Atelier international « Perspectives sur l'analyse des données en dimensions supérieures II »

30 mai – 1 juin, 2012

International Workshop on the Perspectives on High-dimensional Data Analysis II
May 30 – June 1, 2012

A nonparametric optimal decision rule in high dimensional space

Yang Feng *
yangfeng@stat.columbia.edu

In a binary classification problem, the Bayes' rule lays down the oracle decision boundary. Classifiers assuming linear and nonlinear boundaries, e.g., linear regression, logistic regression, LDA, neural network and support vector machines, are abundant in the statistics and machine learning literature. Many of these classifiers have also been modified to handle the high dimensional settings. Despite the good performances, there is much rigidity in assuming a special orale boundary category or class conditional densities. In this paper, instead of modeling directly the class conditional densities, we target on the oracle decision boundary. In theoretical derivations, we assume that the log-likelihood of class conditional densities can be written as a linear combination of log-likelihoods of marginal class conditional densities. This flexible category allows us to incorporate a large collection of density assumptions, including the LDA setting with non-identity covariance matrix. After we plug-in estimates of marginal class conditional densities, to solve the classification problem is equivalent to find the best linear classification rule in a nonparametrically- transformed feature space. Then penalized logistic regression and other established linear classification rule employed to close the analysis. Such a new classification procedure is named Nonparametric Optimal Decision Rule (NOD). Oracle inequalities are developed regarding the classification error. A vast array of simulation and real data analysis demonstrate the outstanding performance of NOD. It is also worth mentioning that although NOD just uses marginal densities, it has incorporated interdependence among features without explicitly modeling the covariance.

*Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, NY 10027, USA.