Inference for Case Probability in High-dimensional Logistic Regression

Zijian Guo

Rutgers University
E-mail: zijguo@stat.rutgers.edu

Abstract: Labeling patients in electronic health records (EHRs) with respect to their statuses of having a disease or condition, i.e. case or control statuses, has increasingly relied on prediction models using high-dimensional variables derived from structured and unstructured EHR data. A major hurdle currently is a lack of valid statistical inference methods for the case probabilities. In this paper, considering high-dimensional logistic regression models for prediction, we propose a bias-corrected estimator for the case probability through integration of linearization and variance enhancement techniques. We establish asymptotic normality and confidence interval construction of the proposed estimator, and propose a hypothesis testing method for patient case-control labelling. The main novelty of our theoretical development is to establish the asymptotic normality of the data-dependent weighted summation of model errors through employing contraction principles, instead of creating independence by sample splitting. This technique can be of independent interest in studying other high-dimensional inference problems in nonlinear models. The validity of our method does not require sparsity conditions on either the loading vector or the precision matrix of the random design. We demonstrate our method via extensive simulation studies and application to a real data set from Penn Medicine EHR.