Probability inequalities are given for the deviation of the resubstitution error estimate from the unknown conditional probability of error. The inequalities are distribution-free and can be applied to linear discrimination rules, to nearest neighbor rules with a reduced sample size, and to histogram rules.

1. Introduction

The discrimination problem may be formulated as follows. The statistician collects data \((X_1,\theta_1), \ldots, (X_n,\theta_n)\), a sequence of independent identically distributed random vectors drawn from the distribution of \((X,\theta)\), a random vector independent of the data. For each \(1 \leq i \leq n\), the observation \(X_i\) takes values in \(\mathbb{R}^m\) and its state \(\theta_i\) takes values in \(\{1, \ldots, M\}\). The discrimination problem is that of estimating the state \(\hat{\theta}\) from the data and the observation \(X\) using procedures which do not require complete knowledge of the distribution of \((X,\theta)\). If \(\hat{\theta}\) denotes the estimate, that is, \(\hat{\theta} = g(X,\hat{\theta})\) where \(g\) is a Borel measurable \((1, \ldots, M)\)-valued function of \(X\) and the data \(V_n = (X_1,\theta_1), \ldots, (X_n,\theta_n)\), then a measure of the performance of the procedure given the data \(V_n\) is \(L_n = P(\hat{\theta} \neq \theta | V_n)\), the conditional probability of error.

Since the distribution of \((X,\theta)\) is unknown, there is in general no way of computing \(L_n\) from the data. Using the data the statistician may try to estimate \(L_n\) by \(\hat{L}_n\). A survey of estimation techniques can be found in Toussaint. One of the oldest estimates is the resubstitution estimate

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

where \(\hat{\theta}_i = g(X_i,\hat{\theta}_i), 1 \leq i \leq n\), are the estimates of the states of \(X_1, \ldots, X_n\) with the given discrimination procedure, and where \(I\) is the indicator function.

In this paper we obtain upper-bounds for \(P(\hat{L}_n - L_n \geq \varepsilon)\) that do not depend upon the distribution of \((X,\theta)\), and that are applicable to three large classes of discrimination rules,

(i) the linear discrimination rules,
(ii) the nearest-neighbor rules with reduced sample size, and
(iii) the histogram decision rules.

The existence of distribution-free bounds with the resubstitution estimate for linear discrimination rules was first noticed by Vapnik and Chervonenkis. The bounds for the class (i) improve the bounds given in Devroye and Wagner, while the results for the rules (ii) and (iii) are new. The possible existence of distribution-free bounds for (ii) was suggested to the authors by Dr. Penrod.

*This work was supported in part by the Air Force under Grant AFOSR 72-2371.

2. Main Results

Let \(\theta_0, \theta_1, \ldots, \theta_m\) be known measurable mappings from \(\mathbb{R}^m\) to \(\mathbb{R}\), where \(m' \geq 1\) and \(m' > 1\), and \(\theta_0 \neq \theta_1\).

Let \(w_0 = (w_{10}, \ldots, w_{1M}), \ldots, w_n = (w_{n0}, \ldots, w_{nM})\) be Borel-measurable \(\mathbb{R}^{m' + 1}\)-valued functions of the data \(V_n\). Then, the rule which assigns the state

\[
\hat{\theta} = j (1 \leq j \leq M)
\]

is called a linear discrimination rule (see Duda and Hart for a survey of the literature on linear discrimination). We emphasize that the \(w_{i0}, \ldots, w_{iM}\) may be picked in an arbitrary fashion, using any method that can or cannot be found in the literature. The functions \(w_i\) are picked in advance. The following bound is proved in the Appendix.

**Theorem 1.** For every \(\varepsilon > 0\) and for all linear discrimination rules with given \(w_0, w_1, \ldots, w_n\), the resubstitution estimate \(\hat{L}_n\) satisfies

\[
P(\hat{L}_n - L_n \geq \varepsilon) \leq 4M(1 + (2n)^m) M^{-1} e^{-nc^2/2}\]

For the interesting case that \(M=2\), we see that

\[
P(\hat{L}_n - L_n \geq \varepsilon) \leq 8(1 + (2n)^m) e^{-nc^2/32}
\]

Using the Borel-Cantelli lemma and Theorem 1, we see that for a given \(m'\) and \(M\), and uniformly over all linear discrimination rules, \(\hat{L}_n \rightarrow L_n\) a.s. with probability one, a result due to Glick. Thus, the statistician could pick the \(w_{00}, \ldots, w_{nM}\) that minimize the resubstitution estimate \(\hat{L}_n\) because he knows from Theorem 1 that the corresponding probability of error \(L_n\) will be very close to \(L_n\), and that for large \(n\), minimizing \(\hat{L}_n\) is nearly equivalent to minimizing \(L_n\) (see Wagner). In the literature special attention has been given to the nearest-neighbor rule with a reduced number of observations where the reduction is a result of editing (see for instance Wilson), condensing (Hart) or any other operation (Tomek). In general, we end up with \(\mathbb{R}^{m'}\) valued random vectors \((Y_1, c_1), \ldots, (Y_K, c_K)\) where \(K\) is an integer-valued random variable with \(1 \leq K \leq n\). The \((Y_i, c_i)\) and \(K\) may depend upon the data in an arbitrary fashion. A new observation \(X\) is assigned the state \(\hat{\theta} = c_j\) whenever \(j\) is the smallest index for which

The U. S. Government is authorized to reproduce and sell this report. Permission for further reproduction by others must be obtained from the copyright owner.
Thus, $\bar{o}$ is the state of the nearest neighbor to $X$ among $Y_1, \ldots, Y_K$. In the Appendix the following Theorem is proved.

**Theorem 2.** For every $\epsilon > 0$, and for all the nearest-neighbor rules with reduced sample size, the resubstitution estimate $L_n$ satisfies

$$P(|L_n - L^o| > \epsilon) < 4M(1 + 2n)^2M_{\epsilon} - e\epsilon^2/8kn_k$$

where $k_n$ is an upper-bound on the reduced sample size.

We remark that this bound is independent of the distribution of $(X, a)$. The bound converges to 0 as $n$ grows large provided that the sequence $k_1, k_2, \ldots$ is picked in such a way that $k_n^2 \log n/n \rightarrow 0$. It is clear that this bound is useless for the well-known nearest-neighbor rule, that is, the rule with $k_1 = k_2 = \ldots$ and $(\hat{y}_i, \hat{a}_i) = (x_i, a_i), 1 \leq i \leq n$. This was to be expected because the resubstitution estimate with the nearest-neighbor rule is overly optimistic. In fact, if the probability measure $\mu$ of $X$ is absolutely continuous with respect to Lebesgue measure, then $L_n = 0$ with probability one, no matter what value $L_n$ takes.

Theorem 2 can be useful for reduced, selective, condensed edited nearest-neighbor rules $k_1, k_2, \ldots$. If $k_n$ is a prespecified number of $(\hat{y}_i, \hat{a}_i)$'s that are to be used in the new nearest-neighbor rule, then the statistician could compute $L_n$ with some selected data $(X_1, a_1), \ldots, (X_i, a_i)$, where $i_1, \ldots, i_k_n$ is a subset of $\{1, \ldots, n\}$, and decide to use that set of indices for which the resubstitution estimate is minimal. Using Theorem 2, we also know how much confidence we can put in our estimate $L_n$ regardless of the selection procedure of the $(\hat{y}_i, \hat{a}_i)$ and without any knowledge of the distribution of $(X, a)$.

The $(\hat{y}_i, \hat{a}_i), i = 1, k_n$, partition $\mathbb{R}^m$ into $k_n$ disjoint sets $A_1, \ldots, A_{k_n}$ where the state of $X$ is estimated by $\hat{o} = \hat{e}_j$ whenever $X$ takes values in $A_j$ (that is, $X$ is closest to $Y_j$). The partition in this case depends on the data because the $Y_i$ depend upon the data. For a given fixed partition of $\mathbb{R}^m$, we can expect to obtain tighter upper-bounds for $P(|L_n - L^o| > \epsilon)$ even if the partition is not generated by a reduced nearest-neighbor rule.

Let $A_1, \ldots, A_{k_n}$ be any fixed partition of $\mathbb{R}^m$ and let $\xi_1, \xi_2, \ldots, \xi_{k_n}$ be random variables where, as before, $\xi_j$ is the state assigned to $X$ whenever $X$ takes values in $A_j$. Such rules will be called histogram decision rules. We prove the following four distribution-free inequalities that are valid no matter how the $\xi_i$ depend upon the data. The inequalities do not imply one another.

**Theorem 3.** For a given $k_n$-member partition of $\mathbb{R}^m$, for any way of specifying $\xi_1, \ldots, \xi_{k_n}$ from the data in a histogram decision rule, and for every $\epsilon > 0$, the resubstitution estimate $L_n$ satisfies

$$P(|L_n - L^o| > \epsilon) < g_{n1}, \quad 1 \leq i \leq 4,$$

where

$$g_{n1} = 4k_n \min\{1 + 2n, M\} \epsilon^{2/8kn_k}$$

$$g_{n2} = 2k_n \epsilon^{2/8kn_k}$$

$$g_{n3} = 4M(4/n^2 + 4n/k_n) \epsilon^{2/8}$$

$$g_{n4} = 2M \epsilon^{2/8kn_k}.$$

We note here that $g_{n1}$ and $g_{n3}$ are useful even if $M = \infty$ (i.e., the $\xi_i$ and $\xi_0$ can take a countably infinite number of values). Clearly, all the $g_{n1}$ are independent of the dimension $m$ and the distribution of $(X, a)$. Fixed partitions are not generated by a reduced nearest-neighbor rule. The closest one can come to the Bayes rule with a fixed partition is to let $\hat{e}_j = j$ if $j$ is the smallest integer such that $N_j > n/2n^2$. Indeed, assume that $M = \infty$ and that $X$ takes values in each $A_1, \ldots, A_{k_n}$. Then the resubstitution estimate does not possess the distribution-free properties that it has with finite partitions of $\mathbb{R}^m$. Assume that $A_1, A_2, \ldots$ is a fixed countably infinite partition of $\mathbb{R}^m$. If the $\xi_i$ are random variables that are independent of the data, then

$$P(|L_n - L^o| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

for any $\epsilon > 0$. However, such rules are impractical. The closest one can come to the Bayes rule with a fixed partition is to let $\hat{e}_j = j$ if $j$ is the smallest integer such that $N_j > n/2n^2$, where $N_j$ is the number of $(x_i, a_i)$'s with $x_i \in A_j$ and $a_i = a_j$. Even with this obvious choice of $\xi_1, \xi_2, \ldots$, we see that for any $m$ and $M \geq 2$, there always exists a distribution of $(X, a)$ such that $|L_n - L^o| > \epsilon$ with probability one. Indeed, assume that $M = 2$, that $a_2$ with probability one, and that $X$ takes values in each $A_1, A_2, A_3$ with equal probability $1/2n$. If the $\xi_i$ are picked as described, then the resubstitution estimate $L_n$ equals 0. Furthermore,

$$L_n = \sum_{i=1}^{2n} P(X = a_i) I(N_{i2} = 0) \geq n/2n = 1/2.$$
3. Appendix

Proof of Theorem 1.

Let \( v \) be the probability measure of \((X,\xi)\) where \( X \) takes values in \( \mathbb{R}^m \) and \( \xi \) takes values in \( \{1,\ldots,M\} \). It is clear that if \( v_n \) is the empirical measure for \((X_1,\xi_1),\ldots,(X_n,\xi_n)\), and if \( A_1,\ldots,A_M \) is the partition of \( \mathbb{R}^m \) that is generated by the linear discrimination rule (that is, \( A_i \) is the set on which we estimate the state of \( X \) by \( i \)), then

\[
L_n = \sum_{i=1}^{M} v(A_i \times \{i\})
\]

and

\[
\hat{L}_n = \sum_{i=1}^{M} v_n(A_i \times \{i\}) .
\]

Thus,

\[
|L_n - \hat{L}_n| = \left| \sum_{i=1}^{M} (v(A_i \times \{i\}) - v_n(A_i \times \{i\})) \right| 
\]

\[
\leq M \sup_{A_i \in \mathcal{W}} \left| v_n(A_i \times \{i\}) - v(A_i \times \{i\}) \right| 
\]

where \( \mathcal{W} \) is the class of all sets that are intersections of \( (M-1) \) linear halfspaces of \( \mathbb{R}^m \). We recall that a linear halfspace of \( \mathbb{R}^m \) is a set of \( x = (x_1,\ldots,x_m) \) for which \( x_1a_1 + \cdots + x_ma_m \geq a_0 \) or \( x_1a_1 + \cdots + x_ma_m < a_0 \) for some \((a_0,a_1,\ldots,a_m) \in \mathbb{R}^{m+1} \). Thus, every \((a_0,a_1,\ldots,a_m) \) defines two linear halfspaces.

By an inequality of Vapnik and Chervonenkis\(^9\),

\[
P(|L_n - \hat{L}_n| \geq \epsilon) \leq 4s(A_n,2n)e^{-n(\epsilon/M)^2/8}
\]

where \( s(\mathcal{A},n) \) is the maximum over all \((x_1,y_1),\ldots,(x_M,y_M)\) in \( \mathbb{R}^m \times (1,\ldots,M) \) of the number of different sets in \( \{(x_1,y_1)\cup\cdots\cup(x_M,y_M)\} \) \( \& \mathcal{A} \in \mathcal{B} \). If \( \mathcal{A} = \mathcal{W} \), then \( s(\mathcal{W},n) \leq 1 + nM^2 \) by a theorem of Cover\(^8\) (see also Vapnik and Chervonenkis\(^9\)). It is clear that if \( \mathcal{W} \) is the class of all intersections of \( M-1 \) or less linear halfspaces and \( M \leq 1 \), then \( s(\mathcal{W},n) \leq 1 + nM^2 \). Indeed, if \( \mathcal{W} \) is the number of different sets in \( \{(x_1,y_1)\cup\cdots\cup(x_M,y_M)\} \), then the number of different sets in \( \{(x_1,y_1)\cup\cdots\cup(x_M,y_M)\} \mathcal{W} \mathcal{A} \) is at most \( M_s + 1 \). Thus we have shown that

\[
P(|L_n - \hat{L}_n| \geq \epsilon) \leq 4M(1+(2n)^m)^{M-1} e^{-n(\epsilon^2/8k^2)} .
\]

Q.E.D.

Proof of Theorem 2.

Let us use the notation of Theorem 1 where we let \( A_1,\ldots,A_M \) be the partition of \( \mathbb{R}^m \) that is generated by the nearest-neighbor rule with \((Y_1,\xi_1),\ldots,(Y_k,\xi_k)\) (i.e., \( A_i \) is the set on which we estimate the state of \( X \) by \( \xi_i \) and for which \( Y_i \) is the nearest neighbor to \( X \) among \( Y_1,\ldots,Y_k \) ), then

\[
L_n = \sum_{i=1}^{K} v(A_i \times \{i\})
\]

and

\[
\hat{L}_n = \sum_{i=1}^{K} v_n(A_i \times \{i\}) .
\]

Thus, arguing as in Theorem 1, we have

\[
P(|L_n - \hat{L}_n| \geq \epsilon) \leq 4M(1+(2n)^m)^{K-1} e^{-n(\epsilon^2/8k^2)} .
\]

Q.E.D.

Proof of Theorem 3.

It is clear that

\[
|L_n - \hat{L}_n| = \left| \sum_{i=1}^{K} (v(U(A_i \times \{i\})) - v_n(U(A_i \times \{i\})) \right| 
\]

\[
\leq K \sup_{A_i \in \mathcal{W}} \left| v_n(U(A_i \times \{i\})) - v(U(A_i \times \{i\})) \right| 
\]

\[
\leq \sum_{i=1}^{K} \sup_{1 \leq i \leq M} |v(A_i \times \{i\}) - v_n(A_i \times \{i\})| .
\]

Thus, if \( \mathcal{A}_i \) is the class of sets of the form \( A_i \), then we know by an inequality of Vapnik and Chervonenkis\(^9\) that

\[
P(|L_n - \hat{L}_n| \geq \epsilon) \leq 4K \sup_{1 \leq i \leq M} s(\mathcal{A}_i,2n) e^{-n(\epsilon/k^2)} .
\]

Also,
by an inequality of Hoeffding. Furthermore,

\[ P(\|L_n - \tilde{L}_n\| \geq \epsilon) \]

\[ \leq P \left( \sum_{i=1}^{k_n} \sum_{t=1}^{M} |v(A_i(x(t))) - v_n(A_i(x(t)))| \geq \epsilon \right) \]

\[ \leq k_n M \sup_{l \leq \epsilon M} P \left( \sum_{i=1}^{k_n} |v(A_i(x(t))) - v_n(A_i(x(t)))| \geq \epsilon/k_n M \right) \]

\[ \leq 2k_n M e^{-2n^2/M^2 k_n^2} \]

where \( \Psi \) is the class of all sets of the form \( U(A_{i_1} \cdots A_{i_k}) \) where \( \{i_1, \ldots, i_k\} \in \Psi = \{1, \ldots, M\} \).

Clearly, \( s(\Psi, 2n) \leq 2n \) for all \( k_n \). However, if \( k_n < 2n \), then \( s(\Psi, 2n) \leq k_n \) and, in general, we must have that \( s(\Psi, 2n) \leq k_n (2n/k_n)^n \). This proves the inequality with \( g_{n3} \).

Finally, notice that

\[ P(\|L_n - \tilde{L}_n\| \geq \epsilon) \]

\[ \leq \sum_{\delta \in \Psi} P \left( \sum_{i=1}^{k_n} |v(U(A_i(x_{i_1}))) - v_n(U(A_i(x_{i_1})))| \geq \epsilon/M k_n \right) \]

\[ \leq 2M k_n e^{-2n^2/M^2 k_n} \]

Q.E.D.

Proof of (1).

Inequality (1) is a corollary of Hoeffding's inequality if we note that \( |L_n - \tilde{L}_n| = |v(C) - v_n(C)| \) where

\[ C = U \left( A_i(x_i)^C \right) \]

Q.E.D.

References

11. N. GLICK: personal communication.
**Title:** DISTRIBUTION-FREE PERFORMANCE BOUNDS WITH THE RESUBSTITUTION ERROR ESTIMATE

**Authors:** L.P. Devroye and T.J. Wagner

**Abstract:**

Probability inequalities are given for the deviation of the resubstitution error estimate from the unknown conditional probability of error. The inequalities are distribution-free and can be applied to linear discrimination rules, to nearest neighbor rules with a reduced sample size, and to histogram rules.