

IPS22: Labeling Issues in Machine Learning Methodology Unweighted estimation based on optimal sample under measurement constraints

Jing Wang¹; HaiYing Wang²; Shifeng Xiong¹

- ¹ Academy of Mathematics and Systems Science, Chinese Academy of Sciences
- ² University of Connecituct

Abstract:

To tackle massive data, subsampling is a practical approach to sift more informative data points. However, when responses are expensive to measure, developing efficient subsampling schemes is challenging, and the optimal sampling approach under measurement constrains was developed to meet this challenge. This method uses the inverses of optimal sampling probabilities to reweight the objective function, which assigns smaller weights on more important data points. Thus the estimation efficiency of the resulting estimator can be improved. In this paper, we propose an unweighted estimating procedure based on optimal subsamples to obtain a more efficient estimator. We obtain the unconditional asymptotic distribution of the estimator via martingale techniques without conditioning on the pilot estimate, which has been less investigated in existing subsampling literature. Both asymptotic results and numerical results show that the unweighted estimator is more efficient in parameter estimation.

Keywords:

generalized linear models; labeling constraints; massive data; unweighted estimator