Implementations of Fisher's linear discriminant analysis from the numerical point of view

Jurjen Duintjer Tebbens (Institute of Computer Science, Academy of Sciences of the Czech Republic) and Pavel Schlesinger (Institute of Formal and Applied Linguistics, Charles University in Prague)

In this contribution we investigate numerical aspects of classification using various implementations of Fisher's linear discriminant analysis (FLDA). FLDA is based on maximizing the ratio of between-group variance to within-group variance of given variables. The maximization problem is frequently solved by reducing it to the symmetric generalized eigenproblem defined by the between-group and the within-group covariance matrices. Unfortunately, these matrices are singular in classification tasks where the number of given variables exceeds the number of objects for training. Consequently, efficient numerical solution of the eigenproblem becomes a challenging problem (see, e.g. [1]) and several techniques to handle the singularity have been proposed in the literature about FLDA (see, e.g. [2]). In comparative studies of these techniques, however, assessment of classification performance more or less discards numerical aspects such as computational or storage costs.

We focus on the relation between implementation and performance of FLDA. We give a comparison of a number of popular FLDA implementations including detailed information about their numerical stability, storage costs, computational costs and estimation of computational error. Moreover, we provide numerical examples by applying the discussed methods to a particular protein classification problem (see, e.g. [3]). The differences of classification performance are striking. An implementation based on reduction to a classical eigenproblem via Moore-Penrose pseudeoinverses outperforms all other strategies, including the Support Vector Machines approach that is generally considered the most powerful for the given data. We discuss some consequences of this observation and possible extension to other types of problems and conclude with a brief overview of relevant software.

References:

[1] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe and H. van der Vorst (eds.) (2000): Templates for the solution of Algebraic Eigenvalue Problems: A Practical Guide. Philadelphia, SIAM

[2] Y.-Q. Cheng, Y.-M. Zhuang and J.-Y. Yang (1992): Optimal Fisher discriminant analysis using the rank decomposition. Pattern recognition, vol. 25, 101–111

[3] L. Edler and J. Grassmann (1999): Protein fold prediction is a new field for statistical classification and regression. In Seillier-Moiseiwitsch F (Ed): Statistics in Molecular Biology and Genetics. IMS Lecture Notes Monograph Series 33, 288–313