TEACHING CONSTRUCTION CLASSIFICATION RULES WITH APPLICATIONS IN SOCIAL SCIENCES

Margarita Díaz, Cecilia Díaz, Patricia Caro and María Ines Stimolo
Universidad Nacional de Córdoba, Argentina
mdiaz@eco.unc.edu.ar

Supervised classification or pattern recognition is a method to solve decision problems in Social Sciences. It is organized on the basis of specific sets of predictor variables and the existence of classes known a priori. Based on a training sample, its main objective is to construct a classification rule in order to predict the class to which a new object belongs. Nowadays, the availability and efficacy of powerful computers have made possible many advances in this field, both in Statistics and Computer Sciences. In this section, different methods will be discussed and illustrated with the results obtained in several applications. The following topics will be dealt with: Parametric Discriminant Analysis, Non-parametric Discriminant Analysis, Logistic Discriminant Analysis, Neuronal Networks, Recursive Partitioning and Estimation of Error Rates.

INTRODUCTION AND OBJECTIVES

A course on Supervised Classification Methods is proposed in this presentation. It is intended for postgraduate students in the field of Social Sciences with previous knowledge of Multivaried Statistics. Linear Discriminant Analysis, derived from the assumption of Multivaried Normality and restricted to continuous data sets, was the first method developed to construct a classification rule based on a training sample. The habitual presence of large databases with mixed variables and the advances in informatics have favored the growth of a great variety of alternative methods, developed by applied researchers in the fields of statistics and informatics and spread in business organizations within data mining tools.

This proposal aims at enabling the course participants to acquire skills to:

1. Recognize the ability of the techniques to increase knowledge regarding a given problem.
2. Identify situations where the application of each method is appropriate.
3. Apply software to carry out the processing task and interpret the results properly.
4. Analyze and interpret results and compare different methods.
5. Report the results to people responsible for decision making who have no previous knowledge of statistics.
6. Understand the literature published about this topic in the most widely read journals.

COURSE DESCRIPTION

The course content has been designed to provide the participants with basic notions of Supervised Classification and the description of some methods. A selection of the methods to develop in this course was necessary because of their great expansion in the last few years. The inclusion criterion used was the application frequency of the different methods.

“Supervised Classification methods” comprise a family of multivaried techniques aimed at obtaining rules which make it possible to assign new objects characterized by a variable vector to one of the various preexisting groups. “Classified object” names an object whose origin group is known. A rule to assign unclassified objects to one of the groups will be referred to as an assignation rule or classifier. Statistically, the problem is set out as follows: beginning with a training sample, i.e., a set of objects known to be members of one of the pre-established groups, a rule is derived which makes it possible to assign each of the observations to one of the mutually exclusive and exhaustive groups, minimizing the chance of classifying the individuals wrongly.

In the first part of the course the basic notions and the types of errors likely to be made are defined and the assignation rule is derived which makes it possible to minimize the chance of classifying wrongly. The classification into two groups is considered first and it is later generalized for more than two. The rules are exemplified below by assuming multivaried normality, a supposition which justifies the use of Parametric Methods in their different variants. If it is not plausible to suppose normality in order to obtain the classification rule, Non-parametric

Methods can be applied, which are appropriate to solve problems with homogeneous variable sets. Finally, Semi-parametric and Recursive Methods are considered, which have the advantage of being applied in mixed data sets.

Therefore, the methods developed to construct classification rules can be grouped as Parametric, Non-parametric, Semi-parametric and Recursive. This makes it possible to set the next thematic blocks in the course, assuming that the participants know the multivaried normal model.

Each is first considered for the classification into two groups and later generalized for more than two.

- In *Parametric* methods, the densities corresponding to the classes are supposed to have a known functional form, and the parameter vector should be estimated. Specifically, in the classic Discriminant Analysis, whether in its linear, quadratic or regularized variants, all variables are supposed to be metric and respond to normal distribution. This assumption becomes a strong barrier for the application of such statistic techniques to typical problems in social sciences.
- *Non-parametric* methods do not establish any assumptions regarding conditional densities and can be applied to homogeneous sets of metric or categorical variables. The two best known techniques in this group are: the *Kernel* method, suitable for metric variables, produces estimations of conditional densities; and the *Nearest Neighbor* method, which gives a direct a posteriori estimation of probabilities and has the important advantage of being applied to categorical data.
- In order to deal with mixed-data problems, *Partially Parametric* methods have been developed, the most widely applied of which is *Logistic Regression*. Only the density quotient is modeled with this approach, without assuming a specific functional form, and only linearity is required for the density quotient logarithm (McLachlan, 1992). Logistics is a particular case in the *Neuronal Network* models, which are processing units or nodes connected forward from the input units (problem variables) toward the output ones (estimated probabilities). A linear combination of the precedents is made in each node, to which a so-called activation function is applied (for example, logistics), arriving at a non linear function of considerable complexity which makes it possible to estimate a posteriori probabilities.
- The result of the *Recursive Procedure* is known as a Decision Tree, which represents a series of consecutive questions and is built by repeated divisions of subsets into descending subsets on the basis of usually binary questions. The initial question or root node is located at the top of the tree, from which descending branches stem where questions are asked regarding the value assumed by other variables.
- Discussion related to non-parametric estimations of *Error Rates*: Apparent, Holdout, Crossvalidated and Jackknife, in order to analyze the performance of the rule and the relative efficiency of the different techniques when they are applied to the same problem.

DIDACTIC METHODOLOGY

This course is designed to be taught in 40 hours, of which 40% will be devoted to practice sessions on computers, where, beginning with simple examples, *Infostat*, *SPSS*, *SPAD-N*, *MATLAB* and *SAS* software will be used, with a focus on the discussion about and the interpretation of results and techniques available to verify the assumptions on which the methods are based.

Additionally, participants will be asked to solve a series of cases set out from the following databases, which were a result of the research done by the team:

1) To illustrate the application of Parametric and Non-parametric Methods: Balance sheets of companies which are quoted on the stock exchange, classified as normal and in crisis.
2) For the comparison of results obtained with Logistic Regression, Neuronal Networks, Decision Trees and Nearest Neighbor: (i) User base of people from the Continuous Household Survey, taking occupational status as an answer. (ii) Customer base of a Mobile Phone company, with customers classified as Active or Cancelled. (iii) Customer base of a financial company, where customers are classified as Responding or Non-

responding to a marketing campaign. There will be a workshop at the end of the course where participants will show the results obtained with the application of these methods to solve problems in their field of work.

FINAL CONSIDERATIONS

A course on multivaried methods is commonly taught in postgraduate degree programs in the field of social sciences and a chapter is included where the main notions of supervised classification are introduced, working basically with Discriminant Linear Analysis. This was the team's experience in teaching a course on Multivaried Methods in the Master of Science in Applied Statistics at the National University of Córdoba to a group of 30 participants in the second year. The students came from different areas of knowledge and they requested greater depth in these topics, as they were of interest to them both for their jobs and their thesis work.

On the other hand, the profusion in the development of supervised classification techniques and the expansion in their application to different areas of knowledge justify a specific course to deal with them. It is worth emphasizing that these techniques are included in data mining, which is being increasingly requested by companies due to the advances of informatics in the use of large databases and the ability of the techniques to create knowledge from them.

REFERENCES

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Biggs, D. B. de Ville and Suen, E. (1991). A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics*, 18, 49-62.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1998). *Classification and Regression Trees*. Boca Raton: Chapman and Hall.

Diaz, M. (2001). *Perfomance del Análisis Discriminante Regularizado en la Predicción de Crisis Financieras*. UN de Córdoba Tesis Maestría en Estadística Aplicada.

Hand, D. J. (1981). *Discrimination and Classification*. New York: Wiley.

Hand, D. J. (1982). *Kernel Discriminant Analysis*. New York: Wiley.

Hand, D. J. (1999). *Construction and Assessment of Classification Rules*. New York: Wiley.

Hosmer, D. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Izenman, A. (1991). Recent developments in nonparametric density estimation. *JASA*, 86(413), 205-224.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*. New York: Prentice-Hall

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data . *Applied Statistics*, 29(2), 119-127.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

SPSS Inc. and Manual del Usuario. (1998). *Answer Tree 2.0.*

Wang, C. and Van Ryzin, J. (1981). A class of smooth estimators for discrete distributions. *Biometrika*, 68, 301-309.

Wilkinson, L. (1992). Tree Structured Data Analysis: AID, CHAID and CART. Paper presented at Joint Software Conference, Sun Valley, ID.