Comp. & Maths. with Appls. Vol. 12A, No. 2, pp. 261–272, 1986
Printed in Great Britain.

0097-4943/86 $3.00 + .00

ASSESSING THE PERFORMANCE OF AN ALLOCATION RULE

G. J. McLACHLAN
Department of Mathematics, University of Queensland, St. Lucia 4067, Australia

Abstract—The problem of estimating the error rates of a sample-based rule on the basis of the same sample used in its construction is considered. The apparent error rate is an obvious nonparametric estimate of the conditional error rate of a sample rule, but unfortunately it provides too optimistic an assessment. Attention is focussed on the formation of improved estimates, mainly through appropriate bias correction of the apparent error rate. In this respect the role of the bootstrap, a computer-based methodology, is highlighted.

1. INTRODUCTION

We consider the problem of estimating the various types of error rates associated with the application of an allocation rule. Suppose there are g possible classes denoted by $C_1, \ldots, C_g$ with prior probabilities $\pi_1, \ldots, \pi_g$, respectively. The elements of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ are nonnegative and sum to 1, so $\boldsymbol{\pi}$ is an element of Δ_g, the unit simplex in R^g. The aim is to assign an unclassified object from one of these classes to its correct class of origin on the basis of the observed value of a p-dimensional random vector $\mathbf{X}$. In class C_i, $\mathbf{X}$ has probability density function $f_i(\mathbf{x})$ (with respect to arbitrary measure) on R^p. Let $\mathbf{Z}$ denote the random vector $(\mathbf{X}, Y)$ with density function $f(\mathbf{z})$ defined on R^{p+1}, where Y is a random variable taking on the values 1 to g so that the value of Y specifies the class to which the object belongs. The discrimination problem can therefore be expressed as attempting to predict the unknown value of Y, y, having observed $\mathbf{X} = \mathbf{x}$.

We let $r(\mathbf{x})$ denote an allocation rule for predicting y, where $r(\mathbf{x}) = i$ implies that an object with $\mathbf{X} = \mathbf{x}$ is assigned to class $C_i (i = 1, \ldots, g)$. In addition to the observed value $\mathbf{x}$, the rule r may also depend on $\boldsymbol{\pi}$, and so it may be thought of as a measurable function

$$r: R^p \times \Delta_g \longrightarrow \{1, \ldots, g\}.$$

The optimal or Bayes rule, which minimizes the error rate averaged with respect to the prior probabilities, is defined to be i, if

$$\xi_i(\mathbf{x}) > \xi_j(\mathbf{x}) \quad (j = 1, \ldots g; j \neq i), \tag{1}$$

where

$$\xi_j(\mathbf{x}) = \pi_j f_j(\mathbf{x}) \Big/ \left\{ \sum_{k=1}^{g} \pi_k f_k(\mathbf{x}) \right\} \tag{2}$$

is the posterior probability that an object with $\mathbf{X} = \mathbf{x}$ belongs to C_j (Anderson[3]). That is, the Bayes rule chooses the class which maximizes the posterior probability. For example, for $g = 2$ normal classes, where

$$\mathbf{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \quad \text{in } C_i (i = 1, 2), \tag{3}$$

we have from (1) that the Bayes rule is 1 or 2, according to whether the linear discriminant function

$$\begin{aligned} L(\mathbf{x}) &= \log\{f_1(\mathbf{x})/f_2(\mathbf{x})\} + \log(\pi_1/\pi_2) \\ &= \left\{ \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \right\} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' + \lambda \end{aligned} \tag{4}$$

is greater or less than zero, and $\lambda = \log(\pi_1/\pi_2)$. The minimax rule is based on (4) with $\lambda = 0$, which also corresponds to the case of equal prior probabilities.

In practice, the class conditional density functions $f_i(\mathbf{x})$ and the prior probabilities π_i are usually unknown, and so the chosen rule of allocation $r(\mathbf{x})$ may not be able to be used in its desired form. For instance, $r(\mathbf{x})$ might be the Bayes rule, which we have seen requires knowledge of all these quantities or, say, the minimax rule which depends on the class conditional densities but not the prior probabilities. For the construction of a suitable sample version of $r(\mathbf{x})$, it is assumed here there are available independent training observations of known origin; that is, there is a set

$$\mathbf{t} = \{\mathbf{z}_1 = (\mathbf{x}_1, y_1), \ldots, \mathbf{z}_n = (\mathbf{x}_n, y_n)\} \tag{5}$$

for which $y_1, \ldots, y_n$ are known and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently distributed. For convenience of notation we relabel $\mathbf{x}_1, \ldots, \mathbf{x}_n$ so that $\mathbf{x}_{ij}$ $(j = 1, \ldots, n_i)$ denote those n_i realisations of $\mathbf{X}$ belonging to C_i $(i = 1, \ldots, g; n_1 + \cdots + n_g = n)$; that is, $\mathbf{t} = \{(\mathbf{x}_{ij}, i), j = 1, \ldots, n_i; i = 1, \ldots, g\}$. The observations $\mathbf{x}_{ij}$ may have been obtained either by sampling separately from each of the classes or from a mixture of the classes in proportions $\pi_1, \ldots, \pi_g$. Under the latter scheme, each case in $\mathbf{t}$ is a realisation of the random variable $\mathbf{Z}$ distributed according to $f(\mathbf{z})$, and so

$$n_i \sim \text{bin}(n, \pi_i) \quad (i = 1, \ldots, g), \tag{6}$$

providing $\hat{\pi}_i = n_i/n$ as an estimate of π_i. With the former scheme, the number of observations from C_i is fixed in advance before sampling, and hence the $\mathbf{x}_{ij}$ $(j = 1, \ldots, n_i)$ constitute a random sample from the ith-class conditional density, $f_i(\mathbf{x})$, $i = 1, \ldots, g$.

We let $r(\mathbf{x}, \mathbf{t})$ denote the sample version of $r(\mathbf{x})$ constructed from $\mathbf{t}$ in a consistent manner so that $r(\mathbf{x}, \mathbf{t}_\infty) = r(\mathbf{x})$ except for sets of probability zero, where $r(\mathbf{x}, \mathbf{t}_\infty)$ represents the rule that would be obtained if the size of the training set were increased to infinity. The sample rule $r(\mathbf{x}, \mathbf{t})$ may be constructed in a nonparametric framework from $\mathbf{t}$ using, say, the kernel method to estimate the class conditional densities $f_i(\mathbf{x})$. A survey of this method may be found in the recent book on the subject by Hand[14]. Alternatively, a known parametric family may be adopted for the $f_i(\mathbf{x})$ or, as with logistic regression, for the posterior probabilities. The allocation rule then has the parametric form $r(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the vector of unknown parameters associated with the parametric formulation. A popular way of proceeding, referred to as the estimative approach, is to take the sample rule to be $r(\mathbf{x}, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the value of some consistent and asymptotically efficient estimator of $\boldsymbol{\theta}$ based on $\mathbf{t}$, for example, the maximum-likelihood (ML) estimate. An example of this approach in the context of logistic regression may be found in McLachlan[26], who considered the bias correction of the ML estimate before its use in forming the estimated posterior probabilities and, hence, the sample rule. Aitchison *et al.*[1] have recommended a Bayesian approach to the estimation of $\boldsymbol{\theta}$, whereby the $f_i(\mathbf{x})$ are replaced by their predictive estimates constructed by adopting some prior distribution for $\boldsymbol{\theta}$.

The estimative approach for model (3) leads to the sample rule $r(\mathbf{x}, \mathbf{t})$, defined to be 1 or 2 according to whether

$$\hat{L}(\mathbf{x}) = \left\{\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)\right\} \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' + \hat{\lambda} \tag{7}$$

is greater or less than zero, where

$$\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij}/n_i,$$

$$\mathbf{S}_i = \left\{\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)\right\} \Big/ (n_i - 1),$$

and

$$\mathbf{S} = \{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2\}/(n - 2).$$

Apart from the cutoff point $\hat{\lambda}$, which is often taken to be zero, $\hat{L}(\mathbf{x})$ is Fisher's linear discriminant function as modified by Anderson[2]. In this example, it can be seen that $r(\mathbf{x}, \mathbf{t})$ is invariant under a permutation of the $\mathbf{x}_{ij}$ for each i; that is, symmetrically defined in $\mathbf{z}_1, \ldots, \mathbf{z}_n$. This will generally be the case in practice and this assumption is adopted here.

2. TYPES OF ERROR RATES

We consider $g = 2$ classes in defining the error rates associated with the sample rule $r(\mathbf{x}, \mathbf{t})$, although the definitions extend in an obvious manner to $g > 2$. The error rates of $r(\mathbf{x}, \mathbf{t})$ averaged over the distribution of $\mathbf{X}$ within a given class are denoted by $ec_i(f_i, \mathbf{t})$, where

$$ec_i(f_i, \mathbf{t}) = \text{pr}\{r(\mathbf{X}, \mathbf{t}) = 3 - i | \mathbf{X} \in C_i; \mathbf{t}\} \quad (i = 1, 2)$$

is the probability that a randomly chosen member of C_i is misallocated. As the notation implies, the errors $ec_i(f_i, \mathbf{t})$ are conditional on the training data $\mathbf{t}$, and the rate with respect to the ith class depends on the density of $\mathbf{X}$ in that class. They are referred to in the literature as the conditional or actual error rates. Their expectations over the sampling distribution of the training data $\mathbf{t}$ give the unconditional or expected error rates,

$$\begin{aligned} eu_i(f) &= E\{ec_i(f_i, \mathbf{T})\} \\ &= \text{pr}\{r(\mathbf{X}, \mathbf{T}) = 3 - i | \mathbf{X} \in C_i\} \quad (i = 1, 2), \end{aligned}$$

where $\mathbf{T}$ is the random quantity with $\mathbf{t}$ as a realisation. The unconditional error with respect to the ith class depends on the density of $\mathbf{X}$, not only in that class but also in the other classes, and on $\boldsymbol{\pi}$ if the prior probabilities are used in the formulation of the rule (McLachlan[18]).

The quantities

$$eo_i(f) = ec_i(f_i, \mathbf{t}_\infty) \quad (i = 1, 2)$$

are the errors associated with the desired rule $r(\mathbf{x})$, since it is assumed that $r(\mathbf{x}, \mathbf{t}_\infty)$ is equivalent to $r(\mathbf{x})$. We shall refer to the $eo_i(f)$ as the optimal error rates, although $r(\mathbf{x})$ may not be optimal in the sense of being the Bayes rule; for instance, $r(\mathbf{x})$ might be the minimax rule.

The overall conditional error rate is given by

$$ec(f, \mathbf{t}) = \pi_1 ec_1(f_1, \mathbf{t}) + \pi_2 ec_2(f_2, \mathbf{t})$$

and, similarly, $eu(f)$ and $eo(f)$ denote the overall unconditional and optimal error rates, respectively.

It can be seen that the conditional errors, as well as the unconditional and optimal rates, depend on the unknown densities $f_i(\mathbf{x})$, and therefore must be estimated. There is a vast literature on the problem of estimating the error rates, and extensive bibliographies may be found in Hills[15], Lachenbruch[16], McLachlan[19] and Toussaint[32], among others.

McLachlan[21] has investigated the relationship between the separate problems of estimating each of the three types of error rate (conditional, unconditional and optimal) associated with the sample linear discriminant function (7). For a given training set $\mathbf{t}$, it is the conditional errors $ec_i(f_i, \mathbf{t})$ which are of prime concern. Interest of the optimal errors is limited in practice to the extent that they represent the errors of the best obtainable version of the given rule. In the subsequent work we concentrate on the estimation of the conditional error rates on the basis of the training set $\mathbf{t}$.

3. APPARENT ERROR RATE

An obvious nonparametric estimator of the conditional error rate, $ec_i(f_i, \mathbf{t})$, is the apparent error rate, A_i, of $r(\mathbf{x}, \mathbf{t})$ when it is applied to the training observations known to belong to C_i. That is, A_i is the proportion of the n_i observations from C_i misallocated by $r(\mathbf{x}, \mathbf{t})$, and so we can write A_i as

$$A_i = \sum_{j=1}^{n_i} Q[i, r(\mathbf{x}_{ij}, \mathbf{t})]/n_i \quad (i = 1, 2),$$

where, for any i and k, $Q[i, k] = 0$ for $i = k$ and 1 for $i \neq k$. This method of estimation was first suggested by Smith[30] in connection with the sample quadratic discriminant function. It is well known that A_i gives too optimistic an estimate of the conditional error, $ec_i(f_i, \mathbf{t})$, as it is based on the same data $\mathbf{t}$ from which $r(\mathbf{x}, \mathbf{t})$ was constructed. Therefore we focus attention now on the bias correction of the apparent error rate. Without loss of generality we consider the bias correction of A_1. On letting $W_1 = A_1 - ec_1(f_1, \mathbf{T})$, we have that the bias of A_1 in estimating the conditional error rate, $ec_1(f_1, \mathbf{t})$, is given by

$$\begin{aligned} \text{bias}(A_1) &= E(W_1) \\ &= b_1, \end{aligned}$$

say, where the dependence of b_1 on the densities $f_i(\mathbf{x})$ has been suppressed.

4. BIAS CORRECTION (CROSS-VALIDATION AND THE JACKKNIFE)

Methods of bias correction of the apparent error rate that have been used include cross-validation, the Quenouille–Tukey jackknife and the recently proposed bootstrap of Efron[6]. An excellent account of these methods has been given by Efron[7], who has exhibited the close theoretical relationship between them.

The cross-validation estimate of the conditional error of the sample rule, $r(\mathbf{x}, \mathbf{t})$, is

$$A_1^{CV} = \sum_{j=1}^{n_1} Q[1, r(\mathbf{x}_{1j}, \mathbf{t}_{(1j)})]/n_1,$$

where $\mathbf{t}_{(1j)}$ denotes $\mathbf{t}$ with the point $(\mathbf{x}_{1j}, 1)$ deleted. Hence, before the sample rule is applied to $\mathbf{x}_{1j}$, it is deleted from the training set and the rule recalculated on the basis of $\mathbf{t}_{(1j)}$; see Lachenbruch and Mickey[17].

There has been confusion in the literature over the roles of cross-validation and the jackknife in correcting the apparent error rate for bias. This is understandable as both methods delete one or more observation at a time in forming the bias corrected estimates. According to Stone[31], "Gray and Schucany ([13], p. 125–136) appear to initiate the confusion in their description of Mosteller and Tukey's sophisticated, simultaneous juggling act with the two concepts." Consider the jackknifed version of the apparent error rate given by

$$A_1^{J0} = A_1 + (n - 1)(A_1 - A_{1(\cdot)}),$$

where

$$A_{1(\cdot)} = \sum_{j=1}^{n_1} A_{1(1j)}/n_1$$

and $A_{1(1j)}$ denotes the apparent error rate of $r(\mathbf{x}, \mathbf{t}_{(1j)})$ when applied to the members of $\mathbf{t}_{(1j)}$ from C_1; that is,

$$A_{1(1j)} = \sum_{k \neq j}^{n_1} Q[1, r(\mathbf{x}_{1k}, \mathbf{t}_{(1j)})]/(n_1 - 1).$$

This jackknifed form of A_1 is appropriate for the estimation of the optimal error $eo_1(f)$, as in this context the bias of A_1 is reduced to the second order with respect to the reciprocal of the size of the training sample. But $A_1^{J_0}$ is frequently used or suggested as an estimate of the conditional error $ec_1(f_1, \mathbf{t})$, as in Crask and Perreault[5]. However, in estimating the conditional error $ec_1(f_1, \mathbf{t})$, the bias of $A_1^{J_0}$ is still of the first order. It follows from Chap. 7 of Efron[7] that the jackknifed version of A_1, which reduces its bias as an estimator of $ec_1(f_1, \mathbf{t})$ to the second order, can be written as

$$A_1^{J_c} = A_1 + (n - 1)(A_1^{\dagger} - A_{1(\cdot)}), \tag{8}$$

where

$$A_1^{\dagger} = \frac{1}{n_1} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \frac{1}{n_1} Q[1, r(\mathbf{x}_{1k}, \mathbf{t}_{(1j)})];$$

see, also, Efron and Gong[9]. Efron[7] noted that the last term on the right-hand side of (8) can be rearranged to give

$$A_1^{J_c} = A_1^{CV} + A_1 - \frac{1}{n_1} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \frac{1}{n_1} Q[1, r(\mathbf{x}_{1k}, \mathbf{t}_{(1j)})],$$

demonstrating the close relationship between the jackknife and the cross-validation methods of bias correction of the apparent error rate in estimating the conditional error of a sample rule. Also, he showed how the jackknife estimate of bias ($A_1 - A_1^{J_c}$ in this instance) can be considered as a quadratic approximation to the nonparametric bootstrap estimate of bias to be defined in the next section. The underlying assumption here that $r(\mathbf{x}, \mathbf{t})$ is symmetrically defined in $\mathbf{z}_1, \ldots, \mathbf{z}_n$ has to be strengthened to $r(\mathbf{x}, \mathbf{t})$ depending on $\mathbf{z}_1, \ldots, \mathbf{z}_n$ through a functional statistic in order to establish the above connection between the bootstrap, cross-validation and the jackknife.

5. BIAS CORRECTION (THE BOOTSTRAP)

The "bootstrap," which is a computer-based methodology, was introduced by Efron[6] for assessing the variability in an estimate on the basis of the data at hand. By resampling the original observations in a way so as to preserve the stochastic structure, pseudodata (bootstrap samples) are obtained on which the estimator of interest can be assessed.

We now consider the application of the bootstrap in the present context of correcting the apparent error rate of a sample allocation rule for bias. The bias correction of A_1 in estimating the conditional error, $ec_1(f_1, \mathbf{t})$, may be implemented according to the bootstrap method as follows.

Step 1. In the case of mixture sampling, a new training set,

$$\mathbf{t}^* = \{\mathbf{z}_1^* = (\mathbf{x}_1^*, y_1^*), \ldots, \mathbf{z}_n^* = (\mathbf{x}_n^*, y_n^*)\},$$

called the bootstrap sample, is generated according to $\hat{f}(\mathbf{z})$, an estimate of the density formed from the original training data $\mathbf{t}$. That is, $\mathbf{t}^*$ consists of the observed values of an independent and identically distributed (i.i.d.) random sample, $\mathbf{Z}_1^*, \ldots, \mathbf{Z}_n^*$, from $\hat{f}(\mathbf{z})$. As with the original observations, we relabel $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ so that $\mathbf{x}_{ij}^* (j = 1, \ldots, n_i^*)$ denote those $\mathbf{x}_j^*$ observations, n_i^* in number, for which $y_j^* = i (i = 1, \ldots, g; n_1^* + \cdots + n_g^* = n)$.

With separate sampling, the bootstrap training set is $\mathbf{t}^* = \{(\mathbf{x}_{ij}^*, i), j = 1, \ldots, n_i; i = 1, \ldots, g\}$, where the class label i is specified before sampling and the $\mathbf{x}_{ij}^* (j = 1, \ldots, n_i)$ are generated then according to an estimate of the ith-class conditional density, $\hat{f}_i(\mathbf{x})$, $i = 1, \ldots, g$. That is, the $\mathbf{x}_{ij}^* (j = 1, \ldots, n_i)$ are the observed values of an i.i.d. sample, $\mathbf{X}_{i1}^*, \ldots, \mathbf{X}_{in_i}^*$, from $\hat{f}_i(\mathbf{x})$.

Step 2. The rule, $r(\mathbf{x}, \mathbf{t}^*)$, is formed from the bootstrap training data $\mathbf{t}^*$ in precisely the same manner as $r(\mathbf{x}, \mathbf{t})$ was from the original set $\mathbf{t}$.

Step 3. The apparent error rate of $r(\mathbf{x}, \mathbf{t}^*)$ with respect to the first class, A_1^*, is computed by noting the proportion of the members in $\mathbf{t}^*$ belonging to C_1 misallocated by $r(\mathbf{x}, \mathbf{t}^*)$. Also, the difference

$$w_1^* = A_1^* - ec_1(\hat{f}_1, \mathbf{t}^*) \tag{9}$$

is computed, where $ec_1(\hat{f}_1, \mathbf{t}^*)$ is the error rate obtained by averaging over $\mathbf{X}$ with respect to the density estimate, $\hat{f}_1(\mathbf{x})$; it is conditional on the bootstrap data $\mathbf{t}^*$.

Step 4. Let W_1^* be the random variable defined according to (9). Then its expectation, the bootstrap bias of the apparent error rate, can be approximated by averaging w_1^* over M repeated independent realisation (say, $M = 50$ or 100) of bootstrap samples $\mathbf{t}_m^*(m = 1, \ldots, M)$. That is,

$$\begin{aligned} E^*(W_1^*) &= E^*\{A_1^* - ec_1(\hat{f}_1, \mathbf{T}^*)\} \\ &\approx \bar{w}_1^*, \end{aligned}$$

where

$$\bar{w}_1^* = \sum_{m=1}^{M} w_{1m}^*/M, \tag{10}$$

and where E^* refers to expectation with respect to the bootstrap distribution of the training data $\mathbf{T}^*$, and w_{1m}^* denotes the value of W_1^* on the mth bootstrap realisation $\mathbf{t}_m^*$. The bootstrap estimate of the bias of A_1, b_1^B, is taken then to be

$$b_1^B = \bar{w}_1^*,$$

and so the apparent error rate corrected for bias according to the bootstrap is given by

$$A_1^B = A_1 - b_1^B.$$

In Step 1 of the above algorithm, the nonparametric version of the bootstrap would under mixture sampling take $\hat{f}(\mathbf{z})$ to be $\hat{f}_0(\mathbf{z})$, the empirical probability function with mass $1/n$ at each original data point $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ in $\mathbf{t}(j = 1, \ldots, n)$. Similarly, under separate sampling, $\hat{f}_i(\mathbf{x})$ would be $\hat{f}_{i0}(\mathbf{x})$, the empirical probability function equal to $1/n_i$ at $\mathbf{x} = \mathbf{x}_{ij}(j = 1, \ldots, n_i)$. Under either sampling scheme with the nonparametric bootstrap, the rate $ec_1(\hat{f}_1, \mathbf{t}^*)$ in (9) is given by

$$ec_1(\hat{f}_1, \mathbf{t}^*) = \sum_{j=1}^{n_1} Q[1, r(\mathbf{x}_{1j}, \mathbf{t}^*)]/n_1.$$

The reader is referred to Efron ([7], p. 30) for ways of smoothing the empirical distribution for use in generating the bootstrap data.

The bootstrap is a very powerful technique and it can be used to assess other sampling properties of the apparent error rate besides its bias. For instance, an estimate of the mean-squared error (MSE) of A_1, in estimating the conditional error $ec_1(f_1, \mathbf{t})$, is provided by

$$\mathrm{MSE}^B(A_1) = \sum_{m=1}^{M} \{A_{1m}^* - ec_1(\hat{f}_1, \mathbf{t}_m^*)\}^2/M, \tag{11}$$

where the right-hand side of (10) is the Monte Carlo approximation to the bootstrap MSE of $ec_1(\hat{f}_1, \mathbf{T}^*)$. Note that the bootstrap sample variance of W_1^*,

$$\sum_{m=1}^{M} (w_{1m}^* - \bar{w}_1^*)^2/(M - 1),$$

suggests a lower bound for the MSE of A_1^B in estimating $ec_1(f_1, \mathbf{t})$. For, the true variance of W_1 can be viewed as the MSE of the "ideal constant" estimator

$$A_1^{IC} = A_1 - b_1,$$

and it would be expected that A_1^B would have MSE at least as large as A_1^{IC}; see Efron and Gong ([9], p. 48).

The bootstrap can be used also to assess the performance of the apparent error rate in its estimation of the other types of error rates. Replacing $ec_1(\hat{f}_1, \mathbf{t}_m^*)$ by $eo_1(\hat{f})$ in (10) and (11) yield the bootstrap estimates of the bias and MSE, respectively, of A_1 in estimating the optimal error $eo_1(f)$. For the nonparametric version of the bootstrap, where each $\hat{f}_i(\mathbf{x})$ is the empirical probability function, $eo_1(\hat{f}) = A_1$ at least if $r(\mathbf{x}, \mathbf{t})$ depends on $\mathbf{t}$ through $\hat{f}_0(\mathbf{z})$. Similarly, an assessment of the MSE of A_1, in estimating the unconditional error $eu_1(f)$, is obtained by replacing $ec_1(\hat{f}_1, \mathbf{t}_m^*)$ with

$$\sum_{m=1}^{M} ec_1(\hat{f}_1, \mathbf{t}_m^*)/M \tag{12}$$

in (11): the Monte Carlo approximation to the bootstrap expectation of $ec_1(\hat{f}_1, \mathbf{T}^*)$; (12) is the bootstrap estimate of the unconditional error, $eu_1(f)$.

For the rule based on the sample linear discriminant function (7) with $\hat{\lambda} = 0$, McLachlan[24] showed that under (3) the bias of the apparent error rate in estimating the conditional error is equal (up to terms of the second order) to

$$b_1 \approx \beta_1(\Delta),$$

where

$$\beta_1(\Delta) = \phi\left(-\frac{1}{2}\Delta\right)\left[\left\{\frac{1}{4}\Delta + (p-1)/\Delta\right\}\Big/ n_1 + \frac{1}{2}(p-1)(\Delta/N)\right],$$

and where

$$\Delta = \{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\}^{1/2}$$

is the Mahalanobis distance between C_1 and C_2, ϕ denotes the standard normal density function, and $N = n_1 + n_2 - 2$. If Δ is now replaced by its sample counterpart,

$$D = \{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\}^{1/2},$$

then $\beta_1(D)$ is the estimate of the bias corresponding to the parametric delta method. It also can be viewed as the bootstrap bias of A_1^*, expanded up to terms of the second order, for the fully parametric version of the bootstrap where, in step 1, the generation of the bootstrap data is undertaken with $\hat{f}_i(\mathbf{x})$ taken to be the multivariate normal density with mean $\bar{\mathbf{x}}_i$ and covariance matrix $\mathbf{S}(i = 1, 2)$. McLachlan[27] carried out some simulations in which he compared the MSE of b_1^B with that of $\beta_1(D)$ in estimating the true bias, b_1, to demonstrate the high efficiency of the nonparametric version of the bootstrap estimator of the bias of the apparent error rate.

6. VARIANTS OF THE BOOTSTRAP

In a recent study, Efron[8] reported some simulation results on the performance of the bootstrap relative to other methods such as cross-validation in their bias correction of the overall apparent error rate,

$$A = \sum_{i=1}^{g} n_i A_i / n,$$

in the context of estimating the overall conditional error rate, $ec(f, \mathbf{t})$, of the sample rule based on (7) with $\hat{\lambda} = 0$, applied under the corresponding normal model (3). It was concluded that cross-validation is nearly unbiased, but that it has often an unacceptably high variability if n is small. The bootstrap estimate of A, A^B, has much less variability, but unfortunately b^B, the bootstrap estimate of the bias of A, is negatively correlated with W, the actual difference between A and $ec(f, \mathbf{t})$. The MSE of A^B can be expresssed as

$$\text{MSE}(A^B) = \text{var}(b^B) + \text{var}(W) + \{E(b^B) - b\}^2 - 2\,\text{cov}(W, b^B). \tag{13}$$

It can be seen that a negative value for the term $\text{cov}(W, b^B)$ in (13) inflates the MSE of A^B, although it is still, in general, less than that of the cross-validated estimator. Also, the bootstrap estimate of the bias tends to underestimate the magnitude of it. Efron[8] therefore has developed more sophisticated variants of his ordinary bootstrap, including the randomized and double bootstraps, and the "0.632 estimator" to be discussed in the next section. These variants were found to clearly outperform cross-validation and the bootstrap.

The double bootstrap corrects the bias of the ordinary bootstrap apparently without increasing its MSE. The bias corrected estimate of A so obtained for the overall conditional error is

$$A^{DB} = A - \text{bias}^B(A^B),$$

where $\text{bias}^B(A^B)$ is the bootstrap estimate of the bias of the ordinary bootstrap estimator A^B. Although it appears that the computation of the estimate, $\text{bias}^B(A^B)$, requires two layers of bootstrapping with a total of M^2 bootstrap replications, Efron[8] has shown that, by using a Monte Carlo "swindle," it can be implemented with just $2M$ replications.

The randomized bootstrap in the case of mixture sampling generates the bootstrap data from the probability function defined over the $2n$ points, $\mathbf{z}_j = (\mathbf{x}_j, y_j)$ and $\bar{\mathbf{z}}_j = (\mathbf{x}_j, 3 - y_j)$ for $j = 1, \ldots, n$, with mass $\nu(\mathbf{z}_j)/n$ and $\nu(\bar{\mathbf{z}}_j)/n$ at $\mathbf{z}_j$ and $\bar{\mathbf{z}}_j$, respectively, and $\nu(\mathbf{z}_j) + \nu(\bar{\mathbf{z}}_j) = 1$. Efron[8] studied the use of

$$\nu(\mathbf{z}_j) = 0.9, \quad \nu(\bar{\mathbf{z}}_j) = 0.1 \tag{14}$$

and a more complicated version, equivalent here to taking

$$\nu(\mathbf{z}_j) = \hat{\xi}_{y_j}(\mathbf{x}_j), \quad \nu(\bar{\mathbf{z}}_j) = 1 - \hat{\xi}_{y_j}(\mathbf{x}_j), \tag{15}$$

with the restriction that $\nu(\mathbf{z}_j)$, and hence $\nu(\bar{\mathbf{z}}_j)$, lie in the range 0.1–0.9. It can be seen that the randomized bootstrap is an attempt to smooth the empirical probability distribution in the y direction. The use of either (14) or (15) in step 1 of the bootstrap algorithm was found to substantially lower the MSE of the ordinary bootstrap estimator of the overall conditional error, with (14) giving almost as much improvement as the more complicated version (15).

7. THE 0.632 ESTIMATOR

We let $A^{0.632}$ be the estimator of the overall conditional error rate, termed the "0.632 estimator" by Efron[8], who reported that it was clearly best in his simulation experiments. It is a weighted sum of the apparent error rate and the quantity ϵ, so that

$$A^{0.632} = 0.368A + 0.632\epsilon,$$

where

$$\epsilon = \sum_{m=1}^{M} \sum_{j=1}^{n} \delta_{mj} Q[y_j, r(\mathbf{x}_j, \mathbf{t}_m^*)]/M,$$

and

$$M_1 = \sum_{m=1}^{M} \sum_{j=1}^{n} \delta_{mj};$$

$\delta_{mj} = 1$, if $\mathbf{x}_j$ is not present in the bootstrap training set $\mathbf{t}_m^*$, and zero otherwise. The quantity ϵ is the bootstrap error rate at an original data point not in the training set.

Efron[8] developed the 0.632 estimator by consideration of the distribution of the distance δ between the point at which the rule is applied and the nearest point in the training set. It was demonstrated that the distribution of δ is quite different in the bootstrap context than in the actual situation. The points which contribute to ϵ have $\delta > 0$ and, as a consequence of the resampling scheme of the nonparametric bootstrap, are about 1/0.632 too far away from the training set than in the actual situation. This led Efron[8] to propose

$$b^{0.632} = 0.632(A - \epsilon)$$

as an estimator of the bias, b, of A in estimating the overall conditional error rate $ec(f, \mathbf{t})$. The bias corrected version of A is therefore

$$\begin{aligned} A^{0.632} &= A - b^{0.632} \\ &= 0.368A + 0.632\epsilon. \end{aligned}$$

Efron[8] showed that $A^{0.632}$ is almost the same as the estimator,

$$0.368A + 0.632A^{HCV},$$

where A^{HCV} is the estimate of the overall error rate after a cross-validation that leaves out half of the observations at a time. Estimators of this type have been considered by McLachlan[25] in the context of choosing the weight τ, so that

$$A_1^{\tau} = (1 - \tau)A_1 + \tau A_1^{GCV}$$

has zero first-order bias as an estimator of the conditional error, $ec_1(f_1, \mathbf{t})$, of the rule based on the sample linear discriminant function (7) with $\hat{\lambda} = 0$; A^{GCV} denotes the estimate after cross-validation is performed removing n/G observations at a time. Under the normal model (3), the desired value of τ, τ_0, was computed as a function of G, Δ, p and the relative size of n_1 and n_2. For $G = 2$, so that $A^{GCV} = A^{HCV}$, McLachlan[25] showed under separate sampling that τ_0 ranged from 0.6 to 0.7 for the combinations of the other parameters ($\Delta = 1, 2$; $p = 4, 8, 16$; $n_1/n_2 = 1/3, 1, 3$). Hence, under (3), the estimator $A_1^{\tau_0}$ is about the same as Efron's 0.632 estimator. The latter therefore should have almost zero first-order bias under (3), at least for the sample rule based on (7). Efron ([8], Table 4) did calculate the first-order bias of $A^{0.632}$ for this rule in the various cases of (3) under which it was applied in his simulations, and it was small. With one exception, the asymptotic bias was in a downward direction, and in the simulations $A^{0.632}$ exhibited a moderate downward bias. The reason for the remarkably low MSE of $A^{0.632}$ in the simulations was the lack of negative correlation between $b^{0.632}$ and $W = A - ec(f, t)$.

8. SMOOTHED MODIFICATION OF THE APPARENT ERROR

Glick[12] has considered ways of smoothing the apparent error rate in order to reduce its variance in estimating the conditional error rate. The smoothed version of the overall apparent error rate, A, is

$$A^S = \sum_{j=1}^{n} K(\mathbf{x}_j)/n;$$

obtained by replacing the zero–one function $Q[y_j, r(\mathbf{x}_j, \mathbf{t})]$ in the definition of A by a smoothing function, $K(\mathbf{x}_j)$, which may take on values between zero and one. It can be seen that a modest perturbation of an $\mathbf{x}_j$ can switch the indicator function Q from zero to one or vice versa, but will cause only small perturbation for a smooth function.

It is well known (Fukunaga and Kessell[10]) that the rate,

$$\sum_{j=1}^{n} \min\{\xi_1(\mathbf{x}_j), \quad \xi_2(\mathbf{x}_j)\}/n, \tag{16}$$

provides an unbiased estimator of the overall error rate, $eo(f)$, of the Bayes rule, with smaller variance than A. An estimator of the optimal error rate with respect to the ith class can be formed in a similar fashion (Schwemer and Dunn[29]). The estimated error rate (16) is sometimes referred to as a posterior probability estimator due to its formation in terms of the posterior probabilities of each $\mathbf{x}_j$ in the training data. It suggests that a possible choice of the smoothing function in the formation of A^S is the minimum of the estimated posterior probabilities, $\hat{\xi}_1(\mathbf{x}_j)$ and $\hat{\xi}_2(\mathbf{x}_j)$, leading to

$$A^{PP} = \sum_{j=1}^{n} \min\{\hat{\xi}_1(\mathbf{x}_j), \quad \hat{\xi}_2(\mathbf{x}_j)\}/n$$

as a smoothed counting estimator of the overall conditional error rate, $ec(f, \mathbf{t})$. The performance of A^{PP} depends on the reliability of the estimates of the posterior probabilities, and so its applicability at least in a nonparametric framework may be limited. Of course A^{PP} gives a biased assessment of $ec(f, \mathbf{t})$, but it can be corrected for bias by using the bootstrap. The asymptotic bias of A^{PP} in estimating $ec(f, \mathbf{t})$ has been studied by Ganesalingam and McLachlan[12].

Since the formation of the estimator A^{PP} does not require the origin of each $\mathbf{x}_j$ in the training set, it has been found to be helpful in a cluster analysis context where there are no training data of known origin. Basford and McLachlan[4] have shown how an estimator of the same form as A^{PP}, after correction for bias according to a parametric version of the bootstrap, can provide a useful assessment of the performance of a clustering rule formed by adopting a mixture model for the training data of unknown origin.

9. PARAMETRIC ESTIMATORS

With any application of a sample rule its apparent error rate with respect to each class and overall would be calculated in the first instance to provide an initial guide to the performance of the rule based on the training data at hand. In the previous sections we have considered how the apparent error rate can be modified to give an improved estimate of the conditional error rate, concentrating on the available nonparametric methods of bias correction.

In the case of a parametric sample rule, we may wish to adopt a parametric approach to the estimation of the conditional error rates, $ec_i(f_i, \mathbf{t})$.

The parametric bias correction term of the apparent error rate A_1 under the normal model (3) was given in Section 4 in the course of commenting on the efficiency of the nonparametric bootstrap correction. Concerning the parametric estimation of the conditional errors themselves, a number of estimators have been proposed and studied over the years; see, for example, Lachenbruch and Mickey[17] and McLachlan[19]. A common approach is to use "plug-in" estimators of the form $eo_i(\hat{f})$ or $eu_i(\hat{f})$, that is, the corresponding optimal or unconditional error with the class conditional densities $f_i(\mathbf{x})$ or their unknown parameters replaced by appropriate estimates. Unfortunately, for most problems, $eu_i(f)$ is unable to be computed exactly, but in some instances the parametric delta method can be used to derive an asymptotic expansion. For example, under (3), the unconditional error $eu_i(\Delta)$ of the rule based on the sample linear discriminant function (7) with $\hat{\lambda}$ fixed depends on the unknown Mahalanobis distance Δ, and Okamoto[28] has derived its asymptotic expansion, $eua_i(\Delta)$, up to terms of the third order with respect to the reciprocals of the sample sizes n_i and N. Lachenbruch and Mickey[17] proposed

using $eua_i(\hat{\Delta})$ as an estimator of $ec_i(\hat{f}_i, \mathbf{t})$, where $\hat{\Delta} = D$ or DS, and

$$DS = \{(N - p - 1)/N\}^{1/2}D.$$

The bias of $eua_i(\hat{\Delta})$ is of the second order, and in the case of $\hat{\lambda} = 0$, McLachlan[20,22] has shown that this bias can be reduced to the third order only by using

$$P_i = \Phi\left(-\frac{1}{2}D\right) + \phi\left(-\frac{1}{2}D\right)[(p - 1)/(Dn_i) + D\{4(4p - 1) - D^2\}/(32N) + (p - 1)(p - 2)/(4Dn_i^2) + (p - 1)\{-D^3 + 8(2p + 1)D + (16/D)\}/(64n_iN)],$$

where Φ denotes the standard normal distribution function.

For the more general model of classes having normal densities with unequal covariance matrices, asymptotic expansions of the unconditional error rates are available only in special cases; for example, with proportional covariance matrices (McLachlan[23]). However, this model can be handled by using the bootstrap in parametric form, where $\hat{f}_i(\mathbf{x})$ is taken to be the multivariate normal density with mean $\bar{\mathbf{x}}_i$ and covariance matrix $\mathbf{S}_i (i = 1, 2)$ in the generation of the bootstrap training data. The bootstrap expectation of $ec_i(\hat{f}_i, \mathbf{T}^*)$ can be approximated by the Monte Carlo approximation

$$\sum_{m=1}^{M} ec_i(\hat{f}_i, \mathbf{t}_m^*)/M,$$

which can be used as an estimate of $ec_i(f_i, \mathbf{t})$.

Note that caution should be exercised with the use of parametric estimators of the error rates as they may not be reliable under departures from the parametric model adopted, even though the sample rule itself may be robust. For example, the rule based on the sample linear discriminant function is known to be fairly robust, but that the normality based estimators of its error rates are not.

REFERENCES

1. J. Aitchison, J. D. F. Habbema and J. W. Kay, A critical comparison of two methods of statistical discrimination. *Appl. Stat.* **26**, 15–25 (1977).
2. T. W. Anderson, Classification by multivariate analysis. *Psychometrika* **16**, 31–52 (1951).
3. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York (1958).
4. K. E. Basford and G. J. McLachlan, Estimation of allocation rates in a cluster analysis context. *J. Am. Stat. Assoc.* **80**, 286–293 (1985).
5. M. R. Crask and W. D. Perreault, Validation of discriminant analysis in marketing research. *J. Marketing Res.* **14**, 60–68 (1977).
6. B. Efron, Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**, 1–26 (1979).
7. B. Efron, *The Jackknife, The Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia (1982).
8. B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.* **78**, 316–331 (1983).
9. B. Efron and G. Gong, A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Stat.* **37**, 36–48 (1983).
10. K. Fukunaga and D. L. Kessell, Nonparametric Bayes error estimation using unclassified samples. *IEEE Trans. Inf. Theory* **IT-19**, 434–440 (1973).
11. S. Ganesalingam and G. J. McLachlan, Error rate estimation on the basis of posterior probabilities. *Pattern Recognition* **12**, 405–413 (1980).
12. N. Glick, Additive estimators for probabilities of correct classification. *Pattern Recognition* **10**, 211–222 (1978).
13. H. L. Gray and W. R. Schucany, *The Generalized Jackkife Statistic*. Marcel Dekker, New York (1972).
14. D. J. Hand, *Kernel Discriminant Analysis*. John Wiley, Chichester (1982).
15. M. Hills, Allocation rules and their error rates. *J. R. Stat. Soc. Ser. B* **28**, 1–31 (1966).
16. P. A. Lachenbruch, *Discriminant Analysis*. Hafner, New York (1975).
17. P. A. Lachenbruch and M. R. Mickey, Estimation of error rates in discriminant analysis. *Technometrics* **10**, 1–11 (1968).
18. G. J. McLachlan, The asymptotic distributions of the conditional error rate and risk in discriminant analysis. *Biometrika* **61**, 131–135 (1974).
19. G. J. McLachlan, Estimation of the errors of misclassification on the criterion of asymptotic mean-square error. *Technometrics* **16**, 255–260 (1974).

20. G. J. McLachlan, An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics* **30**, 239–249 (1974).
21. G. J. McLachlan, The relationship in terms of asymptotic mean square error between the separate problems of estimating each of the three types of error rate of the linear discriminant function. *Technometrics* **16**, 569–575 (1974).
22. G. J. McLachlan, Confidence intervals for the conditional probability of misclassification in discriminant analysis. *Biometrics* **31**, 161–167 (1975).
23. G. J. McLachlan, Some expected values for the error rates of the sample quadratic discriminant function. *Aust. J. Stat.* **17**, 161–165 (1975).
24. G. J. McLachlan, The bias of the apparent error rate in discriminant analysis. *Biometrika* **63**, 239–244 (1976).
25. G. J. McLachlan, A note on the choice of a weighting function to give an efficient method for estimating the probability of misclassification. *Pattern Recognition* **9**, 147–149 (1976).
26. G. J. McLachlan, A note on bias correction in maximum-likelihood estimation with logistic discrimination. *Technometrics* **22**, 621–627 (1980).
27. G. J. McLachlan, The efficiency of Efron's bootstrap approach applied to error rate estimation in discriminant analysis, *J. Stat. Comput. Simul.* **11**, 273–279 (1980).
28. M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Stat.* **34**, 1286–1301 (1963). Correction: *Ann. Math. Stat.* **39**, 1358, 1359 (1968).
29. G. T. Schwemer and O. J. Dunn, Posterior probability estimators in classification simulations. *Commun. Stat. Theor. Meth.* **B9**, 133–140 (1980).
30. C. A. B. Smith, Some examples of discrimination. *Ann. Eugen. London* **13**, 272–282 (1947).
31. M. Stone, Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B* **36**, 111–147 (1974).
32. G. T. Toussaint, Bibliography on estimation of misclassification. *IEEE Trans. Inf. Theory* **IT-20**, 472–479 (1974).