Model-free Subsampling Method Based on Uniform Designs

Yongdao Zhou

Nankai University, China

Abstract

Subsampling or subdata selection is a useful approach in large-scale statistical learning. Most existing studies focus on model-based subsampling methods which significantly depend on the model assumption. In this paper, we consider the model-free subsampling strategy for generating subdata from the original full data. In order to measure the goodness of representation of a subdata with respect to the original data, we propose a criterion, generalized empirical F-discrepancy (GEFD), and study its theoretical properties in connection with the classical generalized L2-discrepancy in the theory of uniform designs. These properties allow us to develop a kind of low-GEFD data driven subsampling method based on the existing uniform designs. By simulation examples and a real case study, we show that the proposed subsampling method is superior to the random sampling method. Moreover, our method keeps robust under diverse model specifications while other popular subsampling methods are under-performing. In practice, such a model-free property is more appealing than the model-based subsampling methods, where the latter may have poor performance when the model is misspecified, as demonstrated in our simulation studies.