The Annals of Statistics
1978, Vol. 6, No, 3, 506-514

A FINITE SAMPLE DISTRIBUTION-FREE PERFORMANCE BOUND FOR LOCAL DISCRIMINATION RULES

BY W. H. ROGERS AND T. J. WAGNER[1]

The Rand Corporation and University of Texas

In the discrimination problem the random variable θ, known to take values in $\{1, \cdots, M\}$, is estimated from the random vector X. All that is known about the joint distribution of (X, θ) is that which can be inferred from a sample $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ of size n drawn from that distribution. A discrimination rule is any procedure which determines a decision $\hat{\theta}$ for θ from X and $(X_1, \theta_1), \cdots, (X_n, \theta_n)$. A rule is called k-local if the decision $\hat{\theta}$ depends only on X and the pairs (X_i, θ_i) for which X_i is one of the k-closest to X from $X_1, \cdots, X_n$. It is shown that for any k-local discrimination rule, the mean-square difference between the probability of error for the rule and its deleted estimate is bounded by A/n where A is an explicitly given small constant which depends only on M and k. Thus distribution-free confidence intervals can be placed about probability of error estimates for k-local discrimination rules.

1. Introduction. In the discrimination problem a statistician makes an *observation* X, a random vector with values in $\mathbb{R}^d$, and wishes to estimate its *state* θ, a random variable known to take values in $\{1, \cdots, M\}$. All that he knows about the distribution of (X, θ) is that which can be inferred from a sample $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ of size n drawn from that distribution. The sample, called *data*, is assumed to be independent of (X, θ). Using X and the data the statistician makes a decision $\hat{\theta}$ for θ where his *rule* is any procedure which determines $\hat{\theta}$ given X and the data.

The rule which serves as a prototype for the class of rules considered in this paper is the k-nearest neighbor rule. Introduced by Cover and Hart (1967), this rule takes $\hat{\theta}$ to be the state which occurs most often among the states of the k closest observations to X from $X_1, \cdots, X_n$. Two types of ties can occur here. In the first case, there may be ties in distance to the observation X so that the k closest observations are not uniquely determined. This case can occur, for example, when the distribution of X is purely atomic. The second case occurs when there are several different states which occur most frequently among the states of the k closest observations to X. To handle both situations, the independent sequence $Z, Z_1, Z_2, \cdots$ of random variables, which are i.i.d. with a uniform distribution on $[0, 1]$, is generated. We will think of Z as being attached to X and Z_i as being attached to X_i for $i = 1, \cdots, n$. Then X_i is closer to X than X_j if

Received October 1974; revised August 1977.

[1] Supported in part by AFOSR Contract F44620-71-C-0091 and AFOSR Grant 72-2371.

AMS 1970 *subject classifications.* Primary 62H30, 62G05; Secondary 62G15.

Key words and phrases. Discrimination, distribution-free bound, local inference, nearest neighbor rules, finite sample bound, deleted estimate, error rate estimate.

(a) $\|X - X_i\| < \|X - X_j\|$ or
(b) $\|X - X_i\| = \|X - X_j\|$ and $|Z - Z_i| < |Z - Z_j|$ or
(c) $\|X - X_i\| = \|X - X_j\|$, $|Z - Z_i| = |Z - Z_j|$ and $i < j$.

The k closest observations to X are now uniquely determined and the description of the rule is completed by taking $\hat{\theta}$ to be the state which occurs most frequently among the states of the k closest observations to X. If several states occur most frequently among the states of the k closest observations to X, the state with the observation closest to X from those tied is chosen. If (X^j, θ^j, Z^j) represents the jth closest observation to X, its state and its attached random variable, we see that the estimate $\hat{\theta}$ of the k-nearest neighbor rule can be written

$$\hat{\theta} = g(X, Z, (X^1, \theta^1, Z^1), \cdots, (X^k, \theta^k, Z^k)) \tag{1.1}$$

for some function g. The class of rules with the representation (1.1) for some g and some k are the rules which we are primarily interested in and will be called *k-local*.

The probability of error L_n, given the data and attached random variables, is

$$L_n = P\{\hat{\theta} \neq \theta \mid D_n\}$$

where

$$D_n = ((X_1, \theta_1, Z_1), \cdots, (X_n, \theta_n, Z_n)) .$$

A frequency interpretation of the random variable L_n is that m new observations whose states are estimated by the rule with D_n will have mL_n discrimination errors on the average. (Each of these new observations will have a new independent "Z" attached to it but the $Z_1, \cdots, Z_n$ stay fixed with the data.) The random variable L_n is economically important to the statistician because it measures the performance of the rule after it is constructed but before it is applied. His immediate need then is an accurate estimate of L_n.

The *deleted estimate* of L_n is given by

$$\hat{L}_n = (1/n) \sum_{i=1}^{n} I_{[\hat{\theta}_i \neq \theta_i]}$$

where $\hat{\theta}_i$ is the estimate of θ_i from X_i, Z_i and D_n with (X_i, θ_i, Z_i) deleted. (This definition makes sense, of course, as long as $k \leqq n - 1$.) Thus $\hat{L}_n$ is the proportion of errors made by the rule on the data that defines it, one observation deleted at a time. Since $I_{[\hat{\theta}_i \neq \theta_i]}$ is a validation on independent data, $\hat{L}_n$ might also be called the cross-validatory estimate of L_n. This estimate has been used in the past as a criterion for selecting a particular discrimination rule among a class of discrimination rules, the one having the smallest estimate being called the cross-validatory choice (e.g., see Stone (1974) for a recent summary of the work done on this problem). Deleted or cross-validatory estimates are not always easy to compute but, in some cases like the k-nearest neighbor rule, the computation is reasonable and the intuitively efficient use of the data can be taken advantage of. The deleted estimate is compared briefly with other estimates at the end of the paper.

In addition to L_n, the current performance of the rule, the statistician is also interested in its ultimate performance. Assume that there exists a constant L such that

$$L_n \to L \quad \text{in probability.} \tag{1.2}$$

The quantity L measures the effectiveness of the rule for an infinite amount of data. Indeed, much previous investigation into the nonparametric discrimination problem centered around exhibiting rules for which (1.2) holds with $L = L^*$, where L^* is the Bayes probability of error. Using such a rule, called asymptotically optimal, at least gives the statistician some limited assurance that he will do as well as possible with large amounts of data. Beyond this, the statistician needs to know the value L since a large value could obviate collecting further data. For the k-nearest neighbor rule, Cover and Hart gave conditions for which L exists with

$$EL_n \to L$$

and where, for $k = 1$,

$$L^* \leqq L \leqq L^*[2 - ML^*/(M-1)]\,. \tag{1.3}$$

Wagner (1970) and Fritz (1975) gave conditions for the distribution of (X, θ) which, for the k-nearest neighbor rule, insured that the convergence in (1.2) was with probability one. Assuming (1.2), every reasonable estimate of L becomes a reasonable estimate of L_n for large n and vice versa. Thus $\hat{L}_n$ has been called a deleted estimate of both L_n and L. For some types of rules which satisfy (1.2), conditions on the distribution of (X, θ) have been given in Wagner (1973) which insure that

$$\hat{L}_n \to L \quad \text{in probability,} \tag{1.4}$$

and, consequently, $L_n - \hat{L}_n \to 0$ in probability. Cover (1969), using (1.3), also has discussed the use of the k-nearest neighbor rule with its deleted estimate to estimate bounds for L^*.

If the statistician has a rule which satisfies (1.2) and (1.4), where L is not necessarily equal to L^*, what beyond these statements does he want? He would undoubtedly like to know, for a given $\varepsilon, \alpha > 0$, an n for which

$$\begin{aligned} &\text{(a)} \quad P\{|\hat{L}_n - L_n| \geqq \varepsilon\} \leqq \alpha \\ &\text{(b)} \quad P\{|\hat{L}_n - L| \geqq \varepsilon\} \leqq \alpha\,. \end{aligned} \tag{1.5}$$

The n must work for all distributions of (X, θ) but could depend on the number of states M and the dimension d. It is unfortunate, but not surprising, that an n cannot be specified for (1.5b). In particular, for $0 < \varepsilon < \frac{1}{2}$ and $d = 1$,

$$\sup P\{|\hat{L}_n - L| \geqq \varepsilon\} = 1$$

for each local rule, where the supremum is taken over all distributions of (X, θ). As pessimistic as this observation seems, it only indicates that knowing

something about the ultimate performance of the rule from a finite amount of data, uniformly in all problems, requires additional a priori information.

Can then anything be said about the finite sample performance of a local rule with regard to estimating the current probability of error L_n? Restating (1.5a), does

$$\sup P\{|L_n - \hat{L}_n| \geqq \varepsilon\} \to 0$$

for each $\varepsilon > 0$? The answer is yes, and $\hat{L}_n$ seems to be a surprisingly good estimate of L_n.

2. Results. The main result of this paper is the following theorem.

THEOREM 2.1. *For all d and any k-local rule*

$$E(L_n - \hat{L}_n)^2 \leqq (2k + (\tfrac{1}{4}))/n + 2k(2k + (\tfrac{1}{4}))^{\frac{1}{2}}/n^{\frac{3}{2}} + k^2/n^2 . \tag{2.1}$$

Before proving Theorem 2.1 we give a result for a somewhat more general class of rules. If the rule can be specified by

$$\hat{\theta} = g_n(X, Z, D_n)$$

then it is termed *symmetric* if any reordering of $(X_1, \theta_1, Z_1), \cdots, (X_n, \theta_n, Z_n)$ leaves the value of $\hat{\theta}$ unchanged with probability one. If (X, θ, Z) and (X_0, θ_0, Z_0) are two independent observations with their states and attached random variables let $\hat{\theta} = g_n(X, Z, D_n)$ and $\hat{\theta}_0 = g_n(X_0, Z_0, D_n)$. If D_{ni} denotes the sequence D_n with (X_i, θ_i, Z_i) deleted then, as before, let

$$\hat{\theta}_i = g_{n-1}(X_i, Z_i, D_{ni}) , \qquad 1 \leqq i \leqq n .$$

THEOREM 2.2. *For a symmetric rule*

$$E(L_n - \hat{L}_n)^2 = P\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0\} - 2P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\} + P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\} + (EL_{n-1} - P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\})/n .$$

PROOF. First, for any g_n

$$\begin{aligned} E(L_n^2) &= E(P\{\hat{\theta} \neq \theta \mid D_n\})^2 \\ &= E(P\{\hat{\theta} \neq \theta \mid D_n\}P\{\hat{\theta}_0 \neq \theta_0 \mid D_n\}) \\ &= E(P\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0 \mid D_n\}) \\ &= P\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0\} . \end{aligned}$$

Next,

$$\begin{aligned} E(L_n\hat{L}_n) &= E\left(L_n \frac{1}{n} \textstyle\sum_1^n I_{[\hat{\theta}_i \neq \theta_i]}\right) = \frac{1}{n} \textstyle\sum_1^n E(L_n I_{[\hat{\theta}_i \neq \theta_i]}) \\ &= \frac{1}{n} \textstyle\sum_1^n E(P\{\hat{\theta} \neq \theta; \hat{\theta}_i \neq \theta_i \mid D_n\}) \\ &= \frac{1}{n} \textstyle\sum_1^n P\{\hat{\theta} \neq \theta; \hat{\theta}_i \neq \theta_i\} \\ &= P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\} \quad \text{(because of symmetry)} . \end{aligned}$$

Finally, again using symmetry,

$$E(\hat{L}_n^2) = \frac{1}{n^2}\sum_1^n P\{\hat{\theta}_i \neq \theta_i\} + \frac{1}{n^2}\sum_{i\neq j} P\{\hat{\theta}_i \neq \theta_i; \hat{\theta}_j \neq \theta_j\}$$

$$= \frac{EL_{n-1}}{n} + \frac{n-1}{n} P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\}$$

and Theorem 2.2 follows immediately.

As an application of Theorem 2.2, consider the k-nearest neighbor rule. Looking first at the term

$$P\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0\} - P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\}, \tag{2.2}$$

let θ' be the k-NNR estimate of θ from $(X_0, \theta_0, Z_0), (X_2, \theta_2, Z_2), \cdots, (X_n, \theta_n, Z_n)$ and let θ_0' be the k-NNR estimate of θ_0 from $(X_2, \theta_2, Z_2), \cdots, (X_n, \theta_n, Z_n)$. Then

$$\begin{aligned}|P\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0\} - P\{\theta' \neq \theta; \theta_0' \neq \theta_0\}| &\leqq P(\{\hat{\theta} \neq \theta; \hat{\theta}_0 \neq \theta_0\} \triangle \{\theta' \neq \theta; \theta_0' \neq \theta_0\}) \\ &\leqq P(\{\hat{\theta} \neq \theta'\} \cup \{\hat{\theta}_0 \neq \theta_0'\}) \\ &\leqq P\{\hat{\theta} \neq \theta'\} + P\{\hat{\theta}_0 \neq \theta_0'\} \leqq \frac{2k}{n+1} + \frac{k}{n} \leqq \frac{3k}{n}.\end{aligned}$$

Since $P\{\theta' \neq \theta; \theta_0' \neq \theta_0\} = P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\}$, it follows that (2.2) is $\leqq 3k/n$. For the term

$$P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\} - P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\}, \tag{2.3}$$

let now θ' be the k-NNR estimate of θ from $(X_1, \theta_1, Z_1), (X_3, \theta_3, Z_3), \cdots, (X_n, \theta_n, Z_n)$ and let θ_1' be the k-NNR estimate of θ_1 from $(X, \theta, Z), (X_3, \theta_3, Z_3), \cdots, (X_n, \theta_n, Z_n)$. Then, as above,

$$\begin{aligned}|P\{\hat{\theta} \neq \theta; \hat{\theta}_1 \neq \theta_1\} - P\{\theta' \neq \theta; \theta_1' \neq \theta_1\}| &\leqq P\{\hat{\theta} \neq \theta'\} + P\{\hat{\theta}_1 \neq \theta_1'\} \leqq \frac{k}{n} + \frac{2k}{n} = \frac{3k}{n}.\end{aligned}$$

Because $P\{\theta' \neq \theta; \theta_1' \neq \theta_1\} = P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\}$ we see that (2.3) is $\leqq 3k/n$. For the last term

$$(EL_{n-1} - P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\})/n, \tag{2.4}$$

let θ_1'', θ_2'' be the k-NNR estimates of θ_1, θ_2 from $(X, \theta, Z), (X_3, \theta_3, Z_3), \cdots, (X_n, \theta_n, Z_n)$. Then, as before,

$$|P\{\hat{\theta}_1 \neq \theta_1; \hat{\theta}_2 \neq \theta_2\} - P\{\theta_1'' \neq \theta_1; \theta_2'' \neq \theta_2\}| \leqq 4k/n$$

so that (2.4) is bounded above by

$$(EL_{n-1} - P\{\theta_1'' \neq \theta_1; \hat{\theta}_2'' \neq \theta_2\})/n + 4k/n^2.$$

But $P\{\theta_1'' \neq \theta_1; \theta_2'' \neq \theta_2\} = EL_{n-1}^2 \geqq (EL_{n-1})^2$ so that the last expression is bounded above by

$$(EL_{n-1} - (EL_{n-1})^2)/n + 4k/n^2 \leqq 1/4n + 4k/n^2.$$

Combining all three bounds yields

$$E(L_n - \hat{L}_n)^2 \leqq (6k + \tfrac{1}{4})/n + 4k/n^2 . \tag{2.5}$$

PROOF OF THEOREM 2.1. Let $L_{n(i)}$ be the probability of error when θ is estimated from X, Z and $D_{n,i}$ and let $\tilde{L}_n = (\sum_1^n L_{n(i)})/n$. Then

$$\sum_{i=1}^n |L_{n(i)} - L_n| \leqq \sum_{i=1}^n P\{X_i \text{ is one of the } k \text{ closest to } X\} = k \tag{2.6}$$

so that, for $s > 0$,

$$|\tilde{L}_n - L_n|^s = |\textstyle\sum_1^n (L_{n(i)} - L_n)|^s/n^s \leqq k^s/n^s$$

and

$$E(\tilde{L}_n - L_n)^2 \leqq k^2/n^2 . \tag{2.7}$$

In addition to the notation used previously, let $\hat{\theta}_{i,j}$ be the estimate of θ_i from the data with both $((X_i, \theta_i, Z_i))$ and (X_j, θ_j, Z_j) deleted and let $L_{n(i,j)}$ denote the probability of error when θ is estimated from the data with both (X_i, θ_i, Z_i) and (X_j, θ_j, Z_j) deleted. Then

$$\begin{aligned} E(\hat{L}_n - \tilde{L}_n)^2 &= \frac{1}{n^2} E\{\textstyle\sum_1^n (I_{[\hat{\theta}_i \neq \theta_i]} - L_{n(i)})\}^2 \\ &= \frac{1}{n^2} \textstyle\sum_1^n E(I_{[\hat{\theta}_i \neq \theta_i]} - L_{n(i)})^2 \\ &\quad + \frac{1}{n^2} \textstyle\sum_{i \neq j} E[(I_{[\hat{\theta}_i \neq \theta_i]} - L_{n(i)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)})] . \end{aligned} \tag{2.8}$$

Because $L_{n(i)}$ is the best mean-square estimate of $I_{[\hat{\theta}_i \neq \theta_i]}$ from the data with (X_i, θ_i, Z_i) deleted we conclude that

$$E(I_{[\hat{\theta}_i \neq \theta_i]} - L_{n(i)})^2 \leqq E(I_{[\hat{\theta}_i \neq \theta_i]} - P[\hat{\theta}_i \neq \theta_i])^2 \leqq \tfrac{1}{4}$$

and the first term of (2.8) is bounded by $1/4n$.

The second term of (2.8) is split up as follows.

$$\frac{1}{n^2} \textstyle\sum_{i \neq j} E[(I_{[\hat{\theta}_i \neq \theta_i]} - L_{n(i)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)})]$$

$$= \frac{1}{n^2} \textstyle\sum_{i \neq j} E(I_{[\hat{\theta}_i \neq \theta_i]} - I_{[\hat{\theta}_{i,j} \neq \theta_i]})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)}) \tag{2.9}$$

$$+ \frac{1}{n^2} \textstyle\sum_{i \neq j} E(I_{[\hat{\theta}_{i,j} \neq \theta_i]} - L_{n(i,j)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)}) \tag{2.10}$$

$$+ \frac{1}{n^2} \textstyle\sum_{i \neq j} E(L_{n(i,j)} - L_{n(i)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)}) . \tag{2.11}$$

For $i \neq j$, we see that each term in (2.9) is bounded in absolute value by the probability that X_j is one of the k-nearest neighbors of X_i. This probability is less than or equal to $k/(n-1)$ so that (2.9) is bounded by k/n. For (2.10) we

see that the expectation of each term is 0 by first taking the conditional expectation given the data with (X_j, θ_j, Z_j) deleted. For example,

$$\begin{aligned} E(I_{[\hat{\theta}_{i,j} \neq \theta_i]} - L_{n(i,j)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)}) \\ = E(E((I_{[\hat{\theta}_{i,j} \neq \theta_i]} - L_{n(i,j)})(I_{[\hat{\theta}_j \neq \theta_j]} - L_{n(j)}) \mid D_{nj})) \\ = E(I_{[\hat{\theta}_{i,j} \neq \theta_i]} - L_{n(i,j)})(L_{n(j)} - L_{n(j)}) = 0 \end{aligned}$$

where we notice that $I_{[\hat{\theta}_{i,j} \neq \theta_i]} - L_{n(i,j)}$ is a function of D_{nj}. Using the same argument that led to (2.6), we see that (2.11) is $\leqq k/n$. Collecting bounds yields

$$E(\hat{L}_n - \tilde{L}_n)^2 \leqq [2k + (\tfrac{1}{4})]/n \,. \tag{2.12}$$

Writing $E(L_n - \hat{L}_n)^2$ as

$$E(L_n - \tilde{L}_n)^2 + 2E(L_n - \tilde{L}_n)(\tilde{L}_n - \hat{L}_n) + E(\tilde{L}_n - \hat{L}_n)^2$$

we see that (2.1) follows from (2.7), Schwarz's inequality, and (2.12).

For a k-local rule let b be the right-hand side of (2.1). We have the following immediate corollaries of Theorem 2.1.

COROLLARY 1. $P[|\hat{L}_n - L_n| \geqq \varepsilon] \leqq b/\varepsilon^2$.

COROLLARY 2. $P[L^* \geqq \hat{L}_n + \varepsilon] \leqq b/\varepsilon^2$.

3. Remarks. The result (2.1) can be weakened to the form A/n by taking A to be the sum of the numerators of the right-hand side of (2.1). So far as we know, our bounds are not sharp, especially for $k > 1$. All possible constant A's satisfy $A \geqq \frac{1}{4}$ as can be seen by considering rules which do not depend on the data. In this case, $k = 0$ and the sum in $\hat{L}_n$ is a Bernoulli sequence whose random variables have an expectation equal to the probability of error of the rule. If p is this probability of error then $E(\hat{L}_n - L_n)^2 = p(1 - p)/n$ and A must be at least $\frac{1}{4}$ to cover the case $p = \frac{1}{2}$.

A Monte Carlo experiment was carried out to ascertain how sharp the bound (2.1) might be in a realistic situation. The experiment involves drawing a sample of size n with $d = 1$, $M = 2$, $P\{\theta = 1\} = P\{\theta = 2\} = \frac{1}{2}$ and $P\{X \leqq x \mid \theta = i\}$ having the densities

$$\begin{aligned} f_1(x) &= 2x \qquad 0 \leqq x \leqq 1 \\ f_2(x) &= 2 - 2x \quad 0 \leqq x \leqq 1 \,. \end{aligned} \tag{3.1}$$

TABLE 1

Monte Carlo results for the nearest neighbor rule with the densities given in (3.1)

n	n avg $(L_n - L)^2$	n avg $(\hat{L}_n - L)^2$	n avg $(\hat{L}_n - L_n)^2$
50	.053	.369	.302
100	.052	.360	.301
250	.051	.418	.356
500	.052	.357	.301

For this problem L_n and $\hat{L}_n$ are both easily calculated for the nearest neighbor rule. The results of 500 samples of size n are shown in Table 1. The best bound presented in this paper for $nE(L_n - \hat{L}_n)^2$ is 2.25 which is considerably larger than the values in the table. In addition, a measure of normality was computed indicating that $(L_n, \hat{L}_n)$ converged to a normal distribution in the Monte Carlo as $n \to \infty$.

Readers may wonder if their favorite method of breaking ties will work. For example, suppose one uses the k-nearest neighbor rule with $k = 7$, but finds, for a particular value x to be classified, that there are fifteen neighbors within the sphere which contains the first seven as previously defined. One reasonable way to try to break this tie now is to take a vote of these fifteen neighbors. One might even modify the concept of a local rule to include this situation; however, the conditions required to obtain comparable results for such modifications are not clear.

Other error rate estimates have been studied (see the review paper by Toussaint (1974)). The resubstitution estimate of L_n (or L) is given by

$$\hat{L}_n = \frac{1}{n} \sum_1^n I_{[\hat{\theta}_j \neq \theta_j]}$$

where $\hat{\theta}_j = g(X_j, \theta_j, D_n)$ is an estimate using *all* the data. $\hat{L}_n$ is frequently an optimistic estimate of L_n (or L) even when it is a consistent estimate of L (Glick (1972), Toussaint (1974)). In particular, when used with the single nearest neighbor rule it always yields an estimate of 0 when $P\{X \leqq x \mid \theta = i\}$ has a probability density for each $i = 1, \cdots, M$. The holdout method has a number l, $0 < l < 1$, and counts the frequency of errors on the last ln observations using the first $n - ln$ observations and g (we have assumed ln is an integer). This method is forever subject to a balancing between the twin perils of a large bias with large l and a large variance with small l. Moreover, if $\hat{L}_n$ denotes the holdout estimate of L_n with l fixed then

$$E(L_n - \hat{L}_n)^2 \geqq (EL_n - E\hat{L}_n)^2$$

a quantity which, for the single nearest neighbor rule, can be shown to go to 0 at an arbitrarily slow algebraic rate following the example in Cover (1968).

We have assumed throughout that θ is a random variable. If we wish to assume, for example, that $\theta, \theta_1, \cdots, \theta_n$ is a deterministic sequence chosen by nature then there are M probabilities of error

$$L_n{}^j = P\{\hat{\theta} \neq j \mid (X_1, \theta_1, Z_1), \cdots, (X_n, \theta_n, Z_n), \theta = j\}, \qquad 1 \leqq j \leqq M.$$

A deleted estimate of $L_n{}^j$ consists of counting the frequency of errors made on the observations with states equal to j by using the data and deleting, one at a time, those observations that are being estimated. The bound (2.5) continues to hold for $E(\hat{L}_n{}^j - L_n{}^j)^2$ with the n on the right-hand side of that inequality replaced by n_j, the number of observations in the data with state j.

Finally Theorem 2.1 holds when X takes values in an arbitrary metric space, where we note that some of the properties of the metric, like the triangle inequality, are not needed.

Acknowledgment. Tom Cover has generously shared with us his ideas on nonparametric discrimination over an extended period of time. The results of this paper are as much his as ours.

REFERENCES

[1] COVER, T. (1968). Rates of convergence of nearest neighbor procedures. *Proc. 1st Annual Hawaii Conf. on System Science* 413–415.

[2] COVER, T. (1969). Learning in pattern recognition. In *Methodologies of Pattern Recognition* (S. Watanabe, ed.) 111–132. Academic Press, New York.

[3] COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-11** 21–27.

[4] FRITZ, J. (1975). Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-21** 552–557.

[5] GLICK, N. (1972). Sample-based classification procedures derived from density estimators. *J. Amer. Statist. Assoc.* **67** 116–122.

[6] STONE, M. (1974). Cross validation choice and assessment of statistical predictors. *J. Roy. Statist. Soc. Ser. B* 111–147.

[7] TOUSSAINT, G. (1974). Bibliography on estimation of misclassification. *IEEE Trans. Information Theory* **IT-20** 472–479.

[8] WAGNER, T. (1971). Convergence of the nearest neighbor rule. *IEEE Trans. Information Theory* **IT-17** 566–571.

[9] WAGNER, T. (1973). Deleted estimates of the Bayes risk. *Ann. Statist.* **1** 359–362.

THE RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF TEXAS
AUSTIN, TEXAS 78712