Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data

Addy Bolívar-Cimé

Centro de Investigación en Matemáticas A.C.

There is an increasing current interest in the statistical analysis of data arising in problems of genomics, medical image analysis, climatology, finance and functional data analysis, where one frequently observes multivariate data with high dimension greater than the sample size. Recently several authors have been interested in the comparison of classical methods for Binary Discrimination Analysis with new ones designed for the High-Dimension, Low Sample Size (HDLSS) context. In this talk we compare the asymptotic behavior of the binary discrimination methods Mean Difference (MD), Support Vector Machine (SVM), Distance Weighted Discrimination (DWD) and Maximal Data Piling (MDP) when the dimension $d$ of the training data set tends to infinity and the sample sizes $m$ and $n$ are fixed. It is worth mentioning that the last two methods are specially designed for the HDLSS context by Marron, et al. [2] and Ahn and Marron [1], respectively. Specifically, we show that when the data sets are spherical Gaussian where one set has mean zero and the other has mean $v_d$, the normal vectors of the separating hyperplanes of the methods tend to be in the same direction as $v_d$ when $\|v_d\| \gg d^{1/2}$, i.e. are consistent; and tend to be orthogonal to $v_d$ when $\|v_d\| \ll d^{1/2}$, i.e. are strongly inconsistent. The case when $\|v_d\| \approx d^{1/2}$ is also considered. We also compare the MD method with the SVM for large $d$. We see in a particular setting that generally the MD method is better than the SVM, in the sense that the angle between the normal vector of the MD hyperplane and the optimal direction $v_d$ is closer to zero than the angle between the normal vector of the SVM hyperplane and $v_d$.

References
