About Fuzzy Discrimination

J.-M. Gautier, C.O.R.E.F., Boulogne-Billancourt, France
G. Saporta, Rene Descartes University, Paris, France

Il arrive souvent en analyse discriminante que l'appartenance des individus aux classes d'une partition de la population ne soit pas connue avec certitude, ou qu'il soit délicat d'attribuer strictement un individu à une catégorie lorsque la partition est définie à partir d'une variable numérique découpée en classes. Il semble alors plus raisonnable de se donner une distribution de probabilité sur les classes qu'une fonction booléenne d'appartenance surtout si un individu est proche de la frontière entre deux classes.

On établit alors les modifications à apporter aux techniques usuelles de discrimination (factorielle et décisionnelle) ainsi que les conséquences de ces modifications sur les indices usuels de qualité d'une discrimination.

Keywords : Fuzzy Discrimination
Discriminant Analysis

Discriminant analysis is the replicative or predictive study of a qualitative variable over a set of predictors that are generally numerical.

In certain situations, one cannot attribute with certainty a category of the variable to be explained to certain (even to all) individuals in the sample. This is particularly the case if :

- The classes are poorly defined, e.g., imprecise nomenclature or else discrete coding of a quantitative variable (1).

- The class to which an individual belongs cannot be determined with certainty, e.g., appearance of a symptom after absorption of a medicine. There is nothing to prove that the symptom could not have appeared naturally.

But it may also be that the determination of the class is too costly, and that one is content with an estimate.

From the formal viewpoint, we will therefore suppose that with each individual, there is associated a probability distribution P_j (i) for the classes $j = 1, \ldots, k$ of the variable to be explained, rather than a set of mutually exclusive indicative variables.

(1) In this case, the quantitative variable being divided up into classes, an individual in the neighborhood of the border between 2 classes probably cannot be attributed to a single one of the classes, if only because of possible measuring errors.

We will denote by X the matrix (n, p) of the p numerical predictors, n being the sample size

Hence we are going to describe how to extend to this situation the various research results of discriminant analysis, i.e., seek the best separation between the classes in the sense of a particular criterion (geometrical methods), or attempt directly to estimate the P_j (i) for every individual for which one sums up the p predictors (Bayesian research).

For geometrical research, this extension is made by introducing weightings into the calculations and by counting each observation i as an element of all of the classes j for which : $P_j(i) \neq 0$. The weighting of observation i in class j is then $P_j(i)$. We note $\alpha_j = \sum_{i=1}^{n} P_j(i)$ the weight of the class j.
It will be shown that in reality, it is not necessary to work on matrices of dimension k.n, but only n (on condition of having written the programs on an ad hoc basis).

The various quality measures of the discrimination are affected by the fuzziness of the classes of the variable to be explained, and for certain ones of these criteria, we will propose a limit calculation that will make it possible to judge the real discriminating power of a variable, or of a set of variables, relative to this limit.

As to the bayesian methods, if one excepts the case in which one considered the $P_j(i)$ as a sample of a random variable, the generalization is carried out in the same fashion as for the geometrical methods, since the only things affected by the fuzziness of the classification are the estimators of the parameters of the conditional distributions.

However, another approach is possible : a direct search for a formula for adjustment of the P_j by means of explanatory variables : logistic regression or linear regression with constraints.

I. GEOMETRICAL METHODS

I.1 Evaluation of the center of gravity of the classes

We will denote by P_j the diagonal matrix (n x n) of the P_j (i) associated with the group j, and $\underline{1}$ the vector of which the n components are equal to 1.

Since the classes are fuzzy, the centers of gravity g_j of each one of them are obtained by taking the average, weighted by the P_j (i), of the coordinates of the observations, hence :

$$\underline{g}_j = \frac{1}{\alpha_j}\left[X' \; P_j \; \underline{1}\right]$$

I.2 Expression of the matrices of variance

The total matrix of variance would then be :

$$V = \frac{1}{n}\sum_{j=1}^{k} X' \; P_j \; X = \frac{1}{n} X'X .$$

if the data are centered.

The matrix of variance of the classe j is written :

$$V_j = \frac{1}{\alpha_j} \left(X' P_j X - \frac{1}{\alpha_j} X' P_j \underline{1}\, \underline{1}' P_j X \right)$$

The intra-class matrix of variance W is then :

$$W = \frac{1}{n} \sum_{j=1}^{k} \alpha_j V_j \quad \text{since} \quad \sum_{j=1}^{k} \alpha_j = n$$

Its current term is worth :

$$U_{l_1 l_2} = \frac{1}{n} \sum_{i_1 i_2} x_{i_1 l_1} x_{i_2 l_2} \left(\delta_{i_1 i_2} - \sum_j \frac{1}{\alpha_j} P_j(i_1) P_j(i_2) \right)$$

The inter-class matrix of variance B is therefore equivalent to :

$$B = \frac{1}{n} X' \left(\sum_j \frac{1}{\alpha_j} P_j \underline{1}\, \underline{1}' P_j \right) X$$

And its current term is :

$$b_{l_1 l_2} = \frac{1}{n} \sum_{i_1 i_2} x_{i_1 l_1} \left(\sum_j \frac{1}{\alpha_j} P_j (i_1) P_j (i_2) \right) x_{i_2 l_2}$$

I.3 Calculation of the distance from a point to the center of gravity of the classes

Let $\underline{e}$ be a point of R^p ; if the matrices V_j are not significantly different, one uses the Mahalanobis metric W^{-1} to calculate these distances :

$$d^2 (\underline{e} ; \underline{g}_j) = (\underline{e}' - \underline{1}' P_j^* X) \; W^{-1} \; (\underline{e} - X' P_j^* \underline{1})$$

$$= \underline{e}' W^{-1} \underline{e}' + \underline{1}' P_j^* X W^{-1} X' P_j^* \underline{1}$$

$$- 2 \underline{e}' W^{-1} X' P_j^* \underline{1}$$

where $P_j^* = \frac{1}{\alpha_j} P_j$

If the V_j are significantly different, one applies V_j^{-1} instead of W^{-1} (Sebestyen's method).

We perceive that these formulas hardly from those of the classic case. Their main interest is that they supply a method of direct calculation not bringing in matrices of dimension kn.

I.4 Criteria of the quality of the discrimination

The first criterion that comes to mind is the one of the percentage of correct classification by the method of reassignment (about which it is known, incidentally, that it yields biased results). In the classic case, this percentage can reach 100 % ; here it is quite obviously limited by :

$$\sum_i \max_j P_j (i) \times \frac{100}{n}$$

In the case of 2 groups, the Mahalanobis distance D^2 between the two centers often serves as a criterion for separability, in particular for the selection of the variables, other criteria such as the F, can be deduced from it for a monotonic transformation :

$$D^2 = (\underline{g}_1 - \underline{g}_2)' \; W^{-1} \; (\underline{g}_1 - \underline{g}_2)$$

In the usual case, this distance is not bounded above and may theoretically be infinite if and only if each group is reduced to a point projected onto straight line $g_1 g_2$. Here this distance is always limited, and this limit can be calculated by the following procedure :

D^2 is a maximum if, in projection on straight line $g_1 g_2$, all the points such tha $p_1 (i) > P_2 (i)$ are confused in a point X_1, and all the others in X_2. Hence one is led back to a uni-dimensional problem on straight line $g_1 g_2$. Let us place ourselves in this case. D^2 being invariant for change of radius and of scale on the variables, one may suppose that on straight line $g_1 g_2$, the total variance is equal to 1, and that the variable is centered. From this, one deduces at the values of X_1 and X_2 and the value of σ^2, which is not zero, since the variances of each group cannot be zero if there is at least one i in which $P_j (i)$ is different from 0 or from 1,

whence $$D^2 \leq \frac{(g_1 - g_2)^2}{\sigma^2}$$

If we denote by P_1 the proportion of individual i affected in group 1, such that $p_1 (i) > p_2 (i)$:

$$x_1 = \sqrt{\frac{1 - P_1}{P_1}} \qquad x_2 = - \sqrt{\frac{P_1}{1 - P_1}}$$

The calculation of $\sigma^2 g_1$ and g_2 depends on the 3 parameters : p_1, $a_1 = \sum p_1$ (i) , $a_2 = \sum p_2$ (i).

In the case of k groups, a usable criterion is the sum of the Mahalanobis distances of the groups taken two by two ; as previously, the sum of the D^2 (g_j ; g_1) is maximum if in the space created by the k centers of gravity, all of the observations such that p_j (i) is a maximum are projected onto the same point $\underline{K}_j$. One may always suppose that the observations are centered and that the matrix of over-all variance is I_{k-1}, which leads to a single configuration of the $\underline{K}_j$ neglecting an isometry. One can then calculate the criterion that supplies the desired increase, which depends only on the distribution of the weightings.

One concrete calculation method consists in takings any set of k points $\underline{z}_1 \ldots \underline{z}_k$ and in carrying out the linear transformation that leads to the $\underline{K}_j$.

II.- PROBABILISTIC METHODS

II.1. Bayesian methods with hypothesis of normality

See Aitchison and Begg (1976).

If one makes the hypothesis of a normal distribution N_p (μ_j ; $\sum$) in each class, the only problem consists in estimating the parameters of the model before applying Bayes' formula (cf. Anderson).

The estimators of the μ_j and of $\sum$ are precisely the g_j and W previously defined.

II.2. Direct estimation of the P_j

Since one has a sample of P_j and explanatory variables X, one may use the regression techniques in the broad sense, or :

a) Logistic regression

This method reduces to supposing that Log $\frac{P_j}{P_k}$ is a linear function of the explanatory variables. The coefficients of these functions being estimated then by the method of maximum likelihood (Cox's model).

b) Linear regression under constraint

One regresses each p_j on the explanatory variable while imposing the constraint $\sum_{P_j \geq 0} p_j = 1$ (which is easy), and the constraints $P_j \geq 0$ $\forall j$, which leads to optimization programs on cones. In other words, it is a question of carrying out the canonical analysis between a convex cone and a vectorial sub-space.

REFERENCES

AITCHISON J. BEGG C.B (1976) Statistical Diagnosis when basic cases are not classified with certainty Biometrika, 63, 1-12.

ANDERSON T.W. (1958) "Introduction to multivariate statistical analysis" - Wiley , New-York.

ANDERSON J.A. (1972) Separate sample logistic discrimination, Biometrika 59, 19-35

COX D.R. (1970) "The analysis of binary data" - Methuen, London

MARTIN J.F. (1980) Le codage flou et ses applications en statistique. These 3è cycle Universite de Pau

SEBESTYEN (1962) "Decision making processes in pattern recognition" Mc Millan, New York.

Acknowledgement :

The authors would like to thank P. Besse for his comments and corrections.