Scale Invariant Optimal Subsampling

Haiying Wang

Department of Statistics, University of Connecticut, USA

Abstract

Subsampling is an effective method to alleviate the computational cost when faced with massive data, and optimal subsampling algorithms aim to achieve a higher estimation efficiency. Existing optimal subsampling probabilities focus on minimizing the asymptotic mean squared error of the subsample parameter estimator. They are scale variant, and their performance changes if the data is scale transformed. We recommend focusing on minimizing the squared prediction error, which results in scale invariant optimal subsampling probabilities. In addition, the resulting probabilities are invariant to model constraints in softmax regression, and they provide a better subsampling strategy than existing methods in terms of balancing the responses among all categories..