Naïve Bayes Variable and Interaction Selection

Jun Liu
Department of Statistics
Harvard University
`jliu@stat.harvard.edu`

Abstract

Suppose we have N individuals and for each individual we observed its response variable Y_i and its p-dimensional categorical-valued covariates $(X_{i1}, \ldots, X_{ip})$. Our goal is to discover which subset of covariates and interactions among them are influential on the response. Assuming that the covariates are discrete, we strive for a more ambitious goal than just finding certain linear relationships. I will start with the two-class prediction problem and the naïve Bayes predictor and present a strategy to allow for a very fast MCMC strategy to do variable and interaction selection. We call the underlying model "the Bayesian partition model." Compared with the standard logistic regression approach, the new strategy is much faster and more flexible. It also does not require one to specify interaction models. We then present extensions of the method to handle continuous responses and multi-dimensional responses through the use of a set of latent indicator vectors. I will illustrate the power of the method mainly using examples in genome-wide genetic association studies and in studies of expression quantitative trait loci (eQTL).