Asymptotic Properties of Nearest Neighbor Rules Using Edited Data

DENNIS L. WILSON, MEMBER, IEEE

Abstract—**The convergence properties of a nearest neighbor rule that uses an editing procedure to reduce the number of preclassified samples and to improve the performance of the rule are developed. Editing of the preclassified samples using the three-nearest neighbor rule followed by classification using the single-nearest neighbor rule with the remaining preclassified samples appears to produce a decision procedure whose risk approaches the Bayes' risk quite closely in many problems with only a few preclassified samples. The asymptotic risk of the nearest neighbor rules and the nearest neighbor rules using edited preclassified samples is calculated for several problems.**

I. Introduction

A BASIC class of decision problems which includes a large number of practical problems can be characterized in the following way. 1) There is a sample to be classified. 2) There are already classified samples from the same distributions as the sample to be classified with which a comparison can be made in making a decision. 3) There is no additional information about the distributions of any of the random variables involved other than the information contained in the preclassified samples. 4) There is a measure of distance between samples. Examples of problems having these characteristics are the problems of handwritten character recognition and automatic decoding of manual Morse. In each of these problems preclassified samples may be provided by a man, and a simple metric can be devised.

"Nearest neighbor rules" are a collection of simple rules which can have very good performance with only a few preclassified samples. We shall develop the *asymptotic performance* of a nearest neighbor rule using editing. The asymptotic performance is the performance when the number of preclassified samples is very large.

Nearest neighbor rules were originally suggested for solution of problems of this type by Fix and Hodges [1] in 1952. Nearest neighbor rules are practically always included in papers which survey pattern recognition, e.g., Sebestyen [2], Nilsson [3], Rosen [4], Nagy [5], and Ho and Agrawala [6]. Analysis of the properties of the nearest neighbor rules was started by Fix and Hodges [1] and continued by Cover and Hart [7] and Whitney and Dwyer [8]. Cover [9] summarizes many of the properties of the nearest neighbor rules. Patrick and Fischer [10] generalize the nearest neighbor rules to include weighting of different types of error and problems "in which the training samples available are not in the same proportions as the *a priori* class probabilities" by using the concept of tolerance regions.

Manuscript received September 16, 1970; revised December 28, 1971.
The author is with the Electronic Systems Group—Western Division, GTE Sylvania, Inc., Mountain View, Calif. 94040.

The nearest neighbor decision procedures use the sample to be classified and the set of preclassified samples in making a decision.

The Sample to Be Classified

Let $X \in E^d$ be a random variable generated as follows. Select $\theta = 1$ with probability η_1 and $\theta = 2$ with probability η_2. Given θ, select X from a population with density $f_1(x)$ when $\theta = 1$ and from a population with density $f_2(x)$ when $\theta = 2$ (E^d is a d-dimensional Euclidean space).

The Preclassified Samples

Let (X_i,θ_i), $i = 1,2,\cdots,N$, be generated independently as follows. Select $\theta_i = 1$ with probability η_1 and $\theta_i = 2$ with probability η_2. Given θ_i, select $X_i \in E^d$ from a population with density $f_1(x)$ when $\theta_i = 1$ and from a population $f_2(x)$ when $\theta_i = 2$. The set $\{(X_i,\theta_i)\}$ constitutes the set of preclassified samples.

Two types of rules will be discussed: nearest neighbor rules and modified nearest neighbor rules.

To make a decision using the K-nearest neighbor rule: Select from among the preclassified samples of the K-nearest neighbors of the sample to be classified. Select the class represented by the largest number of the K-nearest neighbors. Ties are to be broken randomly.

To make a decision using the modified K-nearest neighbor rule:

a) For each i,
   1) find the K-nearest neighbors to X_i among $\{X_1,X_2,\cdots,X_{i-1},X_{i+1},\cdots,X_N\}$;
   2) find the class θ associated with the largest number of points among the K-nearest neighbors, breaking ties randomly when they occur.
b) Edit the set $\{(X_i,\theta_i)\}$ by deleting (X_i,θ_i) whenever θ_i does not agree with the largest number of the K-nearest neighbors as determined in the foregoing.

Make a decision concerning a new sample using the modified K-nearest neighbor rule by using the single-nearest neighbor rule with the reduced set of preclassified samples.

Examples of the Power of the Nearest Neighbor Rules

The nearest neighbor rules can be very powerful rules, useful in many problems. Figs. 1–4 demonstrate the asymptotic performance of the K-nearest neighbor rule and the modified K-nearest neighbor rule in four different problems. Fig. 1 compares the performance of the two types of rules with the performance of Bayes' rule when population one is a logistic distribution centered at -1 and population two is a logistic distribution centered at $+1$. The distribution

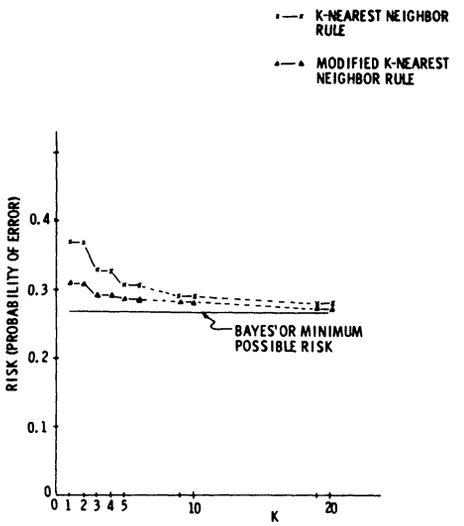

Fig. 1. Asymptotic risk of using K-nearest neighor rule and the modified K-nearest neighbor rule compared to Bayes' risk when $\eta_1 = \eta_2 = 0.5$ and population one is logistic centered at -1 and population two is logistic centered at $+1$.

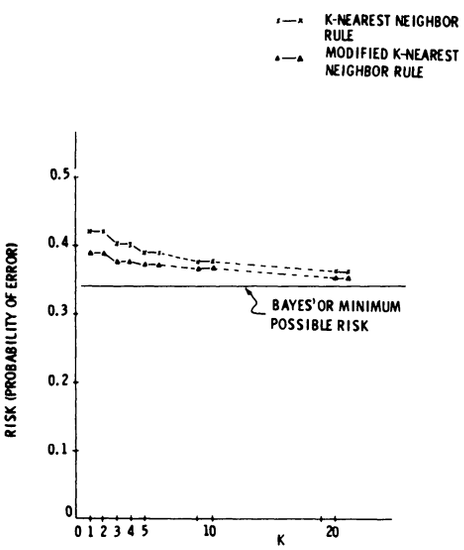

Fig. 3. Asymptotic risk of using K-nearest neighbor rule and modified K-nearest neighbor rule compared to Bayes' risk when $\eta_1 = \eta_2 = 0.5$ and population one is $N(0, 1)$ and population two is $N(0, 4)$.

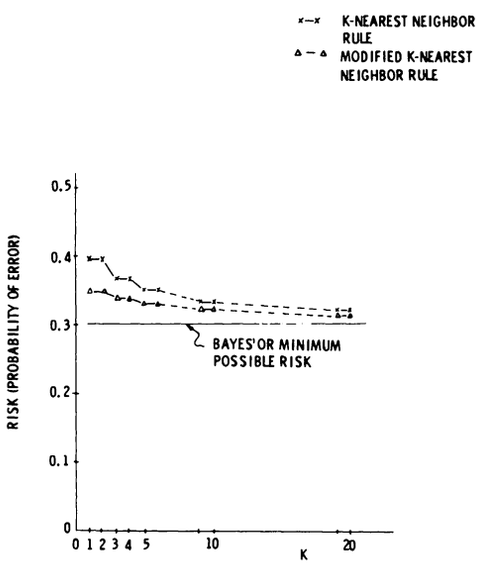

Fig. 2. Asymptotic risk of using K-nearest neighbor rule and modified K-nearest neighbor rule compared to Bayes' risk when $\eta_1 = \eta_2 = 0.5$ and population one is $N(0, 1)$ and population two is $N(1, 1)$.

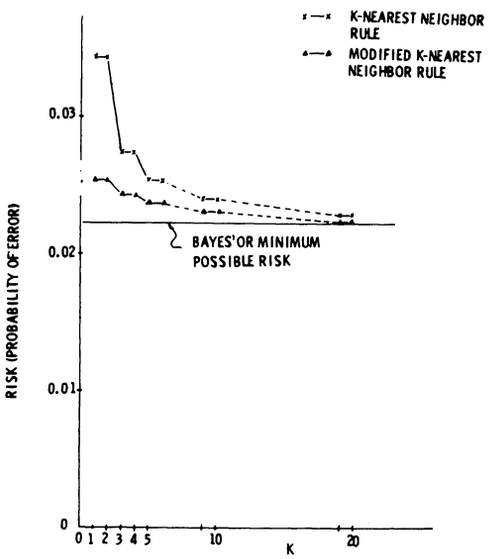

Fig. 4. Asymptotic risk of using K-nearest neighbor rule and modified K-nearest neighbor rule compared to Bayes' risk when $\eta_1 = \eta_2 = 0.5$ and population one is $N(2, 1)$ and population two is $N(-2, 1)$.

functions for these two populations are

$$f_1(x) = \frac{\exp(-(x - 1))}{[1 + \exp(-(x - 1))]^2}$$

$$f_2(x) = \frac{\exp(-(x + 1))}{[1 + \exp(-(x + 1))]^2}.$$

Fig. 2 compares asymptotic performance of the nearest neighbor rules with the performance of Bayes' rule when population one is a normal distribution centered at 0 with variance 1 ($N(0,1)$) and population two is a normal distribution centered at 1 with variance 1 ($N(1,1)$).

Fig. 3 compares the asymptotic performance of the nearest neighbor rules with the performance of Bayes' rule when population one is a normal distribution with mean 0 and variance 1 ($N(0,1)$) and population two is a normal distribution with mean 0 and variance 4 ($N(0,4)$).

In each of these three figures the risk of using the nearest neighbor rules decreases as the number of neighbors used increases. The risk of using the modified nearest neighbor rule is about halfway between the risk of using the nearest neighbor rule with the same number of neighbors and the Bayes' risk.

Fig. 4 presents a more realistic situation. The error rate is on the order of 2 or 3 per 100 trials as compared to the error rate of 2 or 3 out of 10 trials in Figs. 1–3. Most decision makers cannot afford to make 2 or 3 errors in 10 trials. They will search for more data on which to base the decision if the risk level is high. In Fig. 4 population one is a normal distribution centered at -2 with variance 1 ($N(-2,1)$) and

population two is a normal distribution centered at +2 with variance 1 ($N(+2,1)$). For this problem the asymptotic risk of using the single-nearest neighbor rule is large compared to the risk of using the other rules. The risk of using the modified three-nearest neighbor rule is about 10 percent more than the Bayes' risk.

It is interesting to consider how many trials in making a decision would be necessary to determine whether a decision maker was using the Bayes' rule or the modified three-nearest neighbor rule. For the problem where the Bayes' risk is about 0.01 and the risk of the modified nearest neighbor rule is about 10 percent greater, it would be necessary to check the accuracy of about 10 000 decisions before there was enough information to begin to estimate the probabilities of error well enough to tell which rule was being used; to draw a reliable conclusion would require about 100 000 sample decisions.

Some example calculations indicate that the number of preclassified samples required for the risk to be close to the asymptotic risk is on the order of 50 for the single-nearest neighbor rule in the problems of Figs. 1–4. (These results are to be presented in a following paper on conversion rates.) This suggests that roughly K times 50 samples would be required to be close to the asymptotic risk for the K-nearest neighbor rule and for the modified K-nearest neighbor rule.

II. Preliminary Development

An Induced Distribution

The nearest neighbor rules depend only on the distances from the sample to be classified to the preclassified samples, and not on the direction. The induced distribution of the distances from the sample to be classified to a preclassified sample will be useful. This induced distribution is developed as follows.

Let $Z_i(x) = \|X_i - x\|$, $Z(x) = \|X - x\|$, and $Z_i^* = \|X_i - X\|$, where $\|A - B\|$ is the usual Euclidean measure of distance from point A to point B on E^d. The $Z_i(x)$, $i = 1,2,\cdots,N$, are independent and identically distributed; the Z_i^* are not. The induced probability measure conditioned on $X = x$ is specified by the conditional cumulative distribution function (cdf)

$$F_Z^{(1)}(z \mid x) \equiv F_{Z(x)}(z \mid X = x, \theta = 1) = \int_{S(x,z)} f_1(x)\,dx$$

$$F_Z^{(2)}(z \mid x) \equiv F_{Z(x)}(z \mid X = x, \theta = 2) = \int_{S(x,z)} f_2(x)\,dx$$

where the notation $S(x,z)$ indicates that the integral is to be taken over the volume of the hypersphere centered at $X = x$ with radius z. The set $\{(Z_i^*,\theta_i)\}$ constitutes a description of the preclassified samples in terms of their distances from the sample to be classified.

A Posteriori Probabilities of the Class Given the Sample Value

Given $X = x$, the probability that the associated class is class 1 or class 2 is calculated by application of Bayes' rule. For example,

$$p_1(x) \equiv P(\theta = 1 \mid X = x) = \frac{\eta_1 f_1(x)}{\eta_1 f_1(x) + \eta_2 f_2(x)}.$$

Similarly,

$$p_1(x,z) \equiv P(\theta = 1 \mid Z(x) = z) = \frac{\eta_1 f_1(z \mid x)}{\eta_1 f_1(z \mid x) + \eta_2 f_2(z \mid x)}$$

where $f_1(z \mid x)$ is the pdf corresponding to $F_Z^{(1)}(z \mid x)$ and $f_2(z \mid x)$ is the pdf corresponding to $F_Z^{(2)}(z \mid x)$.

Decisions and the Associated Risk

A possible decision rule is described by the probability $\phi(i \mid x)$ of selecting $\theta = i$ conditioned on the value of X. Conditioning on the preclassified sample values will also be used. The risk associated with using a decision rule is given by

$$R = \int_X [p_1(x)\phi(2 \mid x)L(2 \mid 1) + p_2(x)\phi(1 \mid x)L(1 \mid 2)]\,dF(x)$$

where $L(j \mid i)$ is the loss when the decision is $\theta = j$ given that i is the true state and $F(x) = \sum_i \eta_i F_X(x \mid \theta = i)$. When the loss is one for each type of error, the risk is simply the probability of error:

$$R = \int [p_1(x)(1 - \phi(1 \mid x)) + (1 - p_1(x))\phi(1 \mid x)]\,dF(x).$$

The nearest neighbor rules also depend upon the preclassified sample set. Where necessary, the dependence will be made explicit.

Bayes' Rule

The Bayes' rule may be developed by using the expression for the risk. To minimize the risk, minimize the integrand of the risk integral, the local risk, at each point x. To minimize the local risk, select $\phi(1 \mid x) = 1$ whenever $p_2(x)L(1 \mid 2) < p_1(x)L(2 \mid 1)$, select $\phi(2 \mid x) = 1$ whenever $p_2(x)L(1 \mid 2) > p_1(x)L(2 \mid 1)$, and make the decision in an arbitrary way when $p_2(x)L(1 \mid 2) = p_1(x)L(2 \mid 1)$.

III. Asymptotic Properties of Nearest Neighbor Rules

The characteristics of the nearest neighbor rules using very large numbers of preclassified samples have constituted most of the well-known results. (See Fix and Hodges [1], Cover and Hart [7], and Whitney and Dwyer [8].) This section derives new asymptotic results for the modified nearest neighbor rules and incidentally rederives most of the already known asymptotic results for the K-nearest neighbor rules.

The asymptotic results that are to be derived will be in terms of convergence "in probability." According to a standard definition, a random variable Y_N is said to converge in probability to Y ($Y_N \xrightarrow{P} Y$) if, for any $\varepsilon > 0$,

$$P[\|Y_N - Y\| > \varepsilon] \to 0 \text{ as } N \to \infty.$$

A more useful concept of "in probability" has been developed by Pratt [14]. Appendix I presents Pratt's definition of "in probability" and develops several theorems. Two theorems which will be useful in this section are reproduced as follows (proofs are found in Appendix I).

Theorem 1

If $Y_N \xrightarrow{P} Y$, Y is finite with probability one, and $P[Y \in Dg] = 0$, where Dg is the set of discontinuities of the function g, then $g(Y_N) \xrightarrow{P} g(Y)$.

Theorem 1' (Slutsky's Theorem)

If $Y_n \xrightarrow{P} c$ and g is continuous at c, then $g(Y_n) \xrightarrow{P} g(c)$.

We shall use these theorems to show that whenever the neighbors involved in the decision converge in probability to the sample to be classified, the probability that the neighbors come from a given class, the probability of deciding that a given class is the true class, and the local risk will converge to easily calculated asymptotic values.

Let $X^{[i]}(X,N)$ be the neighbor which is the ith distant neighbor from X when there are N preclassified samples. Also, let L_N be a sequence of numbers such that $L_N = o(N)$. (That is, $L_N/N \to 0$. See Appendix I for a careful definition of $o(N)$ and $O(N)$.) We begin showing the convergence properties of the nearest neighbor rules by presenting a theorem demonstrating that $X^{[L_N]}(X,N) \xrightarrow{P} X$. This theorem concerning convergence before editing is included for completeness.

Note that $L_N = i$, where i is a constant independent of N, is a sequence of numbers with the required properties. Those properties which hold for $X^{[L_N]}(X,N)$ will also hold for $X^{[i]}(X,N)$.

Convergence Properties of the Nearest Neighbors

Let $Z_{[i]}^*$ be the ith order statistic of the random variables Z_i^*, $i = 1,2,\cdots,N$. Let $S(x)$ be an open neighborhood of x. The following theorem is suggested by the work of Cover and Hart [7].

Theorem 2

For $L_N = o(N)$, $X^{[L_N]}(X,N) \xrightarrow{P} X$ as $N \to \infty$.

This theorem is proved in Appendix II. A major step in the proof of the theorem was the proof that the nearest neighbors converged to the sample value $X = x$. This fact will be important in following theorems. The conditions under which it holds are stated carefully in the next theorem which has already been proved.

Theorem 2'

If there does not exist a neighborhood $S(x)$ such that $P[S] = 0$, then for $L_N = o(N)$, $X^{[L_N]}(x,N) \xrightarrow{P} x$ as $N \to \infty$.

Convergence After Editing

Editing of the preclassified samples for the modified nearest neighbor rule proceeds by determining whether the indicated decision for the K-nearest neighbor rules agrees with the actual classification for each of the preclassified samples. After all of the preclassified samples are considered, those samples for which the decision does not agree with the true classification are deleted.

Let $X_{EK}^{[1]}(X,N)$ be the sample which is nearest to X after editing. Also, let Df_1 and Df_2 be the set of discontinuities of $f_1(x)$ and $f_2(x)$, respectively.

Theorem 3

If $P[X \in Df_1] = 0$ and $P[X \in Df_2] = 0$, then

$$X_{EK}^{[1]}(X,N) \xrightarrow{P} X \text{ as } N \to \infty.$$

The proof of the theorem is long and tedious, so in spite of its importance it has been relegated to Appendix III. At first glance, the proof of the theorem seems easy, and would be very easy if the editing of the preclassified samples occurred independently. Finding preclassified samples which are edited independently constitutes most of the proof. In the same way that the proof of Theorem 2 involved the proof of Theorem 2' the proof of Theorem 3 involves the proof of a theorem concerning the convergence of the edited nearest neighbor to a sample to be classified.

Theorem 3'

If there does not exist a neighborhood $S(x)$ such that $P[S] = 0$ and if $f_1(x)$ and $f_2(x)$ are continuous at $X = x$, then $X_{EK}^{[1]}(x,N) \xrightarrow{P} x$ as $N \to \infty$.

Most of the important asymptotic properties of the nearest neighbor rules can be developed from the preceding six theorems. The basic asymptotic properties are summarized in the theorems to follow. In order to state the theorem carefully it is necessary to define a few terms.

A Generalized Convergent Sample

The theorem will be stated in terms of a generalized sample $X^*(X,N)$ which converges to the sample to be classified, X. Theorems 2 and 3 have shown that for $L_N = o(N)$, $X^{[L_N]}(X,N) \xrightarrow{P} X$ as $N \to \infty$ and that $X_{EK}^{[1]}(X,N) \xrightarrow{P} X$ as $N \to \infty$. Both $X^{[1]}(X,N)$ and $X_{EK}^{[1]}(X,N)$ qualify as random variables that can be represented by $X^*(X,N)$.

Dependence of the Decision

With each preclassified sample there is associated a class θ_i. The preclassified samples have been viewed in terms of their ordering according to their distance from a particular point x. This ordering led to defining $X^{[i]}(x,N)$, the ith distant sample from x when there are N preclassified samples. Let $\theta^{[i]}(x,N)$ be the classification associated with the sample $X^{[i]}(x,N)$. The nearest neighbor rules can be described in terms of dependence on the sample values of $\theta^{[i]}(x,N)$ whose indices i lie in a set $I_N(X)$.

Definition: A decision is said to depend directly on the values of $\theta^{[i]}(x,N) i \in I_N(x)$ if $X^{[i]}(x,N)$ remains after editing, the decision can be determined when the values of $\theta^{[i]}(x,N) i \in I_N(x)$ are known, and the decision cannot be determined when any of the values of $\theta^{[i]}(x,N) i \in I_N(x)$ are unknown.

Theorem 4

If $P[X \in Df_1] = 0$, $P[X \in Df_2] = 0$, and X is bound with probability one, then for $X^*(X,N)$ such that

$X^*(X,N) \xrightarrow{P} X$ as $N \to \infty$,

a) $f_1(X^*(X,N)) \xrightarrow{P} f_1(X)$
$f_2(X^*(X,N)) \xrightarrow{P} f_2(X)$.

b) $p_1(X^*(X,N)) \xrightarrow{P} p_1(X)$
$p_2(X^*(X,N)) \xrightarrow{P} p_2(X)$.

c) For all rules which depend upon $\theta^{(i)}$, $i \in I_N$ such that for $i \in I_N(X) \|X_i - X\| \leq \|X^* - X\|$,

$$\phi(1 \mid X, X_i, i \in I_N(X)) \xrightarrow{P} \phi_\infty(1 \mid X)$$

where $\phi_\infty(1 \mid X)$ is obtained by substituting $p_1(X)$ for $p_1(X_i, i \in I_N(X))$ wherever necessary in $\phi_N(1 \mid X, X_i, i \in I_N(X))$.

d) $r_N(X) \xrightarrow{P} r_\infty(X)$, where $r_\infty(X)$ is obtained by substituting $p_1(X)$ for $p_1(X_i, i \in I_N(X))$ wherever necessary in the expression for $r_N(X)$ for the rules specified in c).

e) $R_N \to R_\infty$ for the rules specified in c).

Proof:

a) Direct application of Theorem 1.

b) $$p_1(X^{[L_N]}) = \frac{\eta_1 f_1(X^{[L_N]})}{\eta_1 f_1(X^{[L_N]}) + \eta_2 f_2(X^{[L_N]})}$$

by definition. Thus by inspection $p_1(X^{[L_N]})$ is a continuous function of the random variables $f_1(X^{[L_N]})$ and $f_2(X^{[L_N]})$. Direct application of Theorem 1 using the results of a) yields the desired result.

c) Lemma:

$$P[A(x) \mid x, (x^{[1]}(x,N), x^{[2]}(x,N), \cdots, x^{[N]}(x,N))]$$

is a continuous function of

$$P[\Theta^{[i]}(x,N) = \theta^{[i]}(x,N) \mid X^{[i]}(x,N) = x^{[i]}(x,N)].$$

Proof: Conditioned on the values of $X = x$ and $X^{[i]}(x,N) = x^{[i]}(x,N)$ the values of the $\theta^{[i]}(x,N)$ are independent. Therefore,

$$P[A(x) \mid x, (x^{[1]}(x,N) x^{[2]}(x,N), \cdots, x^{[N]}(x,N))] = \int_{\times_{i=1}^{N} \theta^{[i]}} I(A(x)) \prod_i dP[\theta^{[i]}(x,N) \mid x^{[i]}(x,N)]$$

where $I(\cdot)$ is the indicator function. Furthermore, the space $\{1,2\}^N$ has only 2^N members. Thus

$$P[A(X) \mid x, (x^{[1]}(x,N), x^{[2]}(x,N), \cdots, x^{[N]}(x,N))] = \sum_{\times_{i=1}^{N} \theta^{[i]}} I(A(x)) \prod_i P[\theta^{[i]}(x,N) \mid x^{[i]}(x,N)].$$

The probability being examined is seen to be a simple weighted product of the $P[\theta^{[i]}(x,N) \mid x^{[i]}(x,N)]$ which is continuous by a simple exercise in elementary analysis. Q.E.D.

The probability $P[A_N(x) \mid x, (x^{[1]}(x,N), \cdots, x^{[N]}(x,N))]$ is defined as $\phi(1 \mid x, (x^{[1]}(x,N), \cdots, x^{[N]}(x,N)))$. The link between the lemma and the description of the decision is provided. Having proved continuity we can complete the proof of c) by using the result of b) and Theorem 1.

d) The expression for the local risk from Section II is a simple continuous function of random variables which have been shown in a)–c) to converge in probability. Direct application of Theorem 1 yields the desired conclusion.

e) The Asymptotic Risk: The application of the dominated convergence theorem shows that the average of the asymptotic local risk developed in Section II is the same as the asymptotic risk. At each point x the local risk is bounded since all of the components of the local risk except the losses are probabilities which are, of course, less than or equal to one. (We assume that the losses are also finite.) Let U be the bound on the local risk so that $|r_N(X)| < U$. U is integrable:

$$\int U \, dF(x) = U \int dF(x) = U < \infty.$$

We have shown that the local risk $r_N(X) \xrightarrow{P} r_\infty(X)$ as $N \to \infty$. Application of the dominated convergence theorem [11, p. 152] shows that $E(r_N(X))$ converges to $E(r_\infty(X))$. But $E(r_N(x)) = R_N$ and $E(r_\infty(x)) = R_\infty$ for all of the types of rules under discussion. Therefore, it has been shown that $R_N \to R_\infty$. Q.E.D.

A theorem similar to Theorem 4 can be stated showing that for any sample value $X = x$ all of the parameters of the rule will converge in probability.

Theorem 4'

If there does not exist a neighborhood $S(x)$ such that $P[S] = 0$, and if $f_1(X)$ and $f_2(X)$ are continuous at $X = x$, then for $X^*(x,N)$ such that $X^*(x,X) \xrightarrow{P} X$ as $N \to \infty$,

a) $f_1(X^*(x,N)) \xrightarrow{P} f_1(x)$
$f_2(X^*(x,N)) \xrightarrow{P} f_2(x)$.

b) $p_1(X^*(x,N)) \xrightarrow{P} p_1(x)$
$p_2(X^*(x,N)) \xrightarrow{P} p_2(x)$.

c) For all rules which depend upon $\theta^{[i]}$, $i \in I(x,N)$ such that for $i \in I(x,N)$

$$\|X_i - X\| \leq \|X^* - x\|$$

$$\phi(1 \mid x, X_i, i \in I(x,N)) \xrightarrow{P} \phi_\infty(1 \mid x)$$

where $\phi_\infty(1 \mid x)$ is obtained by substituting $p_1(x)$ for $p_1(X_i, i \in I(x,N)$ whenever necessary in $\phi_N(1 \mid x, X_i, i \in I(x,N))$.

d) $r_N(x) \xrightarrow{P} r_\infty(x)$ where $r_\infty(x)$ is obtained by substituting $p_1(x)$ for $p_i(X_i, i \in I(x,N)$ whenever necessary in the expression for $r_N(x)$ for the rules specified in c).

Proof: The proof is the same as the proof of Theorem 4 using Slutsky's theorem (Theorem 1') instead of Theorem 1.

Asymptotic Probability of Deciding that Class 1 is the True Class

For the K-nearest neighbor rule, the asymptotic value of the probability of deciding that class 1 is the correct class, $\phi_\infty{}^K(1 \mid x)$, is given by the following expression:

$$\phi_\infty{}^K(1 \mid x) = \sum_{i=(K+1)/2}^{K} \binom{K}{i} p_1{}^i (1 - p_1)^{K-1}, \qquad \text{for } K \text{ odd}$$

$$= \sum_{i=K/2}^{K} \binom{K}{i} p_1{}^i (1 - p_1)^{K-i} - \frac{1}{2} \binom{K}{K/2} p_1{}^{K/2} (1 - p_1)^{K/2}, \qquad \text{for } K \text{ even}$$

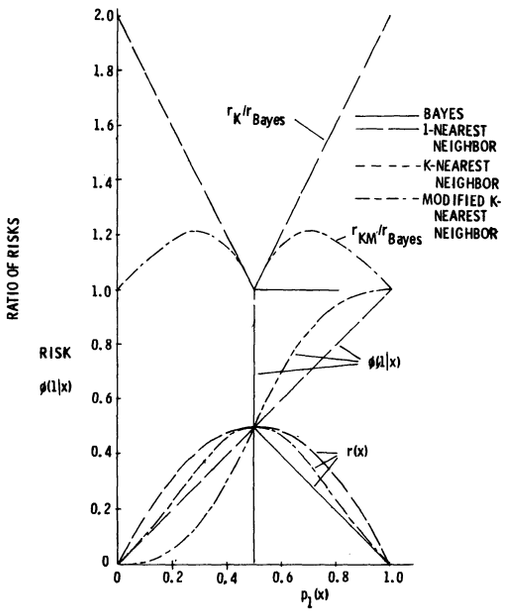

Fig. 5. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 1, 2$.

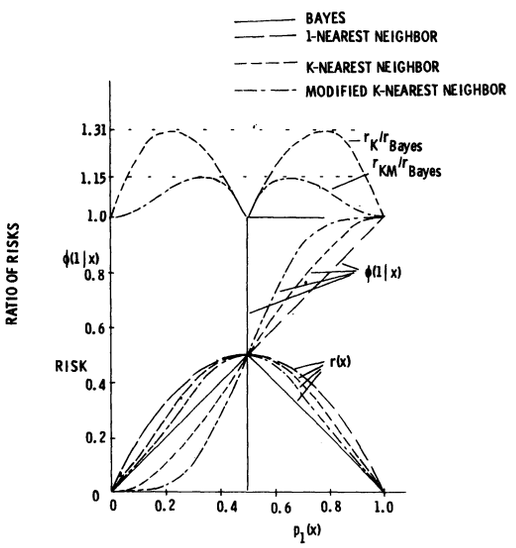

Fig. 6. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 3, 4$.

where $p_1 = p_1(x)$. The probability $\phi_\infty^K(1 \mid x)$ is simply the probability that more than one-half of the K-nearest preclassified samples will be from class 1 when the probability of one of the samples being from class 1 is $p_1(x)$. Ties are broken randomly.

For the modified K-nearest neighbor rule, the probability of deciding that class 1 is the correct class $\phi_\infty^{KM}(1 \mid x)$ is

$$\phi_\infty^{KM}(1 \mid x) = \frac{p_1\phi_\infty^K(1 \mid x)}{p_1\phi_\infty^K(1 \mid x) + (1 - p_1)(1 - \phi_\infty^K(1 \mid x))}$$

$$= \frac{\phi_\infty^K(1 \mid x)}{\phi_\infty^K(1 \mid x) + [(1 - p_1)/p_1](1 - \phi_\infty^K(1 \mid x))}.$$

Applying the K-nearest neighbor rule to each of the preclassified samples results in a probability equal to ϕ_∞^K that a sample from class 1 is retained. The probability that a sample is from class 1 is $p_1(x)$. Normalizing by the probability that the sample was retained regardless of its class yields the probability that any of the nearby preclassified samples is from class 1, given that it is retained. In particular, this probability applies to the nearest remaining neighbor to the sample to be classified.

The asymptotic local risk is obtained by substituting one of the expressions for $\phi(1 \mid x)$ in the expression for the local risk in Section I. From Section I

$$r(x) = L(2 \mid 1)p_1(x)(1 - \phi(1 \mid x)) + L(1 \mid 2)(1 - p_1(x))\phi(1 \mid x).$$

The comparison of the probabilities of deciding that a sample is from class 1 as a function of $p_1(x)$ and the comparison of the local risk as a function of $p_1(x)$ are shown in Figs. 5–10. Several facts should be noted from the comparison.

1) The performance of the K-nearest neighbor rule when K is even is the same as the performance of the K-nearest neighbor rule for the next smallest value of K, an odd value of K. As a consequence, the same result holds true for the modified K-nearest neighbor rule.

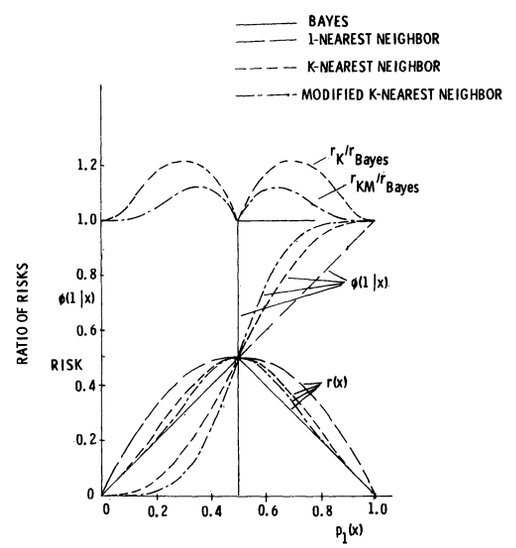

Fig. 7. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 5, 6$.

2) For K small the use of the modified K-nearest neighbor rule instead of the K-nearest neighbor rule reduces the risk by about half of the total amount that it can be reduced. For larger values of K the advantage is not so great.

3) At any value of $p_1(x)$ the probability of deciding that a sample is from class 1 rapidly approaches the Bayes' decision as the number of neighbors used increases. Also, the risk at any value of $p_1(x)$ rapidly approaches the Bayes' risk as the number of neighbors used increases.

4) However, the maximum value of the ratio of the risk of a nearest neighbor rule to the Bayes' risk does not decrease very rapidly. For the modified nearest neighbor

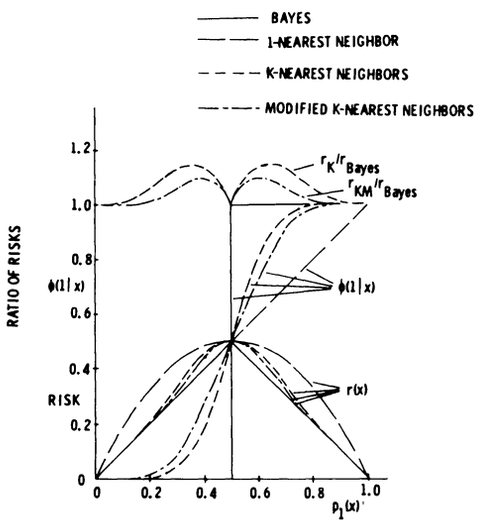

Fig. 8. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 9, 10$.

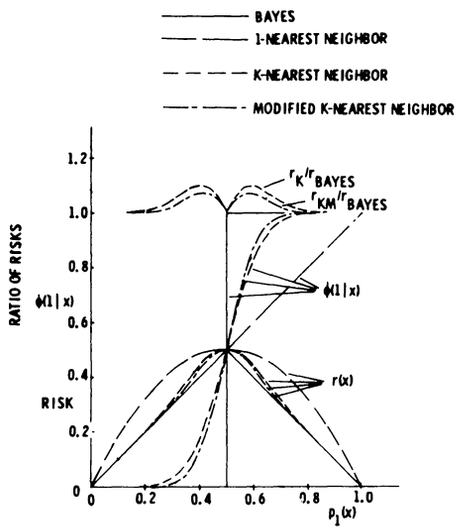

Fig. 9. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 19, 20$.

rule and $K = 1$ the maximum value of the ratio is 1.20. For the modified three-nearest neighbor rule the maximum value of the ratio is 1.149. For the modified twentieth-nearest neighbor rule the maximum value of the ratio has decreased only to 1.066. This result suggests that perhaps the additional complexity required to use a larger number of neighbors than three is not warranted due to the small decrease in the error rate when more than three are used. (Cover and Hart [7] developed an expression for the maximum value of the asymptotic local risk compared to the Bayes' risk for the K-nearest neighbor rule.)

IV. Conclusions

The results presented here have demonstrated that for a large class of problems the nearest neighbor rules form a set of very powerful decision rules. The modified three-nearest neighbor rule which uses the three-nearest neighbor rule to edit the preclassified samples and then uses a single-nearest neighbor rule to make decisions is a particularly attractive rule. The results shown here have indicated that the modified three-nearest neighbor rule has an asymptotic performance which is difficult to differentiate from the performance of a Bayes' rule in many situations. The modified three-nearest neighbor rule improves considerably on the performance of the single-nearest neighbor rule and the modified single-nearest neighbor rule. On the other hand, it has been suggested that only a few preclassified samples are required to approach the asymptotic performance quite closely for the modified three-nearest neighbor rule, many fewer samples than are required to approach the asymptotic performance for using five or more nearest neighbors.

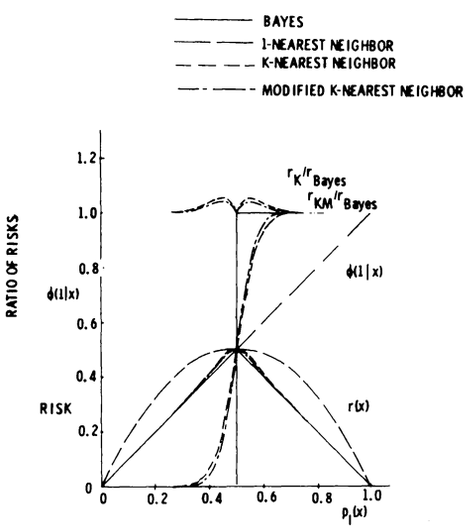

Fig. 10. Comparison of $\phi(1 \mid x)$ and local risk as function of $p_1(x)$ when losses are zero or one and number of preclassified samples is large. $K = 49, 50$.

Appendix I

Convergence in Probability

The convergence properties of random variables is one of the major branches of statistics. Loeve [11] discusses many different kinds of convergence for random variables and random functions. The type of convergence that will be considered here is convergence in probability. Simplification of the concept of "in probability" was begun by Mann and Wald [12] in 1943 with the development of the relationship of the operations that could be performed in determining convergence of sequences to the operations that could be performed in determining convergence of sequences of random variables. Chernoff [13] continued this development in his consideration of large sample problems. Chernoff's ideas were simplified and generalized by Pratt [14]. Using Pratt's concept of "in probability" leads to simple proofs of theorems. In particular, Theorems 1 and 1′ of Section III, which were conveyed to the author along with an outline of the proofs by Chernoff in his classes on large sample theory, are proved in this very simple manner.

Suppose, for $n = 1,2,\cdots$, P_n is the distribution of the random variable X_n in the set $\mathbf{X}_n$. That is, $P_n[X_n \in S_n] = P_n[S_n]$ is a probability measure on the measurable sets S_n

of X_n. If S_n is a measurable subset of X_n, the event $X_n \in S_n$ will be called an "X_n-event" E_n. If S is any subset of the product space $X = \times_{n=1}^{\infty} X_n$, the event $(X_1,X_2,\cdots) \in S$ will be called an "$(X_1,X_2,\cdots)$-event" E.

Definition: The $(X_1,X_2,\cdots)$-event E will be said to occur "in probability," written $\mathcal{P}(E)$, if for every positive ε, there exist X_n-events E_n of probability at least $1 - \varepsilon$ such that E occurs whenever all E_n occur.

Pratt [14] discusses the advantage of this definition and shows the relationship of the foregoing definition to the standard definition of "in probability."

Suppose $\{x_n\},\{r_n\}$ are sequences of points on the extended real line.

Definition: $x_n = o(r_n)$ if, for every positive η, for some N, for every $n > N$, $|x_n/r_n| \leq \eta$.

Definition: $x_n = O(r_n)$ if for some η and N, for every $n > N$, $|x_n/r_n| \leq \eta$.

Using Pratt's definition of "in probability" convergence in probability is defined as follows.

Definition: $X_n = o_p(r_n)$ if $\mathcal{P}(S)$, where $S = \{x: x_n = o(r_n)\}$.

Definition: $X_n = O_p(r_n)$ if $\mathcal{P}(S)$, where $S = \{x: x_n = O(r_n)\}$.

Pratt uses these concepts to prove a number of theorems about convergence in probability. The theorem of interest here is as follows.

Theorem 5 (Pratt [14, theorem 5])

Suppose that

$$f_n^{(j)}(X_n) = O_p(r_n^{(j)}), \qquad j = 1,\cdots,J$$
$$g_n^{(k)}(X_n) = o_p(s_n^{(k)}), \qquad k = 1,\cdots,K$$

and that $h_n(X_n) = O(t_n)$ whenever

$$f_n^{(j)}(x_n) = O(r_n^{(j)}), \qquad j = 1,\cdots,J$$
$$g_n^{(k)}(x_n) = o(s_n^{(k)}), \qquad k = 1,\cdots,K.$$

Then it follows that $h_n(X_n) = O_p(t_n)$. Furthermore, if $O(t_n)$ is replaced by $o(t_n)$ in the hypothesis, the conclusion is $h_n(X_n) = o_p(t_n)$.

With a few additional definitions Theorem 5 can be used to prove two theorems central to the development of the asymptotic properties of nearest neighbor rules.

Definition: A sequence $\{y_n\}$ is restrained from a set D if there is an open set $U \supset D$ such that $y_n \in U^c$ for n sufficiently large.

Definition: Y_n is restrained from D if $\mathcal{P}(S)$, where $S = \{x: y_n = f_n(x)$ is restrained from $D\}$.

Definition: $Y_n = f_n(X_n)$ converges in probability to c ($Y_n \xrightarrow{P} c$) if $\mathcal{P}(S)$, where $S = \{x: f_n(x_n) \to c\}$.

Theorem 1' (Slutsky's Theorem)

If $Y_n = f_n(X_n) \xrightarrow{P} c$ and g is continuous at c, then $g(Y_n) \xrightarrow{P} g(c)$.

Proof: $Y_n \xrightarrow{P} c$ implies $Y_n - c = o_p(1)$ from the definitions. Let $g_n^{(1)}(X_n) = f_n(X_n) - c$. Applying Theorem 5, it is only necessary to show that for a nonrandom sequence $y_n \to c$ implies $g(y_n) \to g(c)$. That $g(y_n) \to g(c)$ for c, a point of continuity of g is a simple proof from elementary analysis. The conclusion of the theorem follows directly.

In the next theorem identify (Y_n', Y_n'') with X_n of Theorem 5, where $Y_n' = f_n(X_n)$, $Y_n'' = f(X)$ establishes the relationship to the original measurable space. Let $Y_n = f_n(X_n)$, $Y = f(X)$.

Theorem 1

If $Y_n \xrightarrow{P} Y$, Y is finite with probability one and $P[Y \in Dg] = 0$, where Dg is the set of discontinuities of the function g, then $g(Y_n) \xrightarrow{P} g(Y)$.

Proof: Y finite with probability one implies $Y_n'' = O_p(1)$. $P[Y \in Dg] = 0$ implies Y_n'' is restrained from the discontinuities of g. It remains to show that $g(y_n') \to g(y_n'')$ at points of continuity of g whenever $y_n' \to y_n''$, with y_n'' bounded and y_n'' is restrained from Dg. For finite values of y_n'' the proof is an exercise in elementary analysis. Counterexamples are easily devised which show that it is not necessary that $g(y_n') \to g(y_n'')$ when y_n'' is not bounded. The conclusion of the theorem follows directly.

Appendix II

Convergence of Nearest Neighbors Before Editing

Let (X_i,θ_i), $i = 1,2,\cdots,N$, be independent random variables identically distributed as in Section I. Let $X^{[i]}(X,N)$ be the neighbor which is the ith distant neighbor from X when there are N preclassified samples. Let $L_N = o(N)$.

Theorem 2

For $L_N = o(N)$, $X^{[L_N]}(X,N) \xrightarrow{P} X$ as $N \to \infty$.

Proof: To show $X^{[L_N]}(X,N) \xrightarrow{P} X$ show for $\varepsilon > 0$ (we have dropped the explicit indication of the dependence of $X^{[L_N]}(X,N)$ on the X and N since the context indicates the dependence),

$$P\|X - X^{[L_N]}\| \geq \varepsilon \to 0 \text{ as } N \to \infty \text{ (definition of } \xrightarrow{P})$$

where $\|x - y\|$ is the distance between x and y. For random variables defined on E^d,

$$P[\|X - X^{[L_N]}\| \geq \varepsilon] = P[Z_{[L_N]}{}^* \geq \varepsilon]$$

by definition of $Z_{[L_N]}{}^*$. Consider a point $X = x$ for which there does not exist a neighborhood $S(x)$ such that $P[S] = 0$. Let

$$Y_i(x) = \begin{cases} 0, & \text{when } Z_i(x) \geq \varepsilon \\ 1, & \text{when } Z_i(x) < \varepsilon. \end{cases}$$

Then

$$P[Z_{[L_N]}{}^* \geq \varepsilon \mid X = x] = P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq \frac{L_N - 1}{N}\right].$$

Let

$$p \equiv F_Z(\varepsilon \mid x) = \eta_1 F_Z{}^{(1)}(\varepsilon \mid x) + \eta_2 F_Z{}^{(2)}(\varepsilon \mid x).$$

That there does not exist a neighborhood S such that $P[S] = 0$ implies $F_Z(\varepsilon \mid x) > 0$. Thus $p > 0$. There exists b, $0 < b < p$ since $p > 0$. The $Y_i(x)$ are independent identically distributed binary random variables. The Chernoff bound [13] for

$$P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq b\right], \qquad \text{when } 0 < b < p$$

has been developed in [12, p. 102]. This reference shows

$$P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq b\right] \leq \exp\left[-N(T_p(b) - H(b))\right]$$

where

$$T_p(b) = -b \ln p - (1 - b) \ln (1 - p)$$

$$H(b) = -b \ln p - (1 - b) \ln (1 - b).$$

Both $T_p(b)$ and $H(b)$ are well-known functions. It is known that $T_p(b) - H(b) \geq 0$. (See [12].) Since $L_N/N \to 0$, for N larger than some N_0,

$$\frac{L_N - 1}{N} \leq b.$$

Therefore,

$$P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq \frac{L_N - 1}{N}\right] \leq P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq b\right]$$

for N greater than N_0. But

$$\exp\left[-N(T_p(b) - H(b))\right] \to 0 \text{ as } N \to \infty.$$

Therefore,

$$P\left[\frac{1}{N}\sum_{i=1}^{N} Y_i \leq \frac{L_N - 1}{N}\right] \to 0 \text{ as } N \to \infty.$$

We have shown that for a sample value $X = x$, the nearest neighbor converges. It remains to show that the random variable X has this property with probability one. We shall do this by showing that the set T of points which do not have this property has probability zero.

Let $S(x,r_x)$ be a sphere of radius r_x centered at x, where r_x is a rational number. Let T be the set of all x for which there exists a rational number r_x sufficiently small that $P[S(x,r_x)] = 0$. The space E^d is certainly a separable space. From the definition of separability of E^d there exists a countable dense subset of A of E^d. For each $x \in T$, there exists $a(x) \in A$ such that $a(x) \in S(x,r_x/3)$ since A is dense. By a simple geometric argument, there is a sphere centered at $a(x)$ with radius $r_x/2$ which is strictly contained in the original sphere $S(x,r_x)$ and which contains x. Thus $P[S(a(x),r_x/2)] = 0$.

The possibly uncountable set T is contained in the countable union of spheres $\bigcup_{x \in T} S(a(x),r_x/2)$. The probability of the countable union of sets of probability zero is zero. Since $T \subset \bigcup_{x \in T} S(a(x),r_x/2)$, $P[T] = 0$, as was to be shown.

Appendix III

Convergence of Nearest Neighbor after Editing

Let (X_i,θ_i), $i = 1,2,\cdots,N$, be independent random variables identically distributed as in Section I. Let (X_i,θ_i) be edited as for the modified nearest neighbor rule. That is,

1) find the K-nearest neighbors to X_i among

$$\{X_1,X_2,\cdots,X_{i-1},X_{i+1},\cdots,X_N\}$$

2) find the class θ associated with the largest number of points among the K-nearest neighbors, breaking ties randomly when they occur;

3) edit the set $\{(X_i,\theta_i)\}$ by deleting (X_i,θ_i) whenever θ_i does not agree with the largest number of K-nearest neighbors as determined in the preceding.

Let there be M classes, $m = 1,2,\cdots,M$. In particular, consider $M = 2$. (The proof is general enough to cover any finite value of M, but for consistency with the remainder of the paper M is considered to be 2.) Then

$$f(x) = \sum_{m=1}^{M} \eta_m f_m(x)$$

$$p_m(x) = P(\theta = m \mid X = x) = \frac{\eta_m f_m(x)}{f(x)}.$$

Let $X_{EK}^{[1]}(x_0,N)$ be the nearest neighbor to x_0 after editing has been performed as outlined in the foregoing.

Theorem 3

If $P[X \in Df_m] = 0$, $m = 1,2,\cdots,M$, where Df_m is the set of discontinuities of $f_m(x)$, then $X_{EK}^{[1]}(X,N) \xrightarrow{P} X$ as $N \to \infty$.

Proof: The proof is carried out by first examining points x_0 such that the $f_m(x)$ are continuous at x_0 and $f(x_0) > 0$. For these points the following statements are proved.

1) There is an f such that $f(x) \geq f > 0$ for all x lying within a hypersphere of radius ε centered at x_0.

2) $N^{1/2}$ nonintersecting hyperspheres of radius $\varepsilon/2N^{1/2d}$ can be placed within the hypersphere of radius ε centered at x_0.

3) As N grows large, the probability that a sample point may be found within a hypersphere of radius $\varepsilon/4N^{1/2d}$ concentric to each of the hyperspheres of radius $\varepsilon/2N^{1/2d}$ for all $N^{1/2}$ such spheres approaches one. (The fact that $N^{1/2}$ may not be an integer will be ignored since it makes no difference to the proof, and the details necessary to find an integer near to $N^{1/2}$ will obscure an already complicated problem.)

4) As N grows large, the probability that at least K neighbors to such a sample point are located within a radius $\varepsilon/4N^{1/2d}$ of the sample point approaches one.

5) When the K neighbors of one sample point are within a hypersphere not intersecting a similar hypersphere containing another sample point with its K neighbors, the probability of retention in the edited set is independent for the two sample points.

6) As N grows large, the probability of at least one point being retained in the edited set approaches one.

7) Finally, the set of points which do not have this property is shown to have probability zero.

Details of Proof: (In what follows $X_{EK}^{[1]}$ will be used for $X_{EK}^{[1]}(x_0,N)$ and $X_{EK}^{[1]}(X,N)$ depending upon the context.) From the definition of convergence in probability $X_{EK}^{[1]} \xrightarrow{P} x_0$ whenever for every $\varepsilon > 0$

$$P[|X_{EK}^{[1]} - x_0| \geq \varepsilon] \to 0 \text{ as } N \to \infty.$$

1) The continuity of $f_m(x)$, $m = 1,2,\cdots,M$ implies that

$$f(x) = \sum_{m=1}^{M} \eta_m f_m(x)$$

is continuous. Also, $p_m(x)$ is continuous since $p_m(x)$ is a simple continuous function of $f_m(x)$ and $f(x)$. (The proof is trivial.) Continuity implies that for every $\hat{\varepsilon} > 0$ there exists

$\delta_j > 0$ such that for $m = 1,2,\cdots,M$,

$$|f_m(x) - f_m(x_0)| < \hat{\varepsilon}, \quad \text{whenever } |x - x_0| < \delta_m$$
$$|p_m(x) - p_m(x_0)| < \hat{\varepsilon}, \quad \text{whenever } |x - x_0| < \delta_{M+m}$$
$$|f(x) - f(x_0)| < \hat{\varepsilon}, \quad \text{whenever } |x - x_0| < \delta_{2M+1}.$$

Select $\delta = \min(\delta_1,\delta_2,\cdots,\delta_{2M+1})$. Then whenever $|x - x_0| < \delta$, all of the preceding quantities are less than $\hat{\varepsilon}$. Select $\hat{\varepsilon}$ such that $0 < \hat{\varepsilon} < f(x_0)$ and $0 < \hat{\varepsilon} < 1/2M$. This selection can be done since $f(x_0)$ and $1/2M$ are positive. Let $\underline{f} = f(x_0) - \hat{\varepsilon}$. Then $|\underline{f}| = \underline{f} > 0$ since $\hat{\varepsilon} < f(x_0)$. But $|f(x) - f(x_0)| < \hat{\varepsilon}$ whenever $|x - x_0| < \delta$ implies that $f(x) > \underline{f}$ whenever $|x - x_0| < \delta$. Thus $\underline{f} > 0$ is a lower bound on $f(x)$ whenever $|x - x_0| < \delta$. Similarly, $\bar{f} = f(x_0) + \hat{\varepsilon}$ is an upper bound. Since

$$|X_{EK}^{[\ ]} - x_0| \geq \varepsilon, \qquad \varepsilon \geq \delta$$

implies that

$$|X_{EK}^{[1]} - x_0| \geq \varepsilon$$

when $\varepsilon < \delta$, proving that

$$P[|X_{EK}^{[1]} - x_0| \geq \varepsilon] \to 0 \text{ as } N \to \infty$$

for every ε such that $0 < \varepsilon < \delta$ is adequate to prove that

$$P[|X_{EK}^{[1]} - x_0| \geq \varepsilon] \to 0 \text{ as } N \to \infty$$

for any ε. The proof is continued on that basis.

2) The region $|x - x_0| < \varepsilon$ defines a hypersphere with a volume in d dimensions of

$$V_d(\varepsilon) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} \varepsilon^d.$$

Lemma: At least $N^{1/2}$ nonintersecting hyperspheres of radius $\varepsilon/2N^{1/2d}$ can be placed within a hypersphere of radius ε.

Proof: When as many nonintersecting hyperspheres as can be packed in randomly have been placed in the hypersphere of radius ε, there is no point in the ε-radius hypersphere such that a small hypersphere cannot be found within a distance equal to the radius of the small sphere. If there were such a point, another small hypersphere could be placed within the large one by centering a new small hypersphere at the point so located. But if there is no point such that a small hypersphere cannot be found within a distance equal to the radius of the small hypersphere, then concentric spheres having twice the radius of the small spheres will cover all of the points of the large hypersphere. These double-radius hyperspheres may be intersecting. However, the total volume covered by the double-radius hyperspheres cannot be more than the sum of the volumes of each of the individual double-radius hyperspheres. The sum of the volumes of $N^{1/2}$ hyperspheres of radius $\varepsilon/N^{1/2d}$ is

$$N^{1/2} \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} \left(\frac{\varepsilon}{N^{1/2d}}\right)^d = \frac{\pi^{d/2}\varepsilon^d}{\Gamma(d/2 + 1)}.$$

The last quantity is identified as the volume of the hypersphere of radius ε. Thus $N^{1/2}$ hyperspheres of radius $\varepsilon/N^{1/2d}$ have a combined volume at most equal to the volume of the hypersphere of radius ε, and at least $N^{1/2}$ hyperspheres of radius $\varepsilon/2N^{1/2d}$ can be placed within the hypersphere of radius ε. Q.E.D.

Let $U^{(j)} = 0$ when there is not a point X_i in the hypersphere with radius $\varepsilon/4N^{1/2d}$ concentric to the jth hypersphere of radius $\varepsilon/2N^{1/2d}$. Let $U^{(j)} = 1$ when there is a point in the hypersphere. When $U^{(j)} = 1$ select one of the points within the hypersphere of radius $\varepsilon/4N^{1/2d}$, and let that point be $X^{(j)}$. Let

$$Y_i^{(j)} = \begin{cases} 0, & \text{when } |X_i - X^{(j)}| \geq \dfrac{\varepsilon}{4N^{1/2d}} \\ 1, & \text{when } |X_i - X^{(j)}| < \dfrac{\varepsilon}{4N^{1/2d}}. \end{cases}$$

Let $E_1 = 1$ whenever

a) for all j, $U^{(j)} = 1$;
b) for all j, $\sum_{i=1}^N Y_i^{(j)} \geq K + 1$; and
c) at least one $X^{(j)}$ has an associated $\theta^{(j)}$ which agrees with the largest number of the K-nearest neighbors to $X^{(j)}$.

Let $E_1 = 0$ otherwise. When $E_1 = 1$ at least one point is retained in the hypersphere of radius ε after the editing process. (There is an i for which $X_i = X^{(j)}$. For that i, $Y_i^{(j)} = 1$, but $X^{(j)}$ is not one of the K-nearest neighbors to $X^{(j)}$ used in the editing process. Thus it is required that $\sum_{i=1}^N Y_i^{(j)} \geq K + 1$ in order that the K-nearest neighbors to $X^{(j)}$ lie within a radius $\varepsilon/4N^{1/2d}$ of the point $X^{(j)}$.)

Also, when

a) for all j, $U^{(j)} = 1$; and
b) for all j, $\sum_{i=1}^N Y_i^{(j)} \geq K + 1$

none of the K neighbors used in editing one point $X^{(j)}$ is used in editing any other point $X^{(j')}$ since the conditions on $U^{(j)}$ and the sum of the $Y_i^{(j)}$ imply that at least K of the nearest neighbors to each point $X^{(j)}$ lie within the hyperspheres of radius $\varepsilon/2N^{1/2d}$ which are nonintersecting.

3) Developing inequalities and using them in the proof:

$$P[|X_{EK}^{[1]} - x_0| \geq \varepsilon] \leq P[E_1 = 0]$$

since the one event implies the other. Continuing:

$$\begin{aligned} P[E_1 = 0] &= P[E_1 = 0 \mid \text{for all } j,\ U^{(j)} = 1] \\ &\quad \cdot P[\text{for all } j,\ U^{(j)} = 1] \\ &\quad + P[E_1 = 0 \mid \text{for some } j,\ U^{(j)} = 0] \\ &\quad \cdot P[\text{for some } j,\ U^{(j)} = 0] \\ &\leq P[E_1 = 0 \mid \text{for all } j,\ U^{(j)} = 1] \\ &\quad + P[\text{for some } j,\ U^{(j)} = 0] \end{aligned}$$

since probabilities are less than or equal to one.

Lemma:

$$P[\text{for some } j,\ U^{(j)} = 0] \to 0 \text{ as } N \to \infty.$$

Proof:

$$P[U^{(j)} = 0] = \left(1 - \int_{S(j,\varepsilon/4N^{1/2d})} f(x)\,dx\right)^N$$

where $S(j, \varepsilon/4N^{1/2d})$ indicates that the integral is taken over the volume of a hypersphere with radius $\varepsilon/4N^{1/2d}$ concentric with the jth hypersphere of radius $\varepsilon/2N^{1/2d}$ ($P[U^{(j)} = 0]$ is the probability that no sample lies within the jth hypersphere from the definition of $U^{(j)}$). This is true since each of N independent samples X_i may be within the specified volume with equal probability. But $f(x) \geq \underline{f}$ in the region occupied by the hypersphere implies that

$$\int_{S(j,\varepsilon/4N^{1/2d})} f(x)\,dx \geq \int_{S(j,\varepsilon/4N^{1/2d})} \underline{f}\,dx.$$

Thus

$$\left(1 - \int f(x)\,dx\right)^N \leq \left(1 - \int \underline{f}\,dx\right)^N = \left(1 - \underline{f} V_d\left(\frac{\varepsilon}{4N^{1/2d}}\right)\right)^N$$

where $V_d(\alpha)$ is the volume of a hypersphere with radius α. Then

$$\begin{aligned} &P[\text{for some } j,\ U^{(j)} = 0] \\ &\quad = P[\text{either } U^{(1)} = 0 \text{ or } U^{(2)} = 0 \text{ or } \cdots \text{ or } U^{(N^{1/2})} = 0] \\ &\quad \leq \sum_{j=1}^{N^{1/2}} P[U^{(j)} = 0] \\ &\quad \leq N^{1/2}\left(1 - \underline{f} V_d\left(\frac{\varepsilon}{4N^{1/2d}}\right)\right)^N. \end{aligned}$$

But

$$V_d\left(\frac{\varepsilon}{4N^{1/2d}}\right) = \frac{\pi^{d/2}\varepsilon^d}{\Gamma(d/2 + 1)4^d N^{1/2}}.$$

Let

$$\underline{C}_d = \underline{f}\,\frac{\pi^{d/2}\varepsilon^d}{\Gamma(d/2 + 1)4^d}.$$

Note that $\underline{C}_d > 0$ since it is the product of strictly positive quantities:

$$\sum_{j=1}^{N^{1/2}} P[U^{(j)} = 0] \leq N^{1/2}\left(1 - \frac{\underline{C}_d}{N^{1/2}}\right)^N$$

and

$$\begin{aligned} &N^{1/2}\left(1 - \frac{\underline{C}_d}{N^{1/2}}\right)^N \\ &\quad = \exp\left(\frac{1}{2}\ln(N) + N\ln\left(1 - \frac{\underline{C}_d}{N^{1/2}}\right)\right) \\ &\quad = \exp\left(\frac{1}{2}\ln(N) + N\left(-\frac{\underline{C}_d}{N^{1/2}} + O\left(\frac{1}{N}\right)\right)\right) \\ &\quad = O(1)\exp\left(\tfrac{1}{2}\ln(N) - N^{1/2}\underline{C}_d\right) \\ &\quad \to 0 \text{ as } N \to \infty. \end{aligned}$$

Therefore, $P[\text{for some } j,\ U^{(j)} = 0] \to 0$ as $N \to \infty$.

Q.E.D.

4) Continuing with the main theorem and recalling that $\sum_{i=1}^N Y_i^{(j)} \geq K + 1$ implies that K neighbors lie within the hypersphere of radius $\varepsilon/4N^{1/2d}$ centered at $X^{(j)}$,

$$\begin{aligned} &P[E_1 = 0 \mid \text{for all } j,\ U^{(j)} = 1] \\ &\quad = P\left[E_1 = 0 \mid \text{for all } j,\ \sum_{i=1}^N Y_i^{(j)} \geq K + 1 \text{ and for all } j,\ U^{(j)} = 1\right] \\ &\qquad \cdot P\left[\text{for all } j,\ \sum_{i=1}^N Y_i^{(j)} \geq K + 1 \mid \text{for all } j,\ U^{(j)} = 1\right] \\ &\qquad + P\left[E_1 = 0 \mid \text{for some } j,\ \sum_{i=1}^N Y_i^{(j)} < K + 1 \text{ and for all } j,\ U_{(j)} = 1\right] \\ &\qquad \cdot P\left[\text{for some } j,\ \sum_{i=1}^N Y_i^{(j)} < K + 1 \mid \text{for all } j,\ U^{(j)} = 1\right] \\ &\quad \leq P\left[E_1 = 0 \mid \text{for all } j,\ \sum_{i=1}^N Y_i^{(j)} \geq K + 1 \text{ and for all } j,\ U^{(j)} = 1\right] \\ &\qquad + P\left[\text{for some } j,\ \sum_{i=1}^N Y_i^{(j)} < K + 1 \mid \text{for all } j,\ U^{(j)} = 1\right] \end{aligned}$$

since probabilities are less than or equal to one.

Lemma:

$$P\left[\text{for some } j,\ \sum_{i=1}^N Y_i^{(j)} < K + 1 \mid \text{for all } j,\ U^{(j)} = 1\right] \to 0 \text{ as } N \to \infty.$$

Proof: Note that $Y_i^{(j)} = 1$ for the point which is $X^{(j)}$. Thus we must show that there are not K other points within a distance $\varepsilon/4N^{1/2d}$ of $X^{(j)}$ to show that $\sum_{i=1}^N Y_i^{(j)} < K + 1$. Note also that since X_i, $i = 1,2,\cdots,N$, are drawn independently from one population, the Y_i for one j are also independent. Considering only one j,

$$P\left[\sum_{i=1}^N Y_i^{(j)} < K + 1 \mid U^{(j)} = 1\right] = P\left[\frac{1}{N-1}\sum_{i=1}^N Y_i^{(j)} < \frac{K+1}{N-1} \mid U^{(j)} = 1\right].$$

Let

$$p = \int_{S(X^{(j)},\varepsilon/4N^{1/2d})} f(x)\,dx$$

where $S(X^{(j)}, \varepsilon/4N^{1/2d})$ indicates that the integral is to be taken over the volume of a hypersphere with radius $\varepsilon/4N^{1/2d}$ centered at $X^{(j)}$. p has an upper and a lower bound:

$$0 < \underline{p} = \int_{S(X^{(j)},\varepsilon/4N^{1/2d})} \underline{f}\,dx \leq p \leq \bar{p} = \int_{S(X^{(j)},\varepsilon/4N^{1/2d})} \bar{f}\,dx$$

since $0 < \underline{f} \leq f(x) \leq \bar{f}$ for x such that $|x - x_0| < \varepsilon$. Evaluating

$$\underline{p} = \int \underline{f}\, dx = \underline{f} \int dx = \underline{f} V_d \left(\frac{\varepsilon}{4N^{1/2d}}\right)$$

and

$$\bar{p} = \int \bar{f}\, dx = \bar{f} \int dx = \bar{f} V_d \left(\frac{\varepsilon}{4N^{1/2d}}\right).$$

For N large enough

$$0 < \frac{K+1}{N-1} < \underline{p} = \frac{\pi^{d/2}\varepsilon^d \underline{f}}{\Gamma(d/2+1)4^d N^{1/2}} < p = E(Y_i^{(j)})$$

except for the i for which $X_i = X^{(j)}$. Thus for N large enough Chernoff's bound can be applied:

$$P\left[\frac{1}{N-1}\sum_{i=1}^{N} Y_i^{(j)} < \frac{K+1}{N-1} \mid U^{(j)} = 1\right] \leq \exp\left[-(N-1)\left(T_p\left(\frac{K}{N-1}\right) - H\left(\frac{K}{N-1}\right)\right)\right]$$

where

$$T_p(a) = -a \ln p - (1-a)\ln(1-p)$$
$$H(a) = -a \ln a - (1-a)\ln(1-a).$$

Let

$$\underline{C}_d = \frac{\pi^{d/2}\varepsilon^d \underline{f}}{\Gamma(d/2+1)4^d}$$

and $\bar{C}_d$ be similarly defined by substituting $\bar{f}$ for $\underline{f}$ in the definition. Using the upper and lower limits developed,

$$\exp\left[-(N-1)\left(T_p\left(\frac{K}{N-1}\right) - H\left(\frac{K}{N-1}\right)\right)\right] \leq \exp\left[(N-1)\left\{\frac{K}{N-1}\left(\ln\left(\frac{\bar{C}_d}{K}\right) + \ln\left(\frac{N-1}{N^{1/2}}\right)\right) + \left(1 - \frac{K}{N-1}\right)\left(\ln\left(1 - \frac{\underline{C}_d}{N^{1/2}}\right) - \ln\left(1 - \frac{K}{N-1}\right)\right)\right\}\right].$$

Observing that for $\gamma > 0$ and γ small $\ln(1-\gamma) = -\gamma + O(\gamma^2)$ and carrying out some algebra, we modify the last expression to

$$\exp\left[-(N-1)\left(T_p\left(\frac{K}{N-1}\right) - H\left(\frac{K}{N-1}\right)\right)\right] = O(1)\exp\left[\frac{K}{2}\ln(N) - N^{1/2}\underline{C}_d\right].$$

Thus

$$P\left[\sum_{i=1}^{N} Y_i^{(j)} < K+1 \mid U^{(j)} = 1\right] \leq O(1)\exp\left[\frac{K}{2}\ln(N) - N^{1/2}\underline{C}_d\right].$$

But

$$\begin{aligned} &P\left[\text{for some } j, \sum_{i=1}^{N} Y_i^{(j)} < K+1 \mid \text{for every } j,\ U^{(j)} = 1\right] \\ &= P[\textstyle\sum Y_i^{(1)} < K+1 \text{ or } \sum Y_i^{(2)} < K+1 \text{ or } \cdots \\ &\quad \text{or } \textstyle\sum Y_i^{(N^{1/2})} < K+1 \mid \text{for all } j,\ U^{(j)} = 1] \\ &\leq \sum_{j=1}^{N^{1/2}} P[\textstyle\sum Y_i^{(j)} < K+1 \mid U^{(j)} = 1] \\ &\leq \sum_{j=1}^{N^{1/2}} O(1)\exp\left[\frac{K}{2}\ln(N) - N^{1/2}\underline{C}_d\right] \\ &= N^{1/2}\, O(1)\exp\left[\frac{K}{2}\ln(N) - N^{1/2}\underline{C}_d\right] \\ &= O(1)\exp\left[\frac{K+1}{2}\ln(N) - N^{1/2}\underline{C}_d\right] \\ &\to 0 \text{ as } N \to \infty. \end{aligned}$$

Therefore,

$$P\left[\text{for some } j, \sum_{i=1}^{N} Y_i^{(j)} < K+1 \mid \text{for every } j,\ U^{(j)} = 1\right] \to 0 \text{ as } N \to \infty. \quad \text{Q.E.D.}$$

5) Continuing with the main proof, let

$$V^{(j)} = \begin{cases} 1, & \text{when } X^{(j)} \text{ is retained} \\ 0, & \text{otherwise.} \end{cases}$$

Let $V = (V^{(1)}, V^{(2)}, \cdots, V^{(N^{1/2})})$

$$\begin{aligned} &P\left[E_1 = 0 \mid \text{for all } j, \sum_{i=1}^{N} Y_i^{(j)} \geq K+1 \text{ and for all } j,\ U^{(j)} = 1\right] \\ &= P[V = (0,0,\cdots,0) \mid \text{for all } j, \textstyle\sum Y_i^{(j)} \geq K+1 \text{ and for all } j,\ U^{(j)} = 1]. \end{aligned}$$

But under the conditioning all of the $V^{(j)}$ are independent since each $V^{(j)}$ depends only on the values of θ associated with sample points within the jth hypersphere of radius $\varepsilon/2N^{1/2d}$, and none of these hyperspheres intersect another such hypersphere. Therefore,

$$\begin{aligned} &P[V = (0,0,\cdots,0) \mid \text{for all } j, \textstyle\sum Y_i^{(j)} \geq K+1 \text{ and for all } j,\ U^{(j)} = 1] \\ &= \prod_{j=1}^{N^{1/2}} P[V^{(j)} = 0 \mid \text{for all } j, \textstyle\sum Y_i^{(j)} \geq K+1 \text{ and for all } j,\ U^{(j)} = 1]. \end{aligned}$$

6) Lemma:

$$P[V^{(j)} = 0 \mid \text{for all } j, \textstyle\sum Y_i^{(j)} \geq K+1 \text{ and for all } j,\ U^{(j)} = 1] < 1 - \gamma, \qquad \gamma > 0.$$

Proof: The lemma states that when there is a sample in every one of the j hyperspheres and when the K-nearest neighbors to each of the samples also lie within the respective hyperspheres, the probability that $X^{(j)}$ is not retained

is less than one. The equivalent statement that the probability of retaining $X^{(j)}$ is greater than zero will be proved.

Let C_j be the event that $\sum Y_i^{(j)} \geq K + 1$ and $U^{(j)} = 1$. Also, let

$$E_2(m) = \begin{cases} 1, & \text{when the plurality of the } K\text{-nearest} \\ & \quad \text{neighbors is from class } m \\ 0, & \text{otherwise} \end{cases}$$

and

$$P[V^{(j)} = 1 \mid C_j] = \sum_{m=1}^{M} P[V^{(j)} = 1 \mid C_j \text{ and } E_2(m) = 1] \cdot P[E_2(m) = 1 \mid C_j].$$

Let $\hat{m}$ be the class that has the largest probability of having a plurality. (Ties are broken randomly.) Then

$$P[V^{(j)} = 1 \mid C_j] \geq P[V^{(j)} = 1 \mid C_j \text{ and } E_2(\hat{m}) = 1] \cdot P[E_2(\hat{m}) = 1 \mid C_j]$$

since the sum is a sum of positive quantities. But

$$P[E_2(\hat{m}) = 1 \mid C_j] \geq \frac{1}{M}.$$

Since probabilities sum to one, there are M classes, and $\hat{m}$ is the class with the greatest probability. Therefore,

$$P[V^{(j)} = 1 \mid C_j] \geq P[V^{(j)} = 1 \mid C_j \text{ and } E_2(\hat{m}) = 1] \cdot \frac{1}{M}.$$

But

$$P[E_2(\hat{m}) = 1 \mid C_j] \geq \frac{1}{M}$$

implies that for some x in the hypersphere of radius $\varepsilon/4N^{1/2d}$ centered at $X^{(j)}$

$$P_{\hat{m}}(x) \equiv P(\theta = m \mid X = x) \geq \frac{1}{M}$$

since $P[E_2(\hat{m}) = 1 \mid C_j]$ is an average of $p_{\hat{m}}(x)$. However, $p_{\hat{m}}(x)$ is a continuous function of x. In particular, the selection of δ in step 1) guarantees

$$|p_{\hat{m}}(x) - p_{\hat{m}}(x_0)| \leq \frac{1}{2M}, \quad \text{for } |x - x_0| < \varepsilon.$$

Therefore, $p_{\hat{m}}(x^{(j)}) > 0$ since

$$\begin{aligned} |p_{\hat{m}}(x) - p_{\hat{m}}(x^{(j)})| &= |p_{\hat{m}}(x) - p_{\hat{m}}(x_0) + p_{\hat{m}}(x_0) - p_{\hat{m}}(x^{(j)})| \\ &\leq |p_{\hat{m}}(x) - p_{\hat{m}}(x_0)| + |p_{\hat{m}}(x_0) - p_{\hat{m}}(x^{(j)})| \\ &< \frac{1}{2M} + \frac{1}{2M} = \frac{1}{M}. \end{aligned}$$

Thus there exists a $\gamma > 0$ such that

$$P[V^{(j)} = 1 \mid C_j \text{ and } E_2(\hat{m}) = 1] \geq M\gamma > 0$$

since this probability is an average of $p_{\hat{m}}(x^{(j)})$ over the possible values of $x^{(j)}$ and $p_{\hat{m}}(x^{(j)}) > 0$. This implies

$$P[V^{(j)} = 1 \mid C_j] \geq M\gamma \cdot \frac{1}{M} = \gamma > 0$$

and finally

$$P[V^{(j)} = 0 \mid C_j] \leq 1 - \gamma. \qquad \text{Q.E.D.}$$

Completing the main proof by using the results of the lemmas,

$$\begin{aligned} P[V &= (0,0,\cdots,0) \mid \text{for all } j,\ C_j] \\ &= \prod_{j=1}^{N^{1/2}} P[V^{(j)} = 0 \mid C_j] \\ &\leq \prod_{j=1}^{N^{1/2}} (1 - \gamma) = (1 - \gamma)^{N^{1/2}} \\ &\to 0 \text{ as } N \to \infty. \end{aligned}$$

Combining all of the inequalities,

$$\begin{aligned} P[|X_{\mathrm{EK}}^{[1]} - x_0| \geq \varepsilon] &\leq P[\text{for some } j,\ U^{(j)} = 0] \\ &\quad + P\left[\text{for some } j,\ \sum_{i=1}^{N} Y_i^{(j)} < K + 1 \mid \text{for all } j,\ U^{(j)} = 1\right] \\ &\quad + P\left[V = (0,0,\cdots,0) \mid \text{for all } j,\ \sum_{i=1}^{N} Y_i^{(j)} < K + 1 \text{ and for all } j,\ U^{(j)} = 1\right] \\ &\to 0 \text{ as } N \to \infty \end{aligned}$$

since each of the three components on the right approaches 0 as $N \to \infty$.

7) Finally, we must argue that the set of points for which the edited nearest neighbor does not converge to the sample to be classified has probability zero. The argument is very similar to the argument of a theorem of Cover and Hart [7]. We have shown that for point of continuity of $f(x)$ for which $f(x)$ is greater than zero the edited nearest neighbor converges in probability. It remains to consider points for which $f(x)$ is not continuous or for which $f(x) = 0$. The set of discontinuities has measure zero by hypothesis. The set for which $f(x) = 0$ is more complicated.

Let $S(x,r_x)$ be a sphere of radius r_x centered at x, r_x a rational number. Let V be the set of all x such that there does not exist an r_x sufficiently small that $P[S(x,r_x)] = 0$, but for which $f(x) = 0$ and $f(x)$ is continuous. For $x \in V$ the edited nearest neighbor converges in probability. Since the set of discontinuities of $f(x)$ has probability zero and x is a point of continuity of $f(x)$, there must be a point t within $\varepsilon/3$ of the point x for which $f(t) > 0$ and which is a point of continuity of f. If not, $P[S(x,r_x)]$ would be zero for some r_x small enough. If there remains a preclassified sample within a distance $\varepsilon/3$ of the point t, there will be a preclassified sample within a distance ε of the point x by a simple geometric argument. We have shown that for points with the property of the point t, the probability of there being a preclassified sample within an arbitrarily small distance $\varepsilon/3$ approaches one as the number of preclassified samples approaches infinity. Therefore, the probability that

there will be at least one preclassified sample within a distance ε of the point x approaches one as the number of preclassified samples approaches infinity. If at least one preclassified sample is within ε, then the nearest preclassified sample is within ε, and the nearest preclassified sample after editing converges in probability to the point x.

Let T be the set of all x for which there exists an r_x sufficiently small so that $P[S(x,r_x)] = 0$. The set T has probability zero. Duplicating the argument of Theorem 2 we begin by observing that the space E^d is a separable space. From the definition of separability, there exists a countable dense subset A of E^d. For each $x \in T$ there exists $a(x) \in A$ such that $a(x) \in S(x,r_x/3)$ since A is dense. By a simple geometric argument there is a sphere centered at $a(x)$ with radius $r_x/2$ which is strictly contained in the original sphere $S(x,r_x)$ and which contains x. Thus $P[S(a(x),r_x/2)] = 0$. The possibly uncountable set T is contained in the countable union of spheres $\bigcup_{x \in T} S(a(x), r_x/2)$. The probability of the countable union of sets of measure zero is zero. Since $T \subset \bigcup_{x \in T} S(a(x), r_x/2)$, $P[T] = 0$, as was to be shown.

Acknowledgment

The author is deeply indebted to Dr. T. Cover and Dr. H. Chernoff for their many helpful suggestions and careful review of the results presented in this paper.

References

[1] E. Fix and J. L. Hodges, Jr., "Discriminatory analysis, nonparametric discrimination: consistency properties," U.S. Air Force Sch. Aviation Medicine, Randolf Field, Tex., Project 21-49-004, Contract AF 41(128)-31, Rep. 4, Feb. 1951.
[2] G. Sebestyen, *Decision-Making Processes in Pattern Recognition.* New York: Macmillan, 1962.
[3] N. Nilsson, *Learning Machines.* New York: McGraw-Hill, 1965.
[4] C. A. Rosen, "Pattern classification by adaptive machines," *Science*, vol. 156, Apr. 7, 1967.
[5] G. Nagy, "State of the art in pattern recognition," *Proc. IEEE*, vol. 56, pp. 836–862, May 1968.
[6] Y. C. Ho and A. K. Agrawala, "On pattern classification algorithms—introduction and survey," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 676–690, Dec. 1968.
[7] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.
[8] A. W. Whitney and S. J. Dwyer, III, "Performance and implementation of K-nearest neighbor decision rule with incorrectly identified training samples," in *Proc. 4th Annu. Allerton Conf. Circuit and System Theory*, 1966.
[9] T. M. Cover, in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic Press, 1969.
[10] E. A. Patrick and F. P. Fischer, III, "A generalized k-nearest neighbor rule," *Inform. Contr.*, vol. 16, pp. 128–152, Apr. 1970.
[11] M. Loeve, *Probability Theory*, 3rd ed. Princeton, N.J.: D. Van Nostrand, 1963.
[12] H. B. Mann and A. Wald, "On stochastic limit and order relationships," *Ann. Math. Statist.*, vol. 14, 1943.
[13] H. Chernoff, "Large sample theory: Parametric case," *Ann. Math. Statist.*, vol. 27, 1956.
[14] J. Pratt, "On a general concept of 'in probability'," *Ann. Math. Statist.*, vol. 30, June 1959.
[15] J. Wozencraft and I. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

End Points, Complexity, and Visual Illusions

DAVID J. PARKER, STUDENT MEMBER, IEEE, AND DOUGLAS J. H. MOORE, MEMBER, IEEE

***Abstract*—One aspect of a new theory of feature perception is considered. An algorithm is presented which can perceive and locate various features of a pattern by analyzing a statistic of the "chords" of the pattern. The procedure is illustrated by applying the algorithm to a pattern containing the Müller-Lyer figures. In measuring the length of the figures it is found that the algorithm has a visual illusion. A machine capable of executing the algorithm is described.**

I. Introduction

A LARGE NUMBER of proposals for computers that, in at least certain senses of the phrase, "recognize patterns" have been published in the past ten years. In spite of this the pattern recognition problem can still be said to be in its infancy. Levine [12], in his survey on feature extraction, stated that "the literature overwhelmingly concentrates on the various aspects of classification," even though David [4], in his review of the book by Sebestyen, raised the objection, "Is not the more significant part of the problem that of characterizing the world by a set of properties that provide the desired discrimination?" It would therefore seem that the problems of feature perception and extraction must be solved before any headway can be made with the pattern recognition problem. Nilsson [16], commenting on the subject of feature extraction, made the point that there exists no general theory which allows us to choose what features are relevant for a particular problem. He also pointed out that the design of feature extractors is empirical and uses many ad hoc strategies. It would seem from these comments that a completely new approach to feature extraction is necessary.

Moore [13] described a theory of feature perception and extraction. It was shown that the features of two-dimensional plane patterns could be perceived and extracted by analyzing the statistics of the chords of a pattern. The types of features that could be extracted included metric, angular, and topological structure. A two-dimensional retinal com-

Manuscript received April 19, 1971; revised November 19, 1971. This work was supported by the Department of Supply.
The authors are with the Department of Electrical Engineering, University of Newcastle, Newcastle, New South Wales, 2308, Australia.