space codes," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 703–712, Nov. 1978.
[4] C-E. Sundberg, T. Aulin, and N. Rydbeck, "Recent results on spectrally efficient constant envelope digital modulation methods," "*M*-ary CPFSK type of signalling with input data symbol pulse shaping-minimum distance and spectrum", "Bandwidth efficient digital FM with coherent phase tree demodulation," in *Conf. Rec. Int. Conf. on Communications*, Boston, MA, June 1979. See also NTC 1978.
[5] D. E. Knuth, *The Art of Computer Programming*, vol. II. Reading, MA: Addison-Wesley, 1969.
[6] D. R. Cox and P. R. Lewis, *The Statistical Analysis of Series of Events*. London: Methuen, 1966, ch. 5 and 6.
[7] J. B. Anderson and K. Srivatsa, "Trellis decoded error events are apparently independent," in *Proc. 1980 Princeton Conf. on Information Sciences and Systems*, Princeton, NJ, Mar. 1980.
[8] R. de Buda, "Coherent demodulation of FSK with low deviation ratio," *IEEE Trans. Commun.*, vol. COM-20, pp. 429–435, 1972.
[9] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding*. New York: McGraw-Hill, 1979.
[10] J. B. Anderson, C.-E. W. Sundberg, T. Aulin, and N. Rydbeck, "Power-bandwidth performance of smoothed phase modulation codes," *IEEE Trans. Commun.*, vol. COM-29, pp. 187–195, Mar. 1981.

The Rate of Convergence of k_n-NN Regression Estimates and Classification Rules

LÁSZLÓ GYÖRFI

***Abstract*—The rate of convergence of k_n-NN regression estimates and the corresponding multiple classification error are calculated without assuming the existence of the density of the observations.**

Introduction

Cover and Hart [1] proved the convergence of the one-nearest-neighbor (1-NN) and k-NN decision rules under some continuity condition on the *a posteriori* probabilities. Under the same conditions Györfi and Györfi [2] showed the convergence of k_n-NN rule. The mere use of a nonparametric, e.g., NN, estimate is normally a consequence of the partial or total lack of information about the underlying distributions. Therefore, it is important to prove their consistency without any condition on the distributions and to prove their rate of convergence under mild conditions. Stone [3] proved the distribution-free consistency of the NN rule.

The rate of convergence was investigated under the assumption that the *a posteriori* probability functions are smooth enough, and the density of the observation exists (Cover [4], Györfi [5], Beck [6], Wagner [10], Fritz [11]). Our purpose is to weaken these conditions.

k_n-NN Regression Estimate

Let (X,Y) be a random vector taking values in $R^d \times R$. If $\boldsymbol{E}|Y| < +\infty$ then the regression function is defined by $m(z) \triangleq \boldsymbol{E}(Y/X=z)$, $z \in R^d$. μ denotes the probability measure of X. An estimate of $m(z)$ is calculated from a sample $Z^n \triangleq \{(X_1,Y_1),\cdots,(X_n,Y_n)\}$, which is a sequence of independent and identically distributed random vectors having the same distribution as that of (X,Y). Assume that (X,Y) and Z^n are independent. For a fixed $z \in R^d$ let $(X_{1,n}^z, Y_{1,n}^z),\cdots,(X_{n,n}^z, Y_{n,n}^z)$ be the nearest neighbor ordering of Z^n according to increasing values of $\|X_i - z\|$. In case of a tie (X_i,Y_i) precedes (X_j,Y_j) if $i<j$. Then the $k_n - NN$ estimate of $m(z)$ is

$$m_n(z) \triangleq \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{i,n}^z.$$

Manuscript received February 25, 1980; revised July 22, 1980.
The author is with the Technical University of Budapest, H-1111 Budapest, Stoczek u. 2, Hungary.

The main idea of previous developments was the following: if m is continuous and we have some values of m at a sphere $S_{z,r}$ centered at z with radius r, then their average is a good approximation of $m(z)$ for small r. However, this average is near to the expectation:

$$\frac{1}{\mu(S_{z,r})} \int_{S_{z,r}} m(u)\mu(du)$$

and by the pointwise density theorem (Federer [7, theorem 2.9.8])

$$\lim_{r \to 0} \frac{1}{\mu(S_{z,r})} \int_{S_{z,r}} m(u)\mu(du) = m(z), \qquad \text{for each } z \in R^d \bmod \mu$$

is true in general for any (measurable) regression function. No continuity condition is required. On the one hand, this is a useful tool for proving the distribution-free (universal) consistency of *NN* rules [8], [9]; on the other hand, we can prove the rate of convergence under weak condition.

Theorem: Suppose that $\boldsymbol{E}|Y|^2 < \infty$ and

$$\lim_{n \to \infty} k_n = \infty, \qquad \lim_{n \to \infty} \frac{k_n}{n} = 0. \tag{1}$$

Assume a function K on R^d and $\alpha > 0$ such that

$$\left| \frac{1}{\mu(S_{z,r})} \int_{S_{z,r}} m(u)\mu(du) - m(z) \right| \le K(z) r^{\alpha} \tag{2}$$

for each $r > 0$ and for each $z \in R^d \bmod \mu$. Then for all sequences $a_n \xrightarrow{n} 0$

$$a_n \min\left\{ \sqrt{k_n}, \left(\frac{n}{k_n} \right)^{\alpha/d} \right\} |m_n(X) - m(X)| \xrightarrow{n} 0 \tag{3}$$

in probability.

One has to emphasize that (2) is a smoothing condition. However, it does not imply that m is continuous mod μ.

If $\sqrt{k_n} = (n/k_n)^{\alpha/d}$, i.e., we choose

$$k_n = n^{1/(1+(d/2\alpha))},$$

then according to (3) the rate of convergence in probability is at least $n^{-1/(2+(d/\alpha))}$.

k_n-NN Decision Rule

For the multiple classification rule the random vector X stands for the observation. Given X, we have to decide on the random variable Θ, which takes values in $\{1,2,\cdots,M\}$. Let

$$P_i(z) \triangleq \boldsymbol{P}(\Theta = i/X = z), \qquad i = 1,2,\cdots,M,\ z \in R^d,$$

denote the *a posteriori* probabilities. The Bayesian decision rule minimizes the probability of misdecision:

$$g_B(X) = i \quad \text{if} \quad \begin{matrix} P_i(X) > P_j(X), & j < i \\ P_i(X) \ge P_j(X), & j > i. \end{matrix}$$

Then

$$\begin{aligned} L^*(X) &\triangleq \boldsymbol{P}(g_B(X) \ne \Theta/X) \\ &= 1 - \max_{1 \le i \le M} P_i(X) \\ &= \inf_{g:\, R^d \to \{1,2,\cdots,M\}} \boldsymbol{P}(g(X) \ne \Theta/X). \end{aligned} \tag{4}$$

P_i might be defined as the regression function of the indicator of the event $\{\Theta = i\}$ $(i = 1,2,\cdots,M)$. Let $P_{i,n}$ denote the k_n-NN estimate of P_i $(i = 1,2,\cdots,M)$. Then the k_n-NN decision rule is

0018-9448/81/0500-0362$00.75

as follows:

$$g_n(X)=i \quad \text{if} \quad \begin{matrix} P_{i,n}(X)>P_{j,n}(X), & j<i \\ P_{i,n}(X)\geqslant P_{j,n}(X), & j>i. \end{matrix}$$

By the bound of Györfi [5, lemma 4]

$$\boldsymbol{P}\left(g_n(X)\neq\Theta/X,Z^n\right)-L^*(X)\leqslant\sum_{i=1}^{M}|P_i(X)-P_{i,n}(X)|.$$

Therefore, our theorem implies the following corollary.

Corollary: Assume that

$$\lim_{n\to\infty}k_n=\infty,\qquad \lim_{n\to\infty}\frac{k_n}{n}=0$$

and the functions P_i satisfy (2) for some functions K_i $(i=1,2,\cdots,M)$ and $\alpha>0$. Then for all sequences $a_n\to 0$

$$a_n\min\left\{\sqrt{k_n},\left(\frac{n}{k_n}\right)^{\alpha/d}\right\}|\boldsymbol{P}\left(g_n(X)\neq\Theta/X,Z^n\right)-L^*(X)|\xrightarrow{n}0$$

in probability.

Proof

If the density of the observation X does not exist, then the basic tool for proving the rate of convergence is a distribution-free (universal) rate of convergence of the distances of the nearest neighbors.

Lemma: For each sequence $a_n\xrightarrow{n}0$, if $\lim_{n\to\infty}(k_n/n)=0$, then

$$a_n\left(\frac{n}{k_n}\right)^{1/d}\|X-X_{k_n,n}^X\|\xrightarrow{n}0 \tag{5}$$

in probability.

Proof: It is sufficient to deal with the case where $1/\delta_n\triangleq a_n(n/k_n)^{1/d}\xrightarrow{n}\infty$, since if $\lim_{n\to\infty}(k_n/n)=0$, then [2]

$$\|X-X_{k_n,n}^X\|\xrightarrow{n}0\quad\text{a.s.}$$

If χ stands for the indicator function, then for each $\epsilon>0$

$$\begin{aligned}\boldsymbol{P}&\left(a_n\left(\frac{n}{k_n}\right)^{1/d}\|X-X_{k_n,n}^X\|\geqslant\epsilon\right)\\&=\int\boldsymbol{P}\left(X_{k_n,n}^z\notin S_{z,\epsilon\delta_n}\right)\mu(dz)\\&=\int\boldsymbol{P}\left(\sum_{i=1}^{n}\chi_{\{X_i\in S_{z,\epsilon\delta_n}\}}<k_n\right)\mu(dz).\end{aligned} \tag{6}$$

The random variable

$$\sum_{i=1}^{n}\chi_{\{X_i\in S_{z,\epsilon\delta_n}\}}$$

has the binomial distribution with expectation $n\mu(S_{z,\epsilon\delta_n})$ and variance $n\mu(S_{z,\epsilon\delta_n})(1-\mu(S_{z,\epsilon\delta_n}))$. By the de Moivre–Laplace theorem

$$\boldsymbol{P}\left(\sum_{i=1}^{n}\chi_{\{X_i\in S_{z,\epsilon\delta_n}\}}<k_n\right)\xrightarrow{n}0 \tag{7}$$

if

$$\frac{n\mu\left(S_{z,\epsilon\delta_n}\right)-k_n}{\sqrt{n\mu\left(S_{z,\epsilon\delta_n}\right)}}\to\infty.$$

Devroye [9] proved that the finite limit

$$\lim_{h\to 0}\frac{h^d}{\mu\left(S_{z,rh}\right)}$$

exists for all $z\in R^d \bmod \mu$ and for all fixed $r>0$. Thus we get

$$\begin{aligned}&\lim_{n\to\infty}\frac{n\mu\left(S_{z,\epsilon\delta_n}\right)-k_n}{\sqrt{n\mu\left(S_{z,\epsilon\delta_n}\right)}}\\&=\lim_{n\to\infty}\sqrt{k_n}\left(\frac{1}{a_n^{d/2}}\sqrt{\frac{\mu\left(S_{z,\epsilon\delta_n}\right)}{\delta_n^d}}-a_n^{d/2}\sqrt{\frac{\delta_n^d}{\mu\left(S_{z,\epsilon\delta_n}\right)}}\right)=\infty,\end{aligned} \tag{8}$$

for all $z \bmod \mu$. Therefore, (7) is true for all $z \bmod \mu$. Applying the Dominated Convergence Theorem, (6) and (7) imply (5).

Proof of Theorem: Introduce the notations

$$\alpha_n(X)\triangleq\|X-X_{k_n+1,n}^X\|$$

and

$$A^*\triangleq\{z;\mu(S_{z,r})>0,\forall r>0\}.$$

Cover and Hart [1] proved that $\mu(A^*)=1$. Thus

$$\begin{aligned}|m_n(X)-m(X)|\leqslant&\left|\frac{1}{k_n}\sum_{i=1}^{k_n}Y_{i,n}^X-\frac{1}{k_n}\sum_{i=1}^{k_n}m\left(X_{i,n}^X\right)\right|\\&+\chi_{\{X\in A^*,\alpha_n(X)>0\}}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}m\left(X_{i,n}^X\right)\right.\\&\left.-\frac{1}{\mu\left(S_{X,\alpha_n(X)}\right)}\int_{S_{X,\alpha_n(X)}}m(u)\mu(du)\right|\\&+\chi_{\{X\in A^*,\alpha_n(X)>0\}}\left|\frac{1}{\mu\left(S_{X,\alpha_n(X)}\right)}\right.\\&\left.\cdot\int_{S_{X,\alpha_n(X)}}m(u)\mu(du)-m(X)\right|\quad\text{a.s.}\end{aligned} \tag{9}$$

For the first term of the right side of (9), introduce the notation

$$\delta^2(z)=\boldsymbol{E}\left((Y-m(X))^2/X=z\right),\qquad z\in R^d.$$

Then

$$\begin{aligned}&\left(\sqrt{k_n}\right)^2\boldsymbol{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{i,n}^X-m\left(X_{i,n}^X\right)\right)\right|^2\\&=k_n\boldsymbol{E}\boldsymbol{E}\left(\left|\frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{i,n}^X-m\left(X_{i,n}^X\right)\right)\right|^2/X,X_1,\cdots,X_n\right)\\&=\boldsymbol{E}\frac{1}{k_n}\sum_{i=1}^{k_n}\delta^2\left(X_{i,n}^X\right)\xrightarrow{n}\boldsymbol{E}\delta^2(X),\end{aligned} \tag{10}$$

where in the last step the weak universal consistency of k_n-NN regression was used (Stone [3], Devroye [9]). For the second term of (9), applying the same argument as that of Devroye [9, lemma 2.1], we get

$$\sup_n k_n\boldsymbol{E}\left|\frac{1}{k_n}\sum_{i=1}^{k_n}m\left(X_{i,n}^x\right)-\frac{1}{\mu\left(S_{X,\alpha_n(X)}\right)}\int_{S_{X,\alpha_n(X)}}m(u)\mu(du)\right|^2<+\infty. \tag{11}$$

For the third term of (9) we have from (2) that

$$a_n\left(\frac{n}{k_n}\right)^{\alpha/d}\left|\frac{1}{\mu\left(S_{X,\alpha_n(X)}\right)}\int_{S_{X,\alpha_n(X)}} m(u)\mu(du) - m(X)\right|$$

$$\leq K(X)\left[a_n^{1/\alpha}\left(\frac{n}{k_n}\right)^{1/d}\|X - X_{k_n+1,n}^X\|\right]^{\alpha} \quad \text{a.s.} \quad (12)$$

and (10)–(12) and the Lemma prove the Theorem.

References

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

[2] L. Györfi and Z. Györfi, "On the nonparametric estimate of *a posteriori* probabilities of simple statistical hypotheses," in *Topics in Information Theory*, I. Csiszár and P. Elias, Eds. Amsterdam, The Netherlands: North-Holland, 1977, pp. 298–308.

[3] C. J. Stone, "Consistent nonparametric regression," *Ann. Stat.*, vol. 5, pp. 595–645, 1977.

[4] T. M. Cover, "Rates of convergence for nearest neighbor procedures," in *Proc. Hawaii Int. Conf. on System Sciences*, 1968, pp. 413–415.

[5] L. Györfi, "On the rate of convergence of nearest neighbor rules," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 509–512, July 1978.

[6] J. Beck, "The exponential rate of convergence of error for $k_n - NN$ nonparametric regression and decision," *Prob. Contr. Inform. Theory*, vol. 8, pp. 303–311, 1979.

[7] H. Federer, *Geometric Measure Theory*. New York: Springer, 1969.

[8] L. Györfi, "Recent results on nonparametric regression estimate and multiple classification," *Prob. Contr. Inform. Theory*, vol. 10, pp. 43–52, 1981.

[9] L. Devroye, "On the almost everywhere convergence of nonparametric regression function estimates," *Ann. Stat.*, vol. 9, 1981 (to appear).

[10] T. J. Wagner, "Convergence of the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566–571, Sept. 1971.

[11] J. Fritz, "Distribution-free exponential error bound for nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 552–557, Sept. 1975.

Asymptotic Efficiency of Classifying Procedures using the Hermite Series Estimate of Multivariate Probability Densities

WŁODZIMIERZ GREBLICKI

Abstract—**Pattern recognition procedures derived from a nonparametric estimate of multivariate probability density functions using the orthogonal Hermite system are examined. For sufficiently regular densities, the convergence rate of the mean integrated square error (MISE) is $O(n^{-1+\epsilon})$, $\epsilon > 0$, where n is the number of observations and is independent of the dimension. As a consequence, the rate at which the probability of misclassification converges to the Bayes probability of error as the length n of the learning sequence tends to infinity is also independent of the dimension of the class densities and equals $O(n^{-1/2+\delta})$, $\delta > 0$.**

I. Introduction

In this correspondence we consider a pattern recognition procedure derived from the Hermite series estimate of a multivariate probability density function (pdf). The usage of an orthogonal series to estimate a density was proposed by Čencov [2], while Schwartz [8] gave conditions for the consistency of the estimate employing a system of uniformly bounded orthogonal functions. The consistency with probability 1 was examined by Bosq [1]. For a recursive version of the estimate we refer to Rutkowski [7]. The rate at which the Hermite series estimate converges to a sufficiently smooth density was given by Schwartz [8] and improved by Walter [11].

We estimate, however, a multivariate density and examine the mean integrated square error (MISE). Thanks to an additional assumption, the rate of the error convergence to zero obtained by us is better than that of Walter [11]. Next we apply the estimate in a pattern recognition procedure and show that, for sufficiently regular class densities, the probability of misclassification converges to the Bayes probability of error as rapidly as $O(n^{-1/2+\delta})$, $\delta > 0$, where n is the length of the learning sequence. The rate is independent of the dimension. To the author's knowledge, this nice and rather surprising property has not been observed for other pattern recognition procedures. Typical results say that the rate of the convergence decreases as the dimension increases, see, e.g., Fukunaga and Hostetler [3], Rosenblatt [6] and Van Ryzin [10].

II. The Estimate and Preliminaries

Let $X_1, \cdots, X_n$ be a sample of independent observations of a random variable X having Lebesgue density f. The random variable $X = (X^{(1)}, \cdots, X^{(p)})$ takes values in the p-dimensional Euclidean space R^p. Let $\{h_j; j = 0, 1, \cdots\}$ be the Hermite orthonormal system defined over R, i.e., let

$$h_j(y) = \left(2^j j! \pi^{1/2}\right)^{-1/2} e^{-y^2/2} H_j(y),$$

where

$$H_j(y) = e^{y^2}\left(d^j/dy^j\right)e^{-y^2}$$

is the jth Hermite polynomial. Since, as is well-known, $\{h_{j_1}(x^{(1)})h_{j_2}(x^{(2)}) \cdots h_{j_p}(x^{(p)}); \; j_1, \cdots, j_p = 0, 1, \cdots\}$ $(x^{(1)}, \cdots, x^{(p)}) = x \in R^p$, constitutes a complete orthonormal system over R^p, the estimate of $f(x)$ considered in this correspondence is of the following form:

$$\hat{f}(x^{(1)}, \cdots, x^{(p)}) = \sum_{j_1=0}^{q} \cdots \sum_{j_p=0}^{q} \hat{a}_{j_1,\cdots,j_p} h_{j_1}(x^{(1)}) \cdots h_{j_p}(x^{(p)}),$$

where q depends on n. The coefficient

$$\hat{a}_{j_1,\cdots,j_p} = n^{-1} \sum_{i=1}^{n} h_{j_1}\left(X_i^{(1)}\right) \cdots h_{j_p}\left(X_i^{(p)}\right)$$

is an estimate of

$$a_{j_1,\cdots,j_p} = E\left\{h_{j_1}(X^{(1)}) \cdots h_{j_p}(X^{(p)})\right\}.$$

In our development we make use of the following three properties of the Hermite system:

$$\max_y |h_j(y)| \leq C(j+1)^{-1/12}, \quad (1)$$

$$\max_{|y| \leq A} |h_j(y)| \leq C_A(j+1)^{-1/4}, \quad (2)$$

for any nonnegative A, and

$$\max_{|y| \geq A} |y^{-1/3} h_j(y)| \leq D_A(j+1)^{-1/4} \quad (3)$$

for any positive A.

All these inequalities follow from Szegö [9, theorem 8.91.3, p. 242]. In this correspondence $a_n \sim b_n$ denotes that a_n/b_n has a nonzero limit as n tends to infinity.

Manuscript received April 14, 1980; revised July 30, 1980.

The author is with the Institute of Engineering Cybernetics, Technical University of Wroclaw, Wroclaw, Poland.

0018-9448/81/0500-0364$00.75