BART MCMC updates

Andrew C. Parnell

March 8, 2019

Overall algorithm details

The equations referred to below are displayed later in this document.

Algorithm 1: BART Markov chain Monte Carlo

Data: Target variables y (length n; standardised), feature matrix X (n rows and p columns

Result: Posterior list of trees T, values of τ, fitted values $\hat{y}$

Initialisation;

Hyper-parameter values of α, β, τ_μ, ν, λ;

Number of trees M;

Number of iterations N;

Initial value $\tau = 1$;

Set trees T_j; $j = 1, \ldots, M$ to stumps;

Set values of μ to 0;

for *iterations i from 1 to N* **do**

  for *trees j from 1 to M* **do**

    Compute partial residuals from y minus predictions of all trees except tree j;

    Grow a new tree T_j^{new} based on grow/prune/change/swap;

    Set l_{new} = log full conditional of new tree T_j^{new} based on Equation 1 plus Equation 4;

    Set l_{old} = log full conditional of old tree T_j based on same equations;

    Set $a = \exp(l_{new} - l_{old})$;

    Generate $u \sim U(0, 1)$;

    if $a > u$ **then**

      Set $T_j = T_j^{new}$;

    end

    Simulate μ values using Equation 3;

  end

  Get predictions $\hat{y}$ from all trees;

  Update τ using Equation 4;

end

1 Updating trees

Suppose there are n observations in a terminal node and suppose that the (partial) residuals in this terminal node are denoted $R_1, \dots, R_n$. The prior distribution for these residuals is:

$$R_1, \dots, R_n | \mu, \tau \sim N(\mu, \tau^{-1})$$

Furthermore the prior on μ is:

$$\mu \sim N(0, \tau_\mu^{-1})$$

Using π to denote a probability distribution, we want to find:

$$\begin{aligned}
\pi(R_1, \dots, R_n | \tau) &= \int \pi(R_1, \dots, R_n | \mu, \tau) \pi(\mu) \partial\mu \\
&\propto \int \prod_{i=1}^{n} \tau^{1/2} e^{-\frac{\tau}{2}(R_i - \mu)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2}\mu^2} \partial\mu \\
&= \int \tau^{n/2} e^{-\frac{\tau}{2}\sum(R_i - \mu)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2}\mu^2} \partial\mu \\
&= \int \tau^{n/2} \tau_\mu^{1/2} e^{-\frac{1}{2}\left[\tau\left\{\sum R_i^2 + n\mu^2 - 2\mu n\bar{R}\right\} + \tau_\mu \mu^2\right]} \partial\mu \\
&= \tau^{n/2} \tau_\mu^{1/2} e^{-\frac{1}{2}\left[\tau \sum R_i^2\right]} \int e^{-\frac{1}{2}Q} \partial\mu
\end{aligned}$$

where

$$\begin{aligned}
Q &= \tau n \mu^2 - 2\tau n \mu \bar{R} + \tau_\mu \mu^2 \\
&= (\tau_\mu + n\tau)\mu^2 - 2\tau n \mu \bar{R} \\
&= (\tau_\mu + n\tau)\left[\mu^2 - \frac{2\tau n \mu \bar{R}}{\tau_\mu + n\tau}\right] \\
&= (\tau_\mu + n\tau)\left[\left(\mu - \frac{2\tau n \bar{R}}{\tau_\mu + n\tau}\right)^2 - \left(\frac{\tau n \bar{R}}{\tau_\mu + n\tau}\right)^2\right] \\
&= (\tau_\mu + n\tau)\left(\mu - \frac{2\tau n \bar{R}}{\tau_\mu + n\tau}\right)^2 - \frac{(n\tau\bar{R})^2}{\tau_\mu + n\tau}
\end{aligned}$$

so therefore:

$$
\begin{aligned}
\int e^{-\frac{1}{2}Q}\partial\mu &= \int \exp\left[-\frac{\tau_\mu + n\tau}{2}\left(\mu - \frac{2\tau n\bar{R}}{\tau_\mu + n\tau}\right)^2 + \frac{(n\tau\bar{R})^2}{2(\tau_\mu + n\tau)}\right]\partial\mu \\
&\propto \exp\left[\frac{1}{2}\frac{(\tau n\bar{R})^2}{\tau_\mu + n\tau}\right](\tau_\mu + n\tau)^{-1/2}
\end{aligned}
$$

And finally:

$$
\begin{aligned}
\pi(R_1,\ldots,R_n|\tau) \propto& (\tau_\mu + n\tau)^{-1/2}\tau^{n/2}\tau_\mu^{1/2}\exp\left[\frac{1}{2}\frac{(\tau n\bar{R})^2}{\tau_\mu + n\tau}\right]\exp\left[-\frac{\tau}{2}\sum R_i^2\right] \\
=& \tau^{n/2}\left(\frac{\tau_\mu}{\tau_\mu + n\tau}\right)^{1/2}\exp\left[-\frac{\tau}{2}\left\{\sum R_i^2 - \frac{\tau(n\bar{R})^2}{\tau_\mu + n\tau}\right\}\right]
\end{aligned}
$$

Including multiple terminal nodes

When we put back in terminal nodes we write R_{ji} where j is the terminal node and i is still the observation, so in terminal node j we have partial residuals $R_{j1},\ldots,R_{jn_j}$. When we have $j=1,\ldots,b$ terminal nodes the full conditional distribution is then:

$$
\prod_{j=1}^{b}\pi(R_{j1},\ldots,R_{jn_j}|\tau) \propto \prod_{j=1}^{b}\left\{\tau^{n_j/2}\left(\frac{\tau_\mu}{\tau_\mu + n_j\tau}\right)^{1/2}\exp\left[-\frac{\tau}{2}\left\{\sum_{i=1}^{n_j}R_{ji}^2 - \frac{\tau(n_j\bar{R}_j)^2}{\tau_\mu + n_j\tau}\right\}\right]\right\}
$$

which on the log scale gives:

$$
\sum_{j=1}^{b}\left\{\frac{n_j}{2}\log(\tau) + \frac{1}{2}\log\left(\frac{\tau_\mu}{\tau_\mu + n_j\tau}\right) - \frac{\tau}{2}\left[\sum_{i=1}^{n_j}R_{ji}^2 - \frac{\tau(n_j\bar{R}_j)^2}{\tau_\mu + n_j\tau}\right]\right\}
$$

This can be simplified further to give:

$$
\frac{n}{2}\log(\tau) + \frac{1}{2}\sum_{j=1}^{b}\log\left(\frac{\tau_\mu}{\tau_\mu + n_j\tau}\right) - \frac{\tau}{2}\sum_{j=1}^{b}\sum_{i=1}^{n_j}R_{ji}^2 + \frac{\tau^2}{2}\sum_{j=1}^{b}\frac{S_j^2}{\tau_\mu + n_j\tau} \tag{1}
$$

where $S_j = \sum_{i=1}^{n_j}R_{ji}$

Updating μ

The full conditional for μ_j (the terminal node parameters for node j) is similar to the above but without the integration:

$$
\begin{aligned}
\pi(\mu_j|\ldots) &\propto \prod_{i=1}^{n_j} \tau^{1/2} e^{-\frac{\tau}{2}(R_{ji}-\mu_j)^2} \tau_\mu^{1/2} e^{-\frac{\tau_\mu}{2}\mu_j^2} \\
&\propto e^{-\frac{\tau}{2}\sum_{i=1}^{n_j}(R_{ji}-\mu_j)^2} e^{-\frac{\tau_\mu}{2}\mu_j^2} \\
&\propto e^{-\frac{\tau}{2}\left[n_j\mu_j^2 - 2\mu_j \sum_{i=1}^{n_j} R_{ji}\right] - \frac{\tau_\mu}{2}\mu_j^2} \\
&\propto e^{-\frac{Q}{2}}
\end{aligned}
$$

Now:

$$
\begin{aligned}
Q &= n_j\tau\mu_j^2 - 2\mu_j\tau S_j + \tau_\mu\mu_j^2 \\
&= (n_j\tau + \tau_\mu)\mu_j^2 - 2\tau\mu_j S_j \\
&= (n_j\tau + \tau_\mu)\left[\mu_j^2 - \frac{2\tau\mu_j S_j}{n_j\tau + \tau_\mu}\right] \\
&\propto (n_j\tau + \tau_\mu)\left[\mu_j - \frac{\tau\mu_j S_j}{n_j\tau + \tau_\mu}\right]^2
\end{aligned}
$$

so therefore:

$$
\mu_j|\ldots \sim N\left(\frac{\tau S_j}{n_j\tau + \tau_\mu}, (n_j\tau + \tau_\mu)^{-1}\right) \tag{2}
$$

Update for τ

I am using the shape/rate parameterisation of the gamma with prior $\tau \sim Ga(\nu/2, \nu\lambda/2)$. Letting μ_i be the prediction of the ith observation we get:

$$
\pi(\tau|\ldots) \propto \prod_{i=1}^{n} \tau^{1/2} e^{-\frac{\tau}{2}(y_i-\mu_i)^2} \tau^{\nu/2-1} e^{-\tau\nu\lambda/2}
$$

Letting $S = \sum_{i=1}^{n} (y_i - \mu_i)^2$ we get:

$$\begin{aligned}
\pi(\tau | \ldots) &\propto \tau^{n/2} e^{-\frac{\tau}{2} S} \tau^{\nu/2-1} e^{-\tau\nu\lambda/2} \\
&= \tau^{(n+\nu)/2-1} e^{-\frac{\tau}{2}(S+\nu\lambda)}
\end{aligned}$$

so

$$\tau | \ldots \sim Ga\left(\frac{n+\nu}{2}, \frac{S+\nu\lambda}{2}\right) \tag{3}$$

Tree prior

The tree prior used by BARTMachine says that the probability of a node being non-terminal is:

$$P(\text{node is non-terminal}) = \alpha(1+d)^{-\beta}$$

So the probability of a node being terminal is 1 minus this. A stump just has probability 1 - α. For Bart Machine $\alpha = 0.95$ and $\beta = 2$

Thus for a tree with k non-terminal nodes and b terminal nodes we have:

$$P = \prod_{i=1}^{b} \left[1 - \alpha(1+d_i^t)^{-\beta}\right] \prod_{i=1}^{k} \left[\alpha(1+d_i^{nt})^{-\beta}\right]$$

where d_i^t is the depth of the ith terminal node and d_i^{nt} is the depth of the ith non-terminal node. On the log scale this gives:

$$\log P = \sum_{i=1}^{b} \left[\log\left(1 - \alpha(1+d_i^t)^{-\beta}\right)\right] + \sum_{i=1}^{k} \left[\log(\alpha) - \beta \log(1+d_i^{nt})\right] \tag{4}$$

2-class classification version

A 2-class classification version can be created by using the latent probit representation of the binomial distribution. We assume the response variable y_i takes values 0 or 1, and depends on a latent variable z_i

upon which the BART model is applied:

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

now $z_i \sim N\left(\sum_{j=1}^m g_j(x_i), 1\right)$ where g_j are the predicted values from the individual trees j. Note that this model implies that $\tau = 1$ and is not updated.

The update for z_i at the end of each set of tree updates is:

$$z_i|y_i = 1 \quad \sim \quad max\left[N\left(\sum_{j=1}^m g_j(x_i), 1\right), 0\right] \tag{5}$$

$$z_i|y_i = 0 \quad \sim \quad min\left[N\left(\sum_{j=1}^m g_j(x_i), 1\right), 0\right] \tag{6}$$

Otherwise all updates (trees and means) are the same. However, finding the predicted probabilities at the end of the algorithm requires $\phi^{-1}(\sum_{j=1}^m g_j(x_i))$ where ϕ is the normal cdf function. These probabilities can be then be used to create misclassification tables, ROC curves, etc.

Mean and variance changing

An alternative, slightly more flexible, model can be created by giving each terminal node it's own precision value. Thus for tree j we have:

$$R_{1j}, \ldots, R_{jn_j}|\mu_j, \tau_j \sim N(\mu_j, \tau_j^{-1})$$

The maths is much simplified by setting the prior on μ_j as:

$$\mu_j \sim N(0, (a\tau_j)^{-1}).$$

Chipman et al (1998) set $a = 1/3$ for a single tree setting, so perhaps $a = 1/(3m)$ might be more appropriate for a multi-tree version with m trees.

We now have a prior for each terminal node:

$$\tau_j \sim Ga(\nu/2, \nu\lambda/2)$$

Chipman et al (1998) suggest making ν and λ functions of tree complexity but we don't do that here.

We don't need the subscripts j for an individual tree so the shortcut to what we require is:

$$\begin{aligned}\pi(R_1,\ldots,R_n|\ldots) &= \int\int \pi(R_1,\ldots,R_n|\mu,\tau)\pi(\mu)\pi(\tau)\partial\mu\,\partial\tau \\ &\propto \int\int \left[\prod_{i=1}^{n} \tau^{1/2}\exp\left(-\frac{\tau}{2}(R_i-\mu)^2\right)\right]\tau^{1/2}\exp\left(-\frac{a\tau}{2}\mu^2\right)\tau^{\nu/2-1}\exp\left(-\frac{\tau\nu\lambda}{2}\right)\partial\mu\,\partial\tau \\ &\propto \int\int \tau^{n/2}\exp\left(-\frac{\tau}{2}\sum(R_i-\mu)^2\right)\tau^{1/2}\exp\left(-\frac{a\tau}{2}\mu^2\right)\tau^{\nu/2-1}\exp\left(-\frac{\tau\nu\lambda}{2}\right)\partial\mu\,\partial\tau\end{aligned}$$

Note that $\sum(R_i-\mu)^2$ can be re-written as $SS_R + n(\mu-\bar{R})^2$ where $SS_R = \sum(R_i-\bar{R})^2$.

We now get

$$\pi(R_1,\ldots,R_n|\ldots) \propto \int\int \tau^{(n+\nu-1)/2}\exp\left(-\frac{\tau}{2}\left\{SS_R + n(\mu-\bar{R})^2 + a\mu^2 + \nu\lambda\right\}\right)\partial\mu\,\partial\tau$$

Simplifying the exponent to separate out the terms with μ gives:

$$n(\mu-\bar{R})^2 + a\mu^2 = (a+n)\left[\mu - \frac{n\bar{R}}{n+a}\right]^2 + \frac{an\bar{R}^2}{n+a}$$

So we can perform the integrand with respect to μ first:

$$\pi(R_1,\ldots,R_n|\ldots) \propto \int \tau^{(n+\nu-2)/2}\exp\left(-\frac{\tau}{2}\left\{SS_R + \nu\lambda + \frac{an\bar{R}^2}{n+a}\right\}\right)\int \tau^{1/2}\exp\left(-\frac{\tau(a+n)}{2}\left[\mu - \frac{n\bar{R}}{n+a}\right]^2\right)\partial\mu\,\partial\tau$$

The integrand wrt μ is proportional to $(a+n)^{-1/2}$ and also provides the full conditional for μ (putting back in the subscripts):

$$\mu_j|\ldots \sim N\left(\frac{n_j\bar{R}_j}{n_j+a}, [\tau_j(a+n_j)]^{-1}\right)$$

Next we have:

$$\pi(R_1,\ldots,R_n|\ldots) \propto \int \frac{\tau^{(n+\nu-2)/2}}{(a+n)^{1/2}}\exp\left(-\frac{\tau}{2}\left\{SS_R + \nu\lambda + \frac{an\bar{R}^2}{n+a}\right\}\right)\partial\tau$$

This also provides the complete conditional for (re-inserting subscripts) τ_j:

$$\tau_j|\ldots \sim Ga\left(\frac{n_j+\nu}{2}, \frac{1}{2}\left[SS_{R_j} + \nu\lambda + \frac{an_j\bar{R}_j^2}{n_j+a}\right]\right)$$

Finally we have (again with subscripts):

$$\pi(R_{1j},\ldots,R_{jn_j}|\ldots) \propto (a+n_j)^{-1/2}\,\Gamma\left(\frac{n_j+\nu}{2}\right)\left[SS_{R_j} + \nu\lambda + \frac{an_j\bar{R}_j^2}{n_j+a}\right]^{-\left(\frac{n_j+\nu}{2}\right)}$$

With multiple terminal nodes for a tree we have:

$$\begin{aligned}\prod_{j=1}^{b}\pi(R_{1j},\ldots,R_{jn_j}|\ldots) &= \prod_{j=1}^{b}(a+n_j)^{-1/2}\,\Gamma\left(\frac{n_j+\nu}{2}\right)\left[SS_{R_j} + \nu\lambda - \frac{an_j\bar{R}_j^2}{n_j+a}\right]^{-(\frac{n_j+\nu}{2})} \\ &= \left[\prod_{j=1}^{b}(a+n_j)^{-1/2}\right]\left[\prod_{j=1}^{b}\Gamma\left(\frac{n_j+\nu}{2}\right)\right]\left\{\prod_{j=1}^{b}\left[SS_{R_j} + \nu\lambda + \frac{an_j\bar{R}_j^2}{n_j+a}\right]^{-\left(\frac{n_j+\nu}{2}\right)}\right\}\end{aligned}$$

which on the log scale is:

$$\log\prod_{j=1}^{b}\pi_j = -\frac{1}{2}\sum_{j=1}^{b}\log(a+n_j) + \sum_{j=1}^{b}\log\Gamma\left(\frac{n_j+\nu}{2}\right) - \sum_{j=1}^{b}\left(\frac{n_j+\nu}{2}\log\left[SS_{R_j} + \nu\lambda + \frac{an_j\bar{R}_j^2}{n_j+a}\right]\right)$$