Optimal Poisson Subsampling from Massive Data

HaiYing Wang

University of Connecticut
E-mail: haiying.wang@uconn.edu

Abstract: Nonuniform subsampling methods are effective to reduce computational burden and maintain estimation efficiency for massive data. Existing methods mostly focus on subsampling with replacement due to its high computational efficiency. If the data volume is too large so that nonuniform subsampling probabilities can not be calculated all at once, then subsampling with replacement is infeasible to implement. This paper solve this problem by using Poisson subsampling. We first derive optimal Poisson subsampling probabilities in the context of quasi-likelihood estimation under the A- and L-optimality criteria. For a practically implementable algorithm with approximated optimal subsampling probabilities, we establish the consistency and asymptotic normality of the resultant estimators. To address the situation that the full data are stored at multiple locations, we develop a distributed subsampling framework, in which statistics are computed simultaneously on smaller partitions of the full data. Properties of resultant estimators are investigated in terms of both mean square errors and asymptotic distributions. The proposed strategies are illustrated and evaluated through numerical experiments on simulated and real data sets.