Reprinted from

S.S. Gupta
J.O. Berger
Editors

Statistical Decision Theory and Related Topics IV

Volume 1

Springer-Verlag
New York Berlin Heidelberg
London Paris Tokyo

BAYESIAN LINEAR PROBABILISTIC CLASSIFICATION

JOSÉ M. BERNARDO

Departamento de Estadistica
Universidad de Valencia
46071-Valencia, Spain

1. THE PROBLEM

Let $\{\delta_1, \ldots, \delta_k\}$ be an *exhaustive* list of k different categories which are defined as possible classes within a given population; without loss of generality, we shall assume that those describe mutually exclusive categories for, otherwise, it would suffice to define as different classes those associations of two or more categories which are considered to be compatible. Furthermore, let $\mathbf{x} = (x_1, \ldots, x_m)^t$ be a vector of observable *attributes* which are judged to be possibly related to the categorization considered. We are interested in a *general* robust procedure which could be used to *predict* the category to which an element of the population belongs, given the values of its relevant attributes and the information provided by the *data bank* $D = \{(\mathbf{x}_j, \delta_{(j)}), j = 1, \ldots, n\}$ which contains the category $\delta_{(j)} \in \{\delta_1, \ldots, \delta_k\}$ and the vector of attributes $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm})^t$ of each of the elements of a sample of size n from the population. Common applications include automatic diagnosis, electoral forecasting and biological taxonomy.

The natural answer to the problem posed, and indeed the only solution which is compatible with rather mild coherence requirements (see, for example, Lindley, 1982, and references therein), is to provide a *probability distribution* over the categories

$$\{p(\delta_i|\mathbf{x}, D, H), \quad i = 1, \ldots, k\} \tag{1}$$

which describes the probability that an element of the population with attributes $\mathbf{x}$ belongs to each of the possible categories, conditional on the available data bank D, and on those assumptions H which one is prepared to make about the structure of the problem. The probability distribution (1) is often referred to as the *diagnostic* (predictive) distribution, (see, for example, Dawid, 1976).

It is important to note that even if *precise* classification is desired, say because legal requirements force to *choose* a particular category, the diagnostic distribution just described still is a *necessary* intermediate step. Indeed, standard coherence arguments show (see, for example, Fishburn, 1981, Bernardo *et al.*, 1985, and references therein) that the optimal *decision* is to choose that

category δ_j which minimizes

$$\sum_{i=1}^{k} \ell(\delta_j|\delta_i,\mathbf{x})p(\delta_i|\mathbf{x},D,H) \tag{2}$$

where $\ell(\delta_j|\delta_i,\mathbf{x})$ is the *loss* suffered if δ_j were to be named as the true category for an element with attributes $\mathbf{x}$ whose actual category is δ_i. It is obvious from (2) that the optimal choice is *not* necessarily the most likely category.

In Section 2 we review the general Bayesian approach to the problem posed, the precise history of which may be traced from Geisser (1982), a review paper by one of its main contributors. Section 3 provides a particular solution based on the sufficiency and multinormality of some function of the attributes rather than on the multinormality of the attributes themselves as in Geisser (1966). Section 4 provides a relatively simple decision-theoretical based approximation to that solution. The paper concludes in Section 5 with a general discussion.

2. THE GENERAL BAYESIAN APPROACH

Let $\mathbf{t} = t(x)$ be a *relevant* (possibly vector valued) function of the attributes which is considered to encapsulate most of the diagnostic information contained in the vector of attributes, in the sense that

$$p(\delta_i|\mathbf{x},D,H) \cong p(\delta_i|\mathbf{t},D,H), \qquad i = 1, \ldots, k \tag{3}$$

The trivial special case $\mathbf{t}(\mathbf{x}) = \mathbf{x}$ demonstrates that this is always possible; we shall see however that the derivation of a non-trivial *approximately sufficient* relevant function is often feasible, and makes the problem far more tractable. By Bayes' theorem,

$$p(\delta_i|\mathbf{t},D,H) \propto p(\mathbf{t}|\delta_i,D,H)p(\delta_i|D,H) \tag{4}$$

where $p(\mathbf{t}|\delta_i,D,H)$ is the (posterior predictive) distribution of $\mathbf{t}$ among the elements of the population which belong to category δ_i and $p(\delta_i|D,H)$ is the (posterior) distribution of the categories in the population, given the information provided by the data bank. We now proceed to analyze in detail those two elements.

It is natural to assume that, within each category, the elements of the population may be considered to be *exchangeable*, i.e. that the joint density of any finite set of $\mathbf{t}$ values, $p(\mathbf{t}_1, \ldots, \mathbf{t}_r|\delta_i)$ is invariant under permutation of the $\mathbf{t}_j$'s; in this case, there exists (DeFinetti, 1975, p. 215) an integral representation of the form

$$p(\mathbf{t}_1, \ldots, \mathbf{t}_r|\delta_i) = \int \prod_{j=1}^{r} p(\mathbf{t}_j|\theta_i)\ p(\theta_i)\ d\theta_i \tag{5}$$

so that the set of vectors $\{\mathbf{t}_{ij}, j = 1, \ldots, n_i\}$ observed in the n_i elements of the data bank which belong to category δ_i may be seen as a random sample from some model $p(\mathbf{t}|\theta_i)$ with unknown parameter vector θ_i. It then follows that the predictive distribution in (4) may be obtained as

$$p(\mathbf{t}|\delta_i,D,H) = \int p(\mathbf{t}|\theta_i,H)\ p(\theta_i|\delta_i,D,H)\ d\theta_i \tag{6}$$

where, by Bayes' theorem,

$$p(\theta_i|\delta_i,D,H) \propto \prod_{j=1}^{n_i} p(\mathbf{t}_{ij}|\theta_i,H)\ p(\theta_i|\delta_i,H). \tag{7}$$

Hence, attention should be centered into the specification of the conditional sampling distributions $\{p(\mathbf{t}|\theta_i,H), i = 1, \ldots, k\}$ of the relevant function $\mathbf{t}$ for the k categories and their corresponding prior distributions $\{p(\theta_i|\delta_i,H), i = 1, \ldots, k\}$. A sensible choice of the relevant function $\mathbf{t}$, and a careful specification of its sampling distribution within each category, are essential to obtain useful results.

Under most circumstances, very little prior information will be available on the values of the θ_i's. Hence, it will often be appropriate to approximate $p(\theta_i|\delta_i,H)$ by the 'non-informative' *reference* prior $\pi(\theta_i|\delta_i)$ (Bernardo, 1979b) which is adequate for prediction of $\mathbf{t}$ within category δ_i. Such an operational prior only depends on the model $p(\mathbf{t}|\theta_i,H)$, and may easily be obtained from the asymptotic behaviour of the posterior distribution θ_i.

To derive the second factor in (4), the probability distribution of the categories within the population, given the information provided by the data bank, $\{p(\delta_i|D,H), i = 1, \ldots k\}$, it is necessary to specify the sampling procedure used to obtain the data. In most applications, the data bank is either obtained by random sampling *within* each category (*retrospective* sampling), or by random sampling from the whole population (*prospective* sampling).

With retrospective sampling, the data cannot obviously provide any information about the prevalence of the categories and, hence

$$p(\delta_i|D,H) = p(\delta_i|H), \quad i = 1, \ldots, k \tag{8}$$

where $p(\delta_i|H)$ is the (prior) probability that an element of the population belongs to category δ_i *before* its attributes have been observed. A reference, uniform distribution $\{\pi(\delta_i|H) = 1/k, i = 1, \ldots, k\}$, may be used if no relevant prior information is available

If the data constitute a random sample form the total population and $\delta_{(j)}$ denotes the category associated to the j-th element in the data bank, one has

the multinomial model

$$p(\delta_{(j)}|\varphi_1,\ldots,\varphi_k)=\varphi_{(j)},\quad \varphi_i>0,\quad \sum_{i=1}^{k}\varphi=1 \tag{9}$$

where φ_i is the (unknown) proportion of elements in the population, which belong to category δ_i. It follows that

$$p(\delta_i|D,H)=\int p(\delta_i|\varphi)\ p(\varphi|D,H)\ d\varphi=E[\varphi_i|D,H] \tag{10}$$

where, by Bayes' theorem,

$$p(\varphi|D,H)\propto\prod_{i=1}^{k}\varphi_i^{n_i}\ p(\varphi) \tag{11}$$

If prior information about φ is diffuse and, hence, may be approximately described by the corresponding reference prior $\pi(\varphi)\propto\prod_i\varphi_i^{-1/2}$, then,

$$p(\delta_i|D,H)=E[\varphi_i|D,H]=(n_i+1/2)/(n+k/2) \tag{12}$$

Combining equations (4), (6), (7) and either (8) or (12) depending on the sampling procedure which was used, and assuming that little relevant information is available beyond that provided by the data bank, the desired diagnostic distributions is obtained as

$$\begin{aligned}
&p(\delta_i|\mathbf{x},D,H)\propto p(\mathbf{t}|\delta_i,D)p(\delta_i|D)\\
&p(\mathbf{t}|\delta_i,D)=\int p(\mathbf{t}|\theta_i)\ p(\theta_i|D)\ d\theta_i\\
&p(\theta_i|D)\propto\prod_{j=1}^{n_1}p(\mathbf{t}_{ij}|\theta_i)\pi(\theta_i|\delta_i)\ d\theta_i
\end{aligned} \tag{13}$$

where, for notational convenience, dependence on the general assumptions H has been omitted, $p(\delta_i|D)$ is either $1/k$ or $(n_i+1/2)/(n+k/2)$, respectively depending on whether the sampling was prospective or retrospective, and $\{\pi(\theta_i|\delta_i),\ i=1,\ldots,k\}$ are the reference priors which are appropriate for prediction of $\mathbf{t}$, within each category, from the assumed models $\{p(\mathbf{t}|\theta_i,H),i=1,\ldots,k\}$.

We shall now illustrate this general approach with a particularly effective choice of the relevant function $\mathbf{t}=\mathbf{t}(\mathbf{x})$.

3. THE SUGGESTED MODEL

3.1. THE LINEAR DISCRIMINANT VECTOR

In his famous pioneer work on classification, Fisher (1936) showed that if $\mathbf{x}_{1*}$ and $\mathbf{x}_{2*}$ are the arithmetic means of the vector of attributes which respectively correspond to categories δ_1 and δ_2 and S is their pooled sample covariance matrix, then the function $\mathbf{t}(\mathbf{x})=\lambda^t\mathbf{x}=\lambda_1x_1+\cdots+\lambda_mx_m$, where

$$\lambda^t=(\lambda_1,\ldots,\lambda_m)=(\mathbf{x}_{1*}-\mathbf{x}_{2*})^t\ S^{-1} \tag{14}$$

is the linear function of the attributes which optimally separates δ_1 and δ_2, in the sense of maximizing the standardized difference between the average values of the $\lambda^t\mathbf{x}$'s obtained for each of the two classes. Goel (1983) has recently pointed out that (14) is in fact optimal for a much larger class of distances. Other derivations are given by Wilks (1962), Dempster (1969) and Geisser (1977).

The idea may be generalized in different ways to the case where there are more than two categories. An attractive possibility is to consider the *set* of linear functions

$$\lambda_i^t\mathbf{x}=(\mathbf{x}_{i*}-\mathbf{x}_{k*})^t\ S^{-1}\mathbf{x},\quad i=1,\ldots,k-1 \tag{15}$$

where $\mathbf{x}_{i*}$ is the (vector) mean of the attributes of the elements in the data bank which belong to category δ_i,

$$S=n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(\mathbf{x}_{ij}-\mathbf{x}_{i*})(\mathbf{x}_{ij}-\mathbf{x}_{k*})^t, \tag{16}$$

and $\mathbf{x}_{ij}$ is the j-th vector of attributes within category δ_i. Thus, $\lambda_i^t\mathbf{x}$ is the linear function which best separates category δ_i from δ_k. The k-th category should be chosen to be a *control* category, for, as we shall later establish, the weight of evidence for each of the δ_i's relative to δ_k may easily be provided, using the $\lambda_i^t\mathbf{x}$ values defined above.

We argue that the (vector) function $\mathbf{t}=\mathbf{t}(\mathbf{x})=\{\lambda_1^t\mathbf{x},\ldots,\lambda_{k-1}^t\mathbf{x}\}$ thus defined is often a very good candidate to the role of relevant function defined in Section 2. Indeed, the geometrically based optimality results mentioned above and the remarkable practical results which have been obtained using discriminant analysis (see, for example, Press and Wilson, 1978, or Titterington *et al.*, 1981) suggest that the linear discriminant functions *do* capture most of the information contained in the data bank which is useful for classification purposes. Thus, as a first *approximation* we shall *assume* that the diagnostic distribution mainly depends on the attributes through the corresponding values of the discriminant functions. Formally,

$$p(\delta_i|\mathbf{x},D,H)\cong p(\delta_i|\mathbf{t},D),\quad i=1,\ldots,k \tag{A1}$$

where $\mathbf{t}=\mathbf{t}(\mathbf{x})=\{\lambda_1^t\mathbf{x},\ldots,\lambda_{k-1}^t\mathbf{x}\}$ is the vector in R^{k-1} whose components were defined in (15).

3.2. CONDITIONAL SAMPLING DISTRIBUTIONS OF THE DISCRIMINANT VECTOR

If the number of attributes is *not very small*, which constitutes the stan-

dard situation in most applications, different forms of central limit theorems may be invoked (see, for example, Diaconis and Freedman, 1984) to guarantee that, within each category, the sampling distribution of the discriminant vector $\mathbf{t}$, a linear combination of possibly standardized random quantities, will be approximately normal. Formally,

$$p(\mathbf{t}|\delta_i) = N_{k-1}(\mathbf{t}|\mu_i, \Sigma_i) \propto |\Sigma_i|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{t}-\mu_i)^t \Sigma_i^{-1}(\mathbf{t}-\mu_i)\} \qquad (A2)$$

where $(\mu_i, \Sigma_i), i = 1, \ldots, k$, are unknown parameters. Note that while the discriminant function $\mathbf{t}$ *may* be expected to be normally distributed, the original vector of attributes $\mathbf{x}$, which typically mixes quantitative and qualitative, highly interdependent variables, *cannot* be expected to have a multivariate normal sampling distribution; yet only the normality of $\mathbf{x}$ would guarantee the optimality of most of the currently used classification schemes.

The normality assumption should obviously be tested in any concrete application and substituted, if necessary, for more appropriate distributional assumptions. It is important to note however, that the problem then posed is far easier than the original, for (A1) has reduced the dimensionality of the problem from m (the number of attributes) to $k-1$ (the number of categories minus one); in standard applications, m is of the order of tens while k is typically very small and often equals two.

3.3. Predictive and Diagnostic Distributions

The prior information on the unknown parameters $\{(\mu_i, \Sigma_i), i = 1, \ldots, k\}$ which control the conditional sampling distributions of $\mathbf{t}$ will typically be very diffuse; thus, it will often be appropriate to use the corresponding reference prior which, for prediction from a multinormal model $N_p(\mathbf{t}|\mu, \Sigma)$ of dimension p, turns out to be $\pi(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}$. Thus, we shall further assume,

$$\pi(\mu_i, \Sigma_i) \propto |\Sigma_i|^{-k/2}, \qquad (A3)$$

Obviously, if a different sampling distribution for $\mathbf{t}$ is specified in (A2), then (A3) should be substituted by the appropriate reference prior for prediction from the model chosen.

Assumptions (A2) and (A3) may now be combined to derive the required predictive distributions of $\mathbf{t}$. Indeed, using (6) and (7),

$$p(\mathbf{t}|\delta_i, D) = \int\int N_{k-1}(\mathbf{t}|\mu_i, \Sigma_i)\, p(\mu_i, \Sigma_i|D)\, d\mu_i\, d\Sigma_i$$

$$p(\mu_i, \Sigma_i|D) \propto \prod_{j=1}^{n_i} N_{k-1}(\mathbf{t}_{ij}|\mu_i, \Sigma_i)\, |\Sigma_i|^{-k/2} \qquad (17)$$

where $\mathbf{t}_{ij} = \{\lambda_1^t \mathbf{t}_{ij}, \ldots, \lambda_{k-1}^t \mathbf{t}_{ij}\}$, the λ_i's were defined in (15), and $\mathbf{t}_{ij}$ is the j-th vector of attributes within category δ_i. The resulting predictive densities

are (see, for example, Geisser, 1964) the multivariate Student densities defined by

$$p(\mathbf{t}|\delta_i, D) = St(\mathbf{t}|\mathbf{m}_i,\ (n_i-k+1)^{-1}\ (n_i+1)\mathbf{V}_i,\ n_i-k+1)$$
$$= C_i|\mathbf{V}_i|^{-1/2}\{1+(n_i+1)^{-1}\ (\mathbf{t}-\mathbf{m}_i)^t\ \mathbf{V}_i^{-1}\ (\mathbf{t}-\mathbf{m}_i)\}^{-n_i/2} \qquad (18)$$

where $\mathbf{m}_i$ and $\mathbf{V}_i$ are respectively the vector of means and covariance matrix of the discriminant vectors $\{\mathbf{t}_{ij}, j = 1, \ldots, n_i\}$ which correspond to those elements in the data bank which belong to category δ_i, and C_i is the (relevant) normalizing constant

$$C_i = \{\Gamma(n_i/2)/\Gamma((n_i-k+1)/2)\}\ \{(n_i+1)\pi\}^{-(k-1)/2} \qquad (19)$$

which only becomes irrelevant if all the n_i's are equal.

It now follows from equations (13) in Section 2 that, under the assumptions (A1) to (A3), the diagnostic distribution which provides a probabilistic automatic classification for an element of the population with attributes $\mathbf{x}$ is given by

$$p(\delta_i|\mathbf{x}, D) \propto C_i|\mathbf{V}_i|^{-1/2}\{1+(n_i+1)^{-1}\ (\mathbf{t}-\mathbf{m}_i)^t\ \mathbf{V}_i^{-1}(\mathbf{t}-\mathbf{m}_i)\}^{-n_i/2}\ p(\delta_i|D) \qquad (20)$$

where $\mathbf{t} = \mathbf{t}(\mathbf{x}) = \{\lambda_1^t\mathbf{x}, \ldots, \lambda_{k-1}^t\mathbf{x}\}$ is his corresponding discriminant vector and $p(\delta_i|D)$ describes the available information about the distribution of the categories within the population, which equals either $(n_i+1/2)/(n+k/2)$ or $1/k$, depending on whether the sampling was prospective or retrospective, if the only available information is that provided by the data bank.

For *large* sample sizes, the student distributions in (18) converge to normals and, hence, (20) simplifies to

$$p(\delta_i|\mathbf{x}, D) \propto |\mathbf{V}_i|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{t}-\mathbf{m}_i)^t\ \mathbf{V}_i^{-1}(\mathbf{t}-\mathbf{m}_i)\}\ p(\delta_i|D). \qquad (21)$$

However, this approximation may be rather misleading with small data banks (see, for example, Aitchison and Dunsmore, 1975, p. 228).

The implementation of the proposed method to obtain the diagnostic distribution requires, for *each* new element to be classified, a series of computations which are often too involved for the numeracy of potential users. In the next section, we derive a linear approximation which may be routinely used by anyone with basic arithmetic.

4. LINEAR APPROXIMATION AS A DECISION PROBLEM

Classification procedures based on logistic regression are popular with practitioners because the resulting diagnostic distributions for new elements only depend on a fixed linear combination of their attributes. Moreover, logistic re-

gression models are often successful because they provide a large, flexible class of diagnostic distributions which may be expected to fit most situations. The logistic (additive) model directly assumes that the diagnostic distribution is of the form

$$\log\{p(\delta_i|\mathbf{x},\omega)/p(\delta_k|\mathbf{x},\omega)\} = \omega_{i0}+\omega_{i1}x_1+\cdots+\omega_{im}x_m, \quad i=1,\ldots,k-1 \quad (22)$$

and uses the data bank D to estimate the $(k-1)\times(m+1)$ matrix of unknown parameters $\{\omega_{ij}, i=1,\ldots,k=1,\ j=0,\ldots,m\}$. Cox (1970) and Anderson (1972) propose to use the corresponding maximum likelihood estimates. It has been demonstrated however that this procedure overfits the data bank and may produce rather misleading diagnostic distributions. Moreover, maximum-likelihood methods are not immediately applicable to the rather ubiquitous retrospective sampling case.

From a Bayesian point of view, estimation of the ω_{ij}'s is a well posed decision problem where the action space is the class of possible values of the matrix ω, the relevant uncertain event is the value $\mathbf{x}$ of the vector of attributes of the next element to be classified, and the loss function $\ell(\omega,\mathbf{x})$ is some measure of the discrepancy between the actual diagnostic distribution $\{p(\delta_i|\mathbf{x},D), i=1,\ldots,k\}$ and its logistic approximation $\{p(\delta_i|\mathbf{x},\omega), i=1,\ldots,k\}$.

The loss suffered when the probability distribution P of the discrete random quantity δ is approximated by another distribution $\mathcal{Q}$ may be expressed as the expected value of the conditional loss, given each value of δ,

$$\sum_{i=1}^{k} p_i\{u(P,\delta_i)-u(\mathcal{Q},\delta_i)\} \quad (23)$$

where $u(P,\delta_i)$ is the *utility* of the (predictive) distribution P, conditional on the true category δ_i of the element considered. Functions of this type are usually referred to as *scoring rules*, for they provide the *score* $u(P,\delta)$ to be accorded to a forecaster who produced a prediction P on the category of an element whose actual category is found to be δ. An obvious requirement for the induced loss function (23) is to be nonnegative and only zero if $P=\mathcal{Q}$; this implies that the function u must verify the condition $\sup_{\mathcal{Q}}\ \Sigma_i\ p_i\ u(\mathcal{Q},\delta_i)=\Sigma_i\ p_i\ u(P,\delta_i)$, which is the definition of a *proper* scoring rule.

Different scoring rules will induce different loss functions; the *quadratic* score $u(P,\delta_i)=2p_i-\Sigma_j p_j^2$, and the *logarithmic* score $u(P,\delta_i)=\log p_i$, are the most popular. It may be shown however (Savage, 1971) that, when there are more than two categories, the logarithmic is the only proper scoring rule whose values only depend on the probability associated to the true category. Moreover, the discrepancy loss induced by the logarithmic score, namely $\Sigma_i p_i \log\{p_i/q_i\}$, is the logarithmic divergence, whose attractive properties (Kullback and Leibler, 1951; Bernardo, 1979a) are well known. With this particular choice, the loss

function of our original decision problem becomes

$$\ell(\omega,\mathbf{x})=\sum_{i=1}^{k} p(\delta_i|\mathbf{x},D)\ \log\{p(\delta_i|\mathbf{x},D)/p(\delta_i|\mathbf{x},\omega)\} \quad (A4)$$

whose *expected* value is minimized if, and only if, one maximizes

$$\int p(\mathbf{x}|D)\sum_{i=1}^{k} p(\delta_i|\mathbf{x},D)\ \log\ p(\delta_i|\mathbf{x},\omega)\ d\mathbf{x} \quad (24)$$

Since the data bank provides random samples of attribute vectors from the population, a Monte Carlo approximation to this integral is provided by one of the following equations,

$$n^{-1}\sum_{j=1}^{n}\sum_{i=1}^{k} p(\delta_i|\mathbf{x}_j,D)\ \log\ p(\delta_i|\mathbf{x}_j,\omega) \quad (25)$$

$$k^{-1}\sum_{i=1}^{k} n_i^{-1}\sum_{j=1}^{n_i}\sum_{\ell=1}^{k} p(\delta_\ell|\mathbf{x}_{\ell j},D)\ \log\ p(\delta_\ell|\mathbf{x}_{\ell j},\omega), \quad (26)$$

depending on whether the sampling was respectively prospective or retrospective. To obtain the value of ω which maximizes either of those expressions, one may use a Newton-Raphson procedure similar to that used by maximum-likelihood logistic regression; as a matter of fact, (25) may be viewed as a *weighted* log-likelihood where each of the elements in the standard log-likelihood is weighted by its predictive probability.

Once the optimum value of ω, say $\hat{\omega}$, has been found, a linear approximation to the diagnostic distribution of a new element with attributes $\mathbf{x}$ is simply provided by

$$\ell_i(\mathbf{x})=\log\{p(\delta_i|\mathbf{x},D)/p(\delta_k|\mathbf{x},D)\}\cong\hat{\omega}_{i0}+\hat{\omega}_{i1}x_1+\cdots+\hat{\omega}_{im}x_m, \quad i=1,\ldots,k-1 \quad (27)$$

so that the linear combination $\ell_i(\mathbf{x})$ directly provides, on a $(-\infty,\infty)$ scale, the relevant log-odds, i.e. the *weight of evidence* (Good , 1950) for category δ_i relative to the category δ_k which was used as control. Naturally, the log-odds $\{\ell_1(\mathbf{x}),\ldots,\ell_{k-1}(\mathbf{x})\}$ may easily be transformed into probabilities; indeed,

$$p(\delta_k|\mathbf{x},D)=1/\{1+\sum_{i=1}^{k-1}\exp(\ell_i(\mathbf{x}))\}$$

$$p(\delta_i|\mathbf{x},D)=p(\delta_k|\mathbf{x},D)\ \exp(\ell_i(\mathbf{x})), \quad i=1,\ldots,k-1. \quad (28)$$

However, experience shows that the use of log-odds is often both more intuitive and more convenient, especially when extreme probabilities arise, as it is often the case in standard problems of probabilistic classification.

5 DISCUSSION

The model described in section 3 is only an example of how the general Bayesian methodology may successfully be applied to automatic classification; indeed, other relevant functions and/or other assumptions on their conditional sampling distributions may turn out to be more appropriate in specific problems. We claim, however, that the model described performs well in a very large class of problems and, as demonstrated in a large number of both simulated and practical examples on which it has been tried, it may be expected to outperform established classification methods.

It should be obvious to the reader that the possible applications of probabilistic classification are virtually endless Indeed, we have successfully used the proposed procedure in medical diagnosis (classification of patients by diseases), geology (land areas by oil potential), biology (animals and plants by species), sociology (citizens by political behaviour), psychology (individuals by abilities) and economics (companies by tax categories).

The proper score functions introduced in Section 4 to derive the linear approximation may also be used both to *evaluate* comparatively different probabilistic classification procedures and to address the problem of the *choice of variables*, therefore reducing the initial (usually very large) number of attributes to those really necessary to obtain good diagnostic distributions. The details are provided in Bernardo and Bermúdez (1985).

ACKNOWLEDGMENTS

The author is grateful to a referee for pointing out some missing, relevant references

BIBLIOGRAPHY

Aitchison, J. and Dunsmore, I. R (1975). *Statistical Prediction Analysis*. Cambridge: University Press.

Anderson, J. A. (1972). Separate sampling logistic classification. *Biometrika* **67**, 217–272.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist* **7**, 686–690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference, *J. Roy. Statist Soc. B* **41**, 113–147 (with discussion).

Bernardo, J. M. and Bermudez, J. D. (1985). The choice of variables in probabilistic classification. *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 67–81 (with discussion) Amsterdam: North-Holland.

Bernardo, J. M., Ferrandiz, J. R. and Smith, A. F. M. (1985). The foundations of decision theory: an intuitive, operational approach with mathematical extensions. *Theory and Decision* **19**, 127–150.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics* **32**, 647–658.

De Finetti, B. (1975). *Theory of Probability, Vol 1*. New York: Wiley.

Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis*. Reading, Mass.: Addison-Wesley.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815

Fisher, R. A. (1936). The use of multiple measurements in taxonomical problems. *Ann. Eugenetics* **7**, 179–188.

Fishburn, P. C. (1981). Subjective expected utility: a review of normative theories. *Theory and Decision* **13**, 139–199.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *J. Roy. Statist. Soc. B* **26**, 69–76.

Geisser, S. (1966). Predictive discrimination *Multivariate Analysis*. (P. Krishnaiah, ed.), 149–163. New York: Academic Press.

Geisser, S. (1977). Discrimination, allocatory and separatory linear aspects. *Classification and Clustering* (J. Van Ryzin, ed.), 301–330. New York: Academic Press.

Geisser, S. (1982). Bayesian discrimination. *Handbook of Statistics 2*, (P. Krishnaiah and L. Kanal, eds), 101–120. Amsterdam: North Holland.

Goel, P. K. (1983). Information measures and Bayesian hierarchical models. *J. Amer Statist. Assoc* **78**, 408–410.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann Math. Statist.* **22**, 525–540

Lindley, D V. (1982). Scoring rules and the inevitability of probability. *Internat. Statist. Rev* **50**, 1–26 (with discussion).

Press, S. J. and Wilson, S. (1978). Chosing between logistic regression and discriminant analysis. *J. Amer. Statist. Assoc.* **73**, 699–705.

Savage, L. J. (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 783–801.

Titterington, D. M. *et al* (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *J. Roy. Statist. Soc. A* **144**, 145–175 (with discussion).

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.