Influential Observations in Bayesian Regression Tree Models

M. T. Pratola*
Department of Statistics
The Ohio State University
Columbus, OH 43065
mpratola@stat.osu.edu

E. I. George
Department of Statistics
The Wharton School
University of Pennsylvania
edgeorge@wharton.upenn.edu

R. E. McCulloch
School of Mathematical and Statistical Sciences
Arizona State University
Robert.E.McCulloch@gmail.com

Abstract

BCART (Bayesian Classification and Regression Trees) and BART (Bayesian Additive Regression Trees) are popular modern regression models. Their popularity is intimately tied to the ability to flexibly model complex responses depending on high-dimensional inputs while simultaneously being able to quantify uncertainties. However, surprisingly little work has been done to evaluate the sensitivity of these modern regression models to violations of modeling assumptions. In particular, we consider influential observations and propose methods for detecting influentials and adjusting predictions to not be unduly affected by such problematic data. We consider two detection diagnostics for Bayesian tree models, one an analogue of Cook's distance and the other taking the form of a divergence measure, and then propose an importance sampling algorithm to re-weight previously sampled posterior draws so as to remove the effects of influential data. Finally, our methods are demonstrated on real-world data where blind application of models can lead to poor predictions.

1 Introduction

In the contemporary approach to data-driven problem solving, statistical models have received increasing attention and popularity as a means for arriving at answers to complex research, science and industry questions. As datasets have increased in size with the transition to the "big-data" era, the complexity and scalability of statistical models have seen rapid advances. Popular models include neural networks (Ghugare et al., 2014), random forests (Breiman, 2001) and localized Gaussian Processes (Gramacy and Apley, 2015). In problems where uncertainty quantification is deemed necessary, Bayesian methods have come to the fore, such as the Bayesian variants of neural networks (MacKay, 1995), Bayesian localized GPs (Liu et al., 2020) and Bayesian Regression Tree models (Chipman et al., 2010; Pratola, 2016; Horiguchi et al., 2021, 2022). However, there has been a conspicuous disconnect in terms of tools that support the application of such complex models when compared to their humble, small-dataset, low-dimensional ancestors.

For example, in linear regression, students are taught an extensive array of tools for validating modeling assumptions in the classical setting, such as residual diagnostics, outlier detection and

*www.matthewpratola.com

NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling.

influence metrics (Weisberg, 2013). Surprisingly, such supporting tools have not received the same attention in the development of modern variants of statistical models. The assumption, it seems, is that in the big-data setting such issues are of lesser concern. We have found this assumption to be incorrect.

Our focus in this paper is on Bayesian classification and regression tree (BCART) models (Chipman et al., 1998), and the Bayesian Additive Regression Tree (BART) model of Chipman et al. (2010) in particular. This class of models is currently receiving much attention in the research community (e.g. Tan and Roy, 2019; Hahn et al., 2020; Pratola and Higdon, 2014; Horiguchi et al., 2022). Our work arose out of a simple curiosity: can BCART or BART models be negatively affected by a problematic observation, i.e. an observation that can be influential or is an outlier (or both)? We find that the answer is yes, and investigate two perspectives to handle such problematic observations:

i. identification of problematic observations;

ii. model (posterior) adjustment given identified problematic observations.

In this work, we propose two approaches for the identification problem (i). First, a direct extension of Cook's distance to the regression tree model setting is outlined, and has the benefit of providing an easy and sensible interpretation. Second, a divergence-based metric is proposed. For the adjustment problem (ii), we explore two alternatives: the simple (but wasteful) dropped-observation approach, and an importance-sampling approach that reweights posterior expectations to account for the problematic observations without going so far as to completely remove them.

2 Influence Diagnostics for Trees

We now outline two diagnostic tests for the detection of problematic observations. The first is a direct application of Cook's distance to Bayesian regression trees, while the second is a divergence-based approach. Throughout, we denote the parameters as $\Theta = \left(\{\mathbf{T}^{(j)}\}_{j=1}^m, \{\mathbf{M}^{(j)}\}_{j=1}^m, \sigma^2\right)$ where $m = 1$ trees simplifies to BCART and $\mathbf{M}^{(j)} = (\mu_{j1}, \ldots)$ are the terminal node parameters of tree j.

2.1 Conditional Cook's Distance for Regression Trees

Conditioning on $\mathbf{T}$, a single regression tree can be expressed in the usual linear form as $g(x; T, M) = \sum_{b=1}^B \mu_b \mathrm{I}_b(x)$ where B is the total number of terminal nodes in the tree and $\mathrm{I}_b(x)$ is the indicator function taking the value 1 when x maps to the hyperrectangle defined by terminal node b, and 0 otherwise. Then, one can express the Cook's distance as

$$D_i \quad = \quad \underbrace{\frac{1}{B}}_{\text{Tree Complexity}} \quad \times \quad \underbrace{\left(\frac{e_i}{\sigma}\right)^2}_{\text{Normalized Residual}} \quad \times \underbrace{\frac{n_{(i)}}{(n_{(i)}-1)^2}}_{\text{Node Purity}} \tag{1}$$

where e_i is the residual for observation i, and $n_{(i)}$ is the number of observations in the terminal node to which observation i maps. Note here that in comparison to the classical Cook's distance, we have replaced $\hat{\sigma}$ with the parameter itself, for which we have samples. This form of D_i provides helpful interpretations. For instance, it is a decreasing function of the number of terminal nodes, B, but on the other hand it increases as node purity increases (i.e. as $n_{(i)}$ becomes small) and in particular will blow-up when $n_{(i)} - 1 = 0$. To arrive at an overall estimate, we take the posterior mean, $\widehat{E[D_i|\mathbf{Y}]} = \frac{1}{N}\sum_{k=1}^N D_i^{(k)}$ where each $D_i^{(k)}$ is the conditional Cook's distance for the kth posterior sample as defined in equation (1). For the sum-of-trees BART model, we report the average D_i across all of the m tree's in BART's sum.

2.2 Kullback-Liebler Divergence Diagnostic

Recall the Kullback-Liebler divergence from distribution Q to P is defined as $D_{KL}(P||Q) := \int_{-\infty}^{\infty} log\left(\frac{dP}{dQ}\right) dP$ where $D_{KL} \geq 0$ with equality iff $P = Q$. In our context, we propose to take the reference distribution to be the posterior involving all the data, $P := \pi(\Theta|\mathbf{Y})$, while the distribution Q is taken to be the posterior when the potentially problematic data is held out. If we consider the

simplest case of holding out a single obseration y_i, then $Q := \pi(\Theta|\mathbf{Y}_{-i})$. Then, the KL divergence diagnostic has a simple Bayesian interpretation when evaluating the potential for observations to be problematic: if $D_{KL} \sim 0$ then observation y_i is not influential, whereas if $D_{KL} >> 0$ then observation y_i is influential. The advantage is the KL diagnostic informs us about the sensitivity of the entire posterior distribution to the problematic observation instead of only the mean function.

3 Adjusting Predictions via Importance Sampling

While one could use the proposed diagnostics to detect problematic observations and then refit the model with such observations removed from the dataset, for Bayesian models this is a computationally wasteful approach. Instead, Bradlow and Zaslavsky (1997) propose to estimate functions of interest, $g(\Theta)$, using importance sampling as

$$E[g(\Theta)|\mathbf{Y}_{-i}] = \int_{\Theta} g(\Theta)\frac{\pi(\Theta|\mathbf{Y}_{-i})}{\pi(\Theta|\mathbf{Y})}\pi(\Theta|\mathbf{Y})d\Theta.$$

Let $w_{(i)}^{(k)} = \frac{\pi(\Theta=\Theta^{(k)}|\mathbf{Y}_{-i})}{\pi(\Theta=\Theta^{(k)}|\mathbf{Y})} \propto \frac{f(\mathbf{Y}_{-i}|\Theta=\Theta^{(k)})}{f(\mathbf{Y}|\Theta=\Theta^{(k)})}$ be the importance sampling weights of interest when observation i is to be dropped and $\Theta^{(k)} \sim \pi(\Theta|\mathbf{Y})$. Then, $E[g(\Theta)|\mathbf{Y}_{-i}] \approx \sum_{k=1}^{N} w_{(i)}^{(k)} g(\Theta^{(k)}) \Big/ \sum_{k=1}^{N} w_{(i)}^{(k)}$, where the renormalization in the denominator removes the dependence on the proportionality constant $\pi(\mathbf{Y})/\pi(\mathbf{Y}_{-i})$. Intuitively, this importance sampling approach adjusts our posterior samples used in predicting $g(\Theta)$ as if we had instead sampled from $\pi(\Theta|\mathbf{Y}_{-i})$.

3.1 Re-weighting BART Predictions

We can extend the idea of reweighting to draws from the additive tree model of BART. Recall that the BART likelihood involving a sum-of-trees mean function can, conditionally, be equivalently described by a single "super-tree" mean function (Horiguchi et al., 2021), that is

$$\begin{aligned}\pi(\Theta|\mathbf{Y}) \propto f\left(\mathbf{Y}|(\mathbf{T}^{(1)},\mathbf{M}^{(1)}),\ldots,(\mathbf{T}^{(m)},\mathbf{M}^{(m)}),\sigma^2\right)\prod_{k=1}^{m}\pi\left((\mathbf{T}^{(k)},\mathbf{M}^{(k)})\right)\pi\left(\sigma^2\right)\\ = f\left(\mathbf{Y}|\mathcal{S},\sigma^2\right)\prod_{k=1}^{m}\pi\left((\mathbf{T}^{(k)},\mathbf{M}^{(k)})\right)\pi\left(\sigma^2\right),\end{aligned}$$

where $\mathcal{S}$ represents the analogous super-tree representation, i.e. $g(x;\mathcal{S}) \equiv \sum_{k=1}^{m} g(x;(\mathbf{T}^{(k)},\mathbf{M}^{(k)}))$. Note that the prior remains the same, even though we reinterpret the likelihood's sum-of-trees as a new, equivalent, single super-tree. Suppose again that y_i is the problematic observation, observed at input x_i and let $\eta_l^{\mathcal{S}}$ be the terminal node in $\mathcal{S}$ to which x_i maps. Let $\mathbb{X}$ represent the hyperrectangle defined by $\eta_l^{\mathcal{S}}$. Then we have the following.

Proposition: Let x be a prediction input of interest with corresponding predictor $g(\Theta) = E[y(x)|\Theta]$. Let $\mathbb{X}_j$ be the hyperrectangles in each tree $j = 1,\ldots,m$ of the BART ensemble such that $x \in \mathbb{X}_j, \forall j$. Let $\mathbb{X} = \cap_{j=1}^{m}\mathbb{X}_j$ be the hyperrectangle defined as the intersection of all the $\mathbb{X}_j$'s, which corresponds to the supertree terminal node $\eta_l^{\mathcal{S}}$ to which x belongs. Suppose also that the input x_i for influential observation y_i also maps to $\eta_l^{\mathcal{S}}$. Then to predict the response $y(x)$ for all $x \in \mathbb{X}$, the weights are

$$w_{(i)}^{(k)}(x) = \begin{cases}\frac{1}{p(y_i|\mu_l^{\mathcal{S},(k)},\eta_l^{\mathcal{S},(k)},P_l^{\mathcal{S},(k)},\sigma^2)} & \text{if } |\eta_{jl}^{(k)}| - 1 \geq n_0 \text{ for all } j = 1,\ldots,m\\ 0 & \text{otherwise}\end{cases}$$

where $\eta_{jl}^{(k)}$ is the lth terminal node in tree j to which x_i maps in the original sum-of-trees representation, n_0 is the minimum # of observations for terminal nodes. We call this method `int`.

Implementation of this proposition results in a different localized region, say $\mathbb{X}^{(k)}$, for each of the $k = 1,\ldots,N$ posterior realizations. This makes predictions more computationally expensive. A practical alternative is to take some sort of "average" localized region as the single region to reweight, simplifying posterior prediction calculations. A natural choice is the union of the individual regions, say $\overline{\mathbb{X}} = \cup_{k=1}^{N}\mathbb{X}^{(k)}$. We refer to this method as `union-int`.

4 Results

Our motivating example comes from a study of biomass fuels and the application of artificial intelligence models to predicting the Higher Heating Value (HHV) of such fuels based on their molecular makeup (Ghugare et al., 2014). Biomass fuels are the fourth largest source of energy, with the most common sources being solid products such as wood and biomass pellets. However, determining the HHV potential of a biomass fuel involves expensive and time-consuming calorimetric experiments. Instead, a popular alternative is to use mathematical models to approximate the HHV potential of a fuel source based on its makeup of key components. Ghugare et al. (2014) consider a dataset involving $n = 536$ observations with covariates being the amount of carbon, hydrogen, oxygen, nitrogen and sulfur present in the fuel, while the response is the HHV value measured in MJ/kg. The dataset consists of 80 test-set observations and 456 training-set observations. We

Table 1: RMSE performance on the HHV test dataset.

	`default`	`drop`	`global`	`int`	`union-int`	GP	MLP
training set	0.63	0.63	0.76	0.68	0.64	1.086	0.867
test set	1.22	0.98	1.21	1.11	1.06	0.942	0.987

applied BART to the training data using $m = 50$ trees. The RMSE performance of BART is summarized in Table 1, where `default` is the regular BART fit, `drop` is the fit obtained by dropping the influentials detected by the proposed diagnostics, `global` is the reweighting method of Bradlow and Zaslavsky (1997), and `int`, `union-int` are the methods proposed in this paper (using the same detected influentials as `drop`). In addition, the RMSE performance of Ghugare et al. (2014)'s Genetic Programming (GP) and Multilayer Perceptron (MLP) models are also noted. The performance of BART's fit on the training dataset is very strong, decreasing somewhat for most of the reweighting methods. We see `union-int` demonstrating the best performance, nearly matching the training sample performance of the `default` BART fit. In comparison, BART's `default` performance on the test data is significantly worse, and trails the GP and MLP models. Again, `union-int` provides the highest reduction in error for BART, bringing it close to the performance of GP and MLP on the test data. The remaining gap here could likely be explained by the smooth, continous fits of the GP and MLP models which would be a favourable characteristic for this dataset. The `drop` test set performance suggests retraining BART with detected influentials removed is best when feasible.

Of particular interest in Ghugare et al. (2014) is the performance of the models at different regimes of HHV. In particular, they note difficulty in predicting high-HHV performance, and breakdown their performance summary into three ranges of HHV values: 0-16MJ/kg (5 observations), 16-25MJ/kg (70 observations) and 25-36MJ/kg (5 observations). The performance in these ranges is summarized in Table 2 for the `default` BART model as well as the `drop` and `union-int` methods. We see that the pattern obtained confirms Ghugare et al. (2014) description of high HHV being particularly hard to predict. Nonetheless, the `union-int` method improves on the `default` BART fit in all three regimes, and in fact beats the `drop` performance in the 16-25MJ/kg range where most of the observations lie. Still, it is hard to match the performance of GP and MLP in the 0-16MJ/kg and 25-36MJ/kg regimes, but in the high-HHV regime the `drop` method dominates.

Table 2: Range-wise RMSE performance on the HHV test dataset.

Range	`default`	`drop`	`union-int`	GP	MLP
0-16MJ/kg	1.88	1.50	1.70	1.16	0.90
16-25MJ/kg	0.95	0.93	0.89	0.84	0.81
25-36MJ/kg	2.80	1.01	2.07	2.55	1.55

5 Conclusion

In this paper we proposed BART diagnostics for detecting influential observations, and devised reweighting procedures that allow posterior BART samples to be reweighted once influential observations are identified. Overall, we have found a suprising amount of gains can be found by addressing influential observations even though conventional wisdom suggests that highly flexible statistical

learning models like BART are not affected by such problematic observations due to their localized fits. In reality, when faced with large datasets and high-dimensional covariate spaces, the notion of 'local' is very much a misnomer. Therefore, careful application of BART should at minimum include a diagnostic step to detect possibly problematic observations, upon which investigation, removal or the reweighting procedures proposed here can be performed.

References

Bradlow, E. T. and Zaslavsky, A. M. (1997). "Case Influence Analysis in Bayesian Inference." *Journal of Computational and Graphical Statistics*, 6, 3, 314–331.

Breiman, L. (2001). "Random Forests." *Machine Learning*, 45, 5–32.

Chipman, H., George, E., and McCulloch, R. (1998). "Bayesian CART Model Search." *Journal of the American Statistical Association*, 93, 443, 935–960.

— (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4, 1, 266–298.

Ghugare, S. B., Tiwary, S., Elangovan, V., and Tambe, S. S. (2014). "Prediction of higher heating value of solid biomass fuels using artificial intelligence formalisms." *BioEnergy Research*, 7, 2, 681–692.

Gramacy, R. B. and Apley, D. W. (2015). "Local Gaussian process approximation for large computer experiments." *Journal of Computational and Graphical Statistics*, 24, 2, 561–578.

Hahn, R. P., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)." *Bayesian Analysis*, 15, 3, 965–1056.

Horiguchi, A., Pratola, M. T., and Santner, T. J. (2021). "Assessing variable activity for Bayesian regression trees." *Reliability Engineering & System Safety*, 207, 107391.

Horiguchi, A., Santner, T. J., Sun, Y., and Pratola, M. T. (2022). "Using BART for Quantifying Uncertainties in Multiobjective Optimization of Noisy Objectives." *arXiv:2101.02558*.

Liu, H., Nattino, G., and Pratola, M. T. (2020). "Sparse Additive Gaussian Process Regression." *arxiv:1908.08864*, 1–33.

MacKay, D. J. C. (1995). "Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks." *Network: computation in neural systems*, 6, 3, 469.

Pratola, M. and Higdon, D. (2014). "Bayesian Regression Tree Calibration of Complex High-Dimensional Computer Models." *Technometrics*.

Pratola, M. T. (2016). "Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models." *Bayesian Analysis*, 11, 885–911.

Tan, Y. V. and Roy, J. (2019). "Bayesian additive regression trees and the General BART model." *Statistics in medicine*, 38, 25, 5048–5069.

Weisberg, S. (2013). *Applied linear regression*. John Wiley & Sons.