Optimal Subdata Selection for Large-scale Multi-class Logistic Regression

Min Yang

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago

Big data presents the unprecedented challenge of analysis due to its immense size. One common solution is to select a subset of the data that can be managed with existing computational resources. Although various subset selection methods exist, the optimal approach, in theory, would be to choose a subset that minimizes the variance-covariance matrix from all possible data subsets. However, this is a classic NP-hard problem. In this paper, we target multi-class logistic regression models and introduce an optimal subset selection algorithm for large-scale datasets. This algorithm aims to derive near-optimal subsets under various setups. Empirical studies show that our proposed algorithm significantly surpasses existing subsampling approaches in statistical efficiency while also reducing computational time