HIROSHIMA MATH. J.
22 (1992), 649–655

An asymptotic optimality of the discriminant procedure based on a nearest neighbor density estimator

Zhuyu LI

(Received September 20, 1991)

1. Introduction

In this paper we consider the discriminant problem with two groups Π_1 and Π_2 in nonparametric case. Let X be a p-dimensional random vector with density function $f_\theta(x)$ given θ. Here the distribution of θ is defined by

$$P(\theta = j) = q_j, \qquad j = 1, 2,$$

where q_j's satisfy $q_1 + q_2 = 1$. Suppose that we want to estimate the state $\theta \in \Theta = \{1, 2\}$ based on a sample $Z^n = \{(X_1, \theta_1), \dots, (X_n, \theta_n)\}$ of size n on (X, θ), where (X_i, θ_i), $i = 1, 2, \dots, n$ are i.i.d., and for each i, X_i has density $f_{\theta_i}(x)$. Let $C(j|i)$ be the cost of misclassifying an observation from Π_i as from Π_j, where $C(i|i) = 0$ and $C(i|j) > 0$ for $i \neq j$. We can write a discriminant procedure as $\hat{\theta}(x)$;

$$\hat{\theta}(x) = \begin{cases} 1 & \text{if } x \in H_1 \\ 2 & \text{if } x \in H_2, \end{cases}$$

where H_i's are two disjoint subsets in R^p satisfying $H_1 \cup H_2 = R^p$, $i = 1, 2$. Let $P(j|i)$ be the probability of misclassifying an observation from Π_i as from Π_j, which is given by

$$P(j|i) = P(\hat{\theta}(X) = j | \theta(X) = i).$$

Then an optimum procedure is defined as the one minimizing the expected cost of misclassification (ECM)

$$\begin{aligned} &\sum_i \sum_{i \neq j} P(j|i) C(j|i) q_i \\ &\quad = \tilde{q}_1 P(2|1) + \tilde{q}_2 P(1|2), \end{aligned} \tag{1.1}$$

where $\tilde{q}_1 = q_1 C(2|1)$ and $\tilde{q}_2 = q_2 C(1|2)$.

If $q_1, q_2, f_1(x)$ and $f_2(x)$ are completely known, we can obtain an optimal discriminant procedure called Bayes procedure (see, e.g., Anderson [1]):

$$\theta^*(x) = \begin{cases} 1 & \text{if } D(x) \geq 0 \\ 2 & \text{if } D(x) < 0, \end{cases}$$

where

$$D(x) = \tilde{q}_1 f_1(x) - \tilde{q}_2 f_2(x). \tag{1.2}$$

Then the minimum ECM, i.e., Bayes ECM, is

$$\begin{aligned} R^* &= \tilde{q}_1 \int_{H^c} f_1(x)\,dx + \tilde{q}_2 \int_H f_2(x)\,dx \\ &= \tilde{q}_1 - \int_{R^p} D(x) I_H(x)\,dx, \end{aligned} \tag{1.3}$$

where

$$H = \{x;\, D(x) \geq 0\}, \qquad H^c = R^p - H.$$

In general, q_1, q_2, $f_1(x)$ and $f_2(x)$ are unknown, and consequently, $D(x)$ is so. Wolverton and Wagner [8] introduced an asymptotical optimal discriminant function based on a nonparametrical kernel density estimation. Z. D. Bai [2] obtained an improvement on the conditions of [8]. However, the conditions in their results are too strong to have practical application.

In this paper, we consider the discriminant procedure based on a nearest neighbor density estimator of $f_i(x)$, $i = 1, 2$. It is shown that the procedure is asymptotically optimal in the sense that under an appropriate condition its ECM converges to the one of Bayes procedure. The new discriminant procedure may be more interesting to the practical application because it is more directly, and possess a weaker condition.

2. Main Result

At first, we give the definition of nearest neighbor estimator. Let X be a p-dimensional random vector with an unknown probability density function $f(x)$. Suppose we want to estimate $f(x)$, based on a random sample $\{X_1, X_2, \ldots, X_n\}$ of size n on X.

Definition: (Rao [7]). Let k_n be a nondecreasing sequence of positive integers such that

$$\lim_{n\to\infty} k_n = \infty \tag{2.1}$$

and

$$\lim_{n\to\infty} k_n/n = 0. \tag{2.2}$$

Then, a nearest neighbor estimator (N. N. estimator) of $f(x)$ is defined by

$$f_n(x) = (k_n/n)/|S_{x,a_n(x)}|, \tag{2.3}$$

where $a_n(x)$ is the random distance from x to the k_n-th nearest among $X_1, X_2, \ldots, X_n$, $S_{x,a_n(x)} = \{y\colon \|y - x\| < a_n(x)\}$, $\|x\| = (\sum_{i=1}^{p} x_i^2)^{1/2}$, and $|S|$ is the volume of S in R^p.

Now suppose that k_{n_i} and $a_{n_i}(x)$ satisfy

$$\lim_{n\to\infty} k_{n_i}/n = 0, \qquad i = 1, 2 \tag{2.4}$$

and

$$\lim_{n\to\infty} k_{n_i}/\log n_i = \infty, \qquad i = 1, 2, \tag{2.5}$$

where $n_1 + n_2 = n$, $2 \le k_{n_1} + k_{n_2} \le n$. Let $q_{ni} = n_i/n$, $i = 1, 2$ be the frequence estimators of q_i based on the n_i observations on the group Π_i. Let

$$f_{ni} = (k_{n_i}/n_i)|S_{x,a_{n_i}}(x)|$$

be the N. N. estimators of $f_i(x)$, $i = 1, 2$. Then we get an estimator $D_n(x)$ of $D(x)$ as

$$D_n(x) = \tilde{q}_{n1} f_{n1}(x) - \tilde{q}_{n2} f_{n2}(x) \tag{2.6}$$

where $\tilde{q}_{n1} = q_{n1} C(2|1)$ and $\tilde{q}_{n2} = q_{n2} C(1|2)$. Now, a discriminant procedure is defined as

$$g(x) = \begin{cases} 1 & \text{if } D_n(x) \ge 0 \\ 2 & \text{if } D_n(x) < 0. \end{cases}$$

The ECM of $g(x)$ with given Z^n is

$$\begin{aligned} L_n(g(x), Z^n) &= \tilde{q}_1 P(2|1, Z^n) + \tilde{q}_2 P(1|2, Z^n) \\ &= \tilde{q}_1 - \int_{R^p} D(x) I_{H_n}(x)\, dx, \end{aligned} \tag{2.7}$$

where

$$H_n = \{x; D_n(x) \ge 0\}, \qquad H_n \cup H_n^c = R^p.$$

THEOREM: *Suppose that* (2.4) *and* (2.5) *hold, and*

$$\int_{R^p} f_i^2(x)\, dx < \infty, \qquad i = 1, 2. \tag{2.8}$$

Then

$$L_n(g(x), Z^n) \longrightarrow R^* \qquad a.s. \tag{2.9}$$

From this theorem, we can see that by using the N. N. estimator in our discriminant procedure, the condition for $f_i(x)$ may be only for the quadratic integrability not for the Lipschitz condition or absolute continuity. The conditions required for the N. N. estimator itself are also mild, and it is also easier for simulation.

3. The Proof of Theorem

For the proof of the theorem, we use the following two lemmas. Let $X_1,\ldots,X_n$ be i.i.d. random vectors with values in R^p and with an unknown density function $f(x)$.

LEMMA 3.1 (Devroye and Wagner [3]). *Let $f_n(x)$ be a* N. N. *density estimator of $f(x)$ and k_n, a positive integer satisfying* (2.1), (2.2) *and*

$$\lim_{n\to\infty} k_n/\log n = \infty. \tag{3.1}$$

Then

$$\sup_x |f_n(x) - f(x)| \longrightarrow 0, \qquad a.s. \tag{3.2}$$

LEMMA 3.2 (Zhao [9]). *Suppose that a density function $f(x)$ satisfies*

$$\int_{R^p} f^k(x)\,dx < \infty, \qquad \text{for some } k \ge 2. \tag{3.3}$$

Further, $k_n \le n$, $n = 1, 2,\ldots,$ satisfy (2.2) *and* (3.1). *Then*

$$\int_{R^p} |f_n(x) - f(x)|^k\,dx \longrightarrow 0 \qquad \text{a.s.} \tag{3.4}$$

where $f_n(x)$ is a N. N. *estimator of $f(x)$ formed as* (2.3).

On the other side, it is seen that if (3.4) *is true, then so are* (3.3), (2.1) *and* (2.2).

The proof of theorem:

For any $\varepsilon > 0$, we can choose a bounded set $B \subset R^p$ such as

$$\int_{B^c} f(x)\,dx \le \varepsilon, \tag{3.5}$$

Then, from (1.3) and (2.7):

$$0 \le L_n(g(x), Z^n) - R^* = \int_{R^p} D(x)[I_H(x) - I_{H_n}(x)]\,dx$$

$$= \int_{R^p} D(x)[I_H(x) - I_{H_n}(x)]I_B(x)\,dx$$
$$+ \int_{R^p} D(x)[I_H(x) - I_{H_n}(x)]I_{B^c}(x)\,dx$$
$$\leq \int_{R^p} (D(x) - D_n(x))(I_H(x) - I_{H_n}(x))\,I_B(x)\,dx$$
$$+ \int_{R^p} D(x)(I_H(x) - I_{H_n}(x))\,I_{B^c}(x)\,dx.$$

Here the last inequality is obtained by using

$$-\int_{R^p} D_n(x)(I_H(x) - I_{H_n}(x))\,I_B(x)\,dx \geq 0.$$

Noting that

$$|D(x)| \leq f(x), \qquad |I_H(x) - I_{H_n}(x)| \leq 1,$$

we have

$$0 \leq L_n(g(x), Z^n) - R^* \leq \int_{R^p} |D_n(x) - D(x)|\,I_B(x)\,dx + \int_{R^p} |D(x)|\,I_{B^c}(x)\,dx$$
$$\leq \int_{R^p} |D_n(x) - D(x)|\,I_B(x)\,dx + \varepsilon.$$

From Hölder inequality, we have

$$\int_{R^p} |D(x) - D_n(x)|\,I_B(x)\,dx \leq \left(\int_{R^p} [D(x) - D_n(x)]^2\,dx\right)^{1/2} (|B|)^{1/2},$$

where $|B|$ is the volume of the bounded set B. Further

$$\int_{R^p} (D(x) - D_n(x))^2\,dx \leq 2(J_{n1} + J_{n2}),$$

where

$$J_{n1} = \int_{R^p} |\tilde{q}_{n1} f_{n1}(x) - \tilde{q}_1 f_1(x)|^2\,dx$$

and

$$J_{n2} = \int_{R^p} |\tilde{q}_{n2} f_{n2}(x) - \tilde{q}_2 f_2(x)|^2 \, dx.$$

Since

$$J_{n1} = \int_{R^p} |\tilde{q}_{n1} f_{n1}(x) - \tilde{q}_1 f_1(x)|^2 \, dx$$
$$\leq 2\left(\tilde{q}_{n1}^2 \int_{R^p} |f_{n1}(x) - f_1(x)|^2 \, dx + (\tilde{q}_{n1} - \tilde{q}_1^2) \int_{R^p} f_1^2(x) \, dx \right),$$

by lemma 3.1, lemma 3.2 and (2.8), we obtain

$$J_{n1} \longrightarrow 0 \qquad a.s.$$

Similarly we can show

$$J_{n2} \longrightarrow 0 \qquad \text{a.s.}$$

Therefore

$$\int_{R^p} (D_n(x) - D(x))^2 \, dx \longrightarrow 0 \qquad \text{a.s.}$$

which implies

$$L_n(g(x), Z^n) \longrightarrow R^* \qquad \text{a.s.}$$

In the proof, we ignore the discussion about the mean and variance of $f_n(x)$ and emphasize a direct establishement of (2.9). As a matter of fact, we are concerned only with the convergence of the discriminant procedure formed from training samples Z^n and the N. N. density estimator of unknown conditional density function $f_i(x)$ to a Bayes discrimination. Consequently, it is not essential whether $E\{f_n(x)\}$ or var $\{f_n(x)\}$ could be established or not as in [5].

This research work was partially supported by NSFGC (P. R. China) under Grant 18971061.

References

[1] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1984.

[2] Z. D. Bai, Asymptotical bayes discriminant, ACTA Mathematic Sinica, **28** (1985), 522–529.

[3] L. P. Devroye and T. J. Wagner, The strong uniform consistency of nearest neighbor density estimates, Ann. Statist. **5** (1977), 536–540.

[4] P. A. Lackenbruch and M. Golelstein, Discriminant analysis, Biometrics, **35** (1979), 69–85.
[5] Y. P. Mack and M. Rosenblatt, Multivariate K-nearest neighbor density estimate, J. Multivariate Anysis, **9** (1979), 1–15.
[6] D. S. Moore and J. W. Yackel, Consistency properties of nearest neighbor density estimates, Ann. Statist., **5** (1977), 143–154.
[7] B. L. S. P. Rao, Nonparametric Functional Estimation, Academic Press, New York, 1983.
[8] C. T. Wolverton and T. J. Wagner, Asymptotically optimal discriminant functions for pattern classification, IEEE Trans. Inform. Theory, **11** (1969), 258–266.
[9] Y. Zhao, The L_p-norm consistency of nearest neighbor density estimates, J. Graduate School of Sinica Academic, **4** (1987), 31–41.

Department of Mathematics,
Faculty of Science,
Hiroshima University
and
Department of Mathematics,
Sichuan University