Scalable logistic regression with crossed random effects

Swarnadip Ghosh
Stanford University

Trevor Hastie
Stanford University

Art B. Owen
Stanford University

Abstract

The cost of both generalized least squares (GLS) and Gibbs sampling in a crossed random effects model can easily grow faster than $N^{3/2}$ for N observations. Ghosh et al. (2020) develop a backfitting algorithm that reduces the cost to $O(N)$. Here we extend that method to a generalized linear mixed model for logistic regression. We use backfitting within an iteratively reweighted penalized least square algorithm. The specific approach is a version of penalized quasi-likelihood due to Schall (1991). A straightforward version of Schall's algorithm would also cost more than $N^{3/2}$ because it requires the trace of the inverse of a large matrix. We approximate that quantity at cost $O(N)$ and prove that this substitution makes an asymptotically negligible difference. Our backfitting algorithm also collapses the fixed effect with one random effect at a time in a way that is analogous to the collapsed Gibbs sampler of Papaspiliopoulos et al. (2020). We use a symmetric operator that facilitates efficient covariance computation. We illustrate our method on a real dataset from Stitch Fix. By properly accounting for crossed random effects we show that a naive logistic regression could underestimate sampling variances by several hundred fold.