Discovering Regression Structure with a Bayesian Ensemble

Edward I. George

A Bayesian ensemble can be used to discover and learn about the regression relationship between a variable of interest \( y \), and \( p \) potential predictor variables \( x_1, \ldots, x_p \). The basic idea is to model the conditional distribution of \( y \) given \( x \) by a sum of random basis elements plus a flexible noise distribution. In my Loeb Research Lecture, I will focus on a Bayesian ensemble approach called BART (Bayesian Additive Regression Trees) that I have developed with my long time collaborators H. A. Chipman and R. E. McCulloch. Based on a basis of random regression trees, BART automatically produces a predictive distribution for \( y \) at any \( x \) (in or out of sample) which automatically adjusts for the uncertainty at each such \( x \). It can do this for nonlinear relationships, even those hidden within a large number of irrelevant predictors. Further, BART opens up a novel approach for model free variable selection. Ultimately, the information provided such a Bayesian ensemble may be seen as a valuable first step towards model building for high dimensional data.

BART is motivated by ensemble methods in general, and boosting algorithms in particular. Essentially, BART approximates the unknown form of \( f(x_1, \ldots, x_p) = E(Y \mid x_1, \ldots, x_p) \) by a “sum-of-trees” model which is coupled with a regularization prior that constrains each tree to be a weak learner. As in 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. Flexibility of the “sum-of-trees” model is enhanced by using a large number of trees which allows BART to approximate a rich class of underlying conditional mean functions thereby enhancing its predictive effectiveness. However, in sharp contrast to boosting, BART is based on an underlying statistical model: a likelihood and a prior. And as opposed to algorithms such as boosting and random forests which sequentially add new trees to the mix, BART begins with a fixed number of trees which are then updated by repeated passes through an iterative backfitting MCMC algorithm. By using an essentially over-complete basis of trees, this algorithm exhibits fast burn-in and good mixing, converging quickly to a stable predictive distribution. A useful byproduct of the algorithm is the repeated generation of posterior draws which can be used for model averaging and uncertainty assessment.

The BART modeling strategy can also be viewed in the context of Bayesian nonparametrics. The key idea there is to use a model rich enough to respond to a variety of signal types, but constrained by a prior from overreacting to noise. The Bayesian ensemble approach provides just such a rich model form, which can expand as needed via the MCMC mechanism, but is kept in check by strong regularization priors. To facilitate their implementation in BART, these priors have been formulated to be interpretable, easy to specify, and to provide results that are robust across a wide range of prior hyperparameter values. In particular, the use of simple regression trees as basis functions enables a stable automatic calibration.
of the prior which yields a ready-to-use default version of BART (publicly available as the R-package BayesTree on CRAN).

Beyond its predictive potential, BART also opens up a novel approach to model-free variable selection, an approach that avoids an a priori parametric assumption of a form for the relationship between $y$ and every subset of $x_1, \ldots, x_p$ as is commonly made. By simply keeping track of the predictor usage frequencies as the MCMC algorithm moves through the model space, those predictors that enter the ensemble model most frequently become the candidates for selection. This strategy is seen to be particularly effective when the number of trees is kept small, creating a bottleneck which forces the variables to compete with each other for entry into the ensemble. Such a strategy is unavailable with other ensemble methods that rely on the sequential addition of trees. It will also be of interest for selection that BART 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.