Abstract:
In massive data analysis, training and testing data often come from very different sources, and their probability distributions are not necessarily identical. A feature example is nonparametric classification in posterior drift model where the conditional distributions of the label given the covariates are possibly different. In this paper, we derive minimax rate of the excess risk for nonparametric classification in posterior drift model in the setting that both training and testing data have smooth distributions, extending a recent work by Cai and Wei (2019) who only impose smoothness condition on the distribution of testing data. The minimax rate demonstrates a phase transition characterized by the mutual relationship between the smoothness orders of the training and testing data distributions. We also propose a computationally efficient and data-driven nearest neighbor classifier which achieves the minimax excess risk (up to a logarithm factor). Simulation studies and a real-world application are conducted to demonstrate our approach.