A UNIVERSAL K-NEAREST NEIGHBOR PROCEDURE IN DISCRIMINATION

LUC P. DEVROYE

Mc Gill University
P.O. Box 6070
Montreal, Canada H3C 3G1

Abstract

The classical k-nearest neighbor rule's performance is scale dependent. The data dependent partitioning rules whose error probabilities are invariant under monotone transformations of the coordinate axes appear unnatural because of the discontinuities inherent to all partitioning rules. In this note a k-nearest neighbor rule based upon the notion of empirical distance is discussed that

(i) is Bayes risk consistent for all possible distributions of the observations,

(ii) has the desired invariance property.

Introduction

In several applications one would like, for physical or economical reasons, to feed the freshly collected data to some machine without preprocessing. In discrimination for instance, a mere rescaling of the coordinate axes has a serious impact on the performance of the popular k-nearest neighbor rule[1-4]. On the other hand, the k-nearest neighbor rule is appealing because of its simple definition, easy implementation and distribution-free properties for small samples[5 6]and large samples[7]. In this note we discuss a variation of the k-nearest neighbor rule whose performance (that is, the probability of error given the data) is not affected by monotone transformations of the coordinate axes. Such a modification would save the user a tedious preprocessing step and would save the manufacturer the trouble of adding input specifications to his discrimination machine.

The data we have collected can be regarded as a sequence of independent identically distributed random vectors : $D_n = (X_1,Y_1),\ldots,(X_n,Y_n)$. Here $X_1 \in R^d$ is called an observation and $Y_1 \in \{1,\ldots,M\}$ is called a state (class). If (X,Y) is distributed as (X_1,Y_1) and is independent of D_n, then in the k-nearest neighbor rule we estimate Y from X and D_n by $\tilde{Y}=g_n(X,D_n)$ where

$$\tilde{Y} \neq i \quad \text{whenever}$$

$$\sum_{1\leq j\leq k} I_{\{Y_j(X)=i\}} < \max_{1\leq \ell\leq M} \sum_{1\leq j\leq k} I_{\{Y_j(X)=\ell\}} , \qquad (1)$$

$(X_1(X),Y_1(X)),\ldots,(X_n(X),Y_n(X))$ is a reordering of the data according to increasing values of $|X_i-X|$ and I is the indicator function. The scaling of the coordinate axes influences the rule through the norm $|.|$. We will write L_n for the probability of error,

$$L_n = P\{\tilde{Y} \neq Y|\ D_n\}$$

and L* for the Bayes probability of error,

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

Stone[7] showed that $L_n \overset{n}{\to} L^*$ in probability for all possible distributions of (X,Y) whenever

$$k \overset{n}{\to} \infty \quad \text{and} \quad k/n \overset{n}{\to} 0. \qquad (2)$$

Thus, the k-nearest neighbor rule behaves asymptotically well without exceptions. To continue our discussion, let us replace the $X_i=(X_i^1,\ldots,X_i^d)$ and $X=(X^1,\ldots,X^d)$ by

$$\psi(X_i) = (\psi_1(X_i^1),\ldots,\psi_d(X_i^d))$$
and
$$\psi(X) = (\psi_1(X^1),\ldots,\psi_d(X^d)) \qquad (3)$$

where $\psi_1,\ldots,\psi_d$ are one-to-one monotone transformations from R to R. Examples are a dilation $\psi_1(u)=au$, a limiter $\psi_2(u)=u/(1+a|u|)$ and an expansion $\psi_3(u)=\exp(au)$. Writing D_n' for $(\psi(X_1),Y_1),\ldots,(\psi(X_n),Y_n)$ it is true that the Bayes probability of error $L^*(\psi)$ is independent of ψ :

$$L^*(\psi) = \inf_{g:R^d\to\{1,\ldots,M\}} P\{g(\psi(X))\neq Y\}= L^* \qquad (4)$$

and that the conditional probability of error with rule (1) and D_n' ,

$$L_n(\psi) = P\{g_n(\psi(X),D_n')\neq Y|D_n'\} ,$$

varies with ψ . Preprocessing of the data is partly concerned with finding a transformation ψ that tends to yield low values for $L_n(\psi)$. If such a ψ is a priori fixed, then Stone's theorem remains valid (that is, $L_n(\psi) \to L^*$ in probability if (2) holds). But if ψ is found after we have peeked at the data, can we still draw the same conclusion?

In essence, we are looking for a discrimination rule with the following properties :

(i) $L_n(\psi) = L_n$ for all monotone transformations (see (3)),

(ii) $L_n \overset{n}{\to} L^*$ in probability for all possible distributions of (X,Y).

The impact of (i) is that preprocessing via scaling or any other monotone transformation does not reduce the probability of error . While for finite n this may or may not be an advantage, (ii) assures that at least for large n the probability of error

Reprinted from *Proceedings of the Conference on Pattern Recognition and Image Processing,* May 1978, pages 142-147.

is close to the Bayes probability of error. In fact, (ii) implies that this asymptotically nice behavior is guaranteed for all possible (unknown) distributions of (X,Y). In the next section we develop a discrimination rule featuring (i-ii) which in form coincides with the k-nearest neighbor rule. A similar discrimination rule was recently suggested by Olshen in a comment on Stone's paper[7].

A Universal K-Nearest Neighbor Discrimination Rule

The k-nearest neighbor rule we are discussing is based upon the notion of empirical distance. Because in general X does not have a density, all sorts of situations involving ties can occur. For instance, the event $X_1^i=X_2^i$ can happen with positive probability. It is for this reason that we propose to enlarge the data by generating $Z,Z_1,\ldots,Z_n$, a sequence of independent random variables, independent of (X,Y,D_n), and all uniformly distributed on (0,1). The data $D_n=(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)$ is then reordered according to increasing values for the empirical distances $\rho((X_i,Z_i),(X,Z))$ between (X_i,Z_i) and (X,Z) defined below. Notice that from here on D_n denotes the enlarged data sequence. The reordered data sequence is $(X_1(X,Z),Y_1(X,Z),Z_1(X,Z))$ $,\ldots,(X_n(X,Z),Y_n(X,Z),Z_n(X,Z))$. Since ρ is integer-valued, we will use the values of $|Z_i-Z|$ to break ties in this reordering process. This has the same effect as reordering D_n according to increasing values of the adjusted empirical distance

$$\rho'((X_i,Z_i),(X,Z))=\rho((X_i,Z_i),(X,Z))+|Z_i-Z|.$$

To define ρ first order the (X_i,Z_i) and (X,Z) according to increasing values of their first components, $X_1^1,X_2^1,\ldots,X_n^1,X^1$. If ties occur, subsequently break them by considering increasing values for the Z_i or Z. Let r_i^1 be the rank of (X_i,Z_i) and let r^1 be the rank of (X,Z). Repeating the same procedure for the remaining d-1 components gives us values for

$$r_i^j\ ,\ r^j\ ,\ 1 \le j \le d,\ 1 \le i \le n.$$

We define

$$\rho((X_i,Z_i),(X,Z)) = \max_{1\le j\le d} |r_i^j-r^j|. \tag{5}$$

In Figure 1 the empirical distances are given in an example with no componentwise ties and d=2. The physical meaning of $\rho((X_i,Z_i),(X,Z))$ in that case is very simple : the empirical distance is the maximal number of lines in the d-dimensional grid generated by the X_i that one can cross or touch if one wants to go from X to X_i along one of the coordinate axes.

It is easy to check that for all transformations (3)

$$\rho((\psi(X_i),Z_i),(\psi(X),Z))=\rho((X_i,Z_i),(X,Z))$$

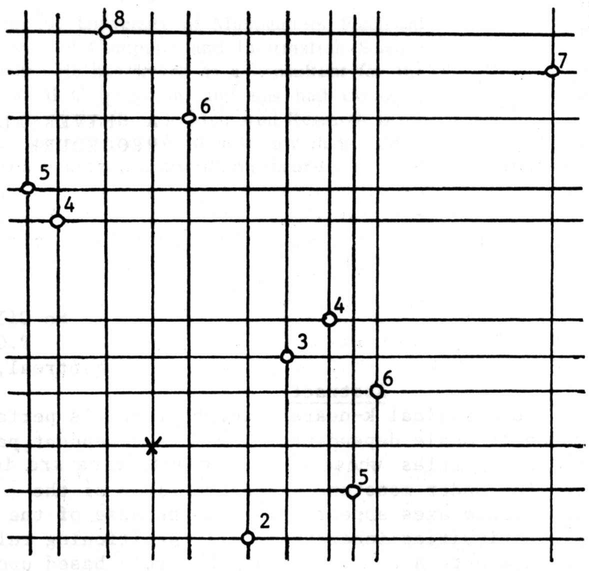

FIG. 1 : Numbers indicate the empirical distances between the X_i (small circles) and X (cross).

and that therefore the desired invariance property (i) holds true for all discrimination rules g_n that are functions of the empirical distances only.

Consider for instance the rules in which Y is estimated by $\tilde{Y} = g_n(X,Z,D_n)$ where

$$\tilde{Y} \ne i \quad \text{whenever} \tag{6}$$

$$\sum_{1\le j\le n} v_j I_{\{Y_j(X,Z)=i\}} < \max_{1\le \ell\le M} \sum_{1\le j\le n} v_j I_{\{Y_j(X,Z)=\ell\}},$$

and the weights $v_1,\ldots,v_n$ vary with n such that :

$$v_1 \ge v_2 \ge \ldots \ge v_n \ge 0\ ;\quad \sum v_i = 1\ ; \tag{7}$$

$$\sup_i\ v_i \xrightarrow{n} 0\ ; \tag{8}$$

$$\sum_{k\le i\le n} v_i \xrightarrow{n} 0 \quad \text{for some}\quad k \text{ with } k/n \xrightarrow{n} 0. \tag{9}$$

Picking the weights $v_1=v_2=\ldots=v_k=1/k$, and $v_i=0$ otherwise,yields a rule that equally weights the k nearest neighbors.

In the next section we prove the following theorem.

Theorem. For all rules satisfying (6-9), $L_n \xrightarrow{n} L^*$ in probability. This property is valid for all distributions of (X,Y).

Olshen, in his comment on Stone's paper[7] , announced a similar theorem for this class of rules under the additional restriction that all the marginal distributions of X are atom-free (which in effect eliminates the possibility of ties). In the proof of our theorem we will follow the train of thought of Stone's proof of his corresponding theorem for the classical k-nearest neighbor rules.

Remark 1. (A generalization). Instead of D_n one can of course use D''_n which is obtained from D_n by considering $(\phi(X_i),Y_i,Z_i)$ instead of (X_i,Y_i,Z_i). Here ϕ is an a priori given but otherwise arbitrary mapping frm R^d to $R^{d'}$. We are thinking of some applications in which some nonlinear functions of X play an important role. If ϕ is a one-to-one mapping, the Bayes probability of error remains unchanged. In particular, this is always the case if $\phi=(\phi_1,\ldots,\phi_{d'})$ is such that $d' > d$, $\phi_1(x)=x^1,\ldots,\phi_d(x)=x^d$ and $\phi_{d+1},\ldots,\phi_{d'}$ are arbitrary real-valued functions on R^d. The theorem stated above remains valid for this interesting case.

Remark 2. (Related work on statistically equivalent blocks). Based on the theory of statistically equivalent blocks (Anderson[8]) several discrimination rules can be found in the literature that also enjoy the invariance property (i)[9-13]. In all these rules R^d is partitioned into rectangles on which $\tilde{Y}$ is constant. The size and location of the rectangles is data-dependent. In the simplest case (d=1) the real line is divided into k parts by the (1/k)-th, (2/k)-th,..., (k-1/k) -th quantiles of $X_1,\ldots,X_n$. On each interval $g_n(x,D_n)$ is constant and its value is determined by a majority vote among the Y_i for which X_i is in the same interval. Gordon and Olshen[13] show that this simple data dependent histogram rule satisfies (ii) if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. The experimental results reported by Gessaman and Gessaman[11] seem to indicate that these rules are strong both where computer time and performance are concerned. Practical ways of generalizing the histogram rules for d > 1 and make the rectangular partition depend upon the Y_i as well as the X_i are proposed by Henrichon and Fu[14], Meisel and Michalopoulos[15], Friedman[12] and Gordon and Olshen[13]. Return to the case d=1. For any new observation X the decision ($\tilde{Y}$) is based upon the approximately n/k states (namely those that correspond to observations in the interval X is in). In general X is not the median of the observations in that interval. In contrast, with the (n/k)-nearest neighbor version of (6), X is the median of those X_i's (X is included) whose states Y_i are taken into consideration in (6). One might therefore expect a slightly superior performance from the rules suggested in this note.

Proof of the Theorem

Consider the regression functions of the indicators $I_{\{Y=i\}}$ on X,

$$p_i(x) = E\{I_{\{Y=i\}}|X=x\} = P\{Y=i|X=x\},\ 1\le i\le M.$$

These functions are not uniquely defined, but it is always possible to find $p_1,\ldots,p_M$ that are Borel measurable and that satisfy

$$0\le p_i(x)\le 1, \text{ all } i,x, \text{ and } \sum_{1\le i\le M} p_i(x)=1, \text{ all } x.$$

Clearly, all discrimination rules of the form $\tilde{Y}= g(X)$ where

$$g(x)\ne i \text{ whenever}$$
$$p_i(x) < \max_{1\le \ell\le M} p_\ell(x) \tag{10}$$

are Bayes (i.e., they achieve L*). If the unknown p_i are estimated from the data by p_{ni}, and if the condition (10) is replaced by

$$g_n(x,Z,D_n)\ne i \text{ whenever}$$
$$p_{ni}(x,Z) < \max_{1\le \ell\le M} p_{n\ell}(x,Z) \tag{11}$$

then, regardless of the estimation procedure of the p_{ni}, the following inequality is valid.

Inequality 1. When g_n satisfies (11) then

$$L_n-L^* \le 2 \sum_{1\le i\le M} E\{|p_i(X)-p_{ni}(X,Z)|\,|D_n\}$$
$$= 2 \sum_{1\le i\le M} \int_0 |p_i(x)-p_{ni}(x,z)|\mu(dx)\,dz. \tag{12}$$

Here μ is the probability measure of X.

Proof. Since $L^*=E\{1-\max_i p_i(X)\}$ and

$$L_n = P\{g_n(X,Z,D_n)\ne Y|D_n\}= E\{1-p_{g_n(X,Z,D_n)}(X)|D_n\}$$

we have

$$|L_n-L^*|= E\{\max_i p_i(X)-p_{g_n(X,Z,D_n)}(X)|D_n\}$$
$$= E\{\max_i p_i(X) - \max_i p_{ni}(X,Z)|D_n\}$$
$$+ E\{\max_i p_{ni}(X,Z) - p_{ng_n(X,Z,D_n)}(X,Z)|D_n\}$$
$$+ E\{p_{ng_n(X,Z,D_n)}(X,Z) - p_{g_n(X,Z,D_n)}(X)|D_n\}$$
$$\le \sum_{1\le i\le M} E\{|p_i(X)-p_{ni}(X,Z)|\,|D_n\} + 0$$
$$+ \sum_{1\le i\le M} E\{|p_i(X)-p_{ni}(X,Z)|\,|D_n\}\ .$$

Q.E.D.

We recall that a discrimination rule (that is, a sequence of mappings g_n) is weakly Bayes risk consistent if $L_n-L^* \to 0$ in probability as $n \to \infty$. From inequality 1 we conclude that this is true whenever for all i,

$$E\{|p_i(X)-p_{ni}(X,Z)|\} \overset{n}{\to} 0.$$

By Jensen's inequality it suffices that for all i

$$E\{(p_i(X)-p_{ni}(X,Z))^2\} \overset{n}{\to} 0.$$

Now, the rules of the form (6) are obtained from (11) if one replaces $p_{ni}(x,Z)$ by

$$\sum_{j:Y_j(x,Z)=i} v_j = \sum_j v_j\, I_{\{Y_j(x,Z)=i\}}\ .$$

Thus, if we define $Y'_j=I_{\{Y_j=i\}}$ (where i is fixed) and if

$$m(x) = p_i(x) = E\{Y'_1|X_1=x\}$$

and

$$m_n(x,Z) = p_{ni}(x,Z) = \sum_{1\le j\le n} v_j Y'_j\ ,$$

then we need only show that

$$E\{(m(X)-m_n(X,Z))^2\} \overset{n}{\to} 0. \qquad (13)$$

The crucial results needed to obtain (13) for <u>all</u> possible distributions of (X,Y) are gathered in the inequalities 2 and 3 below. Inequality 3 is essentially due to Stone[7]. If $a_1,\dots,a_n;a$ are elements of $R^d x [0,1]$ then let

$$c_j(a;a_1,\dots,a_N) = v_i$$

where a_j is the i-th nearest neighbor (in the sense explained above) to a among $a_1,\dots,a_N$.

<u>Inequality 2</u>. If $a,a_1,\dots,a_n$ are different elements from $R^d \times [0,1]$ and if $v_1 \ge v_2 \ge \dots \ge v_n \ge 0$ then

$$\sum_{1\le i\le n} c_i(a_i;a_1,\dots,a_{i-1},a,a_{i+1},\dots,a_n)$$
$$\le \qquad 2^d \sum_{1\le i\le n} v_i \ . \qquad (14)$$

<u>Proof</u>. Let $a_j=(a_j(1),\dots,a_j(d+1))$, $a=(a(1),\dots,a(d+1))$. Partition $R^d \times [0,1]$ into 2^d quadrants centered at a in the following way. If the quadrants are indexed by the elements of $\{+,-\}^d$, then a_1 is in the quadrant $(+,+,+,\dots,+)$ if for all $1\le j\le d$

$$a_1(j)>a(j) \ , \text{ or } a_1(j)=a(j) \text{ and } a_1(d+1)>a(d+1).$$

(Thus, the last components are used to take care of the ties.) If Q_j is the j-th quadrant, we find the points $b_1,\dots,b_N$ from among $a_1,\dots,a_n$ that are in Q_j. The crucial observation is that if b_1 is the k-th nearest neighbor to a (from $b_1,\dots,b_N$) then a is at least the k-th nearest neighbor to b_1 (among $a,b_2,\dots,b_N$). In view of the monotonicity condition on the v_i we have

$$\sum_{i:a_i\in Q_j} c_i(a_i;a_1,\dots,a_{i-1},a,a_{i+1},\dots,a_n)$$
$$\le \sum_{1\le i\le N} c_i(b_i;b_1,\dots,b_{i-1},a,b_{i+1},\dots,b_N)$$
$$\le \sum_{1\le i\le N} c_i(a;b_1,\dots,b_N) = \sum_{1\le i\le N} v_i \le \sum_{1\le i\le n} v_i .$$

This bound does not depend upon $j,a,a_1,\dots,a_n$. Since there are 2^d quadrants, (14) obtains.

Q.E.D.

<u>Inequality 3</u>. Let $(X,Z,Y),(X_1,Z_1,Y_1),\dots,(X_n,Z_n,Y_n)$ be iid random vectors from $R^d x [0,1] xR$ and let Z be independent of (X,Y) and uniformly distributed on $[0,1]$. If $(v_1,\dots,v_n)$ satisfies (7) then

$$E\{|\sum_{1\le i\le n} c_i((X,Z);(X_1,Z_1),\dots,(X_n,Z_n))Y_i|^p\}$$
$$\le E\{\sum_{1\le i\le n} c_i((X,Z);(X_1,Z_1),\dots,(X_n,Z_n))|Y_i|^p\}$$
$$\le \quad 2^d E\{|Y|^p\} \quad , p \ge 1. \qquad (15)$$

<u>Proof</u>. The first inequality follows from Jensen's inequality. Inequality 3 now follows from Inequality 2 and the fact that (X,Z,Y) and the (X_i,Z_i,Y_i) are independent and identically distributed :

$$E\{\sum_{1\le i\le n} c_i((X,Z);(X_1,Z_1),\dots,(X_n,Z_n))|Y_i|^p\}$$
$$= \sum_{1\le i\le n} E\{c_i((X,Z);(Z_1,Z_1),\dots,(X_n,Z_n))|Y_i|^p\}$$
$$= \sum_{1\le i\le n} E\{c_i((X_i,Z_i);(X_1,Z_1),\dots,(X_{i-1},Z_{i-1}),$$
$$(X,Z),\dots,(X_n,Z_n))|Y|^p\}$$
$$\le 2^d \sum_{1\le i\le n} v_i \ E\{|Y|^p\} \qquad = 2^d E\{|Y|^p\}.$$

Q.E.D.

Return to the proof of the Theorem. The regression function m (which is essentially bounded) can for any $\varepsilon > 0$ be approximated by a bounded continuous function h such that

$$E\{|h(X)-m(X)|^p\} = \int\int |h(X)-m(x)|^p \mu(dx) < \varepsilon \ .$$

See Dunford and Schwartz[16],pp. 298. By the c_r - inequality (Loeve[17],pp. 155) we have

$$\int\int \quad |m_n(x,z)-m(x)|^p \mu(dx)dz$$
$$= \int\int \ | \sum_{1\le i\le n} C_{ni}(x,z) \ Y_i' \ -m(x)|^p \mu(dx)dz$$
$$\le 4^{p-1}(U_1 + U_2 + U_3 + U_4) \qquad (16)$$

where

$$C_{ni}(x,z)=c_i((x,z);(x_1,Z_1),\dots,(X_n,Z_n)) \ , \ 1\le i\le n,$$
$$U_1 = \int\int |\sum_{1\le i\le n} C_{ni}(x,z)(m(X_i)-h(X_i))|^p \mu(dx)dz \ ,$$
$$U_2 = \int\int |\sum_{1\le i\le n} C_{ni}(x,z)h(X_i)-h(x)|^p \mu(dx)dz \ ,$$
$$U_3 = \int \ |h(x)-m(x)|^p \mu(dx) \ ,$$
$$U_4 = E\{|\sum_{1\le i\le n} C_{ni}(X,Z)(Y_i'-m(X_i))|^p| \ D_n\}.$$

U_3 is small by choice of h. Next, by Inequality 3,

$$E\{U_1\} \quad \le \quad 2^d \ E\{|m(X)-h(X)|^p\} \ ,$$

which again can be made arbitrarily small. For $p \ge 2$, the term $E\{U_4\}$ is estimated as follows :

$$E\{U_4\} = E\{E\{|\sum_{1\le i\le n} C_{ni}(X,Z)|Y_i'-m(X_i)|^p|D_n\}\}$$
$$\le K(p) \ E\{E\{|\sum_{1\le i\le n} C^2_{ni}(X,Z)(Y_i'-m(X_i))^2|^{p/2}|D_n\}\}$$

(for some $K(p) > 0$ by the inequality of Marcinkiewicz and Zygmund[18] ; see also Petrov[19],pp. 59-60).

$$\le K(p)(\sup_i v_i)^{p/2} E\{|\sum_{1\le i\le n} C_{ni}(X,Z)(Y_i'-m(X_i))^2|^{p/2}\}$$
$$\le K(p)(\sup_i v_i)^{p/2} E\{\sum_{1\le i\le n} C_{ni}(X,Z)|Y_i'-m(X_i)|^p\}$$

(by Jensen's inequality)

$$\leq 2^d\ K(p)\ (\sup_i v_i)^{p/2} E\{|Y_1'-m(X_1)|^p\}$$

(by Inequality 3)

$$\leq 2^d\ K(p)\ (\sup_i v_i)^{p/2} \tag{17}$$

(since here Y_1' is $\{0,1\}$ -valued, and thus $|Y_1'-E\{Y_1'|X_1=x\}|\leq 1$ for almost all x.)

Thus, $E\{U_4\} \overset{n}{\to} 0$ from (8). We need only show that $E\{U_2\}$ tends to 0 for all bounded continuous functions h. If we define

$$W_n = \sum_{1\leq i\leq n} C_{ni}(x,z)|h(X_i)-h(x)|^p$$

then we need only show that $\iint W_n(x,z)\mu(dx)dz \overset{n}{\to} 0$ in probability. Actually, we will show more :

<u>Inequality 4</u>. Let h be a bounded continuous function on R^d and let $(v_1,\ldots,v_n)$ be a probability vector satisfying (9). For every $\varepsilon>0$, $p\geq 1$ there exist $\alpha,\beta,\gamma > 0$ such that

$$P\{\int\int_0^1 W_n(x,z)\mu(dx)dz>\varepsilon\}\leq\alpha\ e^{-\beta n}\ ,\ \text{all}\ n \geq \gamma.$$

<u>Proof</u>. Define the following ordering on $R\times[0,1]$:

$$(x_1,z_1) \prec (x_2,z_2) \quad \text{if}\ x_1<x_2,\ \text{or}\ x_1=x_2, z_1<z_2.$$

The relation $(x_1,z_1) \preceq (x_2,z_2)$ also includes the case $x_1=x_2$, $z_1=z_2$. For any sequence from $R\times[0,1]$, say $(x_0,z_0)\prec(x_1,z_1)\prec \ldots \prec(x_N,z_N)$ the set B_{ji} from $R^d\times[0,1]$ is defined as the set of all $(x^1,\ldots,x^d,$ $z)$ with the property that

$$(x_{i-1},z_{i-1}) \preceq (x^j,z) \prec (x_i,z_i).$$

For a given number $c\in(0,1)$ and for each $j=1,\ldots,d$ we find numbers $(x_0,z_0)\prec(x_1,z_1)\prec\ldots\prec(x_N,z_N)$ with the property that

$$P\{\ (X,Z) \in B_{ji}\} = c/N\ ,\ \text{all}\ i=1,\ldots,n\ .$$

Thus, if

$$B_j = \bigcup_{i=1}^N B_{ji}\ ,\quad B = \bigcap_{j=1}^d B_j\ ,$$

then it follows that $P\{\ (X,Z) \in B_j\} = c$.

In the proof we will group all "bad" cylinder sets B_{ji} together in a set B*. To start with, we include the border strips $B_{j1}, B_{j2}, B_{j\,N-1}$ and B_{jN} for all j. We also include all the B_{ji} whose length (that is, x_i-x_{i-1}) exceeds $\delta >0$. Notice that for fixed j the number of strips B_{ji} with length greater than or equal to δ can not exceed $T/\delta + 1$ where

$$T = \max_{1\leq j\leq d} (x_N-x_0).$$

In addition we include $B_{j\,i-1}$, $B_{j\,i-2}$, $B_{j\,i+1}$ and $B_{j\,i+2}$ whenever $2<i<N-2$ and B_{ji} already is in B*, that is, for every strip whose length exceeds δ , we include in B* the four neighbouring strips as well. Repeating this for all j, we obtain,

$$P\{(X,Z)\in B^*\}\leq d(4+5+5T/\delta)(c/N). \tag{18}$$

A rectangle C is defined as a set of the form $\bigcap_{j=1}^N B_{j\,i_j}$, $(i_1,i_2,\ldots,i_d)\in\{1,\ldots,N\}^d$.

Pick k as in (9) and construct a set B** by taking the union of all the rectangles C for which $P\{(X,Z)\in C\} <\theta + k/n$ where $\theta > 0$ is a number to be determined further on. Clearly,

$$P\{(X,Z)\in B^{**}\} \leq \sum_{\substack{\text{rectangles C with}\\ P\{(X,Z)\ C\}<\theta+k/n}} P\{(X,Z)\in C\}$$
$$\leq N^d(\theta+ k/n).$$

If $H=\sup|h(x)|$ and $(\ )^c$ denotes the complement of a set, then

$$\int_0^1\int W_n(x,z)\mu(dx)dz \leq (2H)^p P\{(X,Z)\in B^c\cup B^*\cup B^{**}\}$$
$$+ \iint_{B\cap B^{*c}\cap B^{**c}} W_n(x,z)\mu(dx)dz \tag{19}$$

The first term on the right hand side of (19) can be estimated by

$$(2H)^p\ d(1-c) + d(9+5T/\delta)c/N + N^d(\theta+ k/n) \tag{20}$$

where we used the fact that $P\{(X,Z)\notin B\}< d(1-c)$. Now, (20) is smaller than $\varepsilon/2$ for all n large enough. First pick c close enough to 1 to take care of d(1-c). This fixes the value of T. If δ is given, then we van pick N large enough to make the second term as small as desired. Finally choose θ so small that $N^d\theta$ is small, and use (9) to conclude that $N^dk/n \to 0$ as $n \to \infty$. The only parameter that can still be assigned an arbitrary value is δ .

The measure induced on the Borel sets of $R^d\times[0,1]$ by $\mu(dx)\ dz$ is called ν . The empirical measure ν_n of a Borel set $S\subseteq R^d\times[0,1]$ is defined by

$$\nu_n(S) = n^{-1} \sum_{1\leq i\leq n} I_{\{(X_i,Z_i)\ S\}}\ .$$

Let E_1 be the event that for all (N^d) rectangles C,

$$|\nu(C)-\nu_n(C)| < \theta$$

and that for all (Nd) cylinder sets B_{ji}

$$|\nu(B_{ji})-\nu_n(B_{ji})| < c/3N.$$

Let us fix $(x,z)\in C\cap B^{*c}\cap B^{**c}$, and assume that E_1 holds for $(X_1,Z_1),\ldots,(X_n,Z_n)$. In that case the empirical distance between (x,z) and any of the $(X_i,Z_i)\in C$ is less than

$$n\ (c/N + c/3N)$$

because by definition $\nu(B_{ji})=c/N$ for all j,i. Next, if $C=B_{1\,i_1}\cap B_{2\,i_2}\cap\ldots\cap B_{d\,i_d}$ and $C'=B_{1\,i_1'}\cap B_{2\,i_2'}\cap\ldots\cap B_{d\,i_d'}$ with $\max_j |i_j-i_j'|>2$, and if $(X_i,Z_i)\in C'$, then its empirical distance to (x,z) is at least equal to

$$2n\ (c/N - c/3N) = n(c/N + c/3N).$$

Therefore, all the (X_i, Z_i) in C are nearer neighbors of $(x,z) \in C$ than any $(X_i, Z_i) \notin C'$. But if E_1 holds, then

$$\nu_n(C) > \nu(C) - \theta \geq k/n + \theta - \theta \quad = k/n,$$

that is, more than k of the (X_i, Z_i) belong to C. This in turn implies that all the k nearest neighbors to (x,z) are in

$$\bigcap_{j=1}^{d} (B_{j\ i_j-2} \cup B_{j\ i_j-1} \cup B_{j\ i_j} \cup B_{j\ i_j+1} \cup B_{j\ i_j+2}).$$

By the construction of B* it is true that under the maximum component norm $|.|$,

$$|X_i(x,z) - x| \leq 3\ \delta, \quad 1 \leq i \leq k.$$

Since we can still pick δ , we pick it so small that

$$\sup_{\substack{(x,z)\in B;\ (y,z)\in B \\ |x-y| < 3\delta}} |h(x) - h(y)| < \varepsilon/4\ ,$$

which is possible by the uniform continuity of h on the closure of the projection of B on R^d, Thus, if E_1 holds,

$$\sup_{(x,z)\in B\cap B^{*c}\cap B^{**c}} W_n(x,z)$$

$$\leq (\varepsilon/4) \sum_{1\leq i\leq k} v_i + (2H)^p \sum_{k<i\leq n} v_i$$

$$< \varepsilon/2 \quad \text{for all n large enough.} \tag{21}$$

From (19) and (21),

$$P\{\iint_0^1 W_n(x,z)\mu(dx)dz \geq \varepsilon\ \}$$

$$\leq P\{\iint_{B\cap B^{*c}\cap B^{**c}} W_n(x,z)\mu(dx)dz \geq \varepsilon/2\}$$

$$\leq P\{E_1^c\} \leq \sum_{\text{rectangles } C} P\{|\nu(C) - \nu_n(C)| \geq \theta\}$$

$$+ \sum_{j,i} P\{|\nu(B_{ji}) - \nu_n(B_{ji})| \geq c/3N\}$$

$$\leq 2\ N^d\ e^{-2n\theta^2} + 2\ Nd\ e^{-2nc^2/9N^2}$$

where we used Hoeffding's inequality for the sum of independent {0,1}-valued random variables[20]. This concludes the proof of Inequality 4.

Q. E. D.

References

1 T.M.COVER,P.E.HART: "Nearest neighbor pattern classification", IEEE Transactions on Information Theory, vol. IT-13, pp. 21-27, 1967.

2 E.FIX,J.L.HODGES: "Discriminatory analysis, nonparametric discrimination, consistency properties", Report No. 4, Project 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

3 E.FIX,J.L.HODGES: "Nonparametric discrimination: small sample performance", Report No. 11, Project 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1952.

4 G.SEBESTYEN: Decision Making Processes In Pattern Recognition, Macmillan, New York, 1962.

5 W.H.ROGERS,T.J.WAGNER: "A finite sample distribution-free performance bound for local discrimination rules", to appear in Annals of Statistics, May 1978.

6 L.P.DEVROYE,T.J.WAGNER: "Distribution-free inequalities for the deleted and the holdout error estimates", to appear in IEEE Transactions on Information Theory, 1978.

7 C.J.STONE: "Consistent nonparametric regression", Annals of Statistics, vol. 5, pp. 595-645, 1977.

8 T.W.ANDERSON: "Some nonparametric multivariate procedures based on statistically equivalent blocks", in : Multivariate Analysis, P.R.Krishnaiah Ed., Academic Press, New York, pp. 5-27, 1966.

9 C.P.QUESENBERRY,M.P.GESSAMAN: "Nonparametric discrimination using tolerance regions", Annals of Mathematical Statistics, vol. 39, pp. 664-673, 1968.

10 J.SONQUIST: "Multivariate model building: the validation of a search strategy", Institute for Social Research, University of Michigan, Ann Arbor, 1970.

11 M.P.GESSAMAN,P.H.GESSAMAN: "A comparison of some multivariate discrimination procedures", Journal of the American Statistical Association, vol. 67, pp. 468-472, 1972.

12 J.H.FRIEDMAN: "A recursive partitioning decision rule for nonparametric classification", IEEE Transactions on Computers, vol. C-26, pp. 404-408, 1977.

13 L.GORDON,R.A.OLSHEN: "Asymptotically efficient solutions to the classification problem", submitted to Annals of Statistics, 1977.

14 E.G.HENRICHON,K.S.FU: "A nonparametric partitioning procedure for pattern classification", IEEE Transactions on Computers, vol. C-18, pp. 614-624, 1969.

15 W.S.MEISEL,D.A.MICHALOPOULOS: "A partitioning algorithm with application in pattern classification and the optimization of decision trees", IEEE Transactions on Computers, vol. C-22, pp. 93-103, 1973.

16 N.DUNFORD,J.T.SCHWARTZ: Linear Operators I, Interscience, New York, 1957.

17 M.LOEVE: Probability Theory, Van Nostrand, Princeton, New Jersey, 1963.

18 J.MARCINKIEWICZ,A.ZYGMUND: "Sur les fonctions independantes", Fund. Math., vol. 29, pp. 60-90, 1937.

19 V.V.PETROV: Sums Of Independent Random Variables, Springer-Verlag, Berlin, 1975.

20 W.HOEFFDING: "Probability inequalities for sums of bounded random variables", Journal of the American Statistical Association, vol. 58, pp. 13-30, 1963.

*Research of author partially sponsored by AFOSR Grant 77-3385.