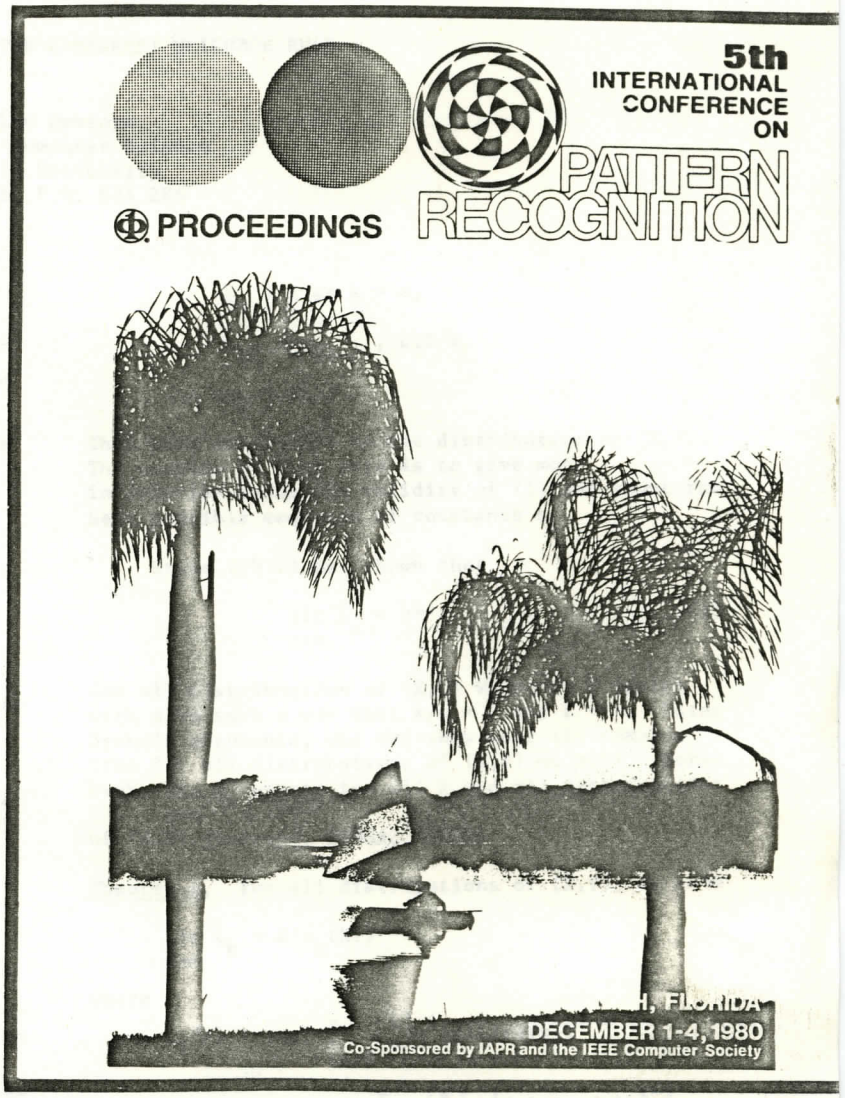

IEEE COMPUTER SOCIETY

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS

Volume 1 of 2

IEEE CATALOG NO. 80CH1499-3

Library of Congress No. 79-90304

SOME PROPERTIES OF THE K-NEAREST NEIGHBOR RULE

Luc Devroye
School of Computer Science
McGill University
Montreal, P.Q. H3A 2K6

Abstract

Let L_n be the probability of error for the k-nearest neighbor rule with a sample of size n, and let R* be the Bayes probability of error. Cover and Hart have show that $\overline{\lim}_{n\to\infty} L_n \le (1+c_k)R^*$ for some sequence of constants c_k. We notice that for all distributions of the data, and all odd $k \ge 5$,

$$c_k \le \alpha\frac{\sqrt{k-1}}{k-3}\left(1+\frac{\beta}{\sqrt{k-3}}\right)$$

where α=0.3399... and β=0.9749... are universal constants. The upper bound behaves as $\alpha/\sqrt{k}$ as $k \to \infty$. This is the best possible behavior because the supremum over all distrubtions of the data of $(\lim_{n\to\infty} L_n/R^* - 1)$ is $\sim \alpha/\sqrt{k}$ as $k \to \infty$. We obtain (numerically) the exact values for c_k, and compare them with the given bound for values of k up to 35.

1. Results

Let the *data* $(X_1,Y_1),\ldots,(X_n,Y_n)$ be independent identically distributed random vectors from $R^d \times \{0,1\}$, and let (X,Y) be independent of the data and distributed as (X_1,Y_1). In *discrimination*, one is asked to construct an estimate $\hat{Y}$ of Y from X and the data. In the *k-nearest neighbor estimate* (or: k-nearest neighbor rule), one first finds the k nearest neighbors of X among $X_1,\ldots,X_n$, and then determines $\hat{Y}$ according to a majority vote among the corresponding Y_i's. Distance ties and voting ties are broken by looking at the indices.

Cover and Hart (1965) have shown that the *probability of error* $L_n = P(\hat{Y} \ne Y)$ satisfies

$$\limsup_{n\to\infty} L_n \le (1+c_k)\, R^* \qquad (1)$$

where R* is the *Bayes probability of error* ($R^* = \inf_{g:R^d\to\{0,1\}} P(g(X)\ne Y)$), and $c_1,c_2,\ldots$ is a sequence of constants satisfying

$$c_k \downarrow 0 \text{ as } k \to \infty,$$

$$c_{2k} = c_{2k-1}, \text{ all } k,$$

$$c_1 = 1.$$

They have conditions on the distribution of (X,Y). The purpose of this note is to give more information about the validity of (1) and about the best possible sequence of constants c_k.

Stone (1977) has shown that

$$\lim_{n\to\infty} L_n = R^* \qquad (2)$$

for all distributions of (X,Y) whenever k varies with n in such a way that $k/n \to 0$ and $k \to \infty$. Using Stone's arguments, one can show that (1) remains true for all distributions of (X,Y) as well. Moreover, $\lim_{n\to\infty} L_n$ exists for all k and all distributions of (X,Y) (Devroye, 1980a, 1980b).

Theorem 1. For all distributions of (X,Y), we have

$$\lim_{n\to\infty} L_n = E(t_k(X))$$

where

$$t_k(x) = \eta(x)\sum_{i\le k/2}\binom{k}{i}\eta^i(x)(1-\eta(x))^{k-i} + (1-\eta(x))\sum_{i>k/2}\binom{k}{i}\eta^i(x)(1-\eta(x))^{k-i},$$

$$k \ge 1,\ k \text{ odd},$$

and

$$t_0(x) = \eta(x)\ ;\ t_{2k}(x) = t_{2k-1}(x), \text{ all } k \ge 1,$$

$$\eta(x) = P(Y=1 \mid X=x),\ x \in R^d.$$

By Theorem 1, it suffices to study the properties of $E(t_k(X))$. One can show that for k odd,

CH1499-3/80/0000-0103$00.75

$$\sup_{\substack{\text{all distributions} \\ \text{of } (X,Y) \text{ with } R^*>0}} \frac{E(t_k(X))}{R^*} = 1$$

$$+ \sup_{0<p\le 1/2} \frac{1-2p}{p} \sum_{i>k/2} \binom{k}{i} p^i (1-p)^{k-i}$$

$$\triangleq 1 + c_k \ . \qquad (3)$$

This supremum is easily determined for k=1 and k=3 : we obtain the values 2 and $\frac{7\sqrt{7}+17}{27}$ respectively. A table with numerically constructed values of c_k is given below. (see table 1). Two questions remain open:

1. How does the sequence c_k behave as $k \to \infty$?
2. Can we find good but simple upper bounds for the c_k's ?

The first question is of theoretical importance. The second question may be relevant in practical situations. Often one is willing to accept rules in discrimination whose probability of error is less than $(1+\varepsilon)\ R^*$ where ε is typically 0.1 ... 0.05 . Let $P_n = P(\hat{Y} \neq Y | X_1, Y_1, \ldots, X_n, Y_n)$. Then

$$P(P_n > R^*(1+\varepsilon)) \le \varepsilon^{-1}[E(P_n)/R^* - 1]$$

$$\to \frac{1}{\varepsilon}[E(t_k(X))/R^* - 1] \quad (\text{as } n \to \infty)$$

$$\le \frac{1}{\varepsilon} c_k \ .$$

Here we used Chebyshev's inequality. Good upper bounds for c_k would enable us to choose k in such a way that the last expression is as small as desired. Unfortunately, we are assuming that n is very large, and even if we had an almost infinite supply of data, Chebyshev's inequality is too crude to be of any practical use.

We will need three universal constants related to the tail of the normal distribution. If

$$\Phi(t) = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \ ,$$

then

$$\alpha = \max_{t>0} 2t\Phi(t) = 0.3399424150... \ ,$$

$$\beta = \max_{t>0} 2t^2\Phi(t)/\alpha = 0.9749687445... \ ,$$

and

$$\delta = 0.7517915241...$$

is the value of t for which $2t\Phi(t)=\alpha$. Let us define the sequence

$$a_k = \frac{\sqrt{k-1}}{k-3}\alpha\left(1 + \frac{\beta}{\sqrt{k-3}}\right) \ , \ k \ge 5.$$

Theorem 2. For all distributions of (X,Y), and for all $k \ge 5$, k odd,

$$E(t_k(X))/R^* \le 1 + a_k \ . \qquad (4)$$

Clearly,

$$a_k \sim \frac{\alpha}{\sqrt{k}} \text{ as } k \to \infty \qquad (5)$$

This is the best possible rate of convergence with respect to k. In fact, c_k is of the form $\frac{\alpha}{\sqrt{k}}(1+o(1))$, and thus, while $c_k \le a_k$, $k \ge 5$, it is also true that $c_k/a_k \to 1$ as $k \to \infty$.

Theorem 3. $c_k \sim \alpha/\sqrt{k}$ as $k \to \infty$.

Consider the following trivial distributions of (X,Y) where Y is independent of X, and $P(Y=1)=p=p(k)$. Clearly, if $p < 1/2$, we have $R^*=p$. Also, by Theorem 1,

$$\lim L_n/R^*$$

$$= 1 + \frac{1-2p}{p} \sum_{i>k/2} \binom{k}{i} p^i (1-p)^{k-i} \ .$$

If we choose

$$p(k) = \frac{1}{2}\left(1 - \frac{\delta}{\sqrt{k}}\right) \ ,$$

then

$$\left(\frac{\lim L_n}{R^*} - 1\right) \sim \frac{\alpha}{\sqrt{k}} \text{ as } k \to \infty \ .$$

In other words, this sequence of p(k)'s is the "worst" possible sequence one can choose with respect to the ratio $\lim L_n/R^*$.

In table 1 below, we give approximate values for

(1) the exact values c_k ;
(2) the p's for which the supremum in (3) is reached ;
(3) the values of a_k ;
(4) the values of $\alpha/\sqrt{k}$.

The proofs of the announced theorems will appear elsewhere (Devroye, 1980a).

k	$\alpha/\sqrt{k}$	c_k	a_k	p
3	0.196	0.316		0.226
5	0.152	0.218	0.574	0.295
7	0.128	0.173	0.310	0.331
9	0.113	1.147	0.224	0.354
11	0.102	0.130	0.181	0.369
13	0.094	0.117	0.154	0.381
15	0.088	0.107	0.136	0.390
17	0.082	0.100	0.122	0.397
19	0.080	0.093	0.112	0.404
21	0.074	0.088	0.104	0.409
23	0.071	0.083	0.097	0.413
25	0.068	0.079	0.091	0.417
27	0.065	0.076	0.087	0.420
29	0.063	0.073	0.082	0.423
31	0.061	0.070	0.079	0.426
33	0.059	0.068	0.076	0.429

Table 1.

2. References

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

L. DEVROYE: "On the asymptotical probability of error in nonparametric discrimination", Manuscript, School of Computer Science, McGill University, Montreal, 1980a.

L. DEVROYE: "On the inequality of Cover and Hart in nearest neighbor discrimination", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 1980b.

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