Neyman-Pearson classification: parametrics and sample size requirement

Lucy Xia

The Hong Kong University of Science and Technology
E-mail: lucyxia@ust.hk

Abstract: The Neyman-Pearson (NP) paradigm in binary classification seeks classifiers that achieve a minimal type II error while enforcing the prioritized type I error controlled under some user-specified level α. This paradigm serves naturally in applications such as severe disease diagnosis and spam detection, where people have clear priorities among the two error types. Recently, Tong, Feng, and Li (2018) proposed a nonparametric order statistics based umbrella algorithm that adapts all scoring-type classification methods (e.g., logistic regression, support vector machines, random forest) to respect the given type I error upper bound α with high probability, without specific distributional assumptions on the features and response. Universal the umbrella algorithm is, it demands an explicit minimum sample size requirement on class 0, which is usually the more scarce class. In this work, we employ the parametric linear discriminant analysis (LDA) model and propose a new parametric thresholding algorithm, which does not need the minimum sample size requirements on class 0 observations and thus is applicable to small sample applications such as rare disease diagnosis. Leveraging both the nonparametric and nonparametric thresholding rules, we propose four LDA based NP classifiers, for both low and high dimensional settings. On the theoretical front, we prove NP oracle inequalities for one proposed classifier. This is the first time such theoretical criteria are established under the parametric model assumption and unbounded feature support. Furthermore, as NP classifiers involve a sample splitting step of class 0 observations, we construct a new adaptive sample splitting scheme that can be applied universally to NP classifiers and this adaptive strategy enhances the accuracy of these classifiers.