PBART: Parallel Bayesian Additive Regression Trees

Matthew T. Pratola
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545
mpratola@lanl.gov

Robert E. McCulloch
IROM Department
University of Texas at Austin
Austin, TX 78712-1175
robert.e.mcculloch@gmail.com

James Gattiker
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545
gatt@lanl.gov

Hugh A. Chipman
Department of Mathematics and Statistics
Acadia University Wolfville, NS B4P 2R6
hugh.chipman@acadiau.ca

David M. Higdon
Statistical Sciences Group
Los Alamos National Laboratory
Los Alamos, NM 87545
dhigdon@lanl.gov

Abstract

The Bayesian Additive Regression Tree (BART) is a statistical Bayesian nonparametric model that represents the observed data as a sum of weak learners. Each tree represents a weak learner, which is accomplished by constraining the tree through a regularization prior. The overall model is constructed as a sum of such trees, and draws from the posterior are sampled using a Markov Chain Monte Carlo (MCMC) algorithm. The method can be viewed as a Bayesian regression approach, where each regressor is formed by a random basis element.

In the current work, we extend the BART model to handle massive datasets using a parellelized MCMC algorithm. The approach scales linearly in the number of processor cores, enabling the practitioner to perform statistical inference on massive datasets. Our approach can also handle datasets too massive to fit on any single data repository. In addition, we develop parallel prediction and sensitivity analysis algorithms using the BART framework. Taken all together, the PBART model allows statistical inference to be scaled to the large datasets arising in modern scientific and applied problems.