Any Discrimination Rule Can Have an Arbitrarily Bad Probability of Error for Finite Sample Size

LUC DEVROYE

Abstract—Consider the basic discrimination problem based on a sample of size n drawn from the distribution of (X, Y) on the Borel sets of $R^d x \{0, 1\}$. If $0 < R^* < \frac{1}{2}$ is a given number, and $\phi_n \to 0$ is an arbitrary positive sequence, then for any discrimination rule one can find a distribution for (X, Y), not depending upon n, with Bayes probability of error R^* such that the probability of error (R_n) of the discrimination rule is larger than $R^* + \phi_n$ for infinitely many n. We give a formal proof of this result, which is a generalization of a result by Cover [1]. Furthermore,

$$\sup_{\substack{\text{all distributions of} \\ (X, Y) \text{ with } R^* = 0}} R_n \geq \frac{1}{2}.$$

Thus, any attempt to find a nontrivial distribution-free upper bound for R_n will fail, and any results on the rate of convergence of R_n to R^* must use assumptions about the distribution of (X, Y).

Index Terms—Bayes risk, consistency, discrimination rule, distribution-free inequalities, probability of error.

I. Introduction

In this paper we will try to clarify the statement that, without assumptions on the distribution of the data, no rate of convergence (to the Bayes probability of error) can be proved for any discrimination rule.

Let $(X, Y), (X_1, Y_1), \cdots, (X_n, Y_n)$ be independent identically distributed $R^d x \{0, 1\}$-valued random vectors. The *discrimination problem* is concerned with the estimation of Y from X and the *data* (the (X_i, Y_i) sequence). A *discrimination rule* is a sequence of Borel measurable mappings $g_1, g_2, \cdots$, where g_n maps $R^d x (R^d x \{0, 1\})^n$ to $\{0, 1\}$. The estimate of Y is

$$\hat{Y} = g_n(X, X_1, Y_1, \cdots, X_n, Y_n)$$

and the associated conditional and unconditional probabilities of error are

$$L_n = P(\hat{Y} \neq Y \mid X_1, Y_1, \cdots, X_n, Y_n)$$

and

$$R_n = E(L_n).$$

In any case we have $L_n \geq R^*$, where R^* is the *Bayes risk* (or Bayes probability of error)

$$R^* = \inf_{g:\, R^d \to \{0,1\}} P(g(X) \neq Y).$$

Manuscript received December 29, 1980; revised September 16, 1981.
The author is with the School of Computer Science, McGill University, Montreal, P.Q., Canada.

A discrimination rule is said to be *Bayes risk consistent* if $\lim_n R_n = R^*$. Some discrimination rules are Bayes risk consistent for some distributions of (X, Y) only (e.g., the linear discrimination rules). It is now known that most nonparametric discrimination rules are Bayes risk consistent for *all* distributions of (X, Y), e.g., the k_n nearest neighbor and related rules [9], the k_n nearest neighbor rule based only on the ranks of all the projections of the X_i's [2], [7], recursive versions of the k_n nearest neighbor rule [4], the kernel rule (or potential function method) [3], [8], recursive versions of the kernel rule [6], and partitioning rules [5].

The next question one might ask is whether there are discrimination rules for which R_n converges to R^* with a certain speed [e.g., $R_n = R^* + 0(1/n)$] regardless of the distribution of (X, Y). The answer is negative in view of Theorems 1 and 2 proved in this paper.

Theorem 1: For any discrimination rule

$$\sup_{\substack{\text{all distributions of} \\ (X, Y) \text{ with } R^* = 0}} R_n \geq \frac{1}{2}.$$

Theorem 2: Let $\phi_n \to 0$ be an arbitrary positive sequence, and let $0 \leq R^* \leq \frac{1}{2}$ be a given number. For any discrimination rule there exists a distribution of (X, Y) with Bayes risk R^* such that

$$R_n \geq \min\left(R^* + \phi_n, \frac{1}{2}\right)$$

for infinitely many n.

Theorem 1 states that no nontrivial distribution-free upper bounds for R_n exist. In particular, if ϕ_n is a given positive number sequence tending to 0, and if $R_n \leq R^* + c\phi_n$, then the constant c must necessarily depend upon the distribution of (X, Y). Theorem 2 is a bit different; it states that no inequality of the form $R_n \leq R^* + c\phi_n$ can be obtained even if c depends upon the distribution of (X, Y)! The reason is that we can always find some distribution of (X, Y) for which $R_n \geq R^* + \sqrt{\phi_n}$ infinitely often. Theorem 2 refines a result by Cover [1]. Notice finally that Theorem 2 also applies to the k-nearest neighbor rules (k fixed) for the special case that $R^* = 0$. For the case $R^* > 0$, Theorem 2 contains no information because the k-nearest neighbor rules are not Bayes risk consistent.

By carefully analyzing the proofs, we see that Theorems 1 and 2 remain valid when X has a fixed (but otherwise arbitrary) density. Thus, putting restrictions on the distribution of X alone will not suffice for the study of the rate of convergence of discrimination rules. One needs at least conditions on the regression functions $\eta_i(x) = P(Y = i \mid X = x)$.

0162-8828/82/0300-0154$00.75

II. Proofs

We will prove both theorems by constructing an example in R^1, in which X has a density with infinite support. One can easily construct similar examples in R^d, or examples in which X is purely atomic, or examples in which X has a uniform distribution on $[0, 1]^d$. The density of X is

$$f(x) = \sum_{i=0}^{\infty} p_i I_{[i, i+1)}(x)$$

where I is the indicator function and $(p_0, p_1, \cdots)$ is a probability vector. In other words, the density of X is histogram-shaped with height p_i on the interval $[i, i+1)$.

In the proof we will need a family of distributions of (X, Y). This family will have a parameter $z \in [0, 1)$, and the distribution $D(z)$ is defined as follows. Let z have binary expansion $0.\ z_0 z_1 z_2 \cdots$, then $Y = z_X$ where X = integer part of X, and X has density f. For all $D(z)$, it is clear that $R^* = 0$ because Y depends upon X. Let us call the probability of error for fixed z $R_n(z)$. The main step in our proof is based upon a randomization argument. Let us introduce an independent uniform $[0, 1]$ random variable Z with binary expansion $0.\ Z_0 Z_1 Z_2 \cdots$. Then we have the following chain of inequalities:

$$\begin{aligned}
&\sup_{z \in [0,1)} R_n(z) \\
&\quad \geq E(R_n(Z)) \qquad \text{(the supremum always exceeds the mean)} \\
&\quad = P(g_n(X, X_1, Z_{X_1}, \cdots, X_n, Z_{X_n}) \neq Z_X) \\
&\quad = E(P(g_n(X, X_1, Z_{X_1}, \cdots, X_n, Z_{X_n}) \\
&\qquad\qquad \neq Z_X \mid X, X_1, \cdots, X_n)) \\
&\quad \geq \frac{1}{2} P\left(\bigcap_{i=1}^{n} [X \neq X_i] \right) \qquad \text{(when } X \neq X_i \text{ for all } i\text{, then } Z_X \text{ is independent of } g_n(X, X_1, \cdots, Z_{X_n}) \text{ and takes the values 0 and 1 with equal probability)} \\
&\quad = \frac{1}{2} \sum_{i=0}^{\infty} p_i (1 - p_i)^n \\
&\quad > \frac{1}{2} \left(1 - \frac{1}{2n}\right)^n \sum_{p_i \leq 1/(2n)} p_i \\
&\quad \geq \frac{1}{4} \sum_{p_i \leq 1/(2n)} p_i. \qquad (1)
\end{aligned}$$

Proof of Theorem 1: Let $p_i = 1/K$, $1 \leq i \leq K$, and $p_i = 0$ elsewhere. Then by (1)

$$\sup_{z \in [0,1)} R_n(z) \geq \frac{1}{2} \sum_{i=1}^{K} p_i (1 - p_i)^n = \frac{1}{2} \left(1 - \frac{1}{K}\right)^n$$

which is arbitrarily close to $\frac{1}{2}$ by the choice of K. Theorem 1 follows since $R^* = 0$ for all distributions $D(z)$, $z \in [0, 1)$.

Proof of Theorem 2: Let us first note that we can assume that ϕ_n is strictly monotone. If we can prove Theorem 2 for all $\psi_n \downarrow 0$ ($\downarrow$ denotes monotone convergence), then we have proved it for all sequences $\phi_n \to 0$. Indeed, $\psi_n = \sup(\phi_i\colon i \geq n) + 1/n \downarrow 0$. Thus, we may find a distribution of (X, Y) with Bayes probability of error R^* such that $R_n \geq \min(R^* + \psi_n, \frac{1}{2}) \geq \min(R^* + \phi_n, \frac{1}{2})$ infinitely often.

We begin by showing that we can find N and $(p_0, p_1, \cdots)$ such that

$$\frac{1}{4} \sum_{p_i \leq 1/(2n)} p_i \geq \phi_n, \quad \text{all } n \geq N.$$

The construction is rather straightforward: let

$$N = \inf(n\colon n \geq 1, \phi_n \leq \tfrac{1}{8}),$$

$$k_0 = k_1 = \cdots = k_{N-1} = 0.$$

Assume that $n \geq N$. Given k_{n-1}, find $k_n > k_{n-1}$ and p_i, $k_{n-1} < i \leq k_n$ as follows:

1) all p_i's are positive;

2) the p_i sequence is nonincreasing (starting with p_1);

3) $p_i \leq 1/(2n)$;

4) $$\sum_{i=k_{n-1}+1}^{k_n} p_i = 4(\phi_n - \phi_{n+1}).$$

Clearly

$$\begin{aligned}
\frac{1}{4} \sum_{p_i \leq 1/(2n)} p_i &\geq \frac{1}{4} \sum_{i=k_{n-1}+1}^{\infty} p_i = \sum_{i=n}^{\infty} \frac{1}{4} \cdot 4(\phi_i - \phi_{i+1}) \\
&= \sum_{i=n}^{\infty} (\phi_i - \phi_{i+1}) = \phi_n, \quad n \geq N.
\end{aligned}$$

To make $(p_0, p_1, \cdots)$ a probability vector, note that

$$\sum_{i=1}^{\infty} p_i = 4\phi_N \leq \frac{1}{2}.$$

Choose p_0 such that the sum of the p_i's is one.

Consider first the case $R^* = 0$. Let

$$\bar{R}_n(z) = \sup_{m \geq n} R_m(z)/\phi_m.$$

We assume first that $\bar{R}_n(z) \to 0$ for all $z \in [0, 1)$. We will reach a contradiction, and must conclude that

$$\sup_{z \in [0,1)} \limsup_{n \to \infty} R_n(z)/\phi_n > 0.$$

Let $D_n = \{z\colon z \in [0, 1), \bar{R}_n(z) \leq 1\}$, D_n^c-complement of D_n. Clearly, $P(Z \in D_n^c) \to 0$ as $n \to \infty$ by the Lebesgue dominated convergence theorem. Let $(p_1, p_2, \cdots)$ be a probability vector chosen as in the previous construction. By Fatou's lemma

$$\begin{aligned}
&\sup_{z \in [0,1)} \limsup_{n \to \infty} R_n(z)/\phi_n \\
&\quad \geq E\left(\limsup_{n \to \infty} (R_n(Z)/\phi_n) I_{[Z \in D_n]}\right) \\
&\quad \geq \limsup_{n \to \infty} E((R_n(Z)/\phi_n) I_{[Z \in D_n]}) \\
&\quad = \limsup_{n \to \infty} \phi_n^{-1} P(Z \in D_n,
\end{aligned}$$

$$
\begin{aligned}
&g_n(X, X_1, Z_{X_1}, \cdots, X_n, Z_{X_n}) \neq Z_X) \\
&= \limsup_{n \to \infty} \phi_n^{-1} E(P(Z \in D_n, g_n(X, X_1, \cdots, Z_{X_n}) \\
&\qquad \neq Z_X \mid X, X_1, X_2, \cdots, X_n)) \\
&= \limsup_{n} \phi_n^{-1} E\Bigg(I_{[\cap_{i=1}^n [X \neq X_i]]} \sum_{j=0}^{1} P(Z_X \\
&\qquad = j, g_n(X, X_1, \cdots, Z_{X_n}) \neq j, \\
&\qquad\qquad Z \text{ (with } j \text{ forced in the } X\text{th position)} \in D_n \\
&\qquad\qquad \mid X, X_1, \cdots, X_n)\Bigg) \\
&\geq \limsup_{n \to \infty} \phi_n^{-1} E(\tfrac{1}{2} I_{[\cap_{i=1}^n [X \neq X_i]]} \\
&\qquad P(Z \text{ (with 0 forced in } X\text{th position)} \\
&\qquad \in D_n, Z \text{ (with 1 forced in } X\text{th position)} \in D_n \mid X)) \\
&= \limsup_{n \to \infty} (2\phi_n)^{-1} P\Bigg(\bigcap_{i=1}^{n} [X \neq X_i], \\
&\qquad \bigcap_{j=0}^{1} [Z \text{ (with } j \text{ forced in } X\text{th position)} \in D_n]\Bigg) \\
&\geq \limsup_{n \to \infty} (2\phi_n)^{-1} \left(\sum_{i=0}^{\infty} p_i(1-p_i)^n\right) (1 - 2P(Z \in D_n^c)) \\
&\geq \tfrac{1}{2}. \qquad (2)
\end{aligned}
$$

By assumption, the left-hand side of (2) is 0, so we have a contradiction. Thus, there exists a $z \in [0, 1)$ and a constant $c > 0$ such that $R_n(z) \geq c\phi_n$ infinitely often. But ϕ_n is arbitrary and can be replaced by $\sqrt{\phi_n}$, and thus there exists a $z \in [0, 1)$ and $c > 0$ such that $R_n(z) \geq c\sqrt{\phi_n} \geq \phi_n$ infinitely often.

In the case $R^* = \frac{1}{2}$, Theorem 2 becomes trivial. We may thus assume that $R^* \in (0, \frac{1}{2})$. We will once again use a family of distributions $D(z)$, $z \in [0, 1)$. As before, $(p_0, p_1, \cdots)$ is a probability vector, but now we set $p_0 = 2R^*$. Let W and X be independent random variables where X has density f and W is Bernoulli, taking the values 0 and 1 with equal probability. Let

$$
Y = \begin{cases} z_X, & X > 0, \\ W, & X = 0. \end{cases}
$$

For each z we have $R^* = \frac{1}{2}P(X = 0) = \frac{1}{2}p_0$, which explains our choice for p_0. Note that the sequence $(X, Y), (X_1, Y_1), \cdots, (X_n, Y_n)$ is known if $(X, W), (X_1, W_1), \cdots, (X_n, W_n)$ is known. Here the couples (X_i, W_i) are independent and distributed as (X, W). The randomization is performed as before, i.e., by introducing a uniform $[0, 1]$ random variable Z, independent of $(X, Y), (X_1, Y_1), \cdots, (X_n, Y_n)$. Arguing as in (1), we have

$$
\begin{aligned}
&\sup_{z \in [0,1)} R_n(z) \geq E(R_n(Z)) \\
&\quad = E(P(g_n(X, X_1, Y_1, \cdots, X_n, Y_n) \\
&\qquad \neq Y \mid X, X_1, \cdots, X_n)) \\
&\quad \geq E(I_{[X=0]} \tfrac{1}{2} + I_{[X \neq 0]} I_{[\cap_{i=1}^n [X \neq X_i]]} \tfrac{1}{2}) \\
&\qquad \text{(if } X \neq X_i, X \neq 0, \text{ we argue as in (1); if } X = 0, \\
&\qquad \text{then } Y = W \text{ is an independent Bernoulli random} \\
&\qquad \text{variable)} \\
&\quad = R^* + \tfrac{1}{2} \sum_{i=1}^{\infty} p_i(1-p_i)^n. \qquad (3)
\end{aligned}
$$

The remainder of the proof can be mimicked from the case $R^* = 0$. Define $N = \inf(n: n \geq 1, \phi_n \leq (1 - 2R^*)/8)$; $k_0 = k_1 = \cdots = k_{N-1} = 1$. Then construct $(p_2, p_3, \cdots)$ as before (nondecreasing from p_2 onwards) so that

$$
\sup_{z \in [0,1)} R_n(z) > R^* + \tfrac{1}{4} \sum_{p_i \leq 1/(2n)} p_i \geq R^* + \phi_n, \quad n \geq N.
$$

Define $p_1 = 1 - 2R^* - 4\phi_N$, and note that $p_1 \geq 0$, and that $p_0 + p_1 + p_2 + \cdots = 2R^* + (1 - 2R^* - 4\phi_N) + 4\phi_N = 1$.

We continue the proof as for the case $R^* = 0$. Let $\bar{R}_n(z) = \sup_{m \geq n} (R_m(z) - R^*)/\phi_m$, and let $(p_0, p_1, \cdots)$ be a probability vector chosen as indicated above. Assume first that $\bar{R}_n(z) \to 0$ for all $z \in [0, 1)$; we will once again reach a contradiction, and must then conclude that for some $z \in [0, 1)$, $\limsup_{n \to \infty} (R_n(z) - R^*)/\phi_n > 0$. This would conclude the proof of the theorem.

Under the said assumption, we have

$$
\begin{aligned}
&\sup_{z \in [0,1)} \limsup_{n \to \infty} (R_n(z) - R^*)/\phi_n \\
&\quad \geq \limsup_{n \to \infty} E((R_n(Z) - R^*) \phi_n^{-1} I_{[Z \in D_n]}) \qquad (4)
\end{aligned}
$$

where D_n is defined as in the proof of (2), and X, Y, Z, and W are distributed as indicated in the example preceding (3). The right-hand side of (4) can be rewritten as

$$
\begin{aligned}
&\limsup_{n \to \infty} \phi_n^{-1} (P(Z \in D_n, g_n(X, X_1, \cdots, Z_{X_n}) \neq Z_X, X \neq 0) \\
&\qquad + P(Z \in D_n, g_n(X, X_1, \cdots, Z_{X_n}) \\
&\qquad\qquad \neq W, X = 0) - R^* P(Z \in D_n)) \\
&\quad \geq \limsup_{n \to \infty} \phi_n^{-1} \Bigg(\tfrac{1}{2} \sum_{i=1}^{\infty} p_i(1-p_i)^n (1 - 2P(Z \in D_n^c)) \\
&\qquad + P(Z \in D_n)(\tfrac{1}{2} P(X = 0) - R^*)\Bigg) \\
&\quad = \limsup_{n \to \infty} \phi_n^{-1} \left(\tfrac{1}{2} \sum_{i=1}^{\infty} p_i(1-p_i)^n\right) (1 - O(1)) \\
&\quad \geq \tfrac{1}{2}.
\end{aligned}
$$

This is the contradiction that we sought.

References

[1] T. Cover, "Rates of convergence for nearest neighbor procedures," in *Proc. Hawaii Int. Conf. on Syst. Sci.*, Honolulu, HI, 1968, pp. 413-415.

[2] L. Devroye, "A universal k-nearest neighbor procedure in discrimination," in *Proc. 1978 IEEE Comput. Soc. Conf. on Pattern Recog. and Image Processing*, Chicago, IL, 1978, pp. 142-147.

[3] L. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Statist.*, vol. 8, pp. 231–239, 1980.
[4] L. Devroye and G. L. Wise, "Consistency of a recursive nearest neighbor regression function estimate," *J. Multivariate Anal.*, vol. 10, pp. 539–550, 1980.
[5] L. Gordon and R. A. Olshen, "Asymptotically efficient solutions to the classification problem," *Ann. Statist.*, vol. 6, pp. 515–533, 1978.
[6] L. Gyorfi, "Recent results on nonparametric regression function estimate and multiple sclassification," in *Problems of Control and Information Theory*, 1981.
[7] R. A. Olshen, "Comment on a paper by C. J. Stone," *Ann. Statist.*, vol. 5, pp. 632–633, 1977.
[8] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Ann. Statist.*, vol. 8, pp. 240–246, 1980.
[9] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 8, pp. 595–645, 1977.

Luc Devroye was born in Tienen, Belgium, on August 6, 1948. He received the Ph.D. degree from the University of Texas, Austin, in 1976.

In 1977 he became an Assistant Professor at the School of Computer Science, McGill University, Montreal, P.Q., Canada. He is interested in various applications of probability theory and mathematical statistics such as nonparametric estimation, probabilistic algorithms, the computer generation of random numbers, and the strong convergence of random processes.

A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise

VICTOR S. FROST, STUDENT MEMBER, IEEE, JOSEPHINE ABBOTT STILES, STUDENT MEMBER, IEEE, K. S. SHANMUGAN, SENIOR MEMBER, IEEE, AND JULIAN C. HOLTZMAN, MEMBER, IEEE

Abstract—Standard image processing techniques which are used to enhance noncoherent optically produced images are not applicable to radar images due to the coherent nature of the radar imaging process. A model for the radar imaging process is derived in this paper and a method for smoothing noisy radar images is also presented.

The imaging model shows that the radar image is corrupted by multiplicative noise. The model leads to the functional form of an optimum (minimum MSE) filter for smoothing radar images. By using locally estimated parameter values the filter is made adaptive so that it provides minimum MSE estimates inside homogeneous areas of an image while preserving the edge structure. It is shown that the filter can be easily implemented in the spatial domain and is computationally efficient. The performance of the adaptive filter is compared (qualitatively and quantitatively) with several standard filters using real and simulated radar images.

Index Terms—Adaptive filtering, image enhancement, minimum mean square error (MMSE), multiplicative noise, radar image modeling, radar image processing, speckle reduction, synthetic aperture radar (SAR).

I. Introduction

A LARGE number of image restoration and enhancement techniques have been proposed in recent years for removing a variety of degradations in recorded images of objects and scenes. These degradations result from the nonideal nature of practical imaging systems. The design of optimum image restoration and enhancement techniques requires a mathematical model of the imaging process. This paper presents a model for the noise in radar images and uses the model to develop an adaptive algorithm to smooth noisy nonstationary images.

Imaging radars, specifically the synthetic aperture radar (SAR), are beginning to make use of the digital techniques, and digitally correlated SAR images are now becoming available. However, optimum techniques for digitally processing radar images are not fully developed due to a lack of understanding of the properties of radar images from a digital image processing perspective. Thus, there is an important need for developing statistical models for radar noise and for using them in deriving appropriate algorithms for processing radar images.

This paper presents a model and a model-based image enhancement technique which is specifically designed for active microwave sensors utilizing coherent imaging techniques. The model portrays the observed radar image as corrupted by multiplicative-convolved noise. That is, the desired information, the terrain backscatter, is multiplied by a stationary random process which represents the effects of coherent fading [1]–[3]. The product signal is then processed (convolved) with the point spread function of the radar system to produce the observed image.

This model can be applied to the design of digital image enhancement algorithms through several approaches. The procedure used here was to develop a minimum mean square error (MMSE) filter to estimate the terrain backscatter from the

Manuscript received September 8, 1980; revised September 21, 1981. This work was supported by the California Institute of Technology President's Fund, NASA under Contract NAS 7-100, and the U.S. Army Research Office under Contract DAAG29-77-G-0075.

The authors are with the Remote Sensing Laboratory, Center for Research, Inc., University of Kansas, Lawrence, KS 66045.

0162-8828/82/0300-0157$00.75