simplifies the computations, up to $k = 18$. These are compared with values obtained from the approximation (36) in the following table:

k	$\ln \lambda_k$	$\ln 1/\sqrt{\pi k}$
2	−1.39	−0.92
4	−1.76	−1.27
6	−1.97	−1.47
8	−2.11	−1.61
10	−2.23	−1.72
12	−2.32	−1.80
14	−2.40	−1.87
16	−2.46	−1.96
18	−2.52	−2.02

An inspection of this table shows that the estimated value of $\ln \lambda_k$ is consistently too large by an amount very close to $\frac{1}{2}$. It is very easy to explain this constant discrepancy. As is well known, in the limit of large l, the binomial distribution has many properties in common with a Gaussian distribution with the same mean $(l\epsilon)$ and variance $(\sigma^2 = l\epsilon(1 - \epsilon))$. Hence the negative entropy of the binomial distribution goes over to that of a Gaussian:

$$f(l; \epsilon) \to -\ln \sqrt{2\pi\sigma^2 e}$$

$$\to \ln \frac{1}{\sqrt{2\pi(1 - \epsilon)}} + \ln \frac{1}{\sqrt{l\epsilon}} - \frac{1}{2},$$

which is just the amount one-half smaller than the estimate (35), and therefore the estimate of $\ln \lambda_k$ in (36) is also too large by this amount. This serves to establish that the discrepancy between the estimate of $\ln \lambda_k$ and the correct value does indeed approach a constant, and the relative difference tends to zero for $k \to \infty$. In the latter sense, (26) is verified. The only modification one might wish to make is that the constant $A/\alpha^{1/2}$ in (26) becomes $\ln [1/\sqrt{2\pi(1 - \epsilon)e}]$ instead of $\ln [1/\sqrt{2\pi(1 - \epsilon)}]$. This constant plays no essential role in subsequent developments of the expression for channel capacity for very large average received signal energy.

Acknowledgment

The author acknowledges several useful discussions with J. A. Cochran, J. P. Gordon, and I. Jacobs.

References

[1] T. E. Stern, *1960 IRE Nat'l Conv. Rec.*, p. 182.
[2] J. P. Gordon, "Quantum effects in communications systems," *Proc. IRE*, vol. 50, pp. 1898–1908, September 1962.
[3] K. Shimoda, H. Takahasi, and C. H. Townes, *J. Phys. Soc. Japan*, vol. 12, p. 686, 1957.
[4] L. P. Bolgiano, Jr., and L. F. Jelsma, "Communication channel model of a photoelectric detector," *Proc. IEEE (Correspondence)*, vol. 52, pp. 218–219, February 1964.
[5] D. E. McCumber, *Phys. Rev.*, vol. 141, p. 306, 1966.

Estimation by the Nearest Neighbor Rule

THOMAS M. COVER, MEMBER, IEEE

Abstract—Let R^* denote the Bayes risk (minimum expected loss) for the problem of estimating $\theta \in \Theta$, given an observed random variable x, joint probability distribution $F(x,\theta)$, and loss function L. Consider the problem in which the only knowledge of F is that which can be inferred from samples $(x_1,\theta_1),(x_2,\theta_2),\ldots,(x_n,\theta_n)$, where the (x_i,θ_i)'s are independently identically distributed according to F. Let the nearest neighbor estimate of the parameter θ associated with an observation x be defined to be the parameter θ_n' associated with the nearest neighbor x_n' to x. Let R be the large sample risk of the nearest neighbor rule.

It will be shown, for a wide range of probability distributions, that $R \leq 2R^*$ for metric loss functions and $R = 2R^*$ for squared-error loss functions. A simple estimator using the nearest k neighbors yields $R = R^*(1 + 1/k)$ in the squared-error loss case. In this sense, it can be said that at least half the information in the infinite training set is contained in the nearest neighbor.

This paper is an extension of earlier work[4] from the problem of classification by the nearest neighbor rule to that of estimation. However, the unbounded loss functions in the estimation problem introduce additional problems concerning the convergence of the unconditional risk. Thus some work is devoted to the investigation of natural conditions on the underlying distribution assuring the desired convergence.

Manuscript received August 12, 1966; revised June 30, 1967. This work was supported under USAF Contract AF 49(638)1517.
The author is with the Dept. of Elec. Engrg., Stanford University, Stanford, Calif.

I. Introduction

LET (x, θ) and $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$ be a collection of $n + 1$ independent identically distributed random variables taking values in $X \times \Theta$, where the observation space X is a metric space with metric ρ, and Θ is an abstract parameter space. The nearest neighbor (NN) estimate of θ on the basis of the knowledge of x and the representative samples $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$ is defined to be θ_n', the parameter associated with x_n', the nearest neighbor to x. Thus $x_n' \in \{x_1, x_2, \cdots, x_n\}$, and $\rho(x_n', x) = \min \rho(x_i, x)$, where the minimum is taken over $i = 1, 2, \cdots, n$. (In case of a tie, we arbitrarily let x_n' be the x_i of lowest index.)

The first analysis of a decision rule of the nearest neighbor type was made in a series of two papers by Fix

and Hodges[1],[2] in which the category of an observation x was decided on the basis of a majority vote of the nearest k neighbors. (It was assumed throughout that the parameter space Θ was finite.) They were able to demonstrate conditions on the underlying probability distributions such that when k and n tend to infinity in such a manner that $k/n \to 0$, the risk of such a rule approaches the Bayes risk. Johns[3] subsequently showed consistency of such a procedure in a slightly different context. Because in many cases the number of samples n is small, it is of interest to know the behavior of decision procedures which are based on a small number of nearest neighbors. This analysis was made by Cover and Hart[4] under slightly more general conditions than those of Fix and Hodges and indicates, perhaps surprisingly, that the large sample risk of a classification procedure using a *single* nearest neighbor is less than twice the Bayes risk. A larger number of nearest neighbors may, of course, be used, with an improvement in the large sample behavior at the expense of the small sample behavior. But there exists no rule, nonparametric or otherwise, for which the large sample risk is decreased by more than a factor of two.

All of the previously cited work has been concerned exclusively with the finite-action (classification) problem in which the statistician must place the observed random variable x into one of a finite number of categories. The purpose of this paper is to investigate the behavior of the nearest neighbor rule in the infinite-action (estimation) problem. This examination involves many new problems due to the more complicated loss structure necessary for an infinite parameter space Θ. However, under suitable conditions we find, almost by coincidence—since the methods of proof are different—that the nearest neighbor risk is still bounded by twice the Bayes risk for both metric and squared-error loss functions.

Perhaps we should make clear that our goal here is not to establish the power of the nearest neighbor estimator among the family of nonparametric estimators, but rather to give explicit bounds on the behavior of the risk of a rule which must be judged, by almost any standard, to be very simple, and thus to provide a reference with which other more sophisticated procedures may be compared.

II. Preliminaries

We shall need the following lemma, proved in earlier work[4],[1]

Lemma 1: Let x and $x_1, x_2, \cdots$ be a sequence of independent, identically distributed random variables taking values in a separable metric space, and let x_n' be the nearest neighbor to x among $\{x_1, x_2, \cdots, x_n\}$. Then $x_n' \to x$ with probability 1.

Let Θ be an abstract parameter space and X a separable metric space. The loss function L, defined on $\Theta \times \Theta$, assigns loss $L(\theta, \hat{\theta})$ to an estimate $\hat{\theta}$ when indeed the true parameter value is θ. An estimator $\hat{\theta} : X \to \Theta$ yields, for each $x \in X$, a conditional risk $E\{L(\theta, \hat{\theta}(x)) \mid x\}$, where the expectation is taken over θ conditioned on x. The estimator minimizing this risk at each x is termed the Bayes estimator θ^*. The resulting conditional and unconditional Bayes risks $r^*(x)$ and R^* are given by

$$r^*(x) = \underset{\theta}{E} \{L(\theta, \theta^*(x) \mid x\} \leq \underset{\theta}{E} \{L(\theta, \hat{\theta}(x)) \mid x\} \tag{1}$$

and

$$R^* = \underset{x}{E}\, r^*(x), \tag{2}$$

where, to be concrete, we note that

$$\underset{\theta}{E} \{L(\theta, \theta^*(x)) \mid x\} = \int_\Theta L(\theta, \theta^*(x)) f(\theta \mid x)\, d\theta \tag{3}$$

and

$$\begin{aligned} \underset{x}{E}\, r^*(x) &= \int_X r^*(x) f(x)\, dx \\ &= \int_\Theta \int_X L(\theta, \theta^*(x)) f(\theta, x)\, d\theta\, dx \end{aligned} \tag{4}$$

in those cases where probability densities $f(x, \theta)$, $f(\theta \mid x)$, $f(x)$ exist.

If $x_n' \in \{x_1, x_2, \cdots, x_n\}$ is the nearest neighbor to x and if θ_n' is the associated parameter of x_n', then the NN estimate θ_n' incurs a loss $L(\theta, \theta_n')$ when θ is the true parameter associated with x. Recall from Section I that (x_i, θ_i), $i = 1, 2, \cdots, n$ is a sequence of mutually independent random variables, each independent of (x, θ), such that for each i the joint distribution of x_i and θ_i is the same as the (unknown) joint distribution of x and θ. We define the conditional n-sample NN risks

$$r_n(x, x_n') = \underset{\theta, \theta_n'}{E} \{L(\theta, \theta_n') \mid x, x_n'\} \tag{5}$$

and

$$r_n(x) = \underset{x_n'}{E}\, r_n(x, x_n'). \tag{6}$$

Thus $r_n(x, x_n')$ is the expected loss in estimating θ by θ_n' when x and x_n' are the observation and nearest neighbor, respectively; and $r_n(x)$ is the n-sample NN risk when x is observed. The asymptotic conditional NN risk is given by

$$r(x) = \lim_{n \to \infty} r_n(x). \tag{7}$$

The unconditional n-sample NN risk R_n is then defined by

$$R_n = \underset{x}{E}\, r_n(x) = \underset{\theta, \theta_n'}{E} L(\theta, \theta_n') \tag{8}$$

and the large-sample NN risk R is defined by

$$R = \lim_{n \to \infty} R_n = \lim_{n \to \infty} \underset{\theta, \theta_n'}{E} L(\theta, \theta_n'). \tag{9}$$

Unfortunately, although we easily obtain bounds on the conditional risk $r(x)$ under mild conditions, certain additional constraints are required on the underlying dis-

[1] We take this opportunity to correct a misprint in the earlier paper by Cover and Hart.[4] The phrase, "since $d(x_k, x)$ is monotonically decreasing in k," which appears immediately below (9), p. 23, should read "since $d(x_k', x)$ is monotonically decreasing in k." Then the convergence of $d(x_k', x)$ to zero in probability implies convergence with probability one because of the monotonicity of $d(x_k', x)$.

tributions (allowing interchange of limit and expectation) in order to conclude that

$$R = \lim_{n\to\infty} E\, L(\theta, \theta_n') = E \lim_{n\to\infty} L(\theta, \theta_n') = E\, r(x) \tag{10}$$

and hence to find bounds on the unconditional nearest neighbor risk R. Finding such natural constraints is one of the primary tasks of this paper.

III. Metric Loss Function

We will term L a *metric loss function* if L is a metric on $\Theta \times \Theta$. Thus L must satisfy the following four conditions:

$$L(\theta_1, \theta_2) + L(\theta_1, \theta_3) \geq L(\theta_2, \theta_3), \quad \theta_1, \theta_2, \theta_3 \in \Theta \tag{11}$$

$$L(\theta_1, \theta_2) \geq 0, \quad \text{all} \quad \theta_1, \theta_2 \tag{12}$$

$$L(\theta_1, \theta_2) = L(\theta_2, \theta_1), \quad \text{all} \quad \theta_1, \theta_2 \tag{13}$$

$$L(\theta_1, \theta_2) = 0 \quad \text{if and only if} \quad \theta_1 = \theta_2. \tag{14}$$

[This definition of a metric is traditional but redundant—for example, (11) implies (12).] For the following theorem, we shall need only the requirement that L satisfies the triangle inequality (11) and that L is symmetric (13).

Remarks: Any norm $\|\cdot\|$ on Θ defines a metric loss under the definition $L(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|$. However, the important squared-error loss function $\|\theta_1 - \theta_2\|^2$ does not, of course, satisfy the triangle inequality and must be treated separately.

Recall that a function f of a random variable x is said to be continuous with probability 1 (continuous wp1) if, with probability 1, x is a point of continuity of f.

Theorem 1: Let L be a metric loss function such that for every $\theta_0 \in \Theta$, $E_\theta\{L(\theta, \theta_0) \mid x\}$ is a continuous function of x wp1. Then

$$r^*(x) \leq r(x) \leq 2r^*(x), \text{ wp1.} \tag{15}$$

That is, the conditional NN risk is less than twice the conditional Bayes risk wp1.

Proof of Theorem 1: As before, let (x, θ), (x_1, θ_1), (x_2, θ_2) $\cdots \in X \times \Theta$ be a sequence of independent identically distributed random variables defined on $(\Omega, \mathfrak{F}, P)$. By the triangle inequality and the symmetry of L

$$L(\theta, \theta_n') \leq L(\theta, \theta^*) + L(\theta^*, \theta_n'). \tag{16}$$

Conditioning on x and x_n', we have

$$\begin{aligned} &r_n(x, x_n') \\ &\leq E_\theta\{L(\theta, \theta^*(x)) \mid x, x_n'\} + E_{\theta_n'}\{L(\theta^*(x), \theta_n') \mid x, x_n'\} \\ &= E_\theta\{L(\theta, \theta^*(x)) \mid x\} + E_{\theta_n'}\{L(\theta^*(x), \theta_n') \mid x_n'\}, \end{aligned} \tag{17}$$

where we have used the conditional independence of θ (given x) and θ_n' (given x_n'). By Lemma 1 and the assumed continuity properties of L, we see that $x_n' \to x$ wp1, and

$$E_{\theta_n'}\{L(\theta^*(x), \theta_n') \mid x_n'\} \to E_\theta\{L(\theta^*(x), \theta) \mid x\} = r^*(x), \quad \text{wp1.} \tag{18}$$

Hence, by (1) and (17)

$$\lim_{n\to\infty} r_n(x, x_n') \leq 2r^*(x) \tag{19}$$

and

$$r(x) \leq 2r^*(x), \text{ wp1,} \tag{20}$$

as was to be shown.

By the dominated convergence theorem we obtain the following corollary.

Corollary 1 of Theorem 1: If L is bounded, then

$$R^* \leq R \leq 2R^*. \tag{21}$$

In search of natural conditions under which $R = Er$, we are motivated to make the following definition.

Definition 1: Let X be a metric space with metric ρ, let $(\Omega, \mathfrak{F}, P)$ be a probability space, and let $x, x' \in X$ be independent random variables defined on $(\Omega, \mathfrak{F}, P)$. Then $E_{x,x'}\rho^r(x, x')$ is defined to be the rth moment of the space X with respect to P.

Remarks: Note that the space X has finite rth moment if and only if there exists $x_0 \in X$ such that $E\rho^r(x, x_0) < \infty$. Of course, if X is bounded, then X has finite rth moment for all $r \geq 0$.

Let $(x_1, \theta_1), (x_2, \theta_2) \in X \times \Theta$ be independent, identically distributed random variables defined on $(\Omega, \mathfrak{F}, P)$.

Corollary 2 of Theorem 1: If X has finite first moment, and if the conditional distributions of θ_1 (given x_1) and θ_2 (given x_2) are close in the sense that there exists a constant A such that, for every $\theta_0 \in \Theta$ and $x_1, x_2 \in X$,

$$|E_{\theta_1}\{L(\theta_1, \theta_0) \mid x_1\} - E_{\theta_2}\{L(\theta_2, \theta_0) \mid x_2\}| \leq A\rho(x_1, x_2), \tag{22}$$

then

$$R^* \leq R \leq 2R^*. \tag{23}$$

Proof of Corollary 2: From (17),

$$\begin{aligned} r_n(x, x_n') &\leq r^*(x) + E_{\theta_n'}\{L(\theta_n', \theta^*(x)) \mid x_n'\} \\ &\leq 2r^*(x) + A\rho(x, x_n'). \end{aligned} \tag{24}$$

Since the nearest of the first n samples drawn is certainly at no greater distance from x than is the first sample, we have $\rho(x, x_n') \leq \rho(x, x_1)$, wp1. But $E\rho(x, x_1) \leq \infty$ by the finite first moment of X. Thus the sequence $r_n(x, x_n')$ is dominated by the integrable random variable $2r^*(x) + A\rho(x, x_1)$ and, again by the dominated convergence theorem,

$$R = \lim_n E\, r_n(x, x_n') = E \lim_n r_n(x, x_n') \leq E\, 2r^*(x) = 2R^*. \tag{25}$$

IV. Squared-Error Loss Function

In this section let Θ be the real line (or any finite-dimensional inner-product space). We shall need the definitions of the conditional mean of θ given x

$$\mu_1(x) = E(\theta \mid x), \tag{26}$$

the conditional second moment

$$\mu_2(x) = E(\theta^2 \mid x), \tag{27}$$

and the conditional variance

$$\sigma^2(x) = \mu_2(x) - \mu_1^2(x). \tag{28}$$

As is well known in the case of the squared-error loss function $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$, the Bayes estimator θ^* is the conditional mean

$$\theta^*(x) = \mu_1(x), \tag{29}$$

and the resulting conditional Bayes risk $r^*(x)$ and Bayes risk R^* are given, respectively, by the conditional variance

$$r^*(x) = \sigma^2(x) \tag{30}$$

and the unconditional variance

$$R^* = \underset{x}{E}\, \sigma^2(x). \tag{31}$$

Theorem 2: Let $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$ and let X be a separable metric space. If $\mu_1(x)$ and $\mu_2(x)$ are continuous wp1, then

$$r = 2r^*, \text{ wp1}. \tag{32}$$

Proof of Theorem 2: Conditioning on x and x_n',

$$\begin{aligned} r_n(x, x_n') &= \underset{\theta, \theta_n'}{E} \{(\theta - \theta_n')^2 \mid x, x_n'\} \\ &= \underset{\theta}{E} \{\theta^2 \mid x, x_n'\} - 2 \underset{\theta, \theta_n'}{E} \{\theta\theta_n' \mid x, x_n'\} \\ &\qquad + \underset{\theta_n'}{E} \{\theta_n'^2 \mid x, x_n'\}. \end{aligned} \tag{33}$$

By the conditional independence of θ and θ_n' (conditioned on x and x_n'),

$$\begin{aligned} &r_n(x, x_n') \\ &= \underset{\theta}{E} \{\theta^2 \mid x\} - 2 \underset{\theta}{E} \{\theta \mid x\} \underset{\theta_n'}{E} \{\theta_n' \mid x_n'\} + \underset{\theta_n'}{E} \{\theta_n'^2 \mid x_n'\} \\ &= \mu_2(x) - 2\mu_1(x)\mu_1(x_n') + \mu_2(x_n'). \end{aligned} \tag{34}$$

Since $x_n' \to x$ wp1, and since μ_1 and μ_2 are continuous wp1,

$$r(x) = \lim_{n\to\infty} r_n(x, x_n') = 2\mu_2(x) - 2\mu_1^2(x) = 2r^*(x). \tag{35}$$

Thus $r = 2r^*$ wp1.

Remarks: The following example makes it clear that come additional conditions are required in order to find the unconditional risks. Let $\mu_1(x) = 1/x$ and $\sigma^2(x) = \sigma^2 = R^* < \infty$. Let x be drawn according to any probability density on the real line which is continuous and nonzero at the origin. Note that $\mu_1(x)$ is continuous wp1, since the point of discontinuity $x = 0$ has probability zero. In this case, the limiting conditional nearest neighbor risk $r(x)$ is $2r^*(x)$, as expected. However, $R_n = \infty$, for all n. In other words, for almost every x, the NN estimate $E\{(\theta - \theta_n')^2 \mid x\}$ converges to $2R^*$ as $n \to \infty$; but for fixed sample size n the unconditional risk $E(\theta - \theta_n')^2$ is infinite no matter how large n may be. Loosely speaking, the infinite contribution to the n-sample risk R_n occurs when x and x_n' have different signs, thus causing θ_n' to be a poor estimate of θ because of the discontinuity of $E\{\theta \mid x\}$. On the other hand, fixing x first ensures that x_n' ultimately will be close enough to x so that θ_n' will be a good estimate of θ.

Corollary 1 of Theorem 2: Let X have finite second moment M and let there exist constants A and B such that

$$|\mu_1(x_1) - \mu_1(x_2)|^2 \le A\rho^2(x_1, x_2) \tag{36}$$

and

$$|\sigma^2(x_1) - \sigma^2(x_2)| \le B\rho^2(x_1, x_2) \tag{37}$$

for all $x_1, x_2 \in X$. Then, for $L(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$,

$$R = 2R^*. \tag{38}$$

Remarks: The multivariate normal distribution satisfies these conditions with $B = 0$ and Euclidean metric ρ.

Proof of Corollary 1: Since $\theta - \mu_1(x)$ (conditioned on x) and $\theta_n' - \mu_1(x_n')$ (conditioned on x_n') are conditionally independent zero-mean random variables,

$$\begin{aligned} r_n(x, x_n') &= \underset{\theta, \theta_n'}{E} \{([\theta - \mu_1(x)] + [\mu_1(x) - \mu_1(x_n')] \\ &\qquad + [\mu_1(x_n') - \theta_n'])^2 \mid x, x_n'\} \\ &= \sigma^2(x) + (\mu_1(x) - \mu_1(x_n'))^2 + \sigma^2(x_n') \\ &\le 2\sigma^2(x) + A\rho^2(x, x_n') + B\rho^2(x, x_n') \qquad (39) \\ &\le 2\sigma^2(x) + (A + B)\rho^2(x, x_1), \quad \text{for all } n. \qquad (40) \end{aligned}$$

Since

$$\begin{aligned} &E\{2\sigma^2(x) + (A + B)\rho^2(x, x_1)\} \\ &\qquad < 2R^* + (A + B)M < \infty, \end{aligned} \tag{41}$$

we have, by the dominated convergence theorem and $\rho(x, x_n') \to 0$ wp1,

$$R = \lim_{n\to\infty} E\, r_n(x, x_n') = E \lim_{n\to\infty} r_n(x, x_n') = E\, 2\sigma^2(x) = 2R^*, \tag{42}$$

which was to be shown.

The following corollary to the proof of Theorem 2 yields bounds for the risk of the k nearest neighbor rule. In this rule, the k nearest neighbors $x_n^{[1]}, x_n^{[2]}, \cdots, x_n^{[k]}$ are inspected, where

$$(x - x_n^{[1]})^2 \le (x - x_n^{[2]})^2 \le \cdots \le (x - x_n^{[n]})^2, \tag{43}$$

and the associated parameters $\theta_n^{[1]}, \theta_n^{[2]}, \cdots, \theta_n^{[k]}$ are combined in the natural way (for the quadratic loss criterion) to give an estimate

$$\bar{\theta}_n^{(k)} = \frac{1}{k} \sum_{i=1}^{k} \theta_n^{[i]}. \tag{44}$$

Thus θ_n' becomes $\theta_n^{[1]}$ under this new definition.

Corollary 2 of Theorem 2: The k nearest neighbor rule has conditional risk $r^{(k)}(x) = (1 + 1/k)r^*(x)$ under the

assumptions of Theorem 2, and has unconditional risk $R = (1 + 1/k)R^*$ under the additional assumptions of Corollary 1 of Theorem 2.

Proof of Corollary 2: The conditional k *NN* risk is given by

$$\begin{aligned} r_n^{(k)}(x, x_1, \cdots, x_n) &= E\{(\theta - \bar{\theta}_n^{(k)})^2 \mid x, x_1, x_2, \cdots, x_n\} \\ &= E\left\{\left(\theta - \frac{1}{k}\sum_{i=1}^{k}\theta_n^{[i]}\right)^2 \,\middle|\, x, x_n^{[1]}, x_n^{[2]}, \cdots, x_n^{[k]}\right\} \\ &= E\left\{\left([\theta - \mu_1(x)] + \frac{1}{k}\sum_{i=1}^{k}[\mu_1(x) - \mu_1(x_n^{[i]})]\right.\right. \\ &\quad \left.\left. + \frac{1}{k}\sum_{i=1}^{k}[\mu_1(x_n^{[i]}) - \theta_n^{[i]}]\right)^2 \,\middle|\, x, x_n^{[1]}, \cdots, x_n^{[k]}\right\} \\ &= \left[\mu_1(x) - \frac{1}{k}\sum_{i=1}^{k}\mu_1(x_n^{[i]})\right]^2 + \sigma^2(x) + \frac{1}{k^2}\sum_{i=1}^{k}\sigma^2(x_n^{[i]}). \end{aligned} \tag{45}$$

For $i = 1, 2, \cdots, k$, $x_n^{[i]} \to x$, wp1 as $n \to \infty$, and hence

$$\mu_1(x_n^{[i]}) \to \mu_1(x) \quad \text{and} \quad \sigma^2(x_n^{[i]}) \to \sigma^2(x); \tag{46}$$

from which we obtain $r = (1 + 1/k)r^*$ wp1.

Passage to the unconditional *NN* risk $R = (1 + 1/k)R^*$ follows from (45), the finite second moment of X, the dominated convergence theorem, and the inequalities

$$\begin{aligned} \left[\mu_1(x) - \frac{1}{k}\sum_{i=1}^{k}\mu_1(x_n^{[i]})\right]^2 &\le \left[\frac{1}{k}\sum_{i=1}^{k}A^{1/2}\rho(x, x_n^{[i]})\right]^2 \\ &\le A\rho^2(x, x_1) \end{aligned} \tag{47}$$

and

$$\begin{aligned} \frac{1}{k^2}\sum_{i=1}^{k}\sigma^2(x_n^{[i]}) &\le \frac{\sigma^2(x)}{k} + \frac{B}{k^2}\sum_{i=1}^{k}\rho^2(x, x_n^{[i]}) \\ &\le \frac{1}{k}[\sigma^2(x) + B\rho^2(x, x_1)]. \end{aligned} \tag{48}$$

Extended Application: Let $\theta = \{\theta(t) : 0 \le t \le T\}$ and $x = \{x(t) : 0 \le t \le T\}$ be stochastic processes. Presumably θ is a random input signal and x is the received signal at the output of a random channel. We wish to minimize $E\{\int_0^T (\theta(t) - \hat{\theta}(t))^2\, dt\}$ over all estimates $\hat{\theta}$ depending on x. Suppose in the past n independent uses of the channel that the input–output pairs $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$ have been observed. A k *NN* procedure, using a metric

$$\rho(x_1, x_2) = \left(\int_0^T (x_1(t) - x_2(t))^2\, dt\right)^{1/2} \tag{49}$$

yields an estimate

$$\hat{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_n^{[i]} \tag{50}$$

which results in a large sample risk

$$R = (1 + 1/k)R^* \tag{51}$$

under suitable regularity conditions on the joint distribution of the random processes. We shall not develop these regularity conditions here.

V. The Jointly Gaussian Case

For (x, θ) jointly multivariate normal, an explicit formula for the n-sample nearest neighbor risk R_n and the n-sample k *NN* risk $R_n^{(k)}$ can be obtained. We shall treat only the univariate case. Let $N(\mu, \sigma^2)$ denote a univariate normal density with mean μ and variance σ^2. Let

$$f(\theta) = N(\mu_1, \sigma_1^2) \tag{52}$$

$$f(x \mid \theta) = N(\theta, \sigma_2^2) \tag{53}$$

from which we find

$$f(\theta \mid x) = N\left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}x + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right) \tag{54}$$

and

$$\theta^*(x) = \mu_1(x) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}x + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\mu_1 \tag{55}$$

and

$$R = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{56}$$

Now, by the conditional independence of the θ_i's,

$$\begin{aligned} E\{(\theta - \bar{\theta}_n^{(k)})^2 \mid x, x_1, x_2, \cdots, x_n\} &= E\left\{\left(\theta - \frac{1}{k}\sum_{i=1}^{k}\theta_n^{[i]}\right)^2 \,\middle|\, x, x_n^{[1]}, x_n^{[2]}, \cdots, x_n^{[k]}\right\} \\ &= \left(\mu_1(x) - \frac{1}{k}\sum_{i=1}^{k}\mu_1(x_n^{[i]})\right)^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} + \frac{1}{k^2}\sum_{i=1}^{k}\sigma^2(x_n^{[i]}) \\ &= \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 E\left(x - \frac{1}{k}\sum_{i=1}^{k}x_n^{[i]}\right)^2 + \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\left(1 + \frac{1}{k}\right). \end{aligned} \tag{57}$$

Hence

$$R_n^{(k)} = R^*\left(1 + \frac{1}{k}\right) + \left(\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}\right)^2 E(x - \bar{x}_n^{(k)})^2 \tag{58}$$

where

$$\bar{x}_n^{(k)} = \frac{1}{k}\sum_{i=1}^{k}x_n^{[i]} \tag{59}$$

is the sample average of the nearest k neighbors to x. Realizing that $x \sim N(\mu_1, \sigma_1^2 + \sigma_2^2)$, we can write (58) in normalized form as

$$R_n^{(k)} = \left(1 + \frac{1}{k} + \frac{\sigma_1^2}{\sigma_2^2}\delta_n^2(k)\right)R^*, \tag{60}$$

where $\delta_n^2(k)$ is defined to be the expected squared distance from an $N(0, 1)$ random variable z to the average of the k nearest neighbors of n independent identically distributed $N(0, 1)$ random variables $z_1, z_2, \cdots, z_n$.

Clearly one wishes k to be large in order that the contribution of the $1/k$ term be small, while one wants k to be small in order that $\delta_n^2(k)$ be small. This sort of trade-off is always present in problems of the nearest neighbor type. Note that, in general, $R_n^{(k)} \to R^*$ as $k \to \infty$

and $n \to \infty$ in such a manner that $k/n \to 0$. We have not attempted an investigation of the precise nature of this convergence.

We remark that if we knew that x and θ were jointly Gaussian, we could certainly use knowledge of this fact to develop a more sophisticated estimator than the *NN* estimator. However, the *NN* estimator has the advantage that $R_n^{(k)} \leq 2\sigma_1^2 < \infty$, for all n and k. Thus even a single sample yields finite risk. In contrast, consider a standard linear regression technique utilizing the fact that, in the Gaussian case, $\mu(x)$ is of the form $ax + b$ to form estimates $\hat{a}$ and $\hat{b}$ on the basis of the minimum mean-squared error line fitting the points $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$. The resulting estimator $\hat{\theta} = \hat{a}x + \hat{b}$ has infinite risk for sample sizes of $n = 1, 2, 3$. Thus the nonparametric *NN* estimator is actually better for sufficiently small sample size.

VI. Conclusions

It has been shown that the large sample risk for the *NN* decision rule is no greater than twice the Bayes risk for both the squared-error loss function and the metric loss function. In particular, it has been shown under mild continuity conditions that the conditional risk $r(x)$ of the *NN* estimation rule in the infinite sample case satisfies the inequalities

$$r(x) \leq 2r^*(x), \text{ for the metric loss case, and}$$
$$r(x) = 2r^*(x), \text{ for the squared-error loss case.}$$

Under certain additional moment conditions the unconditional risk R of the *NN* estimate satisfies

$$R \leq 2R^*, \quad \text{for metric loss,}$$
$$R = 2R^*, \quad \text{for squared-error loss, and}$$
$$R = (1 + 1/k)R^*, \quad \text{for squared-error loss with a } k \text{ NN estimate.}$$

These conclusions are complemented by those of earlier work,[4],[5] in which it is shown that $R \leq 2R^*$ for the classification problem with a probability of error loss criterion. Thus the most sophisticated decision rule, based on the entire sample set, may reduce the risk by at most a factor of two. In this sense it may be concluded that at least half the decision information in an infinite set of classified samples is contained in the nearest neighbor.

Acknowledgment

The author wishes to thank B. Efron and P. Hart for their helpful comments during the preparation of this paper.

References

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination, consistency properties," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 4, Contract AF41(128)-31, February 1951.

[2] ——, "Discriminatory analysis: small sample performance," USAF School of Aviation Medicine, Randolph Field, Tex., Project 21-49-004, Rept. 11, August 1952.

[3] M. V. Johns, "An empirical Bayes approach to non-parametric two-way classification," in *Studies in Item Analysis and Prediction*, Herbert Solomon, Ed. Stanford, Calif.: Stanford University Press, 1961.

[4] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Information Theory*, vol. IT-13, pp. 21–27, January 1967.

[5] P. E. Hart, "An asymptotic analysis of the nearest-neighbor decision rule," Stanford Electronics Labs., Stanford University, Stanford, Calif., Tech. Rept. 1828-2, May 1966.

On the Mean Accuracy of Statistical Pattern Recognizers

GORDON F. HUGHES, MEMBER, IEEE

Abstract—**The overall mean recognition probability (mean accuracy) of a pattern classifier is calculated and numerically plotted as a function of the pattern measurement complexity n and design data set size m. Utilized is the well-known probabilistic model of a two-class, discrete-measurement pattern environment (no Gaussian or statistical independence assumptions are made). The minimum-error recognition rule (Bayes) is used, with the unknown pattern environment probabilities estimated from the data relative frequencies. In calculating the mean accuracy over all such environments, only three parameters remain in the final equation: n, m, and the prior probability p_c of either of the pattern classes.**

With a fixed design pattern sample, recognition accuracy can first increase as the number of measurements made on a pattern increases, but decay with measurement complexity higher than some optimum value. Graphs of the mean accuracy exhibit both an optimal and a maximum acceptable value of n for fixed m and p_c. A four-place tabulation of the optimum n and maximum mean accuracy values is given for equally likely classes and m ranging from 2 to 1000.

The penalty exacted for the generality of the analysis is the use of the mean accuracy itself as a recognizer optimality criterion. Namely, one necessarily always has some particular recognition problem at hand whose Bayes accuracy will be higher or lower than the mean over all recognition problems having fixed n, m, and p_c.

Manuscript received November 3, 1966; revised July 19, 1967. This work was supported in part by RADC under Contract AF 30(602)-3976.

The author is with Autonetics, a division of North American Rockwell, Inc., Anaheim, Calif. 92803

I. Introduction

SOME consequences of the statistical model of pattern recognition will be presented.[1]-[5] It will be shown that certain useful numerical conclusions can be drawn from rather few assumptions. Basically, the only