BART:
Finding Low Dimensional Structure in High Dimensional Data

Hugh A. Chipman, Edward I. George, Robert E. McCulloch

Abstract

Consider the canonical regression setup where one wants to model the relationship between y, a variable of interest, and $x_1, \ldots, x_p$, p potential predictor variables. For this general problem we propose BART (Bayesian Additive Regression Trees), a new approach to discover the form of $f(x_1, \ldots, x_p) \equiv E(Y \mid x_1, \ldots, x_p)$ and draw inference about it. BART approximates f by a Bayesian "sum-of-trees" model where each tree is constrained by a prior to be a weak learner as in boosting. Fitting and inference are accomplished via an iterative backfitting MCMC algorithm. By using a large number of trees, which yields an overcomplete basis for f, we have found BART to be remarkably effective at finding highly nonlinear relationships hidden within a large number of irrelevant potential predictors.

BART is motivated by ensemble methods in general, and boosting algorithms in particular. Like boosting, each weak learner (i.e., each weak tree) contributes a small amount to the overall model, and the training of a weak learner is conditional on the estimates for the other weak learners. The differences from boosting algorithms are just as striking as the similarities: BART is defined by a statistical model: a prior and a likelihood, while boosting is defined by an algorithm. MCMC is used both to fit the model and to quantify inferential uncertainty through the variation of the posterior draws.

The BART modelling strategy can also be viewed in the context of Bayesian nonparametrics. The key idea is to use a model which is rich enough to respond to a variety of signal types, but constrained by the prior from overreacting to weak signals. The ensemble approach provides for a rich base model form which can expand as needed via the MCMC mechanism. The priors are formulated so as to be interpretable, relatively easy to specify, and provide results that are stable across a wide range of prior hyperparameter values. The MCMC algorithm, which exhibits fast burn-in and good mixing, can be readily used for model averaging and for uncertainty assessment.

After introducing BART, we proceed to illustrate how it opens up a new approach to variable selection when one wants to model the relationship between y and a subset of $x_1, \ldots, x_p$, but there is uncertainty about which subset to use. This selection problem is typically treated by assuming that the relationship between y and $x_1, \ldots, x_p$ belongs to a parametric family such as the normal linear models. If incorrect, however, such an assumption can at the outset defeat the ultimate goal; subsets of $x_1, \ldots, x_p$ may be excluded simply because their relationship to y is far outside the assumed parametric family. To avoid this limitation, we show how BART may be used to discover the nature of the relationship between y and $x_1, \ldots, x_p$ before attempting to find relevant variables and a suitable parametric form.

To begin with, BART automatically screens for relevant predictors. As the BART algorithm moves through the model space, different potential predictors enter the model with different frequencies. Those that enter rarely or not at all are candidates for elimination, and those that enter frequently are candidates for inclusion. Based on such information, we consider various strategies for rerunning BART on subsets of $x_1, \ldots, x_p$ which lead to a stable subset for selection. Note that BART also provides an omnibus test: the absence of any relationship between y and any subset of $x_1, \ldots, x_p$ is suggested when BART posterior intervals for f reveal no signal.

Going further, let $\hat{f}$ be a BART estimate of f based on the selected subset of $x_1, \ldots, x_p$. Intuitively, $\hat{f}$ may be regarded as a sufficient statistical summary of the systematic relationship between y and $x_1, \ldots, x_p$. Thus $\hat{f}$ and the selected subset can be used, instead of the raw data, to find a parametric model for this relationship. For example, let $\mathcal{M}_1, \ldots, \mathcal{M}_M$ be M different parametric model classes under consideration such as the normal linear models or other exponential family models. Partial dependence plots applied to $\hat{f}$ may be useful for suggesting the form of such model classes as well as useful transformations of the predictors. Basically, the goal is to find the model within any of these model classes that is "best supported" by $\hat{f}$. For this purpose, we consider the strategy of selecting the model corresponding to the projection of $\hat{f}$ onto the nearest model class with respect to a utility criterion such as the Kullback-Leibler discrepancy. Yet another strategy is to construct a likelihood over the model space based on the probability distribution of $\hat{f}$ for each model. This opens the door to $\hat{f}$ based Bayesian approaches for model selection and averaging over $\mathcal{M}_1, \ldots, \mathcal{M}_M$.

KEY WORDS: Bayesian backfitting; Boosting; CART; MCMC; Sum-of-trees model; Weak learner; Variable selection.

Reference

Chipman, H. A., George, E. I. and McCulloch, R. E. (2005). BART: Bayesian Additive Regression Trees. *Submitted for Publication.*

Five Relevant Publications

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

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

Chipman, H. A., George, E. I. and McCulloch, R. E. (2002). Bayesian Treed Models, *Machine Learning* **48**, 299–320.

Chipman, H. A., George, E. I. and McCulloch, R. E. (2001). The Practical Implementation of Bayesian Model Selection (with discussion). In *Model Selection* (P. Lahiri, ed.) IMS Lecture Notes – Monograph Series, **38**, 65–134.

Clyde, M. and George, E. I. (2004). Model Uncertainty, *Statistical Science*, **19**, 81–94.