Pre-Modeling Via BART

Edward I. George
University of Pennsylvania

Consider the canonical regression setup where one wants to learn about the relationship between y, a variable of interest, and $x\backslash, \ldots , x_p$, p potential predictor variables. Although one may ultimately want to build a parametric model to describe and summarize this relationship, preliminary analysis via flexible nonparametric models may provide useful guidance. For this purpose, we propose BART (Bayesian Additive Regression Trees), a flexible nonparametric ensemble Bayes approach for estimating $f(x_i, \ldots , x_p) = E\{Y \mid x_i, \ldots , x_p\}$, for obtaining predictive regions for future y, for describing the marginal effects of subsets of $x_i, \ldots , x_p$ and for model-free variable selection. Essentially, BART approximates f by a Bayesian "sum-of-trees" model where fitting and inference are accomplished via an iterative backfitting MCMC algorithm. By using a large number of trees, which yields a redundant basis for f, BART is seen to be remarkably effective at finding highly nonlinear relationships hidden within a large number of irrelevant potential predictors. BART also provides an omnibus test: the absence of any relationship between y and any subset of $X\backslash, \ldots , Xp$ is indicated when BART posterior intervals for f reveal no signal. (This is joint work with Hugh Chipman and Robert McCulloch).