14:50-15:40, Dec. 11th, 2015
Session (2)

Generalized *t*-statistic and AUC for binary classification

Osamu Komori

University of Fukui

Abstract

In binary classification, the Fisher linear discriminant analysis has been widely used, where the ratio of the variance between the classes to the variance within the classes is maximized to derive the linear predictor. This simple method has shown good performances in real data analyses in medical and clinical sciences; in some cases it outperforms more sophisticated methods such as machine learning methods. From this viewpoint, we proposed generalized *t*-statistic and AUC (area under the ROC curve) to extend the Fisher linear discriminant analysis based on a generator function U, and investigated the statistical properties in terms of estimation as well as classification accuracies. The generalized *t*-statistic assumes that the probability distribution of one population is homogeneous, such as a normal distribution; on the other hand, the probability distribution of the other population could be highly heterogeneous. This is the typical situation in case-control studies in clinical trials. The generalized AUC assumes heterogeneity for probability distributions for both populations. For these cases, we derived the optimal generator function U to derive the best linear predictor in terms of asymptotic variances. In order to be applicable to high dimensional data analysis, Lasso-type method is also proposed by imposing L_1 penalty on the objective function. The performances of the proposed methods are illustrated in simulation studies as well as real data analyses.