Biometrika (1975), 62, 2, p. 403
Printed in Great Britain

G_1-minimax procedures for the case of prior distributions in discriminant analysis

By T. A. DeROUEN and Y. R. SARMA[†]

Departments of Health Measurement Sciences and Mathematics, Tulane University, New Orleans

Summary

In discriminant analysis, let $q' = (q_1, \ldots, q_m)$ denote the vector of prior probabilities associated with an observation coming from populations $\Pi_1, \ldots, \Pi_m$, respectively. Present approaches either consider q fixed and known, and use the corresponding Bayes procedure, or, if q is unknown, assume that the q_i are equal, and use the Bayes procedure based on that assumption. In this paper, consideration is given to the idea of a prior distribution on q, and procedures are developed which are optimal for the class G_1 of all priors on q with specified first moment. In addition, a procedure is suggested which would incorporate an estimate of the prior mean into the discrimination procedure.

Some key words: Classification; Discrimination; G-minimax; Minimax; Partial prior information.

1. Introduction

In the problem of classification, a vector of p measurements $x' = (x_1, \ldots, x_p)$ is taken on an individual. It is desired to classify the individual into one of the populations $\Pi_1, \ldots, \Pi_m$, on the basis of these measurements, using a decision procedure that minimizes the expected cost of misclassification. Let $q' = (q_1, \ldots, q_m)$, with

$$q_m = 1 - \sum_{i=1}^{m-1} q_i,$$

be the vector of *a priori* probabilities associated with an observation coming from $\Pi_1, \ldots, \Pi_m$, respectively (Anderson, 1958, p. 142). Present approaches consider optimal solutions for two situations: (i) $q_1, \ldots, q_m$ are fixed and known, in which case the corresponding Bayes procedure is optimal; and (ii) $q_1, \ldots, q_m$ are unknown, in which case they are assumed to be equal, and the corresponding admissible Bayes procedure used. In many situations, instead of assuming that q is fixed, it is more reasonable to assume that q has some prior density, and that we are merely obtaining observations at one point in that prior. Also, discriminant analysis often requires the use of a 'training sample' containing observations known to be from $\Pi_1, \ldots, \Pi_m$ in order to estimate their respective densities $p_1(x), \ldots, p_m(x)$. In many cases, the proportion of observations from population Π_i in the training sample is an unbiased estimate of the prior probability q_i, and this information should be utilized in obtaining an optimal discrimination procedure. The purpose of this paper is to incorporate the ideas of a prior distribution on q and estimates for a fixed value of q into the discrimination procedure.

Consideration of this problem was prompted by discussions of how to use epidemiological

[†] Present address: Indian Statistical Institute, Barrackpore Trunk Road, Calcutta.

information collected from one hospital on patients who have undergone one of several possible treatments for a particular disease. It is desirable to use this information to help the physician choose for an incoming patient the treatment with the best prognosis for recovery, with categorical outcomes of recovery, e.g. excellent, good, fair or poor. The *a priori* probabilities $\{q_i\}$ associated with these categories may reasonably be assumed to be fixed for that particular hospital; however there is no reason to assume that the proportions in these categories will be the same in other hospitals. In developing a procedure which can be extended beyond that particular hospital to the general population, the assumption of a prior distribution for $\{q_i\}$, with observations obtained at one point in that prior, seems appropriate.

2. G_1-minimax procedures

2·1. *Two populations*

Suppose an individual is an observation from either population Π_1 or population Π_2. The classification of an individual depends on the vector of measurements $x' = (x_1, \ldots, x_p)$ and on the classification function $\Phi'(x) = \{\Phi_1(x), \Phi_2(x)\}$. This classification function has elements satisfying the conditions

$$0 \leqslant \Phi_i(x) \leqslant 1 \quad (i = 1, 2), \quad \Phi_1(x) + \Phi_2(x) = 1,$$

for all x, and assigns an observation x to population Π_1 with probability $\Phi_1(x)$, and to population Π_2 with probability $\Phi_2(x)$.

Suppose that the populations Π_1 and Π_2 have distributions with density functions $p_1(x)$ and $p_2(x)$ respectively. For convenience we only treat the case where the densities are continuous. With usual modifications the results hold in other cases. Let q be the probability that an individual comes from population Π_1 and $(1-q)$ the probability of coming from population Π_2. Let $\xi(q)$ be the prior distribution of q over $[0, 1]$ with mean value μ. Also let $C(i|j)$ denote the cost of misclassifying an individual from population Π_j as belonging to Π_i.

From the definition of the classification function, for any function $\Phi(x)$,

$$L_1(\Phi) = \int \Phi_2(x)\, p_1(x)\, dx$$

is the probability of misclassifying an individual from Π_1 as belonging to Π_2; and $L_2(\Phi)$, similarly defined, is the probability of misclassifying an individual from Π_2 as belonging to Π_1. Thus the average loss from costs of misclassification as a function of q and Φ is given by

$$R(\Phi, q) = qC(2|1)\, L_1(\Phi) + (1-q)\, C(1|2)\, L_2(\Phi).$$

The expected risk $\bar{R}(\Phi, \xi)$ is then

$$\int_0^1 R(\Phi, q)\, d\xi(q) = \mu C(2|1)\, L_1(\Phi) + (1-\mu)\, C(1|2)\, L_2(\Phi). \tag{1}$$

Thus, the expected risk, as a function of the prior distribution, depends only on the first moment, μ. It is now necessary to find the Bayes solution for this prior, which will be called the μ-Bayes procedure.

THEOREM 1. *The classification function $\Phi^*(x)$ for which*

$$\Phi_1^*(x) = \begin{cases} 1 & \text{if} \quad \mu C(2|1)\, p_1(x) \geqslant (1-\mu)\, C(1|2)\, p_2(x), \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

$$\Phi_2^*(x) = 1 - \Phi_1(x)$$

minimizes $\bar{R}(\Phi, \xi)$.

Proof. The proof is on the same lines as in the Neyman–Pearson lemma. Let $\Phi^0(x)$ be any other classification function. Consider

$$\begin{aligned} \bar{R}(\Phi^*, \xi) - \bar{R}(\Phi^0, \xi) = & \int \{\Phi_1^0(x) - \Phi_1^*(x)\}\{\mu C(2|1)\, p_1(x) - (1-\mu)\, C(1|2)\, p_2(x)\}\, dx \\ & - \int \{\Phi_1^0(x) - \Phi_1^*(x)\}\, \mu C(2|1)\, p_1(x)\, dx \\ & + \int \{\Phi_2^*(x) - \Phi_2^0(x)\}\, \mu C(2|1)\, p_1(x)\, dx. \end{aligned} \tag{3}$$

The sum of the last two integrals in (3) is identically zero and

$$\Phi_1^0(x) - \Phi_1^*(x) \begin{cases} < 0 & \text{if} \quad \mu C(2|1)\, p_1(x) - (1-\mu)\, C(1|2)\, p_2(x) > 0, \\ \geqslant 0 & \text{otherwise}. \end{cases}$$

Hence it follows that $\bar{R}(\Phi^*, \xi) \leqslant \bar{R}(\Phi^0, \xi)$ for any other classification function Φ^0. Thus Φ^* is the μ-Bayes procedure.

It can also be shown that $\Phi^*(x)$ is admissible in the class of all classification functions in the sense that there is no other classification function $\Phi^0(x)$ such that $L_1(\Phi^0) \leqslant L_1(\Phi^*)$ and $L_2(\Phi^0) \leqslant L_2(\Phi^*)$, with the strict inequality holding in at least one case.

It has been demonstrated that Φ^* defined in (2) is a Bayes solution with respect to any prior $\xi(q)$ with mean μ. Examination of (1) indicates that the expected risk $\bar{R}(\Phi^*, \xi)$ is constant for any prior $\xi \in G_1$, where G_1 is the class of all priors with specified first moment μ. It then follows that Φ^* minimizes the maximum expected risk over the class G_1, and such an optimal procedure has previously been designated G_1-minimax, as developed by Robbins (1964). Further discussions of the G_1-minimax criterion are given by Blum & Rosenblatt (1967), and DeRouen & Mitchell (1974).

2·2. *Several populations*

For the general case suppose there are m populations $\Pi_1, \ldots, \Pi_m$ with the ith population having probability density $p_i(x)$ and prior probability q_i. Assume that q is a random vector with prior distribution $\xi(q)$ which is assumed to belong to a class of distributions with fixed mean vector $\mu' = (\mu_1, \ldots, \mu_m)$. Extending the definitions of § 2·1, $\Phi'(x) = \{\Phi_1(x), \ldots, \Phi_m(x)\}$, with the usual restrictions on the $\Phi_i(x)$, is the classification function and

$$L_{ij}(\Phi) = \int \Phi_j(x)\, p_i(x)\, dx \quad (i \neq j = 1, \ldots, m)$$

is the probability of misclassifying an individual from population Π_i as belonging to population Π_j. With the same notation for $C(j|i)$ as before, the risk, or average loss from costs of misclassification, as a function of q and Φ is given by

$$R(\Phi, q) = \sum_{i=1}^{m} q_i \sum_{j \neq i=1}^{m} C(j|i)\, L_{ij}(\Phi).$$

The expected risk is given by

$$\bar{R}(\Phi, \xi) = \sum_{i=1}^{m} \mu_i \left\{ \sum_{j \neq i=1}^{m} C(j|i) L_{ij}(\Phi) \right\},$$

where

$$\mu_i = \int q_i d\xi(q) = E(q_i) \quad (i = 1, \ldots, m).$$

It can be shown as in the case of two populations that the classification function $\Phi^*(x)$ defined for $k = 1, \ldots, m$ by $\Phi_k^*(x) = 1$ if

$$\sum_{i \neq k=1}^{m} \mu_i C(k|i) p_i(x) \leqslant \sum_{i \neq j=1}^{m} \mu_i C(j|i) p_i(x)$$

for all $j = 1, \ldots, m$, and $j \neq k$, with $\Phi_k^*(x) = 0$ otherwise, minimizes the expected risk $\bar{R}(\Phi, \xi)$. Again, since $\bar{R}(\Phi^*, \xi)$ is a function of only the first moments $\mu_1, \ldots, \mu_m$ of the prior distribution of q, it is clear that $\Phi^*(x)$ is the Bayes solution for the class G_1 of all prior distributions ξ with specified first moments $\mu_1, \ldots, \mu_m$; that is $\Phi^*(x)$ is the μ-Bayes or G_1-minimax solution.

Thus for the case of prior distributions of q, the optimal classification function is of the same form as that for fixed prior probabilities, but with the $q_1, \ldots, q_m$ replaced by the first moments $\mu_1, \ldots, \mu_m$ of the distribution of prior probabilities. Use of this procedure thus does not require the knowledge of the prior distributions completely, but only the values of the means $\mu_1, \ldots, \mu_m$. If the discriminant is estimated from a training sample and if this can be considered to be a random sample from the appropriate population, then it is clear that the proportions of the different subpopulations in the training sample can be used as unbiased estimates of these means.

The authors are grateful to the referee for his helpful comments.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Blum, J. R. & Rosenblatt, J. (1967). On partial *a priori* information in statistical inference. *Ann. Math. Statist.* 38, 1671–8.

DeRouen, T. A. & Mitchell, T. J. (1974). A G_1-minimax estimator for a linear combination of binomial probabilities. *J. Am. Statist. Assoc.* 69, 231–3.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* 35, 1–20.

[*Received July* 1974. *Revised February* 1975]