ACKNOWLEDGMENT

The author is grateful to Prof. S. J. Mason of M.I.T. for his interest in this work, and for his many helpful suggestions. The author also wishes to thank Prof. K. N. Stevens and Prof. M. Eden for some very helpful discussions, and Prof. D. E. Troxel for his help in designing the sensory display.

REFERENCES

[1] I. Pollack, "The information of elementary auditory displays," *J. Acoust. Soc. Am.*, vol. 24, pp. 745–749, November 1952.
[2] ——, "The information of elementary auditory displays II," *J. Acoust. Soc. Am.*, vol. 25, pp. 765–769, July 1953.
[3] I. Pollack and L. Ficks, "Information of multidimensional auditory displays," *J. Acoust. Soc. Am.*, vol. 26, pp. 155–158, March 1954.
[4] W. R. Garner, "An informational analysis of absolute judgements of loudness," *J. Exper. Psychol.*, vol. 46, pp. 373–380, November 1953.
[5] C. W. Ericsen, "Multidimensional stimulus differences and the accuracy of discrimination," Wright Air Development Center, Wright-Patterson AFB, Dayton, Ohio, Tech. Rept., 1954.
[6] N. S. Andersen, and P. S. Fitts, "Amount of information gained during brief exposures to numerals and colors," *J. Exper. Psychol.*, vol. 56, pp. 362–369, October 1958.
[7] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, pp. 81–97, March 1956.
[8] F. Attneave, *Applications of Information Theory to Psychology.* New York: Holt, 1959.
[9] D. E. Troxel, "Comparison of tactile and visual reading rates," M.I.T. Electronics Research Lab., Cambridge, Mass. Quart. Progress Rept. 67, pp. 267–272, October 1962.
[10] R. W. Donaldson, "Approximate formulas for the information transmitted by a discrete communication channel," M.I.T. Electronics Research Lab., Cambridge, Mass. Quart. Progress Rept. 77, pp. 335–342, April 1965.
[11] A. de Saint-Exupery, *The Little Prince.* New York: Harcourt, Brace, and World, 1943.
[12] F. A. Geldard, *Cutaneous Channels of Communication*, W. A. Rosenblith, Ed. New York: Wiley 1961.

Nearest Neighbor Pattern Classification

T. M. COVER, MEMBER, IEEE, AND P. E. HART, MEMBER, IEEE

Abstract—**The nearest neighbor decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. This rule is independent of the underlying joint distribution on the sample points and their classifications, and hence the probability of error R of such a rule must be at least as great as the Bayes probability of error R^*—the minimum probability of error over all decision rules taking underlying probability structure into account. However, in a large sample analysis, we will show in the M-category case that $R^* \leq R \leq R^*(2 - MR^*/(M-1))$, where these bounds are the tightest possible, for all suitably smooth underlying distributions. Thus for any number of categories, the probability of error of the nearest neighbor rule is bounded above by twice the Bayes probability of error. In this sense, it may be said that half the classification information in an infinite sample set is contained in the nearest neighbor.**

I. INTRODUCTION

IN THE CLASSIFICATION problem there are two extremes of knowledge which the statistician may possess. Either he may have complete statistical knowledge of the underlying joint distribution of the observation x and the true category θ, or he may have no knowledge of the underlying distribution except that which can be inferred from samples. In the first extreme, a standard Bayes analysis will yield an optimal decision procedure and the corresponding minimum (Bayes) probability of error of classification R^*. In the other extreme, a decision to classify x into category θ is allowed to depend only on a collection of n correctly classified samples $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$, and the decision procedure is by no means clear. This problem is in the domain of nonparametric statistics and no optimal classification procedure exists with respect to all underlying statistics.

If it is assumed that the classified samples (x_i, θ_i) are independently identically distributed according to the distribution of (x, θ), certain heuristic arguments may be made about good decision procedures. For example, it is reasonable to assume that observations which are close together (in some appropriate metric) will have the same classification, or at least will have almost the same posterior probability distributions on their respective classifications. Thus to classify the unknown sample x we may wish to weight the evidence of the nearby x_i's most heavily. Perhaps the simplest nonparametric decision procedure of this form is the *nearest neighbor* (NN) rule, which classifies x in the category of its nearest neighbor. Surprisingly, it will be shown that, in the large sample case, this simple rule has a probability of error which

Manuscript received February 23, 1966; revised April 29, 1966. This work has been supported at Stanford University by U. S. Army Electronics Command under Contract DA28-043-AMC-01764(E) and by USAF under Contract AF49(638)1517; and at the Stanford Research Institute, Menlo Park, Calif., by RADC under Contract AF30(602)-3945.

T. M. Cover is with the Department of Electrical Engineering, Stanford University, Stanford, Calif.
P. E. Hart is with the Stanford Research Institute, Menlo Park, Calif.

is less than twice the Bayes probability of error, and hence is less than twice the probability of error of any other decision rule, nonparametric or otherwise, based on the infinite sample set.

The first formulation of a rule of the nearest neighbor type and primary previous contribution to the analysis of its properties, appears to have been made by Fix and Hodges [1] and [2]. They investigated a rule which might be called the k_n-nearest neighbor rule. It assigns to an unclassified point the class most heavily represented among its k_n nearest neighbors. Fix and Hodges established the consistency of this rule for sequences $k_n \to \infty$ such that $k_n/n \to 0$. In reference [2], they investigate numerically the small sample performance of the k_n-NN rule under the assumption of normal statistics.

The NN rule has been used by Johns [3] as an example of an empirical Bayes rule. Kanal [4], Sebestyen [5] (who calls it the proximity algorithm), and Nilsson [6] have mentioned the intuitive appeal of the NN rule and suggested its use in the pattern recognition problem. Loftsgaarden and Quesenberry [7] have shown that a simple modification of the k_n-NN rule gives a consistent estimate of a probability density function. In the above mentioned papers, no analytical results in the nonparametric case were obtained either for the finite sample size problem or for the finite number of nearest neighbors problem.

In this paper we shall show that, for any number n of samples, the single-NN rule has strictly lower probability of error than any other k_n-NN rule against certain classes of distributions, and hence is admissible among the k_n-NN rules. We will then establish the extent to which "samples which are close together have categories which are close together" and use this to compare in Section VI the probability of error of the NN-rule with the minimum possible probability of error.

II. The Nearest Neighbor Rule

A set of n pairs $(x_1, \theta_2), \cdots, (x_n, \theta_n)$ is given, where the x_i's take values in a metric space X upon which is defined a metric d, and the θ_i's take values in the set $\{1, 2, \cdots, M\}$. Each θ_i is considered to be the index of the category to which the ith individual belongs, and each x_i is the outcome of the set of measurements made upon that individual. For brevity, we shall frequently say "x_i belongs to θ_i" when we mean precisely that the ith individual, upon which measurements x_i have been observed, belongs to category θ_i.

A new pair (x, θ) is given, where only the measurement x is observable by the statistician, and it is desired to estimate θ by utilizing the information contained in the set of correctly classified points. We shall call

$$x'_n \in \{x_1, x_2, \cdots, x_n\}$$

a nearest neighbor to x if

$$\min d(x_i, x) = d(x'_n, x) \quad i = 1, 2, \cdots, n. \tag{1}$$

The nearest neighbor rule decides x belongs to the category θ'_n of its nearest[1] neighbor x'_n. A mistake is made if $\theta'_n \neq \theta$. Notice that the NN rule utilizes only the classification of the nearest neighbor. The $n - 1$ remaining classifications θ_i are ignored.

III. Admissibility of Nearest Neighbor Rule

If the number of samples is large it makes good sense to use, instead of the single nearest neighbor, the majority vote of the nearest k neighbors. We wish k to be large in order to minimize the probability of a non-Bayes decision for the unclassified point x, but we wish k to be small (in proportion to the number of samples) in order that the points be close enough to x to give an accurate estimate of the posterior probabilities of the true class of x.

The purpose of this section is to show that, among the class of k-NN rules, the single nearest neighbor rule (1-NN) is admissible. That is, for the n-sample problem, there exists no k-NN rule, $k \neq 1$, which has lower probability of error against all distributions. We shall show that the single NN rule is undominated by exhibiting a simple distribution for which it has strictly lower probability of error P_e. The example to be given comes from the family of distributions for which simple decision boundaries provide complete separation of the samples into their respective categories. Fortunately, one example will serve for all n.

Consider the two category problem in which the prior probabilities $\eta_1 = \eta_2 = \frac{1}{2}$, and the conditional density f_1 is uniform on the unit disk D_1 centered at $(-3, 0)$, and the conditional density f_2 is uniform on the unit disk D_2 centered at $(3, 0)$ as shown in Fig. 1. In the n-sample problem, the probability that j individuals come from category 1, and hence have measurements lying in D_1, is $(\frac{1}{2})^n\binom{n}{j}$. Without loss of generality, assume that the unclassified x lies in category 1. Then the NN rule will make a classification error only if the nearest neighbor x'_n belongs to category 2, and thus, necessarily, lies in D_2. But, from inspection of the distance relationships, if the nearest neighbor to x is in D_2, then each of the x_i must lie in D_2. Thus the probability $P_e(1; n)$ of error of the NN rule in this case is precisely $(\frac{1}{2})^n$—the probability that $x_1, x_2, \cdots, x_n$ all lie in D_2. Let $k = 2k_0 + 1$. Then the k-NN rule makes an error if k_0 or fewer points lie in D_1. This occurs with probability

$$P_e(k; n) = (\tfrac{1}{2})^n \sum_{j=0}^{k_0} \binom{n}{j}. \tag{2}$$

Thus in this example, the 1-NN rule has strictly lower P_e than does any k-NN rule, $k \neq 1$, and hence is admissible in that class. Indeed

[1] In case of ties for the nearest neighbor, the rule may be modified to decide the most popular category among the ties. However, in those cases in which ties occur with nonzero probability, our results are trivially true.

$$P_e(k; n) \uparrow \tfrac{1}{2} \quad \text{in} \quad k, \quad \text{for any} \quad n,$$
$$P_e(k; n) \downarrow 0 \quad \text{in} \quad n, \quad \text{for any} \quad k > 0, \tag{3}$$

and

$$P_e(k_n; n) \to 0, \quad \text{if} \quad 0 < \frac{k_n}{n} \leq \alpha < 1, \quad \text{for all} \quad n.$$

In general, then, the 1-NN rule is strictly better than the $k \neq 1$-NN rule in those cases where the supports of the densities $f_1, f_2, \cdots, f_M$ are such that each in-class distance is greater than any between-class distance.

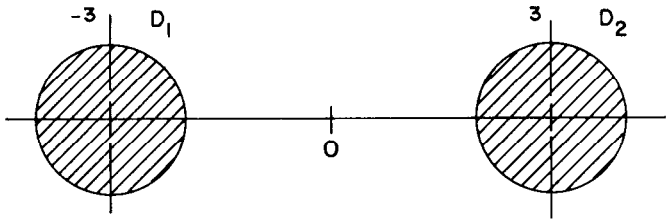

Fig. 1. Admissibility of nearest neighbor rule.

IV. Bayes Procedure

In this section we shall present the simplest version of the Bayes decision procedure for minimizing the probability of error in classifying a given observation x into one of M categories. All the statistics will be assumed known. Bear in mind, however, that the NN rule is nonparametric, or distribution free, in the sense that it does not depend on any assumptions about the underlying statistics for its application. The Bayes risk serves merely as a reference—the limit of excellence beyond which it is not possible to go.

Let x denote the measurements on an individual and X the sample space of possible values of x. We shall refer to x as the observation. On the basis of x a decision must be made about the membership of the individual in one of M specified categories.

For the purposes of defining the Bayes risk, we assume $f_1(x), f_2(x), \cdots, f_M(x)$, probability densities at x with respect to a σ-finite measure ν, such that an individual in category i gives rise to an observation x according to density f_i. Let $L(i, j)$ be the loss incurred by assigning an individual from category i to category j.

Let $\eta_1, \eta_2, \cdots, \eta_M, \eta_i \geq 0, \sum \eta_i = 1$, be the *prior* probabilities of the M categories. The conditional probability $\hat{\eta}_i(x)$ of an individual with measurements x belonging to category i is, by the Bayes theorem,

$$\hat{\eta}_i = \frac{\eta_i f_i}{\sum \eta_i f_i}, \qquad i = 1, 2, \cdots, M. \tag{4}$$

Thus the random variable x transforms the prior probability vector η into the *posterior* probability vector $\hat{\eta}(x)$. If the statistician decides to place an individual with measurements x into category j, the conditional loss is

$$r_j(x) = \sum_{i=1}^{M} \hat{\eta}_i(x) L(i, j). \tag{5}$$

For a given x the conditional loss is minimum when the individual is assigned to the category j for which $r_j(x)$ is lowest. Minimizing the conditional expected loss obviously minimizes the unconditional expected loss. Thus the minimizing decision rule δ^*, called the Bayes decision rule with respect to η, is given by deciding the category j for which r_j is lowest. Using δ^*, the conditional Bayes risk $r^*(x)$ is

$$r^*(x) = \min_j \left\{ \sum_{i=1}^{M} \hat{\eta}_i(x) L(i, j) \right\}, \tag{6}$$

and the resulting overall minimum expected risk R^*, called the Bayes risk, is given by

$$R^* = E r^*(x), \tag{7}$$

where the expectation is with respect to the compound density

$$f(x) = \sum_{i=1}^{M} \eta_i f_i(x). \tag{8}$$

V. Convergence of Nearest Neighbors

Most of the properties of the NN rules hinge on the assumption that the conditional distributions of θ_n' and θ approach one another when $x_n' \to x$. In order to put bounds on the NN risk for as wide a class of underlying statistics as possible, it will be necessary to determine the weakest possible conditions on the statistics which guarantee the above convergence.

Lemma (Convergence of the Nearest Neighbor)

Let x and $x_1, x_2, \cdots$ be independent identically distributed random variables taking values in a separable metric space X. Let x_n' denote the nearest neighbor to x from the set $\{x_1, x_2, \cdots, x_n\}$. Then $x_n' \to x$ with probability one.

Remark: In particular, $x_n' \to x$ with probability one for any probability measure in Euclidean n-space. We prove the lemma in this generality in order to include in its coverage such standard pathological candidates for counterexamples as the Cantor ternary distribution function defined on X the real line.

Since the convergence of the nearest neighbor to x is independent of the metric, the bounds on the risks of the NN rule will be independent of the metric on X.

Proof: Let $S_x(r)$ be the sphere $\{\tilde{x} \in X: d(x, \tilde{x}) \leq r\}$ of radius r centered at x, where d is the metric defined on X.

Consider first a point $x \in X$ having the property that every sphere $S_x(r), r > 0$, has nonzero probability measure. Then, for any $\delta > 0$,

$$P\{\min_{k=1,2,\cdots,n} d(x_k, x) \geq \delta\} = (1 - P(S_x(\delta))^n \to 0 \tag{9}$$

and therefore, since $d(x_k, x)$ is monotonically decreasing in k, the nearest neighbor to x converges to x with probability one.

It remains to argue that the random variable x has this property with probability one. We shall do so by proving that the set N of points failing to have this property has probability measure zero. Accordingly, let N be the set of all x for which there exists some r_x sufficiently small that $P(S_x(r_x)) = 0$.

By the definition of the separability of X, there exists a countable dense subset A of X. For each $x \in N$ there exists, by the denseness of A, a_x in A for which $a_x \in S_x(r_x/3)$. Thus, there exists a small sphere $S_{a_x}(r_x/2)$ which is strictly contained in the original sphere $S_x(r_x)$ and which contains x. Thus $P(S_{a_x}(r_x/2)) = 0$. Then the possibly uncountable set N is contained in the countable union (by the countability of A) of spheres $\bigcup_{x \in N} S_{a_x}(r_x)$. Since N is contained in the countable union of sets of measure zero, $P(N) = 0$, as was to be shown.

VI. Nearest Neighbor Probability of Error

Let $x_n' \in \{x_1, x_2, \cdots, x_n\}$ be the nearest neighbor to x and let θ_n' be the category to which the individual having measurement x_n' belongs. If θ is indeed the category of x, the NN rule incurs loss $L(\theta, \theta_n')$. If $(x, \theta), (x_1, \theta_1), \cdots, (x_n, \theta_n)$ are random variables, we define the n-sample NN risk $R(n)$ by the expectation

$$R(n) = E[L(\theta, \theta_n')] \tag{10}$$

and the (large sample) *NN risk* R by

$$R = \lim_{n \to \infty} R(n). \tag{11}$$

Throughout this discussion we shall assume that the pairs $(x, \theta), (x_1, \theta_1), \cdots, (x_n, \theta_n)$ are independent identically distributed random variables in $X \times \Theta$. Of course, except in trivial cases, there will be some dependence between the elements x_i, θ_i of each pair.

We shall first consider the $M = 2$ category problem with probability of error criterion given by the $0 - 1$ loss matrix

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{12}$$

where L counts an error whenever a mistake in classification is made. The following theorem is the principal result of this discussion.

Theorem

Let X be a separable metric space. Let f_1 and f_2 be such that, with probability one, x is either 1) a continuity point of f_1 and f_2, or 2) a point of nonzero probability measure. Then the NN risk R (probability of error) has the bounds

$$R^* \leq R \leq 2R^*(1 - R^*). \tag{13}$$

These bounds are as tight as possible.

Remarks: In particular, the hypotheses of the theorem are satisfied for probability densities which consist of any mixture of δ-functions and piecewise continuous density functions on Euclidean d-space. Observe that $0 \leq R^* \leq R \leq 2R^*(1 - R^*) \leq \frac{1}{2}$; so $R^* = 0$ if and only if $R = 0$, and $R^* = \frac{1}{2}$ if and only if $R = \frac{1}{2}$. Thus in the extreme cases of complete certainty and complete uncertainty the NN probability of error equals the Bayes probability of error. Conditions for equality of R and R^* for other values of R^* will be developed in the proof.

Proof: Let us condition on the random variables x and x_n' in the n-sample NN problem. The conditional NN risk $r(x, x_n')$ is then given, upon using the conditional independence of θ and θ_n', by

$$\begin{aligned} r(x, x_n') &= E[L(\theta, \theta_n') \mid x, x_n'] = P_r\{\theta \neq \theta_n' \mid x, x_n'\} \\ &= P_r\{\theta = 1 \mid x\}P_r\{\theta_n' = 2 \mid x_n'\} \\ &\quad + P_r\{\theta = 2 \mid x\}P_r\{\theta_n' = 1 \mid x_n'\} \end{aligned} \tag{14}$$

where the expectation is taken over θ and θ_n'. By the development of (4) the above may be written as

$$r(x, x_n') = \hat{\eta}_1(x)\hat{\eta}_2(x_n') + \hat{\eta}_2(x)\hat{\eta}_1(x_n'). \tag{15}$$

We wish first to show that $r(x, x_n')$ converges to the random variable $2\hat{\eta}_1(x)\hat{\eta}_2(x)$ with probability one.

We have not required that f_1, f_2 be continuous at the points x of nonzero probability measure $\nu(x)$, because these points may be trivially taken into account as follows. Let $\nu(x_0) > 0$; then

$$P_r\{x_0 \neq x_n'\} = (1 - \nu(x_0))^n \to 0. \tag{16}$$

Since x_n', once equalling x_0, equals x_0 thereafter,

$$r(x, x_n') \to 2\hat{\eta}_1(x_0)\hat{\eta}_2(x_0) \tag{17}$$

with probability one.

For the remaining points, the hypothesized continuity of f_1 and f_2 is needed. Here x is a continuity point of f_1 and f_2 with conditional probability one (conditioned on x such that $\nu(x) = 0$). Then, since $\hat{\eta}$ is continuous in f_1 and f_2, x is a continuity point of $\hat{\eta}$ with probability one. By the lemma, x_n' converges to the random variable x with probability one. Hence, with probability one,

$$\hat{\eta}(x_n') \to \hat{\eta}(x) \tag{18}$$

and, from (15), with probability one,

$$r(x, x_n') \to r(x) = 2\hat{\eta}_1(x)\hat{\eta}_2(x), \tag{19}$$

where $r(x)$ is the limit of the n-sample conditional NN risk.

As shown in (6) the conditional Bayes risk is

$$\begin{aligned} r^*(x) &= \min\{\hat{\eta}_1(x), \hat{\eta}_2(x)\} \\ &= \min\{\hat{\eta}_1(x), 1 - \hat{\eta}_1(x)\}. \end{aligned} \tag{20}$$

Now, by the symmetry of r^* in $\hat{\eta}_1$, we may write

$$\begin{aligned} r(x) = 2\hat{\eta}_1(x)\hat{\eta}_2(x) &= 2\hat{\eta}_1(x)(1 - \hat{\eta}_1(x)) \\ &= 2r^*(x)(1 - r^*(x)). \end{aligned} \tag{21}$$

Thus as a by-product of the proof, we have shown in the large sample case, that with probability one a randomly chosen x will be correctly classified with probability $2r^*(x)(1 - r^*(x))$. For the overall NN risk R, we have, by definition,

$$R = \lim_n E[r(x, x'_n)] \tag{22}$$

where the expectation is taken over x and x'_n. Now L, and hence r, is bounded by one; so applying the dominated convergence theorem,

$$R = E[\lim_n r(x, x'_n)]. \tag{23}$$

The limit, from (19) and (21), yields

$$\begin{aligned} R &= E[r(x)] \\ &= E[2\hat{\eta}_1(x)\hat{\eta}_2(x)] \\ &= E[2r^*(x)(1 - r^*(x))]. \end{aligned} \tag{24}$$

Since the Bayes risk R^* is the expectation of r^*, we have

$$R = 2R^*(1 - R^*) - 2 \operatorname{Var} r^*(x). \tag{25}$$

Hence

$$R \leq 2R^*(1 - R^*), \tag{26}$$

with equality iff $\operatorname{Var} r^* = 0$, which holds iff $r^* = R^*$ with probability one. Investigating this condition we find that for $R = 2R^*(1 - R^*)$ it is necessary and sufficient that

$$\frac{\eta_1 f_1(x)}{\eta_2 f_2(x)} = R^*/(1 - R^*) \quad \text{or} \quad (1 - R^*)/R^* \tag{27}$$

for almost every x (with respect to the probability measure ν).

Rewriting (24), we have

$$\begin{aligned} R &= E[r^*(x) + r^*(x)(1 - 2r^*(x))] \\ &= R^* + E[r^*(x)(1 - 2r^*(x))] \\ &\geq R^* \end{aligned} \tag{28}$$

with equality if and only if $r^*(x)(1 - 2r^*(x)) = 0$ almost everywhere (with respect to ν). Thus the lower bound $R = R^*$ is achieved if and only if r^* equals 0 or $\frac{1}{2}$ almost everywhere and $Er^* = R^*$. Examples of probability distributions achieving the upper and lower bounds will be given at the end of this section following the extension to M categories.

Consider now the M-category problem with the probability of error criterion given by the loss function $L(i, j) = 0$, for $i = j$, and $L(i, j) = 1$, for $i \neq j$. The substitution trick of (21) can no longer be used when $M \neq 2$.

Theorem (Extension of Theorem 1 to $M \neq 2$)

Let X be a separable metric space. Let $f_1, f_2, \cdots, f_M$ be probability densities with respect to some probability measure ν such that, with probability one, x is either 1) a continuity point of $f_1, f_2, \cdots, f_M$, or 2) a point of nonzero probability measure. Then the NN probability of error R has the bounds

$$R^* \leq R \leq R^*\left(2 - \frac{M}{M - 1} R^*\right). \tag{29}$$

These bounds are as tight as possible.

Proof: Since $x'_n \to x$ with probability one, the posterior probability vector $\hat{\eta}(x'_n) \to \hat{\eta}(x)$ with probability one. The conditional n-sample NN risk $r(x, x'_n)$ is

$$r(x, x'_n) = E[L(\theta, \theta'_n) \mid x, x'_n] = \sum_{i \neq j}^{M} \hat{\eta}_i(x)\hat{\eta}_j(x'_n) \tag{30}$$

which converges with probability one to the large sample conditional risk $r(x)$ defined by

$$r(x) = \sum_{i \neq j} \hat{\eta}_i(x)\hat{\eta}_j(x) = 1 - \sum_{i=1}^{M} \hat{\eta}_i^2(x). \tag{31}$$

The conditional Bayes risk $r^*(x)$, obtained by selecting, for a given x, the maximum $\hat{\eta}_i(x)$, say $\hat{\eta}_k(x)$, is given by

$$r^*(x) = 1 - \max_i \{\hat{\eta}_i(x)\} = 1 - \hat{\eta}_k(x). \tag{32}$$

By the Cauchy-Schwarz inequality

$$\begin{aligned} (M - 1) \sum_{j \neq k}^{M} \hat{\eta}_j^2(x) &\geq \left[\sum_{j \neq k}^{M} \hat{\eta}_j(x)\right]^2 \\ &= [1 - \hat{\eta}_k(x)]^2 = (r^*(x))^2. \end{aligned} \tag{33}$$

Adding $(M - 1)\hat{\eta}_k^2(x)$ to each side,

$$\begin{aligned} (M - 1) \sum_{j=1}^{M} \hat{\eta}_j^2(x) &\geq (r^*(x))^2 + (M - 1)\hat{\eta}_k^2(x) \\ &= (r^*(x))^2 + (M - 1)(1 - r^*(x))^2 \end{aligned} \tag{34}$$

or

$$\sum_{j=1}^{M} \hat{\eta}_j^2(x) \geq \frac{(r^*(x))^2}{M - 1} + (1 - r^*(x))^2 \tag{35}$$

Substituting (35) into (31),

$$r(x) \leq 2r^*(x) - \frac{M}{M - 1} (r^*(x))^2. \tag{36}$$

Taking expectations, and using the dominated convergence theorem as before,

$$R = 2R^* - \frac{M}{M - 1} (R^*)^2 - \frac{M}{M - 1} \operatorname{Var} r^*(x). \tag{37}$$

Hence

$$R \leq R^*\left(2 - \frac{M}{M - 1} R^*\right) \tag{38}$$

with equality if and only if $\operatorname{Var} r^*(x) = 0$. Of course, $\operatorname{Var} r^* = 0$ implies $r^*(x) = R^*$ with probability one.

The upper bound is attained for the no-information experiment $f_1 = f_2 = \cdots = f_M$, with $\eta_1 = 1 - R^*$, and $\eta_i = R^*/(M-1)$; $i = 2, \cdots, M$. The lower bound $R = R^*$ is attained, for example, when $\eta_i = 1/M$, $i = 1, 2, \cdots, M$, and

$$f_i(x) = \begin{cases} 1, & 0 \le x \le \dfrac{MR^*}{M-1} \text{ or } i \le x \le i + 1 - \dfrac{MR^*}{M-1} \\ 0, & \text{elsewhere.} \end{cases} \tag{39}$$

VII. Example

Let the real valued random variable x have triangular densities f_1 and f_2 with prior probabilities $\eta_1 = \eta_2 = \frac{1}{2}$, as shown in Fig. 2. The density $f = \eta_1 f_1 + \eta_2 f_2$ on x is uniform on [0, 1], thus facilitating calculation of the distribution of the nearest neighbor x_n'.

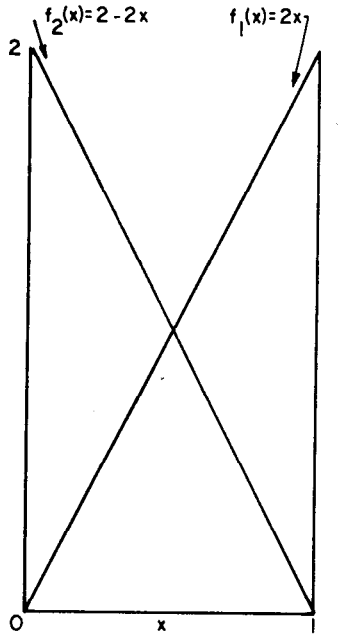

Fig. 2. Triangle densities for example.

The probability of error for this example in the n-sample single NN case is

$$\begin{aligned} R(n) &= E[\eta_1\eta_2 f_1(x)f_2(x_n') + \eta_1\eta_2 f_2(x)f_1(x_n')] \\ &= E[x(1-x_n') + (1-x)x_n']. \end{aligned} \tag{40}$$

Upon performing a lengthy but straightforward calculation, we obtain

$$R(n) = \frac{1}{3} + \frac{1}{(n+1)(n+2)}. \tag{41}$$

Thus

$$R = \lim_{n\to\infty} R(n) = \tfrac{1}{3}. \tag{42}$$

The NN risk R is to be compared to the Bayes risk

$$\begin{aligned} R^* &= \int \min\{\eta_1 f_1, \eta_2 f_2\}\, d\nu \\ &= \int_0^1 \min\{x, 1-x\}\, dx = \tfrac{1}{4}. \end{aligned} \tag{43}$$

Exhibiting corresponding terms we have

$$R^* \le R \le 2R^*(1-R^*)$$

or

$$\tfrac{1}{4} \le \tfrac{1}{3} \le \tfrac{3}{8}. \tag{44}$$

In this example we have found an exact expression for the NN risk $R(n)$ for any finite sample size. Observe that $R(1) = \frac{1}{2}$, in agreement with simpler considerations, and that $R(n)$ converges to its limit approximately as $1/n^2$.

VIII. The k-NN Rule

From Section V it is also possible to conclude that the kth nearest neighbor to x converges to x with probability one as the sample size n increases with k fixed. Since each of the nearest neighbors casts conditionally independent votes as to the category of x, we may conclude, in the 2-category case for odd k, that the conditional k-NN risk $r_k(x)$ is given in the limit (with probability one) as n increases, by

$$\begin{aligned} r_k(x) = \hat{\eta}_1(x) &\sum_{j=0}^{(k-1)/2} \binom{k}{j} \hat{\eta}_1^j(x)(1-\hat{\eta}_1(x))^{k-j} \\ &+ (1-\hat{\eta}_1(x)) \sum_{j=(k+1)/2}^{k} \binom{k}{j} \hat{\eta}_1^j(x)(1-\hat{\eta}_1(x))^{k-j}. \end{aligned} \tag{45}$$

Note that the conditional NN risks $r_k(x)$ are monotonically decreasing in k (to $\min\{\hat{\eta}_1(x), 1-\hat{\eta}_1(x)\}$), as we might suspect. Thus the least upper bounds on the unconditional NN risks R_k will also be monotonically decreasing in k (to R^*).

Observe that in (45) r_k is symmetric in $\hat{\eta}_1$ and $1-\hat{\eta}_1$. Thus r_k may be expressed solely in terms of $r^* = \min\{\hat{\eta}_1, 1-\hat{\eta}_1\}$ in the form

$$\begin{aligned} r_k &= \rho_k(r^*) \\ &= r^* \sum_{j=0}^{(k-1)/2} \binom{k}{j} (r^*)^j (1-r^*)^{k-j} \\ &\quad + (1-r^*) \sum_{j=(k+1)/2}^{k} \binom{k}{j} (r^*)^j (1-r^*)^{k-j}. \end{aligned} \tag{46}$$

Now let $\tilde{\rho}_k(r^*)$ be defined to be the least concave function greater than $\rho_k(r^*)$. Then

$$r_k = \rho_k(r^*) \le \tilde{\rho}_k(r^*), \tag{47}$$

and, by Jensen's inequality,

$$R_k = Er_k = E\rho_k(r^*) \le E\tilde{\rho}_k(r^*) \le \tilde{\rho}_k(Er^*) = \tilde{\rho}_k(R^*). \tag{48}$$

So $\tilde{\rho}_k(R^*)$ is an upper bound on the large sample k-NN risk R_k. It may further be shown, for any R^*, that $\tilde{\rho}_k(R^*)$ is the least upper bound on R_k by demonstrating simple statistics which achieve it. Hence we have the bounds

$$R^* \leq R_k \leq \tilde{\rho}_k(R^*) \leq \tilde{\rho}_{k-1}(R^*) \leq \cdots \leq \tilde{\rho}_1(R^*) = 2R^*(1 - R^*) \quad (49)$$

where the upper and lower bounds on R_k are as tight as possible.

IX. Conclusions

The single NN rule has been shown to be admissible among the class of k_n-NN rules for the n-sample case for any n. It has been shown that the NN probability of error R, in the M-category classification problem, is bounded below by the Bayes probability of error R^* and above by $R^*(2 - MR^*/(M - 1))$. Thus *any* other decision rule based on the infinite data set can cut the probability of error by at most one half. In this sense, half of the available information in an infinite collection of classified samples is contained in the nearest neighbor.

References

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination," 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*, H. Solomon, Ed. Stanford, Calif.: Stanford University Press, 1961.

[4] L. N. Kanal, "Statistical methods for pattern classification," Philco Rept., 1963; originally appeared in T. Harley et al., "Semi-automatic imagery screening research study and experimental investigation," *Philco Reports*, V043-2 and V043-3, Vol. I, sec. 6, and Appendix H, prepared for U. S. Army Electronics Research and Development Lab. under Contract DA-36-039-SC-90742, March 29, 1963.

[5] G. Sebestyen, *Decision Making Processes in Pattern Recognition*. New York: Macmillan, 1962, pp. 90–92.

[6] Nils Nilsson, *Learning Machines*. New York: McGraw-Hill, 1965, pp. 120–121.

[7] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Annals Math Stat.*, vol. 36, pp. 1049–1051, June 1965.

A Generalized Form of Price's Theorem and Its Converse

JOHN L. BROWN, JR., SENIOR MEMBER, IEEE

Abstract—The case of n unity-variance random variables $x_1, x_2, \cdots, x_n$ governed by the joint probability density $w(x_1, x_2, \cdots x_n)$ is considered, where the density depends on the (normalized) cross-covariances $\rho_{ij} = E[(x_i - \bar{x}_i)(x_j - \bar{x}_j)]$. It is shown that the condition

$$(*) \quad \frac{\partial}{\partial \rho_{ij}} \{E[f(x_1, x_2, \cdots, x_n)]\} = E\left[\frac{\partial^2}{\partial x_i \, \partial x_j} f(x_1, x_2, \cdots, x_n)\right] \quad (i \neq j)$$

holds for an "arbitrary" function $f(x_1, x_2, \cdots, x_n)$ of n variables if and only if the underlying density $w(x_1, x_2, \cdots, x_n)$ is the usual n-dimensional Gaussian density for correlated random variables. This result establishes a generalized form of Price's theorem in which: 1) the relevant condition (*) subsumes Price's original condition; 2) the proof is accomplished without appeal to Laplace integral expansions; and 3) conditions referring to derivatives with respect to diagonal terms ρ_{ii} are avoided, so that the unity variance assumption can be retained.

Manuscript received February 10, 1966; revised May 2, 1966.
The author is with the Ordnance Research Laboratory, Pennsylvania State University, State College, Pa.

Introduction

PRICE'S THEOREM and its various extensions ([1]–[4]) have had great utility in the determination of output correlations between zero-memory nonlinearities subjected to jointly Gaussian inputs. In its original form, the theorem considered n jointly normal random variables, $x_1, x_2, \cdots x_n$, with respective means $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_n$ and nth-order joint probability density,

$$P(x_1, x_2, \cdots, x_n) = (2\pi)^{-n/2} |M_n|^{-1/2} \cdot \exp\left\{-\tfrac{1}{2} \sum_r \sum_s \frac{M_{rs}}{|M_n|} (x_r - \bar{x}_r)(x_s - \bar{x}_s)\right\}, \quad (1)$$

where $|M_n|$ is the determinant of $M_n = [\rho_{rs}]$,

$$\rho_{rs} = E[(x_r - \bar{x}_r)(x_s - \bar{x}_s)] = \overline{x_r x_s} - \bar{x}_r \bar{x}_s$$

is the correlation coefficient of x_r and x_s, and M_{rs} is the cofactor of ρ_{sr} in M_n.

From [1], the theorem statement is as follows:

"Let there be n zero-memory nonlinear devices specified by the input-output relationship $f_i(x)$, $i = 1, 2, \cdots, n$. Let each x_i be the single input to a corresponding $f_i(x)$