Bayesian Forests

Matt Taddy

The University of Chicago Booth School of Business
faculty.chicagobooth.edu/matt.taddy

Abstract: We interpret random forests, and bagging ensemble estimators in general, via the framework of nonparametric Bayesian (npB) analysis. The ensemble strategies are revealed as approximations to posterior mean inference for complex summaries of the population data generating process. This insight motivates a class of fully Bayesian forest algorithms that provide gains in interpretability (from a Bayesian perspective) and predictive performance over their classically bagged predecessors. The npB framework is then used to reinterpret common distributed algorithms for efficient forest estimation on Big Data. From this, we propose a novel blocking strategy for fitting tree ensemble predictors on internet-scale data stored in a distributed file system (such as HDFS).

1 Introduction

Decision trees, and ensembles of decision trees,

For prediction problems with relatively small input dimensions, or in conjunction with dimension reduction strategies, our experience is that ensembles of trees will work out-of-the-box nearly as well as a carefully tuned, application-specific alternative.

Straightforward but not well-recognized nonparametric Bayesian interpretation of forests.

Other Bayesian studies of bagging have focused on its interpretation as *model averaging*, which arises when $\mathcal{T}$ is a *parameter* of the data generating process. We will also compare against a set of Bayesian procedures that interprets the tree in this way – as an unknown model parameter. All of these models are fit with MCMC, rather than through a bootstrap algorithm.