

Appendix for Bayesian additive regression trees and the General BART model

Yaoyuan Vincent Tan and Jason Roy

July 5, 2019

A Parameters for BART

The parameters for continuous outcomes BART that needs to be set are: α , β , μ_μ , σ_μ , ν , and λ . These parameters are constructed as a mix of apriori fixed and data-driven. For α and β , the default values of $\alpha = 0.95$ and $\beta = 2$ provide a balanced penalizing effect for the probability of a node splitting (Chipman et al., 2010). For μ_μ and σ_μ , they are set such that $E[y|x] \sim N(m\mu_\mu, m\sigma_\mu^2)$ assigns high probability to the interval $(\min(y), \max(y))$. This can be achieved by defining v such that $\min(y) = m\mu_\mu - v\sqrt{m}\sigma_\mu$ and $\max(y) = m\mu_\mu + v\sqrt{m}\sigma_\mu$. For ease of posterior distribution calculation, y is transformed to become $\tilde{y} = \frac{y - \frac{\min(y) + \max(y)}{2}}{\max(y) - \min(y)}$. This results in $\tilde{y} \in (-0.5, 0.5)$ where $\min(y) = -0.5$ and $\max(y) = 0.5$. This has the effect of allowing the hyperparameter μ_μ to be set as 0 and σ_μ to be determined as $\sigma_\mu = \frac{0.5}{v\sqrt{m}}$ where v is to be chosen. For $v = 2$, $N(m\mu_\mu, m\sigma_\mu^2)$ assigns a prior probability of 0.95 to the interval $(\min(y), \max(y))$ and is the default value. Finally for ν and λ , the default value for ν is 3 and λ is the value such that $P(\sigma^2 < s^2; \nu, \lambda) = 0.9$ where s^2 is the estimated variance of the residuals from the multiple linear regression with y as the outcomes and x as the covariates.

For binary outcomes, the α and β parameters are the same but the μ_μ and σ_μ parameters are specified differently from continuous outcomes BART. To set the parameters for μ_μ and σ_μ , we set $\mu_\mu = 0$ and $\sigma_\mu = \frac{3}{v\sqrt{m}}$ where $v = 2$ would result in an approximate 95% probability that draws of $f(x) = \sum_{j=1}^m g(x; T_j, M_j)$ will be within $(-3.0, 3.0)$. No transformation of the latent variable z would be needed although it should be noted that this setup shrinks $f(x)$ toward 0 (See main paper Section 2.2).

B Posterior distributions for μ_{ji} and σ^2 in BART

B.1 $p(\mu_{ji}|T_j, \sigma, r_j)$

Let $r_{ji} = (r_{ji1}, \dots, r_{jin_i})^T$ be a subset from r_j where n_i is the number of r_{jih} s allocated to the terminal node with parameter μ_{ji} and h indexes the subjects allocated to the terminal

node with parameter μ_{ji} . We note that $r_{ji}|T_j, \mu_{ji}, \sigma \sim N(\mu_{ji}, \sigma^2)$ and $\mu_{ji}|T_j \sim N(\mu_\mu, \sigma_\mu^2)$. Then the posterior distribution of μ_{ji} is given by

$$\begin{aligned}
p(\mu_{ji}|T_j, \sigma, r_j) &\propto p(r_{ji}|T_j, \mu_{ji}, \sigma)p(\mu_{ji}|T_j) \\
&\propto \exp\left[-\frac{\sum_h (r_{jih} - \mu_{ji})^2}{2\sigma^2}\right] \exp\left[-\frac{(\mu_{ji} - \mu_\mu)^2}{2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{(n_i\sigma_\mu^2 + \sigma^2)\mu_{ji}^2 - 2(\sigma_\mu^2 \sum_h r_{jih} + \sigma^2\mu_\mu)\mu_{ji}}{2\sigma^2\sigma_\mu^2}\right] \\
&\propto \exp\left[-\frac{(\mu_{ji} - \frac{\sigma_\mu^2 \sum_h r_{jih} + \sigma^2\mu_\mu}{n_i\sigma_\mu^2 + \sigma^2})^2}{2\frac{\sigma^2\sigma_\mu^2}{n_i\sigma_\mu^2 + \sigma^2}}\right]
\end{aligned}$$

where $\sum_h (r_{jih} - \mu_{ji})^2$ is the summation of the squared difference between the parameter μ_{ji} and the r_{jih} s allocated to the terminal node with parameter μ_{ji} .

B.2 $p(\sigma^2|(T_1, M_1), \dots, (T_m, M_m), y)$

Let $y = (y_1, \dots, y_n)^T$ with $\sigma^2 \sim IG(\frac{\nu}{2}, \frac{\nu\lambda}{2})$. We obtain the posterior draw of σ as follows

$$\begin{aligned}
p(\sigma^2|(T_1, M_1), \dots, (T_m, M_m), y) &\propto p(y|(T_1, M_1), \dots, (T_m, M_m), \sigma)p(\sigma^2) \\
&= \left\{ \prod (\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(y - f(x))^2}{2\sigma^2}\right] \right\} (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp\left(-\frac{\nu\lambda}{2\sigma^2}\right) \\
&= (\sigma^2)^{-(\frac{\nu+n}{2}+1)} \exp\left[-\frac{\nu\lambda + \sum (y - f(x))^2}{2\sigma^2}\right].
\end{aligned}$$

C Metropolis-Hastings ratio for the grow and prune step

This section is modified from Appendix A of Kapelner and Bleich (Kapelner and Bleich (2016)). Note that

$$\alpha(T_j, T_j^*) = \min \left\{ 1, \frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} \frac{P(r_j|x, T_j^*, M_j)}{P(r_j|x, T_j, M_j)} \frac{P(T_j^*)}{P(T_j)} \right\}.$$

where $\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)}$ is the transition ratio, $\frac{P(r_j|x, T_j^*, M_j)}{P(r_j|x, T_j, M_j)}$ is the likelihood ratio, and $\frac{P(T_j^*)}{P(T_j)}$ is the tree structure ratio of Kapelner and Bleich, Appendix A. We now present the explicit formula for each ratio under the grow and prune proposal.

C.1 Grow proposal

C.1.1 Transition ratio

$q(T_j^*, T_j)$ indicates the probability of moving from T_j to T_j^* i.e. selecting and terminal node and growing two children from T_j . Hence,

$$\begin{aligned} P(T_j^*|T_j) &= P(\text{grow})P(\text{selecting terminal node to grow from}) \times \\ &\quad P(\text{selecting covariate to split from}) \times \\ &\quad P(\text{selecting value to split on}) \\ &= P(\text{grow}) \frac{1}{b_j} \frac{1}{p} \frac{1}{\eta}. \end{aligned}$$

In the above equation, $P(\text{grow})$ can be decided by the researcher although the default provided is 0.25. b_j is the number of available terminal nodes to split on in T_j , p is the number of variables left in the partition of the chosen terminal node, and η is the number of unique values left in the chosen variable after adjusting for the parents splits.

$q(T_j, T_j^*)$ on the other hand indicates a pruning move which involves the probability of selecting the correct internal node to prune on such that T_j^* becomes T_j . This is given as

$$\begin{aligned} P(T_j|T_j^*) &= P(\text{prune})P(\text{selecting the correct internal node to prune}) \\ &= P(\text{prune}) \frac{1}{w_2^*} \end{aligned}$$

where w_2^* denotes the number of internal nodes which have only two children terminal nodes.

This gives a transition ratio of

$$\frac{q(T_j^*, T_j)}{q(T_j, T_j^*)} = \frac{P(T_j^*|T_j)}{P(T_j|T_j^*)} = \frac{P(\text{prune})}{P(\text{grow})} \frac{b_j p \eta}{w_2^*}.$$

If there are no variables with two or more unique values, this transition ratio will be set to 0.

C.1.2 Likelihood ratio

Since the rest of the tree structure will be the same between T_j^* and T_j except for the terminal node where the two children are grown, we need only concentrate on this terminal node. Let l be the selected node and l_L and l_R be the two children of the grow step. Then

$$\begin{aligned} \frac{P(r_j|x, T_j^*, M_j)}{P(r_j|x, T_j, M_j)} &= \frac{P(r_{l_{(L,1)},j}, \dots, r_{l_{(L,n_L)},j}|\sigma^2)P(r_{l_{(R,1)},j}, \dots, r_{l_{(R,n_R)},j}|\sigma^2)}{P(r_{1,j}, \dots, r_{n_l,j}|\sigma^2)} \\ &= \sqrt{\frac{\sigma^2(\sigma^2 + n_l\sigma_\mu^2)}{(\sigma^2 + n_L\sigma_\mu^2)(\sigma^2 + n_R\sigma_\mu^2)}} \exp \left[\frac{\sigma_\mu^2}{2\sigma^2} \left(\frac{(\sum_{k=1}^{n_L} r_{l_{(L,k)},j})^2}{\sigma^2 + n_L\sigma_\mu^2} \right. \right. \\ &\quad \left. \left. + \frac{(\sum_{k=1}^{n_R} r_{l_{(R,k)},j})^2}{\sigma^2 + n_R\sigma_\mu^2} - \frac{(\sum_{k=1}^{n_l} r_{l_{(l,k)},j})^2}{\sigma^2 + n_l\sigma_\mu^2} \right) \right]. \end{aligned}$$

C.1.3 Tree structure ratio

Because the T_j can be specified using three aspects, we let $P_{SPLIT}(\theta)$ denote the probability that a selected node θ will split and $P_{RULE}(\theta)$ denote the probability that a certain variable and value is selected. Then based on $P_{SPLIT}(\theta) \propto \frac{\alpha}{(1+d_\theta)^\beta}$ and because T_j and T_j^* only differs at the children nodes, we have

$$\begin{aligned}
\frac{P(T_j^*)}{P(T_j)} &= \frac{\prod_{\theta \in H_{terminals}^*} (1 - P_{SPLIT}(\theta)) \prod_{\theta \in H_{internals}^*} P_{SPLIT}(\theta) \prod_{\theta \in H_{internals}^*} P_{RULE}(\theta)}{\prod_{\theta \in H_{terminals}} (1 - P_{SPLIT}(\theta)) \prod_{\theta \in H_{internals}} P_{SPLIT}(\theta) \prod_{\theta \in H_{internals}} P_{RULE}(\theta)} \\
&= \frac{[1 - P_{SPLIT}(\theta_L)][1 - P_{SPLIT}(\theta_R)]P_{SPLIT}(\theta)P_{RULE}(\theta)}{1 - P_{SPLIT}(\theta)} \\
&= \frac{(1 - \frac{\alpha}{(1+d_{\theta_L})^\beta})(1 - \frac{\alpha}{(1+d_{\theta_R})^\beta})\frac{\alpha}{(1+d_\theta)^\beta}\frac{1}{p}\frac{1}{\eta}}{1 - \frac{\alpha}{(1+d_\theta)^\beta}} \\
&= \alpha \frac{(1 - \frac{\alpha}{(2+d_\theta)^\beta})^2}{[(1 + d_\theta)^\beta - \alpha]p\eta}
\end{aligned}$$

because $d_{\theta_L} = d_{\theta_R} = d_\theta + 1$.

C.2 Prune proposal

Since prune is the direct opposite of the grow proposal, the explicit formula of $\alpha(T_j, T_j^*)$ will just be the inverse of the grow proposal.

References

- Chipman, H., George, E., and McCulloch, R. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics* **4**, 266–298.
- Kapelner, A. and Bleich, J. (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software* **70**, 1–40.