Sufficient dimension reduction with additional information

Hung Hung
National Taiwan University

Abstract

Sufficient dimension reduction is widely applied to help model building between the response $Y$ and covariate $X$. In some situations, we also collect additional covariate $W$ that has better performance in predicting $Y$, but has a higher obtaining cost, than $X$. While constructing a predictive model for $Y$ based on $(X, W)$ is straightforward, this strategy is not applicable since $W$ is not available for future observations in which the constructed model is to be applied. As a result, the aim of the study is to build a predictive model for $Y$ based on $X$ only, where the available data is $(Y, X, W)$. A naive method is to conduct analysis using $(Y, X)$ directly, but ignoring $W$ can cause the problem of inefficiency. On the other hand, it is not trivial to utilize the information of $W$ to infer $(Y, X)$, either. In this article, we propose a two-stage dimension reduction method for $(Y, X)$, that is able to utilize the information of $W$. In the breast cancer data, the risk score constructed from the two-stage method can well separate patients with different survival experiences. In the Pima data, the two-stage method requires fewer components to infer the diabetes status, while achieving higher classification accuracy than conventional method.