(3), pp. 885–911.

Pratola, MT, Chipman, HA, Gattiker, JR, Higdon, DM, McCulloch, R & Rust, WN (2014), 'Parallel Bayesian Additive Regression Trees,' *Journal of Computational and Graphical Statistics*, **23**(3), pp. 830–52.

Ripley, BD (2007), *Pattern recognition and neural networks*, Cambridge University press.

Sparapani, R, Logan, BR, McCulloch, RE & Laud, PW (2019), 'Nonparametric competing risks analysis using Bayesian Additive Regression Trees (BART),' *Statistical Methods in Medical Research*, **(in press)**, [https://doi.org/10.1177/0962280218822140].

Sparapani, R, Rein, L, Tarima, S, Jackson, T & Meurer, J (2018), 'Nonparametric recurrent events analysis with BART and an application to the hospital admissions of patients with diabetes,' *Biostatistics*, **(in press)**, [https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy032/5061112].

Sparapani, RA, Logan, BR, McCulloch, RE & Laud, PW (2016), 'Nonparametric survival analysis using Bayesian Additive Regression Trees (BART),' *Statistics in medicine*, **35**(16), pp. 2741–53.

Venables, WN & Ripley, BD (2013), *Modern applied statistics with S-PLUS*, Springer Science & Business Media, New York, [`https://cran.r-project.org/package=nnet`].

Walker, DW & Dongarra, JJ (1996), 'MPI: a standard message passing interface,' *Supercomputer*, **12**, pp. 56–68.

Wu, Y, Tjelmeland, H & West, M (2007), 'Bayesian CART: Prior specification and posterior simulation,' *Journal of Computational and Graphical Statistics*, **16**(1), pp. 44–66.

Xu, D, Daniels, MJ & Winterstein, AG (2016), 'Sequential BART for imputation of missing covariates,' *Biostatistics*, **17**(3), pp. 589–602.

Zhang, H, Crowley, J, Sox, HC & Olshen-Jr., RA (2014), 'Tree-structured statistical methods,' *Wiley StatsRef: Statistics Reference Online*, [`https://doi.org/10.1002/9781118445112.stat05678`].

Zhang, S, Shih, YCT & Müller, P (2007), 'A spatially-adjusted Bayesian Additive Regression Tree model to merge two datasets,' *Bayesian Analysis*, **2**(3), pp. 611–633.

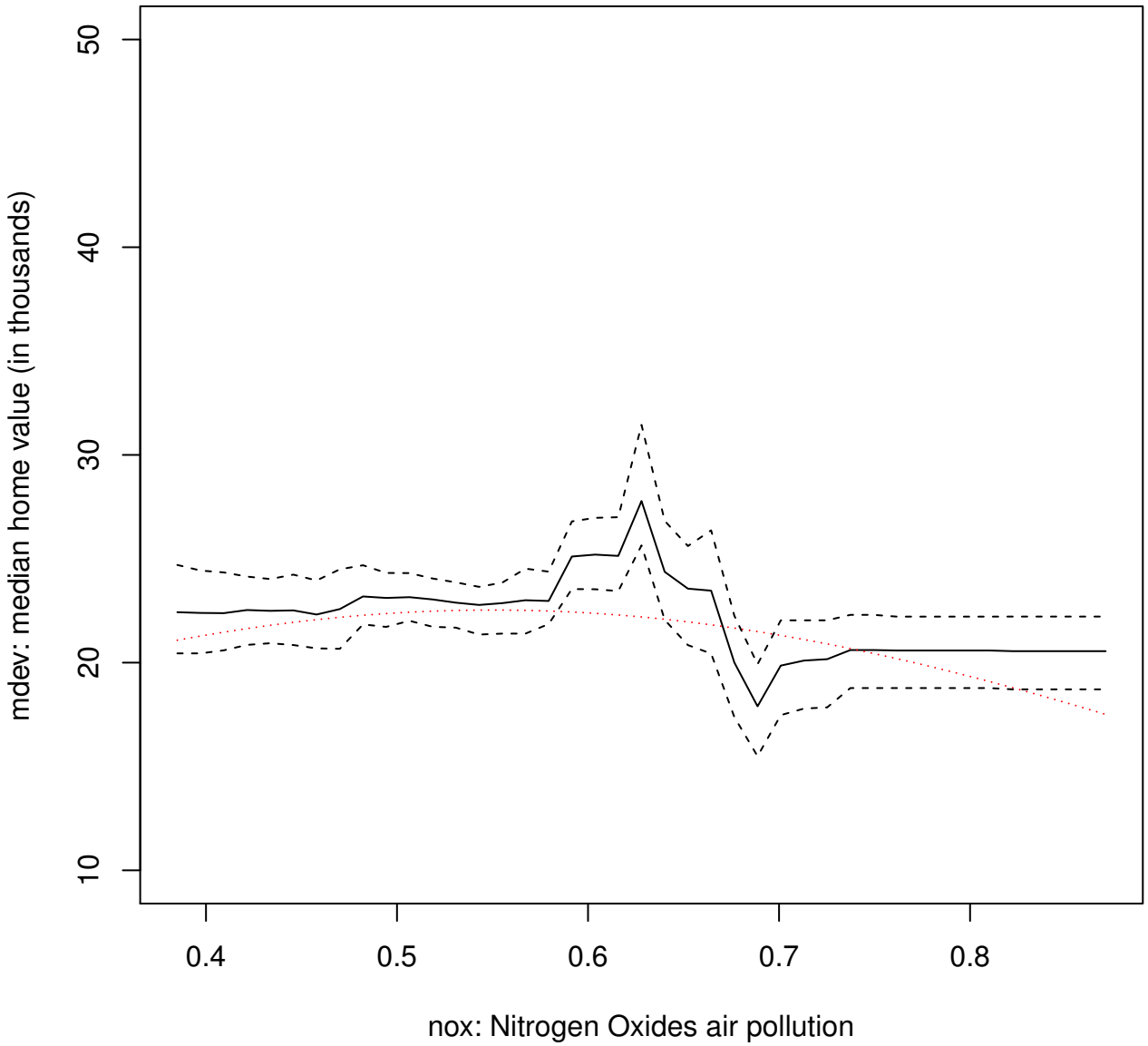

Figure 1. The Boston housing data was compiled from the 1970 US Census where each observation represents a Census tract in Boston with owner-occupied homes. For each tract, we have the median value of owner-occupied homes (in thousands of dollars), `mdev`, and thirteen other covariates including a localized air pollution estimate, the concentration of nitrogen oxides `nox`, which is our primary interest. We summarize the marginal effect of `nox` on `mdev` while aggregating over the other covariates with Friedman's partial dependence function. The marginal estimate and its 95% credible interval are shown. The red line with short dashes comes from the linear regression model of Harrison Jr & Rubinfeld (1978) where a quadratic effect of `nox` with respect to the logarithm of `mdev` is assumed.