The Annals of Statistics
1973, Vol. 1, No. 2, 359-362

DELETED ESTIMATES OF THE BAYES RISK[1]

BY T. J. WAGNER

University of Texas, Austin

Consider the usual decision theoretic situation where one observes a random vector X from which an estimate of its classification $\theta \in \{0, 1\}$ is to be made. If one knows the a priori probabilities for θ and the conditional densities of X given θ then the smallest probability of error which can be achieved is called the Bayes risk and denoted by R^*. Assuming that the a priori probabilities and conditional densities are unknown we consider the problem of estimating R^* from the independent observations $(X_1, \theta_1), \cdots, (X_n, \theta_n)$. Suppose X has an unknown classification θ where (X, θ) is independent of the observations $(X_1, \theta_1), \cdots, (X_n, \theta_n)$. If $\{\delta_n\}$ is a sequence of decision procedures, where δ_n determines the estimate of θ from X and $(X_1, \theta_1), \cdots, (X_n, \theta_n)$, then the notion of a deleted estimate of R^* with δ_n is introduced and, under mild assumptions, is shown to be a consistent estimate of R^*.

Consider the usual decision theoretic situation where one observes an m-dimensional random vector X from which an estimate of its true classification $\theta \in \{0, 1\}$ is to be made. It is assumed that $P[\theta = i] = \pi_i$ and that $P[X \leqq x \mid \theta = i]$ has a continuous density f_i, $i = 0, 1$. If a decision function $\delta: R^m \to \{0, 1\}$ is used then the probability of misclassification is given by

$$P[\delta(X) \neq \theta] = \int_{R^m} [\delta(x)\pi_0 f_0(x) + (1 - \delta(x))\pi_1 f_1(x)]\, dx\,. \tag{1}$$

(1) is minimized by δ^* where $\delta^*(x) = 1$ if $\pi_1 f_1(x) - \pi_0 f_0(x) \geqq 0$ and $\delta^*(x) = 0$ otherwise. The resulting probability of misclassification is called the Bayes Risk and is denoted by R^* so that from (1)

$$R^* = \int_{R^m} \min\{\pi_0 f_0(x), \pi_1 f_1(x)\}\, dx\,. \tag{2}$$

Suppose that π_0, π_1, f_0, f_1 are unknown. A nonparametric classification problem frequently considered can be stated: how does one use a sequence of independent, classified observations $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ to estimate the classification θ of an independent, unclassified observation X. Here, θ_j is the classification (label) of X_j and it is assumed that (X_j, θ_j) has the same distribution as (X, θ). Specifically, if $\delta_n: R^m \times (R^m \times \{0, 1\})^n \to \{0, 1\}$ represents a decision function for θ based on X and $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ then

$$L_n = P[\theta(n) \neq \theta \mid (X_1, \theta_1), \cdots, (X_n, \theta_n)]\,,$$

where $\theta(n) = \delta_n(X, (X_1, \theta_1), \cdots, (X_n, \theta_n))$, is the probability of misclassification for δ_n given the observations $(X_1, \theta_1), \cdots, (X_n, \theta_n)$. L_n is a random variable whose value is just the frequency of errors that one would obtain if δ_n and $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ were used to classify a large number of independent observations. The

Received December 17, 1971; revised August 1, 1972.

[1] Work supported by Air Force Office of Scientific Research under Grant No. AFOSR 72-2371.

nonparametric classification problem above is concerned with finding a sequence $\{\delta_n\}$ which insures that L_n is close to R^* with high probability for large n.

We here go one step beyond the above nonparametric classification problem and ask: how do we estimate R^* from the data $(X_1, \theta_1), \cdots, (X_n, \theta_n)$? One reason for desiring an estimate of R^* is that an estimate of R^* should be close to L_n with reasonable procedures and a direct estimate of L_n is impossible without additional independent observations. A second reason is best explained by an example. Suppose one wants to automate the checking of chest X-rays for tuberculosis. For this problem one would naturally have available a large number of X-rays of people already diagnosed as tuberculin or nontuberculin. The data, at this point, is not a sequence of vectors with their labels but a sequence of X-rays with their labels. One then has the initial problem of reducing an X-ray to a vector if processing along the above lines is to be done. Which reduction to use depends on several factors but, with all other factors being equal, the reduction with the smallest R^* is preferable. It is desirable then to have good estimates of R^* just for comparisons of this type.

As before let $\{\delta_n\}$ be a sequence of decision functions for the data (X_1, θ_1), $(X_2, \theta_2), \cdots$ and let

$$\theta_j(n) = \delta_{n-1}(X_j, (X_1, \theta_1), \cdots, (X_{j-1}, \theta_{j-1}), (X_{j+1}, \theta_{j+1}), \cdots, (X_n, \theta_n)),$$

that is, $\theta_j(n)$ is the estimate of θ_j with δ_{n-1} based on X_j and $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ *with* (X_j, θ_j) *deleted*. If $S_n = \sum_1^n I_{[\theta_j(n) \neq \theta_j]}$ is the number of times that $\theta_j(n) \neq \theta_j$, $1 \leqq j \leqq n$, then $E(S_n/n) = EL_{n-1}$. This type of estimate was suggested by Cover [1] who termed it a *deleted estimate* of EL_{n-1} and who also proposed it as an estimate of the nearest neighbor risk where now δ_n represents the nearest neighbor decision function. (These ideas as well as the framework of the nonparametric classification problem are discussed in the highly recommended paper of Cover [1].) To achieve a somewhat greater generality we will assume that the asymptotic risk of the decision rules δ_n is R, where R is not necessarily equal to R^* (see (3a) below). If another sequence of rules δ_n' has an asymptotic risk R' then, subject to the conditions of the theorem below, the deleted estimates of R and R' allow a comparison of $\{\delta_n\}$ and $\{\delta_n'\}$.

THEOREM. *If δ_n is symmetric in $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ and if*

(3a) $$L_n \to_n R \quad \textit{in probability}$$

(3b) $$P[\theta(n) \neq \theta(n+1)] \to 0$$

then $S_n/n \to_n R$ in probability.

EXAMPLE 1. Take $\delta_n = 1$ if the vector from $X_1, \cdots, X_n$ which is closest to X has a label equal to 1 and take $\delta_n = 0$ otherwise (ties are broken arbitrarily). Then [2], [4] it is known that

$$L_n \to_n R \quad \text{in probability}$$

where $R = \int \{2\pi_0 f_0(x)\pi_1 f_1(x)/f(x)\}\,dx$, $f(x) = \pi_0 f_0(x) + \pi_1 f_1(x)$ and $R^* \leqq R \leqq$

$2R^*(1 - R^*)$. In addition $P[\theta(n) \neq \theta(n+1)] \leqq 1/(n+1)$ so that (3a) and (3b) are both satisfied.

EXAMPLE 2. Consider choosing a different rule δ_n. Let $r_n = n^{-p}$, where $0 < p < \frac{1}{2}m$ is fixed and let $N_1^n(N_0^n)$ be the number of labels θ_i equal to 1(0) which have each coordinate of X_i within a distance r_n of the corresponding coordinate of X. Next choose $\delta_n = 1$ if $N_1^n \geqq N_0^n$ and choose $\delta_n = 0$ if $N_0^n < N_1^n$. Then δ_n corresponds to the decision function

$$\begin{aligned} \delta_n(X) &= 1 \quad \text{if} \quad \hat{\pi}_1\hat{f}_1(X) - \hat{\pi}_0\hat{f}_0(X) \geqq 0 \\ &= 0 \quad \text{if} \quad \hat{\pi}_1\hat{f}_1(X) - \hat{\pi}_0\hat{f}_0(X) < 0 \end{aligned} \tag{4}$$

where

$$\hat{\pi}_1 = \textstyle\sum_1^n \theta_i/n\,, \qquad \hat{\pi}_0 = \sum_1^n (1-\theta_i)/n\,,$$

$$\hat{f}_1(x) = \textstyle\sum_1^n \theta_i r_n^{-m} K(r_n^{-1}(x - X_i))/\sum_1^n \theta_i\,,$$

$$\hat{f}_0(x) = \textstyle\sum_1^n (1-\theta_i) r_n^{-m} K(r_n^{-1}(x - X_i))/\sum_1^n (1-\theta_i)\,, \qquad \text{and}$$

$$\begin{aligned} K(x) &= 1 \quad \text{if} \quad ||x|| = \max\{x_1, \cdots, x_m\} \leqq 1\,, \quad x = (x_1, \cdots, x_m) \\ &= 0 \quad \text{if otherwise.} \end{aligned}$$

Thus δ_n is just the decision function obtained if the estimates of π_0, π_1, f_0, f_1 from $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ are used in the discriminant function $\pi_1 f_1 - \pi_0 f_0$. Van Ryzin [3] has shown that $L_n \to_n R^*$ in probability for this sequence $\{\delta_n\}$ if the set of points where f_0 and f_1 are discontinuous has Lebesgue measure 0. To show that $P[\theta(n+1) \neq \theta(n)] \to_n 0$ note that

$$[\theta(n+1) \neq \theta(n)] \subset A \cup A^c[\theta(n+1) \neq \theta(n)]$$

where A is the event that each coordinate of X_{n+1} is within r_n of the corresponding coordinate of X. Now

$$A^c[\theta(n+1) \neq \theta(n)] \subset \textstyle\bigcup_j [r_{n+1} \leqq ||X - X_j|| \leqq r_n]$$

and it follows easily that $P(A) \to_n 0$ and $P\{\bigcup_j [r_{n+1} \leqq ||X - X_j|| \leqq r_n]\} \to_n 0$, and thus that $P[\theta(n+1) \neq \theta(n)] \to_n 0$.

PROOF OF THE THEOREM. The technique is essentially that contained in the proof of Theorem 2 of [3]. We show that $E(S_n/n - R)^2 \to 0$. A simple computation shows that $E(S_n/n - R)^2 \to 0$ if, and only if, $P[\theta_j(n) \neq \theta_j;\, \theta_k(n) \neq \theta_k] \to (R)^2$ for each $1 \leqq j < k \leqq n$. Fix $j < k$ and let $\theta_j'(n)$ and $\theta_k'(n)$ be the estimates of θ_j and θ_k, respectively, with δ_{n-2} from $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ with *both* (X_j, θ_j) and (X_k, θ_k) deleted. Then

$$\begin{aligned} &[\theta_j(n) \neq \theta_j;\, \theta_k(n) \neq \theta_k] \triangle [\theta_j'(n) \neq \theta_j;\, \theta_k'(n) \neq \theta_k] \\ &\qquad\qquad \subset [\theta_j'(n) \neq \theta_j(n)] \cup [\theta_k'(n) \neq \theta_k(n)]\,. \end{aligned}$$

Because $P[\theta_j'(n) \neq \theta_j(n)] = P[\theta_k'(n) \neq \theta_k(n)] = P[\theta(n) \neq \theta(n-1)] \to 0$, it suffices to show that $P[\theta_j'(n) \neq \theta_j;\, \theta_k'(n) \neq \theta_k] \to (R)^2$. Now

$$\begin{aligned} P[\theta_j'(n) \neq \theta_j;\, \theta_k'(n) \neq \theta_k] &= E\{P[\theta_j'(n) \neq \theta_j;\, \theta_k'(n) \neq \theta_k \mid \mathscr{B}_n'']\} \\ &= E\{P[\theta_j'(n) \neq \theta_j \mid \mathscr{B}_n'']P[\theta_k'(n) \neq \theta_k \mid \mathscr{B}_n'']\} \end{aligned}$$

where $\mathscr{B}_n''$ is the σ-algebra generated by $(X_1, \theta_1), \cdots, (X_n, \theta_n)$ with (X_j, θ_j) and (X_k, θ_k) deleted. Using (3a) with θ replaced by θ_j and θ_k respectively, it follows that

$$P[\theta_j'(n) \neq \theta_j \mid \mathscr{B}_n'']P[\theta_k'(n) \neq \theta_k \mid \mathscr{B}_n''] \rightarrow_n (R)^2 \quad \text{in probability}$$

so that

$$E\{P[\theta_j'(n) \neq \theta_j \mid \mathscr{B}_n'']P[\theta_k'(n) \neq \theta_k \mid \mathscr{B}_n'']\} \rightarrow_n (R)^2$$

by the Lebesgue Dominated Convergence Theorem and the theorem is proved.

Acknowledgment. I appreciate the reviewer's helpful suggestions. Both he and Tom Cover pointed out an improvement in an earlier version of the theorem given here.

REFERENCES

[1] Cover, T. (1969). Learning in pattern recognition. *Methodologies of Pattern Recognition*, ed. S. Watanabe. Academic Press, New York, 111-132.

[2] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Information Theory* **IT-13** 21-27.

[3] Van Ryzin, J. (1967). Non-parametric Bayesian decision procedures for (pattern) classification with stochastic learning. *Trans. Fourth Prague Conf. on Information Theory, Statistical Decision Functions and Random Processes* (Prague, 1965), Academia, Prague, 479-494.

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

Department of Electrical Engineering
University of Texas
Austin, Texas 78712