

再论最近邻判别分析*

白志东 陈希孺 陈桂景
(中国科学技术大学) (安徽大学)

一、引言:

设 (X, θ) 为在 $\mathbb{R}^d \times \{1, \dots, M\}$ 中取值的随机向量, 称 X 为指标变量, θ 为类别变量。假定 X 的值已知, 据此要对与其匹配的 θ 值进行判定, 这就是判别分析问题。在 (X, θ) 的分布未知的情况下, 为进行判别, 除利用 X 的当前观察值外, 还要借助于历史知识。设 $(X_i, \theta_i), i = 1, \dots, n$, 是 (X, θ) 的一组 iid 样本, 常称为训练样本。此时我们考虑的问题是, 利用 $(X_i, \theta_i), i = 1, \dots, n$ 及 X 值对 θ 进行判别。假定在 \mathbb{R}^d 中引进了某一距离函数 $\rho(x, y)$, $x, y \in \mathbb{R}^d$ 对给定的 X , 我们可以按照 $\{\rho(X, X_j), j = 1, \dots, n\}$ 由小到大的秩序将指标样本 X_1, \dots, X_n 重新排列为 X_{R_1}, \dots, X_{R_n} 。当 X 的分布 Q 无原子时, 这种排序方法几乎是确定的。^{*}但当 Q 含有原子时, 这种排序法就不确定了。为解决这一问题, 可采取如下处理方法。

设 $\rho(X, X_{i_1}) = \dots = \rho(X, X_{i_N}) \neq \rho(X, X_j)$, 其中 $1 < i_1 < \dots < i_N < n$, $j \in \{1, \dots, n\} - \{i_1, \dots, i_N\}$ 。对于 $\{i_1, \dots, i_N\}$ 的每的置换 $\{\sigma_1, \dots, \sigma_N\}$, 赋予一个概率 $P_{\sigma_1, \dots, \sigma_N}$, 使 $\sum_{(\sigma_1, \dots, \sigma_N)} P_{\sigma_1, \dots, \sigma_N} = 1$ 。于是, 我们可以以概率 $P_{\sigma_1, \dots, \sigma_N}$ 随机地选取 (i_1, \dots, i_N) 的排列 $(\sigma_1, \dots, \sigma_N)$, 把 $\{X_{i_1}, \dots, X_{i_N}\}$ 按照 $X_{\sigma_1}, \dots, X_{\sigma_N}$ 的次序排在一起, 关于概率 $P_{\sigma_1, \dots, \sigma_N}$ 的定义方法, 常可采用下面三种:

(i) 下标升序法, 即取 $P_{i_1, \dots, i_N} = 1$, 其余 $P = 0$ 。

(ii) 下标降序法, 即取 $P_{i_N, \dots, i_1} = 1$, 其余 $P = 0$ 。

(iii) 等权随机化法, 即取 $P_{\sigma_1, \dots, \sigma_N} = \frac{1}{N!}$, 对 (i_1, \dots, i_N) 的任一排列 $(\sigma_1, \dots, \sigma_N)$ 。

把 X_1, \dots, X_n 排好次序 X_{R_1}, \dots, X_{R_n} 以后, $\theta_1, \dots, \theta_n$ 也排成了相应的次序 $\theta_{R_1}, \dots, \theta_{R_n}$ 。所谓 k -近邻 (简记为 k -NN) 判别法, 就是对给定的正整数 $k (< n)$, 考察 $(\theta_{R_1}, \dots, \theta_{R_k})$, 如果有某一 $j \in \{1, \dots, M\}$, 它在 $(\theta_{R_1}, \dots, \theta_{R_k})$ 中出现的次数最多, 则取对 θ 的判别值为 j , 如果有两个或更多的整数, 它们在 $(\theta_{R_1}, \dots, \theta_{R_k})$ 中同样都是出现的最多, 则以等权的方式随机地从中选取一个作为 θ 的判别值, 如此定义的 k -NN 判别值记为 $\theta^{(k)}(X)$ 。当 $k=1$ 时, 简称为 NN 判别。

对于 X 的分布 Q 无原子的情况, Devroye^[3], Wagner^[1], Fritz^[2], 陈希孺^[4], 白志东^[7]等做出了较好的研究。但是在实际问题中, 常会遇到指标变量 X 的分布含有原子乃至为纯原子的情况。这样就引起了人们对 X 的分布 Q 含有原子的情况的研究兴趣。陈桂景、孔繁超^[5],
^[6]及陈桂景、陈刚^[8]考虑了这种情况下的下标升序法和下标降序法 NN 判别问题, 得到了如下一般结果: 记 $X^n = (X_1, \dots, X_n)$, $Z^n = ((X_i, \theta_i), i = 1, \dots, n)$,

$$R = 1 - \sum_{i=1}^M EP^2(\theta = i | X) \quad (1.1)$$

$$R_n = P(\theta \neq \theta'_n), \quad T_n = P(\theta \neq \theta'_n | X^n) \quad (1.2)$$

* 1983年7月12日收到。

$$L_n = P(\theta \neq \theta'_n | Z^n) \quad (1.3)$$

其中 θ'_n 为采用下标升序法或下标降序法的 NN-判别值。

定理 1.1 在 (X, θ) 的任一分布下，总有

$$R_n \xrightarrow{P} R \quad (n \rightarrow \infty), \quad (1.4)$$

$$T_n \xrightarrow{P} R \quad (n \rightarrow \infty). \quad (1.5)$$

并且

$$L_n \xrightarrow{P} R \quad (n \rightarrow \infty) \quad (1.6)$$

的充要条件是，对 X 的分布 Q 的每一个原子 x ，有

$$P(\theta = i, X = x) P(\theta = j, X = x) | P(\theta = i, X = x) - P(\theta = j, X = x) | = 0 \quad (1.7)$$

对 $1 \leq i < j \leq M$ 。

又，当 (1.7) 式不成立时，若采用下标升序法，则 $L_n \xrightarrow{P} L$ ，其中 L 为某一随机变量，若采用下标降序法，则 $L_n a.s.$ 发散。

再，当 Q 为纯原子时，上述结论中关于 T_n, L_n 的 P -收敛均可改为 $a.s.$ 收敛。

从这一结果中可以看到，当 Q 含有原子时，不管采用下标升序法，还是采用下标降序法，除去一些不足道的情况外，后验错判概率 L_n 不收敛于常数值 R 。这说明，利用比较下标的方法进行 NN 判别，至少对于大样本来说有一定的不合理性，此时造成了后验错判概率的不稳定性。其原因在于处理在不同位置的样本，在使用中没有得到平等的待遇。针对这一问题，本文研究了等权随机化最近邻判别分析的性质，得到了后验错判概率 P -收敛于 R 的指数限，从而从大样本理论上论证了等权随机化法的合理性。

以下二、三节中仍采用记号 (1.1)~(1.3)，不过其中 θ'_n 是表示等权随机化 NN 判别。

二、离散指标变量的情形

假定指标变量 X 的分布 Q 为纯原子的，其原子全体为 x_1, x_2, \dots 。又记 $p_k = P(X = x_k)$ ， $k = 1, 2, \dots$ 。此时我们有

定理 2.1 当 Q 为纯原子的分布，则对任何给定的 $\varepsilon > 0$ ，存在 $b = b(\varepsilon) > 0$ ， $c = c(\varepsilon) < \infty$ ，使得有

$$P(|L_n - R| \geq \varepsilon) \leq ce^{-bn}. \quad (2.1)$$

其中 L_n 为等权随机 NN 判别的后验错判概率。

由 Borel-Cantelli 引理及有界控制收敛定理，由此定理立即可得

系 2.1 在定理 2.1 的条件下，有

$$L_n \rightarrow R, a.s., T_n \rightarrow R, a.s., R_n \rightarrow R \quad (n \rightarrow \infty) \quad (2.2)$$

在证明定理 2.1，我们引进记号：

$$A_{n,k} = \{j : X_j = x_k, j \leq n\}, \quad N_{n,k} = |A_{n,k}| \quad (2.3)$$

其中 $|A|$ 表示集合 A 中所含元素的个数。

引理 2.1 对每一 $p_k > 0$ ，有

$$P(N_{n,k} \leq \frac{1}{2} p_k n) \leq e^{np_k/10} \quad (2.4)$$

证：记 $Y_j = I(X_j = x_k)$ ， $j = 1, \dots, n, \dots$ ，则 $\{Y_j\}$ 为 iid 序列， $Y_j \sim B(1, p_k)$ 于是由 Hoeffding [10] 不等式，有

$$\begin{aligned} P(N_{n,k} \leq \frac{1}{2} p_k n) &= P\left(\frac{1}{n} \sum_{j=1}^n Y_j - p_k \leq -\frac{1}{2} p_k\right) \leq \exp\left\{-n\left(\frac{p_k}{2}\right)^2 / \left(2p_k + \frac{p_k}{2}\right)\right\} \\ &= \exp\{-np_k/10\}, \end{aligned}$$

即 (2.4) 式得证。

引理 2.2 在定理 2.1 的条件下, 对每一 $p_k > 0$, $\varepsilon > 0$, 存在 $b = b(k, \varepsilon)$, 使得对 $i = 1, \dots, M$, 有

$$P\left\{\left|\tilde{P}(\theta = i, \theta'_n = i, X = x_k) - \frac{(\eta_i p_{i,k})^2}{p_k}\right| \geq \varepsilon\right\} \leq 3e^{-bn} \quad (2.5)$$

其中记 $\eta_i = P(\theta = i)$, $p_{i,k} = P(X = x_k | \theta = i)$, $\tilde{P}(\cdot) = P(\cdot | Z^n)$ 。

证 由等权随机法及 θ'_n , $A_{n,k}$, $N_{n,k}$ 的定义知

$$\begin{aligned} P\left\{\left|\tilde{P}(\theta = i, \theta'_n = i, X = x_k) - \frac{(\eta_i p_{i,k})^2}{p_k}\right| \geq \varepsilon\right\} &\leq P\left(N_{n,k} \leq \frac{1}{2} p_k n\right) \\ &+ E\left\{\left|\tilde{P}\left[N_{n,k} \geq \frac{1}{2} p_k n; \left|\frac{1}{N_{n,k}} \sum_{j \in A_{n,k}} I_{(\theta_j=i)} \eta_j p_{j,k} - \frac{(\eta_i p_{i,k})^2}{p_k}\right| \geq \varepsilon\right]\right\} \\ &\triangleq I_{n1} + I_{n2} \end{aligned} \quad (2.6)$$

其中记 $\hat{P}(\cdot) = \hat{P}(\cdot | X^n)$ 。由引理 2.1, 有

$$I_{n1} \leq e^{-p_k n/10} \quad (2.7)$$

注意, 当 X^n 给定时, $A_{n,k}$, $N_{n,k}$ 也就给定了, 在这条件下, $I_{(\theta_j=i)}$, $j \in A_{n,k}$ 为 $N_{n,k}$ 个条件 iid 随机变量, 其条件分布与 $I_{(\theta=i)} | X = x_k$ 相同, 因为

$$P(\theta = i | X = x_k) = \frac{\eta_i p_{i,k}}{p_k},$$

应用 Hoeffding 不等式得

$$\begin{aligned} &\hat{P}\left\{\left|\frac{1}{N_{n,k}} \sum_{j \in A_{n,k}} I_{(\theta_j=i)} \eta_j p_{j,k} - \frac{(\eta_i p_{i,k})^2}{p_k}\right| \geq \varepsilon\right\} \\ &\leq 2\exp\left\{-N_{n,k} \left(\frac{\varepsilon}{\eta_i p_{i,k}}\right)^2 / \left(\frac{2p_{i,k}\eta_i}{p_k} + \frac{\varepsilon}{\eta_i p_{i,k}}\right)\right\} \\ &\leq 2\exp\left\{-n\left[\frac{1}{2}p_k \left(\frac{\varepsilon}{\eta_i p_{i,k}}\right)^2 / \left(\frac{2p_{i,k}\eta_i}{p_k} + \frac{\varepsilon}{\eta_i p_{i,k}}\right)\right]\right\} \end{aligned} \quad (2.8)$$

其中最后一步在 $N_{n,k} \geq \frac{1}{2} p_k n$ 时成立。于是由上式得

$$I_{n2} \leq 2\exp\{-nb(k, \varepsilon, i)\}, \quad (2.9)$$

其中 $b(k, \varepsilon, i) = \frac{1}{2} p_k \left(\frac{\varepsilon}{\eta_i p_{i,k}}\right)^2 / \left(\frac{2p_{i,k}\eta_i}{p_k} + \frac{\varepsilon}{\eta_i p_{i,k}}\right)$ 。因此若取 $b = \min\left\{\frac{p_k}{10}, b(k, \varepsilon, i)\right\}$, 将 (2.9), (2.7) 代入 (2.6) 中, (2.5) 式得证。

定理 2.1 的证明 由 R 及 L_n 的定义有

$$R = 1 - \sum_{k=1}^{\infty} \sum_{i=1}^M (P^2(\theta = i | X = x_k) p_k) = 1 - \sum_{k=1}^{\infty} \sum_{i=1}^M \frac{(\eta_i p_{i,k})^2}{p_k},$$

$$L_n = 1 - \tilde{P}(\theta'_n = \theta) = 1 - \sum_{k=1}^{\infty} \sum_{i=1}^M \tilde{P}(\theta'_n = i, \theta = i, X = x_k),$$

故有

$$\begin{aligned} |L_n - R| &\leq \sum_{k=1}^{k_0} \sum_{i=1}^M \left| \tilde{P}(\theta'_n = i, \theta = i, X = x_k) - \frac{(\eta_i p_{i,k})^2}{p_k} \right| \\ &+ \sum_{k=k_0+1}^{\infty} \sum_{i=1}^M \tilde{P}(\theta'_n = i, \theta = i, X = x_k) + \sum_{k=k_0+1}^{\infty} \sum_{i=1}^M \frac{(\eta_i p_{i,k})^2}{p_k} \triangleq J_{n1} + J_{n2} + J_{n3}. \end{aligned} \quad (2.10)$$

取 k_0 充分大，使得对给定的 $\varepsilon > 0$ ，有 $\sum_{k=k_0+1}^{\infty} p_k < \frac{\varepsilon}{6}$ 。于是由 Z^n 与 (X, θ) 的独立性知，

$$J_{n2} \leq \sum_{k=k_0+1}^{\infty} \sum_{i=1}^M P(\theta = i, X = x_k) = \sum_{k=k_0+1}^{\infty} P(X = x_k) < \frac{\varepsilon}{6} \quad (2.11)$$

注意到 $\eta_i p_{i,k} / p_k = P(\theta = i | X = x_k) < 1$ ，从而

$$\sum_{i=1}^M \frac{(\eta_i p_{i,k})^2}{p_k} < \sum_{k=1}^M P(\theta = i | X = x_k) = p_k,$$

故有

$$J_{n3} < \frac{\varepsilon}{6} \quad (2.12)$$

于是由 (2.10) ~ (2.12)，便得

$$\begin{aligned} P(|L_n - R| > \varepsilon) &\leq P\left(\sum_{k=1}^{k_0} \sum_{i=1}^M \left| \tilde{P}(\theta'_n = i, \theta = i, X = x_k) - \frac{(\eta_i p_{i,k})^2}{p_k} \right| > \frac{2}{3}\varepsilon\right) \\ &\leq \sum_{k=1}^{k_0} \sum_{i=1}^M P\left(\left| \tilde{P}(\theta'_n = i, \theta = i, X = x_k) - \frac{(\eta_i p_{i,k})^2}{p_k} \right| > \frac{2\varepsilon}{3k_0 M}\right) \end{aligned}$$

应用引理 2.2，并取

$$b = \min_{1 \leq k \leq k_0} b\left(k, \frac{2\varepsilon}{3k_0 M}\right), \quad c = 3k_0 M,$$

即得 $P(|L_n - R| > \varepsilon) < ce^{-bn}$ ，定理 2.1 得证。

三、一般指标变量的情形

设指标变量 X 为一般的 d -维随机向量，其分布 Q 具有 Lebesgue 分解： $Q = \lambda Q_1 + \mu Q_2$ ， $\lambda \in [0, 1]$ ， $\mu = 1 - \lambda$ 其中 Q_2 无原子，而 Q_1 为纯原子，其原子为 $x_k, 0 < p_k = Q_1(\{x_k\})$ ， $k = 1, 2, \dots$ 。设 $f(x)$ 为某一定义于 \mathbb{R}^d 上的 Borel 可测函数，称其为 KQ 连续的，如果存在一个 a.s.， x, Q_2 连续的函数 $\tilde{f}(x)$ ，使得有

$$f(x) = \tilde{f}(x), \text{ 当 } x \neq x_k, k = 1, 2, \dots. \lim_{k \rightarrow \infty} (f(x_k) - \tilde{f}(x_k)) = 0.$$

例如，如果 Q 的支撑为 $[0, 1]$ ，在每个有理点上有正测度，且总测度为 $\frac{1}{2}$ ，其余部分为均匀分布，总测度为 $\frac{1}{2}$ ，那么，在 $[0, 1]$ 上的 Riemann 函数是 KQ 连续的，但 Dirichlet 函数则不是。

以下记

$$\eta_i(x) = P(\theta = i | X = x), x \in \mathbb{R}^d, i = 1, \dots, M \quad (3.1)$$

定理 3.1 如果每个 $\eta_i(x)$ 都是 KQ 连续的，则对任一给定的 $\varepsilon > 0$ ，存在 $b = b(\varepsilon) > 0$ ， $c = c(\varepsilon) < \infty$ ，使在等权随机 NN 判别下有

$$P(|L_n - R| > \varepsilon) < ce^{-b\sqrt{n}} \quad (3.2)$$

证 因在 Q 的 Lebesgue 分解中, 当 $\lambda = 0$ 时, (3.2) 是 Fritz^[2]的结果, 而当 $\lambda = 1$ 时, 我们在定理 2.1 中证得了更强的结果。故以下我们仅考虑 $\lambda \in (0, 1)$ 的情形。对任给的 $\varepsilon' > 0$, 存在正整数 k_0 , 使得

$$(i) |\eta_i(x_k) - \tilde{\eta}_i(x_k)| < \varepsilon', \text{ 对 } k \geq k_0, i = 1, \dots, M \quad (3.3)$$

$$(ii) \sum_{k=k_0+1}^{\infty} p_k < \varepsilon' \quad (3.4)$$

记 $\Omega_0 = \{x_1, \dots, x_{k_0}\}$, $\Omega_1 = \{x_{k_0+1}, x_{k_0+2}, \dots\}$, $\Omega_2 = \mathbb{R}^d - (\Omega_0 \cup \Omega_1)$ 。则有

$$\tilde{P}(\theta = \theta'_n, X \in \Omega_1) \leq P(X \in \Omega_1) = \lambda \sum_{k=k_0+1}^{\infty} p_k < \varepsilon' \quad (3.5)$$

其中记号 \tilde{P} 仍表示在 Z^n 条件下的条件分布。

考虑事件

$E_{n1} = \{\text{对每一 } k < k_0, \text{ 至少有一个 } X_j = x_k, j \leq n\}$ 。则有

$$P(E_{n1}) \leq k_0(1-p)^n \leq ce^{-bn} \quad (3.6)$$

其中 $p = \min_{1 \leq k \leq k_0} p_k$, $c = k_0$, $b = -\log(1-p)$ 。以下我们以 c 和 b 表示与 n 无关 (可与 ε' 有关) 的常数, 但每次出现时其值不必相同。

仍用 (2.3) 定义的记号 $A_{n,k}$, $N_{n,k}$, 那么当 E_{n1} 发生时, $N_{n,k} \geq 1$, $k = 1, \dots, k_0$ 。因此对 $k < k_0$, 当 $X = x_k$ 时, 必有 $X'_n = x_k$, 此处 X'_n 表示用等权随机 NN 法其判别 θ'_n 所对应的指标值。故由等权随机 NN 法定义知,

$$\tilde{P}(\theta = \theta'_n, X \in \Omega_0) = \lambda \sum_{k=1}^{k_0} \sum_{i=1}^M \left\{ \left(\frac{1}{N_{n,k}} \sum_{j \in A_{n,k}} I_{(\theta_j=i)} \right)^2 p \eta_i(x_k) \right\}$$

类似定理 2.1 的证明可得

$$P \left\{ E_{n1}, \left| \tilde{P}(\theta = \theta'_n, X \in \Omega_0) - \lambda \sum_{k=1}^{k_0} \sum_{i=1}^M \eta_i^2(x_k) p_k \right| \geq \varepsilon' \right\} \leq ce^{-bn} \quad (3.7)$$

由 (3.4) 式, 有

$$\lambda \sum_{k=k_0+1}^{\infty} \sum_{i=1}^M \eta_i^2(x_k) p_k \leq \lambda \sum_{k=k_0+1}^{\infty} p_k < \varepsilon' \quad (3.8)$$

故由 (3.5)~(3.8), 得

$$P(|\tilde{P}(\theta = \theta'_n, X \in \Omega_0 \cup \Omega_1) - \lambda \sum_{i=1}^M \sum_{k=1}^{\infty} \eta_i^2(x_k) p_k| \geq 3\varepsilon') \leq ce^{-bn} \quad (3.9)$$

对给定的 X^n , 记

$$B_n = \{j : X_j \in \Omega_2, j \leq n\}, B'_n = \{1, \dots, n\} - B_n \quad (3.10)$$

又对每一 $j \in B'_n$, 记

$$D_{n,j} = \{i : X_i = X_j, i \in B'_n\}, M_{n,j} = |D_{n,j}| \quad (3.11)$$

$j = 1, \dots, n$ 。又记

$$V_{n,j} = \{x : \rho(X_j, x) = \min_{1 \leq i \leq n} \rho(X_i, x), x \in \mathbb{R}^d\} \quad (3.12)$$

$j = 1, \dots, n$ 。注意当 $j \neq j'$ 时, $V_{n,j} \cap V_{n,j'} = \emptyset$ 或 $V_{n,j} = V_{n,j'}$ 。但因 Ω_2 中无原子, 故不失一般性可设对任一 $j \in B_n$, $j \neq j'$, 有 $V_{n,j} \cap V_{n,j'} = \emptyset$ 。但若 $j \in B'_n$, 且 $j' \in D_{n,j}$ 时, 有 $V_{n,j} = V_{n,j'}$ 。由这些记号及等权随机 NN 判别的定义, 我们有

$$\tilde{P}(\theta = \theta'_n, X \in \Omega_2) = \mu \sum_{i=1}^M \left\{ \sum_{j \in \mathbf{B}_n} I_{(\theta_j=i)} \int_{V_{nj}} \eta_i(x) Q_2(dx) + \sum_{j \in \mathbf{B}'_n} \frac{1}{M_{n,j}} I_{(\theta_j=i)} \int_{V_{nj}} \eta_i(x) Q_2(dx) \right\} \quad (3.13)$$

记 $\hat{E}(\cdot) = E(\cdot | X^n)$, 并令

$$H_n = \hat{E}\{\tilde{P}(\theta = \theta'_n, X \in \Omega_2)\} \\ = \mu \sum_{i=1}^M \left\{ \sum_{j \in \mathbf{B}_n} \eta_i(X_j) \int_{V_{nj}} \eta_i(x) Q_2(dx) + \sum_{j \in \mathbf{B}'_n} \frac{1}{M_{n,j}} \eta_i(X_j) \int_{V_{nj}} \eta_i(x) Q_2(dx) \right\} \quad (3.14)$$

为继续进行定理 3.1 的证明, 我们给出下面一个事实, 其证明不难, 故略去。

引理 3.1 设 ξ_1, \dots, ξ_n 的均值为零, 以 1 为界且相互独立的随机变量, 又 a_1, \dots, a_n 为非负常数, $\sum_{i=1}^n a_i < 1$, 则对任一 $\varepsilon > 0$, 有

$$P\left(\left|\sum_{i=1}^n a_i \xi_i\right| \geq \varepsilon\right) \leq 2 \exp\left\{-\varepsilon^2 / 4 \sum_{i=1}^n a_i^2\right\} \quad (3.15)$$

仍记 $\hat{P}(\cdot) = P(\cdot | X^n)$ 。注意当 X^n 给定时, $\{I_{(\theta_j=i)} - \eta_i(X_j), j=1, \dots, n\}$ 为条件独立、均值为 0, 以 1 为界的随机变量。又注意

$$\sum_{j \in \mathbf{B}_n} \int_{V_{nj}} \eta_i(x) Q_2(dx) + \sum_{j \in \mathbf{B}'_n} \frac{1}{M_{n,j}} \int_{V_{nj}} \eta_i(x) Q_2(dx) \leq 1 \\ \sum_{j \in \mathbf{B}_n} \left(\int_{V_{nj}} \eta_i(x) Q_2(dx) \right)^2 + \sum_{j \in \mathbf{B}'_n} \left(\frac{1}{M_{n,j}} \int_{V_{nj}} \eta_i(x) Q_2(dx) \right)^2 \leq \max_{1 \leq j \leq n} \{U_{nj}\} \triangleq U_n.$$

其中记

$$U_{nj} = \begin{cases} Q_2(V_{nj}) & \text{当 } j \in \mathbf{B}_n \\ \frac{1}{M_{n,j}} Q_2(V_{nj}) & \text{当 } j \in \mathbf{B}'_n \end{cases}$$

于是由引理 3.1 可得

$$\hat{P}(|\tilde{P}(\theta = \theta'_n, X \in \Omega_2) - H_n| \geq \varepsilon') \\ \leq 2M \exp\left\{-(\varepsilon')^2 / 4MU_n\right\} \leq 2M \sum_{j=1}^n \exp\left\{-\tilde{\varepsilon}/U_{nj}\right\} \quad (3.16)$$

其中记 $\tilde{\varepsilon} = (\varepsilon')^2 / 4M$, 注意 $U_{nj}, j=1, \dots, n$ 分布相同。于是由上式可得

$$\hat{P}(|\tilde{P}(\theta = \theta'_n, X \in \Omega_2) - H_n| \geq \varepsilon') \leq n2M \exp\left\{-\tilde{\varepsilon}/U_{n1}\right\} \quad (3.17)$$

记 $N_n = |\mathbf{B}_n|$, 并以 Y_1, \dots, Y_{N_n} 表示落入 Ω_2 中的诸 X_j , 由 Hoeffding 不等式, 有

$$P\left\{N_n < \frac{1}{2}\mu n\right\} \leq \exp\left\{-n\left(\frac{\mu}{2}\right)^2 / \left(2\mu + \frac{\mu}{2}\right)\right\} = ce^{-bn} \quad (3.18)$$

对 $j = 1, \dots, N_n$, 再记

$$\tilde{V}_{nj} = \{x : \rho(Y_j, x) = \min_{1 \leq i \leq N_n} (\rho(X_i, x), \rho(Y_i, x))\}$$

因为总有 $\tilde{V}_{n1} \subset \tilde{V}_{nj}$. 当 $1 \in \mathbf{B}_n$ 时, $Y_1 = X_1$, 于是由 Fritz^[2] 的方法, 并利用 (3.18) 式可得

$$E[\exp\{-\tilde{\varepsilon}/U_{n1}\} I_{(1 \in \mathbf{B}_n)}] \leq P(1 \in \mathbf{B}_n) e^{-\delta\sqrt{N_n}} \leq ce^{-b\sqrt{n}} \quad (3.19)$$

但当 $1 \in \mathbf{B}'_n$ 时, X_1 为 Q_2 的非支撑点, 不难看出 Fritz^[2] 方法仍有效, 仍可得

$$E[\exp\{-\tilde{\varepsilon}/U_{n1}\} I_{(1 \in \mathbf{B}'_n)}] \leq ce^{-b\sqrt{n}} \quad (3.20)$$

把 (3.19), (3.20) 代入 (3.17), 便得

合 (3.21), (3.28) 及上式, 便得

$$P\left(\left|\tilde{P}(\theta'_n = \theta, X \in \Omega_2) - \mu \sum_{i=1}^M \int \eta_i^2(x) Q_2(dx)\right| \geq 5\varepsilon'\right) < ce^{-b\sqrt{n}} \quad (3.30)$$

由上式及 (3.9) 式, 便得

$$P\left(\left|\tilde{P}(\theta'_n = \theta) - \sum_{i=1}^M E\eta_i^2(X)\right| \geq 8\varepsilon'\right) < ce^{-b\sqrt{n}}$$

此式等价于

$$P(|L_n - R| \geq 8\varepsilon') < ce^{-b\sqrt{n}}$$

由 $\varepsilon' > 0$ 的任意性, 定理 3.1 得证。

最后, 我们再给出两个定理, 其证明不难由综合陈、孔^[6]的方法及如上使用的方法而得到, 故略去。

定理 3.2 在 (X, θ) 的任一分布下, 若使用等权随机化 NN 判别法, 总有

$$R_n \xrightarrow{p} R, \quad T_n \xrightarrow{p} R, \quad L_n \xrightarrow{p} R \quad (n \rightarrow \infty) \quad (3.31)$$

定理 3.3 在定理 3.1 的条件下, 若采用下标升序 NN 判别法, 则有

$$\lim_{n \rightarrow \infty} L_n = L(Z) \quad a.s. \quad (3.32)$$

其中 $Z = ((X_i, \theta_i), i = 1, 2, \dots)$,

$$L(Z) = 1 - \sum_{i=1}^M E[P^2(\theta = i | X) I_{(X \in \Omega_i)}] - \sum_{i=1}^M \sum_{k=1}^{\infty} I_{(\theta_{j_k} = i)} P(\theta = i, X = x_k),$$

$$j_k = \min\{j : X_j = x_k\}$$

并且, $L = R a.s.$ 的充要条件是对 Q 的每一原子 x_k , 有

$$P(\theta = i, X = x_k) P(\theta = j, X = x_k) [P(\theta = i, X = x_k) - P(\theta = j, X = x_k)]^2 = 0,$$

对 $|i-j| \leq M$ 。

四、补充与讨论

(1) 以上结果是对随机化 NN 判别而言的, 若采用等权随机 k -NN 判别法, 同样可以证明如下结果:

定理 4.1 不论 (X, θ) 的分布如何, 都有

$$L_n^{(k)} \xrightarrow{p} R^{(k)} \quad \text{当 } n \rightarrow \infty \quad (4.1)$$

而当 (X, θ) 的分布满足定理 3.1 中关于 KQ 连续性要求时, 则有

$$L_n^{(k)} \xrightarrow{a.s.} R^{(k)}, \quad \text{当 } n \rightarrow \infty \quad (4.2)$$

其中 $L_n^{(k)} = P(\theta_n^{(k)}(X) \neq \theta | Z^n)$ 为在等权随机 k -NN 判别 $\theta_n^{(k)}$ 下的条件错判概率, 而常数 $R^{(k)}$ 定义如下: 记 $U = \{1, \dots, M\}$, $C_i(u) = C_i(u_1, \dots, u_k) = \{u_j : u_j = i, j = 1, \dots, k\}$, $u \in U$

$$\text{令 } G_i(x) = \sum_{u \in T_{k_i}} P(\theta = u | x) \cdots P(\theta = u_k | x)$$

其中

$$T_{k_i} = \{u : u \in U^k, C_j(u) < C_i(u) > C_l(u), \text{ 当 } j = 1, \dots, i-1 \text{ 及 } l = i+1, \dots, k\}$$

则

$$R^{(k)} = 1 - E\left(\sum_{i=1}^M P(\theta = i | X) G_i(X)\right) \quad (4.3)$$

(2) 在对 (X, θ) 的分布不加任何限制下, 我们又得到如下结果:

$$P\{|\tilde{P}(\theta'_n = \theta, X \in \Omega_2) - H_n| \geq \varepsilon'\} \leq ce^{-b\sqrt{n}} \quad (3.21)$$

记 $S_r = \bigcap_{k=1}^{k_0} \{x : \rho(x, x_k) \geq r, x \in \mathbb{R}^d\}$ 。取闭球 S 充分大以及 $r > 0$ 充分小，使得
 $Q_2((S \cap S_r)^c) < \varepsilon'$ (3.22)

并且有

$$\inf_{x \in S^c} \min_{1 \leq k \leq k_0} \rho(x, x_k) > r \quad (3.23)$$

又记 Q 的支撑为 F ，定义

$$F_n^1 = \sup_{x \in F \cap S} \left\{ \inf_{1 \leq j \leq n} \rho(X_j, x) \right\} \quad (3.24)$$

用 Wagner^[3]的方法可以证明

$$P(F_n^1 \geq \frac{1}{2}r) \leq ce^{-bn} \quad (3.25)$$

注意，当 $F_n^1 \leq \frac{1}{2}r$ 时，若 $X_j = x_k$ ，对某 $k \leq k_0$ ，则必有 $S \cap S_r \cap V_{nj} = \emptyset$ 。由此及 (3.3) 式，得

$$\begin{aligned} & \left| \sum_{i=1}^M \left\{ \sum_{j \in B_n} \eta_i(X_j) \int_{V_{nj} \cap S \cap S_r} \eta_i(x) Q_2(dx) + \sum_{j \in B_n} \frac{1}{M_{n,j}} \eta_i(X_j) \int_{V_{nj} \cap S \cap S_r} \eta_i(x) Q_2(dx) \right\} \right| \\ & - \sum_{i=1}^M \left\{ \sum_{j \in B_n} \tilde{\eta}_i(X_j) \int_{V_{nj} \cap S \cap S_r} \eta_i(x) Q_2(dx) + \sum_{j \in B_n} \frac{1}{M_{n,j}} \tilde{\eta}_i(X_j) \int_{V_{nj} \cap S \cap S_r} \tilde{\eta}_i(x) Q_2(dx) \right\} \Big| \\ & = \left| \sum_{i=1}^M \left\{ \sum_{j \in B_n} (\eta_i(X_j) - \tilde{\eta}_i(X_j)) \int_{V_{nj} \cap S \cap S_r} \tilde{\eta}_i(x) Q_2(dx) \right. \right. \\ & \left. \left. + \sum_{j \in B_n} \frac{1}{M_{n,j}} (\eta_i(X_j) - \tilde{\eta}_i(X_j)) \int_{V_{nj} \cap S \cap S_r} \tilde{\eta}_i(x) Q_2(dx) \right\} \right| < \varepsilon' \end{aligned} \quad (3.26)$$

在上面的推导中，注意 $\tilde{\eta}_i$ 与 η_i 仅在可列个点上可能取不同的值，而 Q_2 无原子，故有

$$\int_B \eta_i(x) Q_2(dx) = \int_B \tilde{\eta}_i(x) Q_2(dx), \text{ 对任意 } B.$$

由 (3.22)，(3.26)，便得

$$\begin{aligned} & \left| \sum_{i=1}^M \left\{ \sum_{j \in B_n} \eta_i(X_j) \int_{V_{nj}} \eta_i(x) Q_2(dx) + \sum_{j \in B_n} \frac{1}{M_{n,j}} \eta_i(X_j) \int_{V_{nj}} \eta_i(x) Q_2(dx) \right\} \right| \\ & - \sum_{i=1}^M \left\{ \sum_{j \in B_n} \tilde{\eta}_i(X_j) \int_{V_{nj}} \tilde{\eta}_i(x) Q_2(dx) + \sum_{j \in B_n} \frac{1}{M_{n,j}} \tilde{\eta}_i(X_j) \int_{V_{nj}} \tilde{\eta}_i(x) Q_2(dx) \right\} \Big| < 3\varepsilon' \end{aligned} \quad (3.27)$$

故有

$$P(|H_n - H_{n1}| \geq 3\varepsilon') \leq P(F_n^1 \geq \frac{1}{2}r) \leq ce^{-bn} \quad (3.28)$$

其中记

$$H_{n1} = \mu \sum_{i=1}^M \left\{ \sum_{j \in B_n} \tilde{\eta}_i(X_j) \int_{V_{nj}} \tilde{\eta}_i(x) Q_2(dx) + \sum_{j \in B_n} \frac{1}{M_{n,j}} \tilde{\eta}_i(X_j) \int_{V_{nj}} \eta_i(x) Q_2(dx) \right\}$$

再用 Fritz^[2]同样方法不难证明

$$P(|H_{n1} - \mu \sum_{i=1}^M \int \tilde{\eta}_i^2(x) Q_2(dx)| \geq \varepsilon' | N_n) \leq ce^{-bN_n}$$

$$\text{由此及 (3.18)，便得 } P(|H_{n1} - \mu \sum_{i=1}^M \int \tilde{\eta}_i(x) Q_2(dx)| \geq \varepsilon') \leq ce^{-nb} \quad (3.29)$$

定理 4.2 不论 (X, θ) 的分布如何, 对任给的 $\varepsilon > 0$, 及一串满足条件

$$C_n \uparrow \infty, \quad C_n/n \rightarrow 0 \quad (4.4)$$

的常数列 $\{C_n\}$, 必存在一个只与给定的 ε 有关而与 n, C_n 和 (X, θ) 的分布无关的常数 $C > 0$, 以及判别法 \tilde{Q}_n , 使

$$P(|P(\tilde{Q}_n \neq \theta | Z^n) - R^{(k)}| > \varepsilon) = O(e^{-Cn}) \quad (4.5)$$

证 需要下面熟知的不等式 (见^[10]): 设随机变量 ξ_1, \dots, ξ_n 独立, 均值都为 0, 且存在常数 $b > 0$, 使 $|\xi_i| < b, i = 1, \dots, n$, 则

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n \xi_i\right| > \varepsilon\right) < 2 \exp\left\{-\frac{n\varepsilon^2}{2(b^2 + b\varepsilon)}\right\} \quad (4.6)$$

记 $R_n^{(k)} = P(\theta_n^{(k)} \neq \theta)$, 由定理 4.1 及控制收敛定理,

$$\lim_{n \rightarrow \infty} R_n^{(k)} = R^{(k)} \quad (4.7)$$

取 n_0 充分大, 使当 $n > n_0$ 时, 有 $|R_n^{(k)} - R^{(k)}| < \varepsilon/2$, 再找 n_1 充分大, 使当 $n > n_1$ 时, $n/C_n > 2n_0$ 。

现设有了样本 Z^n , 并给定了 X 。构造一个判别法 $\tilde{\theta}_n$ 如下: 标出 $b_n = [n/\lfloor c_n \rfloor]$, 此处 $\lfloor a \rfloor$ 为不超过 a 的最大整数, 且暂记 $c'_n = \lfloor c_n \rfloor$, 把 X_1, \dots, X_n 排列为一个矩阵如下:

$$\begin{matrix} X_1, & X_2, & \cdots & \cdots & X_{b_n} \\ X_{b_n+1}, & X_{b_n+2}, & \cdots & \cdots & X_{2b_n} \\ \vdots & \vdots & & & \vdots \\ X_{(c'_n-1)b_n+1}, & X_{(c'_n-1)b_n+2}, & \cdots, & X_{c'_n b_n} \end{matrix}$$

上述各行的 X 值, 连同各自的 θ 匹配值, 依次记为 $Z_{n1}, \dots, Z_{nc'_n}$ 。例如, $Z_{n1} = ((X_1, \theta_1), \dots, (X_{b_n}, \theta_{b_n}))$, 等等。现以等概率分布从 $1, 2, \dots, c'_n$ 这 c'_n 个数字中随机选择一个, 若选出的是 i , 利用 Z_{ni}, X , 采用等权随机 k -NN 法去判别 θ , 记其判别值为 $\tilde{\theta}_n$ 。于是有

$$P(\tilde{\theta}_n \neq \theta | Z^n) = \frac{1}{c'_n} \sum_{i=1}^{c'_n} P(\tilde{\theta}_n \neq \theta | Z_{ni}) = \frac{1}{c'_n} \sum_{i=1}^{c'_n} L_{b_n}^{(k)}(Z_{ni})$$

因为 $EL_{b_n}^{(k)}(Z_{ni}) = R_{b_n}^{(k)}$, 当 $n > n_1$ 时, $n/c'_n > n/c_n > 2n_0$ 。故 $b_n = [n/c'_n] > n_0$, 因而 $|R_{b_n}^{(k)} - R^{(k)}| < \varepsilon/2$, 故有

$$P(|P(\tilde{\theta}_n \neq \theta | Z^n) - R^{(k)}| > \varepsilon) < P\left(\left|\frac{1}{c'_n} \sum_{i=1}^{c'_n} L_{b_n}^{(k)}(Z_{ni}) - R_{b_n}^{(k)}\right| > \varepsilon/2\right) \text{ 当 } n > n_1 \quad (4.8)$$

因为 $L_{b_n}(Z_{ni}), i = 1, \dots, n$, 是 iid 序列, 取值于 $[0, 1]$, 且均值为 $R_{b_n}^{(k)}$, 依不等式 (4.6), 有

$$P\left(\left|\frac{1}{n} \sum_{i=1}^{c'_n} L_{b_n}(Z_{ni}) - R_{b_n}\right| > \frac{\varepsilon}{2}\right) < 2 \exp\left(-\frac{c'_n (\varepsilon/2)^2}{2 + 2\varepsilon/2}\right) < 2 \exp\left(-\frac{c_n \varepsilon^2}{16 + 8\varepsilon}\right)$$

由此式及 (4.8) 式, 取 $c = \varepsilon^2/(16 + 8\varepsilon)$, (4.5) 式成立。定理 4.2 得证。

(3) 在上述定理 4.2 中, 若选择 $\{c_n\}$, 使 $c_n/n \rightarrow 0$ 且 $c_n/\log n \rightarrow \infty$, 将有

$$P(\tilde{\theta}_n \neq \theta | Z^n) \rightarrow R^{(k)}, \quad a.s. \quad (n \rightarrow \infty) \quad (4.9)$$

而且收敛速度可任意接近指数型 $O(e^{-cn})$ 。当然, 判别 $\tilde{\theta}_n$ 的构造在应用上并不吸引人, 因为它实质上只利用了样本 Z^n 中的一小部分。从直观上看, 等权随机 k -NN 判别 $\theta_n^{(k)}$ 要优于

$\tilde{\theta}_n$ 。因此就引起了一个猜测：在任一 (X, θ) 的分布下，可能都有 $P(\theta_n^{(k)} \neq \theta | Z^n) \rightarrow R^{(k)}$, a.s. ($n \rightarrow \infty$)，甚至有指型收敛速度

$$P(|P(\theta_n^{(k)} \neq \theta | Z^n) - R^{(k)}| > \varepsilon) = O(e^{-b(\varepsilon)n})$$

(4) 我们已证得，在定理 3.1 条件下，采用等权随机 k -NN 法，必有 $P(\theta_n^{(k)} \neq \theta | Z^n) \rightarrow R^{(k)}$, a.s. 那么，另一个有兴趣的问题是，在定理 3.1 的条件下，是否存在另一种 k -NN 法 $\hat{\theta}_n$ ，使 $P(\hat{\theta}_n \neq \theta | Z^n)$ 以概率 1 收敛于一个与 $R^{(k)}$ 不等的另一常数？我们可以证明，只要 k 近邻的选择不依赖于 X_1, \dots, X_n 的匹配值 $\theta_1, \dots, \theta_n$ ，则这是不可能的。但是当 k -近邻的选择中可以与 θ 的观察值 $\theta_1, \dots, \theta_n$ 有关时，则情况有所不同。

(5) 最后需要说明的是，在 \mathbb{R}^d 空间中距离函数 $\rho(x, y)$ 的选择，需要使其满足下述不强的条件，即在此距离下，使 \mathbb{R}^d 能分解成有限个“ ω 锥”，这是因为 Fritz^[2] 需要这一条件，而我们在论证中引用了他的结果。不过 \mathbb{R}^d 中常见的距离都满足这一要求，例如 $\|x - y\|^2 = \sum_{i=1}^M a_i (x_i - y_i)^2$, $\|x - y\| = \max_{1 \leq i \leq d} |a_i (x_i - y_i)|$ 等，其中 $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d)$, $a_i > 0$, $i = 1, \dots, d$ 。

参 考 文 献

- [1] Wagner, T.J., IEEE Trans. Inform. Theory, 1971, 566—570.
- [2] Fritz, J., IEEE Trans. Inform. Theory, 1975, 552—557.
- [3] Devroye, L., Ann. Statist., 1981, 1320—1327.
- [4] 陈希孺， k -近邻判别法后验错判概率的指数限（待发表）。
- [5] 陈桂景、孔繁超，安徽大学学报（自然科学版），1983, 1:18—25.
- [6] Chen Gui-jing & Kong Fan-chao, Sufficient and Necessary Condition For Covergence of Conditional Error Probability In NN-Pattern Discrimination, 《数学研究与评论》，本期（Vol.6, No.1）。
- [7] Bai Zhidong, Strong Consistency of Error Probability Estimates in NN Discrimination (已投《数学年刊》)。
- [8] 陈桂景、陈刚，下标降序法最近邻判别分析（已投《数学进展》）。
- [9] 陈希孺，平方损失下的最近邻预测理论，系统科学与数学，1983, 3:213—219.
- [10] Hoeffding, W., J. Amer. Statist. Assoc. 58(1963), 13—50.