On the Rate of Convergence of Error Estimates for the Partitioning Classification Rule *

Márta Pintér †

Dept. of Computer Science and Information Theory,

Budapest University of Technology and Economics

1521 Stoczek u. 2, Budapest, Hungary.

marti@szit.bme.hu

Abstract

The error probability of the partitioning classification rule is shown to converge to the Bayes error faster than $1/\sqrt{n}$ under certain conditions. The resubstitution and the deleted error estimates for the partitioning classification rule from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ are studied. The random part of the resubstitution estimate is shown to be small for arbitrary partition and for any distribution of (X, Y). If we assume that X has a density f and the partitions consist of rectangles, then the difference between the expected value of the estimate and the Bayes error restricted to the partition is less than a constant times $1/\sqrt{n}$. The main result of the paper is that, under the same conditions, for both estimates the difference between the estimate and the real error probability of the classification rule is asymptotically normal with 0 mean and variance $L^*/2$, where L^* is the Bayes error.

1 Introduction

Let X be the d-dimensional feature vector with distribution μ, and let Y be the binary valued label. Denote the posterior probabilities by

$$P_i(x) = \mathbf{P}\{Y = i | X = x\},\ i = 0, 1.$$

In pattern recognition the value of the label Y is to be predicted upon observing the feature vector X. The prediction rule or classifier g is a function $\mathcal{R}^d \to \{0, 1\}$, whose performance is measured by the probability of error

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

*Parts of the paper were presented at EUROCOLT'99.

†The research of the author was supported by the Computer and Automation Institute of the Hungarian Academy of Sciences (MTA SZTAKI).

The Bayes classifier

$$g^*(x) = \begin{cases} 0, & \text{if } P_1(x) \le 1/2 \\ 1, & \text{otherwise.} \end{cases}$$

is well-known to have minimal probability of error among all possible classifiers. Its error probability $L(g^*)$ is called the Bayes error, and is denoted by L^*.

$$L^* = \mathbf{E}\{\min(P_0(X), P_1(X))\}.$$

Assume that n independent copies of (X, Y) form the available data sequence:

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

These data may be used to design the classification rule $g_n(x)$, whose probability of error is the random variable

$$L_n = L(g_n) = \mathbf{P}\{g_n(X) \ne Y | D_n\}.$$

Many important classification rules partition $\mathcal{R}^d$ into disjoint cells and classify in each cell according to the majority vote among the labels of the sample points X_i falling in the same cell. Let $\mathcal{P}_n = \{A_{nj}, j = 1, 2, \ldots\}$ be a partition of $\mathcal{R}^d$ and let $A_n(x)$ denote the cell in the partition that includes x. Then the partitioning classification rule:

$$g_n(x) = \begin{cases} 0, & \text{if } \sum_{i=1}^n I_{\{Y_i=1\}} I_{\{X_i \in A_n(x)\}} \\ & \quad \le \sum_{i=1}^n I_{\{Y_i=0\}} I_{\{X_i \in A_n(x)\}} \\ 1, & \text{otherwise.} \end{cases}$$

For cubic partitions consistency was established under some additional conditions by Glick (1973). Universal consistency follows from the result of Gordon and Olshen (1978), (1980). The usual conditions needed for the consistency are that the diameter of the cells tends to zero and the number of cells in a bounded region is small compared to the sample size. Strong universal consistency of the partitioning classification rule was proved by Devroye and Györfi (1983). We will show in Section 2 that, although universal rate of convergence doesn't exist, the error probability of the partitioning classification rule converges to the Bayes error faster than $1/\sqrt{n}$ under certain conditions. In Section 3 we investigate two error estimates, the resubstitution and the deleted estimate (or cross validation). Asymptotic normality of the resubstitution estimate was earlier studied in Györfi and Horváth (1998), there it is shown that the difference between the resubstitution estimate and the Bayes error restricted to the partition is asymptotically normal. We prove that the difference between the estimate and the true error probability is asymptotically normal and that the same holds also for the deleted estimate.

We will need in all results the assumption, that the labels in a cell are not uniformly chosen, more precisely we will assume that there is a constant c such that

$$|\mu(A_{nj}) - 2\nu(A_{nj})| > c\mu(A_{nj}) \tag{1}$$

for all n and j, where $\nu(A) = \mathbf{E}\left\{YI_{\{X\in A\}}\right\}$.

The necessity of such a condition is demonstrated in Proposition 1, where we show that if the label is uniform in every cell then the difference between the resubstitution estimate and the true error probability is too big and asymptotic normality doesn't hold.

Partitioning estimates, which were introduced by Tukey (1947) for regression estimation, are present also in density estimation, their behaviour has been studied in several papers, and they are known to be consistent. Asymptotic normality of the L_1- and L_2-errors in partitioning density estimation (histogram) was proved by Beirlant, Györfi and Lugosi (1994) and Berlinet, Devroye and Györfi (1995). Beirlant and Mason (1995) developed a general method for deriving asymptotic normality of the L_p-errors of empirical functionals, which make use of the neighboring data at any point x of interest. Beirlant and Györfi (1998) continued this line of research, and established the asymptotic normality of L_2-error in partitioning regression estimation. To obtain asymptotic normality all these authors used the idea of Poissonization in conjunction with Bartlett's (1938) method of partial inversion for obtaining characteristic functions of conditional distributions.

To lighten notations, throughout the paper we assume that X takes its value in a bounded region $\mathcal{X} \subset \mathcal{R}^d$.

2 Rate of Convergence

It is wellknown that no universal rate of convergence can be guaranteed for any classifiers if nothing is assumed about the distribution. There is no sequence of numbers a_n converging to zero such that there is a classification rule with error probability below $L^* + a_n$ for all distributions (Devroye (1982), Sec. 7.2 in Devroye, Györfi and Lugosi (1996)). In this section we show that the error probability of the partitioning classification rule converges to the Bayes error faster than $1/\sqrt{n}$ under certain conditions.

Theorem 1 *Let R_n^* denote the Bayes error restricted to the partition $\mathcal{P}_n$, i.e.*

$$R_n^* = \sum_{j=1}^{m_n} \min\{\nu(A_{nj}), \mu(A_{nj}) - \nu(A_{nj})\},$$

Then

(i) $L_n \geq R_n^*$.

(ii) Assume that the partitions are finite, let m_n denote the number of cells in the partition. If (1) is satisfied, then

$$\mathbf{E}(L_n - R_n^*) \leq \frac{12}{ce}\frac{m_n}{n}.$$

PROOF

(i): Clearly $L_n \geq R_n^*$, because R_n^* is the Bayes error corresponding to the pair of random variables $(T_n(X), Y)$, where $T_n(x) = j$ if $x \in A_{nj}$, and since the partitioning classification

rule $g_n(x)$ can be written as a function of $T_n(x)$, its error probability L_n cannot be smaller than R_n^*.
(ii): The decision of the partitioning rule in a cell is different from the Bayes decision exactly if either there are more ones in the cell than zeros, but $\nu(A_{nj})$, the probability that a label in the cell is 1, is less than $\mu(A_{nj}) - \nu(A_{nj})$, the probabilty that the label is 0, or vica versa. And in that case the difference between L_n and the Bayes error is the difference between the two probabilities: $|\mu(A_{nj}) - \nu(A_{nj}) - \nu(A_{nj})| = |\mu(A_{nj}) - 2\nu(A_{nj})|$.

$$\mathbf{E}(L_n - R_n^*) = \mathbf{E}\left(\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| I_{\{sign(\mu_n(A_{nj})-2\nu_n(A_{nj}))\neq sign(\mu(A_{nj})-2\nu(A_{nj}))\}}\right)$$

where

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\in A\}}Y_i \quad \text{and} \quad \mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\in A\}}.$$

Thus

$$\begin{aligned}
&\mathbf{E}(L_n - R_n^*)\\
&= \mathbf{E}\left(\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| I_{\{sign(\mu_n(A_{nj})-2\nu_n(A_{nj}))\neq sign(\mu(A_{nj})-2\nu(A_{nj}))\}}\right)\\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \mathbf{P}\left(sign(\mu_n(A_{nj}) - 2\nu_n(A_{nj})) \neq sign(\mu(A_{nj}) - 2\nu(A_{nj}))\right)\\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \Big[\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) > 0\right) I_{\{\mu(A_{nj})-2\nu(A_{nj})\leq 0\}}\\
&\qquad +\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) \leq 0\right) I_{\{\mu(A_{nj})-2\nu(A_{nj})>0\}}\Big]\\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})|\\
&\qquad [\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) > -(\mu(A_{nj}) - 2\nu(A_{nj}))\right)\\
&\qquad\quad I_{\{\mu(A_{nj})-2\nu(A_{nj})\leq 0\}}\\
&\qquad +\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) \leq -(\mu(A_{nj}) - 2\nu(A_{nj}))\right)\\
&\qquad\quad I_{\{\mu(A_{nj})-2\nu(A_{nj})>0\}}]\\
&\leq \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})|\\
&\qquad [\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) > |\mu(A_{nj}) - 2\nu(A_{nj})|\right)\\
&\qquad +\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) \leq -|\mu(A_{nj}) - 2\nu(A_{nj})|\right)]\\
&\leq 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})|\\
&\qquad \mathbf{P}\left(|\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj}))| > |\mu(A_{nj}) - 2\nu(A_{nj})|\right) \qquad (2)
\end{aligned}$$

We can apply Lemma A.1 of the Appendix with $\epsilon = |\mu(A) - 2\nu(A)|$ and $Z_i = I_{\{X_i\in A\}}(1 - 2Y_i) - (\mu(A) - 2\nu(A))$, because they are independent identically distributed zero mean random variables and $|Z_i| \le 2$:

$$\mathbf{P}\left(|\mu_n(A) - 2\nu_n(A) - (\mu(A) - 2\nu(A))| > |\mu(A) - 2\nu(A)|\right)$$
$$\le 2\exp\left(\frac{-n|\mu(A) - 2\nu(A)|^2}{2(\sigma_A^2 + 2|\mu(A) - 2\nu(A)|)}\right),$$

where $\sigma_A^2 = \mathbf{E}(Z_i^2) = \mu(A) - (\mu(A) - 2\nu(A))^2$.

From this and (2)

$$\begin{aligned}
\mathbf{E}(L_n - R_n^*) &\le 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{\frac{-n|\mu(A_{nj})-2\nu(A_{nj})|^2}{2(\mu(A_{nj})-(\mu(A_{nj})-2\nu(A_{nj}))^2+2|\mu(A_{nj})-2\nu(A_{nj})|)}} \\
&\le 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{\frac{-n|\mu(A_{nj})-2\nu(A_{nj})|^2}{2(3\mu(A_{nj}))}} \\
&\le 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{-\frac{c}{6}n|\mu(A_{nj})-2\nu(A_{nj})|} \\
&\quad \text{(because of condition (1))} \\
&= 2\frac{6}{cn}\sum_{j=1}^{m_n} \frac{c}{6} n|\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{-\frac{c}{6}n|\mu(A_{nj})-2\nu(A_{nj})|} \\
&\le 4\frac{6}{cn} m_n \max_z z e^{-z} \\
&= \frac{12}{ce}\frac{m_n}{n}
\end{aligned}$$

□

Corollary 1 *If in addition to condition (1)*

$$\lim_{n\to\infty} \frac{m_n^2}{n} = 0, \tag{3}$$

then

$$\sqrt{n}\,(L_n - R_n^*) \to 0$$

in probability as $n \to \infty$.

PROOF

It is well-known that if, for a nonnegative random variable Z_n, $\mathbf{E}Z_n \to 0$, then $Z_n \to 0$ in probability. Thus we have to prove that

$$\sqrt{n}\mathbf{E}(L_n - R_n^*) \to 0$$

as $n \to \infty$. From Theorem 1

$$\sqrt{n}\mathbf{E}\left(L_n - R_n^*\right) \le \frac{12e^{-1}}{c}\frac{m_n}{\sqrt{n}} \to 0$$

because of condition (3). □

3 Error Estimation

To derive two important corollaries of Theorem 1, introduce two error estimates: the resubstitution error estimate and the deleted estimate (or cross-validation).

Estimating the error probability of a classification rule g_n is of great importance. The designer always wants to know what performance can be expected from a classifier. Since the distribution of the data is unknown, it is important to find and analyze error estimation methods that work well independently from the distribution of (X, Y).

One natural way of error estimation is to split the data into two parts, a training sample and a testing sample, and use the first for designing a classifier and the latter for the estimation. But when data are expensive, this constitutes a waste. There are well-known methods that use the same sample for training and testing: the resubstitution estimate, the deleted estimate etc.

3.1 Resubstitution Estimate

The resubstitution estimate $\hat{L}_n$ counts the number of errors committed on the training sequence $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ by the classification rule, i.e. for a classifier g_n it is defined as

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{g_n(X_i) \neq Y_i\}},$$

which for the partitioning classification rule can be written in the following form:

$$\hat{L}_n = \sum_{j=1}^{m_n} \min\{\nu_n(A_{nj}), \mu_n(A_{nj}) - \nu_n(A_{nj})\}, \tag{4}$$

where

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in A\}} Y_i \quad \text{and} \quad \mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in A\}}.$$

$\hat{L}_n$ is an estimate of the Bayes error R_n^* restricted to the partition $\mathcal{P}_n$.

Concerning the resubstitution error estimate for partitioning rule, the following inequalities are known (see Sec. 23.2 in Devroye, Györfi and Lugosi (1996)): for an arbitrary partition $\mathcal{P}_n$,

$$Var(\hat{L}_n) \leq \frac{1}{n}$$

and

$$\mathbf{E}\hat{L}_n \leq R_n^*.$$

For a finite partition of size m_n

$$R_n^* - \mathbf{E}\hat{L}_n \leq \sqrt{\frac{2m_n}{n}}. \tag{5}$$

Combining this with the result of Theorem 1, we get

$$\mathbf{E}(L_n - \hat{L}_n) \leq \sqrt{\frac{2m_n}{n}} + \frac{12}{ce} \frac{m_n}{n}.$$

The resubstitution estimate for the partitioning rule is asymptotically normal under certain conditions:

Theorem 2 *Consider the partitions where A_{nj} are d-dimensional rectangles. Let $a_n^i(x)$, $i = 1, 2, ..., d$, denote the sidelengths of $A_n(x)$. Assume that for all x there exists a $K(x)$ so that $\frac{a_n^i(x)}{a_n^j(x)} \leq K(x)$ for all $1 \leq i, j \leq d$ and n. Assume that*

$$\lim_{n\to\infty} \sup_{A_{nj}} diam(A_{nj}) = 0$$

and

$$\lim_{n\to\infty} \frac{\log n}{n\lambda(A_n(x))} = 0,$$

where λ is the Lebesgue measure. If μ has a density and conditions (1) and (3) are satisfied, then

$$n^{1/2}\left(L_n - \hat{L}_n\right) / \sqrt{L^*/2} \xrightarrow{\mathcal{D}} N(0,1).$$

Thus the resubstitution estimate is a good estimate of the true error probability of the partitioning classification rule, and using the result of the theorem a good estimate can be given to the probability that the difference between the resubstitution estimate and the error probability is greater than $\epsilon/\sqrt{n}$. The result is a consequence of Corollary 1 and Lemma A.2 of the Appendix.

Remark.
For cubic partitions with size h_n these conditions mean $h_n \to 0$ and $nh_n^{2d} \to \infty$ as $n \to \infty$. Although the second condition is stronger than the condition $nh_n^d \to \infty$ for the consistency of the partitioning rule, it does allow the usual good choices for h_n, i.e. choices that minimize the rate of convergence, e.g. $h_n = n^{-1/3}$. The condition might be weakened with a longer and more complicated proof.

Under the conditions of Theorem 2, we can remove m_n from the upper bound in (5).

Theorem 3 *For any distribution of (X, Y) and for n large enough for the estimate $\hat{L}_n$ of the error probability of a partitioning rule satisfying the conditions of Theorem 2*

$$R_n^* - \mathbf{E}\hat{L}_n \leq \frac{1.18}{\sqrt{n}}.$$

PROOF
From the result on the asymptotic normality

$$\mathbf{P}\{n^{1/2}|\hat{L}_n - R_n^*| > \epsilon\} \to 1 - 2\Phi\left(\frac{\epsilon}{\sqrt{L^*/2}}\right),$$

where Φ is the standard normal distribution function. Since $R_n^* - \mathbf{E}\hat{L}_n$ is not random

$$\begin{aligned} I_{\{\sqrt{n}|R_n^*-\mathbf{E}\hat{L}_n|>\epsilon\}} &= \mathbf{P}\{\sqrt{n}|R_n^* - \mathbf{E}\hat{L}_n| > \epsilon\} \\ &\le \mathbf{P}\{\sqrt{n}(|\hat{L}_n - \mathbf{E}\hat{L}_n| + |\hat{L}_n - R_n^*|) > \epsilon\} \\ &\le \mathbf{P}\{\sqrt{n}|\hat{L}_n - \mathbf{E}\hat{L}_n| > \epsilon/2\} \\ &\quad +\mathbf{P}\{\sqrt{n}|\hat{L}_n - R_n^*| > \epsilon/2\}. \end{aligned}$$

For the first term we can apply McDiarmid's inequality (Lemma A.3) with $Z_i = (X_i, Y_i)$ and $A = \mathcal{R}^d \times \{0,1\}$. It can be easily seen that in the case of the resubstitution estimate for partitioning rule $c_i = \frac{1}{n}$. Thus for all n and $\epsilon > 0$,

$$\mathbf{P}\{n^{1/2}|\hat{L}_n - \mathbf{E}\hat{L}_n| > \epsilon\} \le 2e^{-2\epsilon^2}.$$

Therefore

$$\limsup_{n\to\infty} I_{\{\sqrt{n}|R_n^*-\mathbf{E}\hat{L}_n|>\epsilon\}} \le 2e^{-\epsilon^2/2} + 1 - 2\Phi\left(\frac{\epsilon}{\sqrt{2L^*}}\right).$$

Obviously there exists such ϵ_0 for which the righthand side of the inequality is strictly smaller than 1, and then

$$I_{\{\sqrt{n}|R_n^*-\mathbf{E}\hat{L}_n|>\epsilon_0\}} = 0$$

for sufficiently large n. Using the trivial upper bound

$$\Phi(\epsilon/\sqrt{2L^*}) \ge 1/2,$$

$\epsilon_0 = \sqrt{2\ln 2} \approx 1.18$ is a valid choice. □

It is necessary to have some condition such as (1) that ensures that labels are far from being uniformly chosen.

Proposition 1 *Assume that $\mu(A_{nj}) = 2\nu(A_{nj})$ for all cells A_{nj} of the partition, then*

$$\sqrt{n}\mathbf{E}(L_n - \hat{L}_n) \ge \frac{1}{4}\sqrt{m_n}.$$

Thus in that case $\sqrt{n}\mathbf{E}(L_n - \hat{L}_n) \to \infty$ as $n \to \infty$.

Proof Since $L_n \ge R_n^*$, and under the conditions of the proposition $R_n^* = 1/2$, we have to show, that

$$\mathbf{E}\hat{L}_n \le \frac{1}{2} - \frac{1}{4}\sqrt{\frac{m_n}{n}}.$$

$$\begin{aligned} \mathbf{E}\hat{L}_n &= \mathbf{E}\sum_{j=1}^{m_n} \min\{\nu_n(A_{nj}), \mu_n(A_{nj}) - \nu_n(A_{nj})\} \\ &= \mathbf{E}\sum_{j=1}^{m_n} \left(\frac{\mu_n(A_{nj})}{2} + \min\{\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}, \frac{\mu_n(A_{nj})}{2} - \nu_n(A_{nj})\}\right) \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\sum_{j=1}^{m_n}\left(\frac{\mu_n(A_{nj})}{2} - \left|\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}\right|\right) \\
&= \frac{1}{2} - \mathbf{E}\sum_{j=1}^{m_n}\left|\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}\right| \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}\mathbf{E}\left|n\nu_n(A_{nj}) - \frac{n\mu_n(A_{nj})}{2}\right| \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}\sum_{k=0}^{n}\mathbf{E}\left(\left|n\nu_n(A_{nj}) - \frac{n\mu_n(A_{nj})}{2}\right| | n\mu_n(A_{nj}) = k\right)\mathbf{P}(n\mu_n(A_{nj}) = k) \quad (6)
\end{aligned}
$$

If $n\mu_n(A_{nj}) = k$, then $n\nu_n(A_{nj})$ is binomial with parameters k and $1/2$, therefore we are interested in the expected value $\mathbf{E}\left|Binom(k, \frac{1}{2}) - \frac{k}{2}\right|$. Let Z_i, $i = 1, \dots, k$ be random variables with distribution $\mathbf{P}(Z_i = 1) = \mathbf{P}(Z_i = -1) = 1/2$. Then $\left|Binom(k, \frac{1}{2}) - \frac{k}{2}\right| = \frac{1}{2}\left|\sum_{i=1}^{k} Z_i\right|$, thus applying Lemma A.4 to these Z_i and $p = 1$, we get

$$
\frac{1}{2\sqrt{2}}\sqrt{k} \le \mathbf{E}\left|Binom\left(k, \frac{1}{2}\right) - \frac{k}{2}\right| = \frac{1}{2}\mathbf{E}\left|\sum_{i=1}^{k} Z_i\right| \le \frac{1}{2}\sqrt{k}.
$$

Thus from (6)

$$
\begin{aligned}
\mathbf{E}\hat{L}_n &\le \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}\sum_{k=0}^{n}\frac{1}{2\sqrt{2}}\sqrt{k}\binom{n}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-k} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}n\mu(A_{nj})\sum_{k=1}^{n}\frac{1}{2\sqrt{2}}\frac{1}{\sqrt{k}}\binom{n-1}{k-1}\mu(A_{nj})^{k-1}(1-\mu(A_{nj}))^{(n-1)-(k-1)} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\frac{1}{n}n\mu(A_{nj})\sum_{k=0}^{n-1}\frac{1}{\sqrt{k+1}}\binom{n-1}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-1-k} \\
&\le \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\mu(A_{nj})\frac{1}{\sqrt{\sum_{k=0}^{n-1}(k+1)\binom{n-1}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-1-k}}} \\
&\quad \text{(from Jensen's inequality)} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\mu(A_{nj})\frac{1}{\sqrt{(n-1)\mu(A_{nj})+1}} \\
&\le \frac{1}{2} - m_n\frac{1}{2\sqrt{2}}\frac{\frac{1}{m_n}\sum_{j=1}^{m_n}\mu(A_{nj})}{\sqrt{(n-1)\frac{1}{m_n}\sum_{j=1}^{m_n}\mu(A_{nj})+1}} \\
&\quad \text{(from Jensen's inequality)} \\
&= \frac{1}{2} - \frac{1}{2\sqrt{2}}\frac{1}{\sqrt{(n-1)/m_n+1}} \\
&\le \frac{1}{2} - \frac{1}{4}\sqrt{\frac{m_n}{n}}
\end{aligned}
$$

$\square$

3.2 Deleted Estimate

The deleted estimate or cross-validation attempts to avoid the bias present in the resubstitution estimate. The method deletes (X_i, Y_i) from the training data and creates a classifier g_{n-1} using the remaining $n-1$ pairs. It tests for an error on (X_i, Y_i), and repeats this procedure for all n pairs of the training data D_n. Formally denote the training set with (X_i, Y_i) deleted by

$$D_{n,i} = ((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)).$$

Then define

$$\bar{L}_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{g_{n-1}(X_i, D_{n,i}) \neq Y_i\}}.$$

Clearly, the deleted estimate is almost unbiased in the sense that

$$\mathbf{E}\bar{L}_n = \mathbf{E}L_{n-1}.$$

Let g_{n-1} be the partitoning classification rule for $n-1$ samples, using partition $\mathcal{P}_n$. Concerning the deleted error estimate for partitioning rule the following is known (see Sec. 24.5 in Devroye, Györfi and Lugosi (1996)):

$$\mathbf{E}\left\{\left(\bar{L}_n - L_n\right)^2\right\} \leq \frac{1 + 6/e}{n} + \frac{6}{\sqrt{\pi(n-1)}}.$$

The resubstitution estimate for any partitioning classification rule is smaller than the deleted estimate:

$$\bar{L}_n \geq \hat{L}_n, \tag{7}$$

since if (X_i, Y_i) is a mistake w.r.t. $\hat{L}_n$, i.e. $Y_i \neq g_n(X_i)$, then the label Y_i is in the minority among the labels of the data falling into its cell. Then, of course, Y_i is in the minority of its cell w.r.t. $D_{n,i} = D_n \setminus \{X_i, Y_i\}$, which implies that (X_i, Y_i) is a mistake w.r.t. $\bar{L}_n$.

The deleted estimate for partitioning classification rule is asymptotically normal under similar conditions as the resubstitution estimate:

Theorem 4 *Under the conditions of Theorem 2,*

$$n^{1/2}\left(L_n - \bar{L}_n\right) / \sqrt{L^*/2} \xrightarrow{\mathcal{D}} N(0,1).$$

PROOF

It is easy to see that if U_n and V_n are random variables and $U_n \xrightarrow{\mathcal{D}} N(0,1)$ and $V_n \to 0$ in probability, then $U_n + V_n \xrightarrow{\mathcal{D}} N(0,1)$.

So, since the conditions are the same as for Theorem 2, it suffices to show that

$$\sqrt{n}\left(\bar{L}_n - \hat{L}_n\right) \to 0$$

in probability as $n \to \infty$.

If for a nonnegative random variable Z_n, $\mathbf{E}Z_n \to 0$, then $Z_n \to 0$ in probability. Thus, since from (7) $\bar{L}_n - \hat{L}_n \geq 0$, we have to prove that

$$\sqrt{n}\mathbf{E}(\bar{L}_n - \hat{L}_n) \to 0$$

as $n \to \infty$.

(X_i, Y_i) is a mistake w.r.t. $\bar{L}_n$, i.e. $g_{n-1}(X_i, D_{n,i}) \neq Y_i$ if Y_i is in the minority of its cell w.r.t. $D_{n,i} = D_n \setminus \{X_i, Y_i\}$. If (X_i, Y_i) is a mistake, then each pair with the same label in that cell is also a mistake. The number of ones in a cell is $\nu_n(A_{nj})$ and the number of zeros is $\mu_n(A_{nj}) - \nu_n(A_{nj})$. Thus, for the partitioning classification rule the deleted estimate can be written in the following form:

$$\begin{aligned}\bar{L}_n &= \sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})-1\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\\ &\qquad +(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1<n\nu_n(A_{nj})\}}\Big).\end{aligned}$$

The resubstitution estimate can be written in a similar form using (4)

$$\begin{aligned}\hat{L}_n &= \sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\\ &\qquad +(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))<n\nu_n(A_{nj})\}}\Big).\end{aligned}$$

Therefore

$$\begin{aligned}&\sqrt{n}\mathbf{E}(\bar{L}_n - \hat{L}_n)\\
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})-1\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\right.\\
&\quad +(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1<n\nu_n(A_{nj})\}}\\
&\quad -\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\\
&\quad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))<n\nu_n(A_{nj})\}}\Big)\Big)\right\}\\
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})\left(I_{\{n\nu_n(A_{nj})-1\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}} - I_{\{n\nu_n(A_{nj})\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\right)\right.\\
&\quad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))\left(I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1<n\nu_n(A_{nj})\}} - I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))<n\nu_n(A_{nj})\}}\right)\Big)\right\}\\
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})-1\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))<n\nu_n(A_{nj})\}}\right.\\
&\qquad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1<n\nu_n(A_{nj})\leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\Big)\right\}\\
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})=n(\mu_n(A_{nj})-\nu_n(A_{nj}))+1\}}\right.\\
&\qquad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))=n\nu_n(A_{nj})\}}\Big)\right\}\\
&= \sqrt{n}\sum_{j=1}^{m_n}\Big(\mathbf{E}\left\{\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})=n(\mu_n(A_{nj})-\nu_n(A_{nj}))+1\}}\right\}\end{aligned}$$

$$
\begin{aligned}
&\qquad +\mathbf{E}\left\{(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))=n\nu_n(A_{nj})\}}\right\}\Big) \\
&= \sqrt{n}\sum_{j=1}^{m_n}(T_{1,nj}+T_{2,nj}) \qquad (8)
\end{aligned}
$$

Consider the first term of the sum for the cell $A = A_{nj}$:

$$
\begin{aligned}
T_{1,n} &= \mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}\right\} \\
&= \sum_{k=0}^{n}\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=k\right\}\mathbf{P}\{n\mu_n(A)=k\} \\
&= \sum_{l=0}^{n/2}\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=2l+1\right\}\mathbf{P}\{n\mu_n(A)=2l+1\}
\end{aligned}
$$

Since if the number of pairs in a cell is even, then it cannot contain one more zeros than ones, so the indicator is 0. $n\mu_n(A)$ is binomial with parameters n and $\mu(A)$, so

$$
\mathbf{P}\{n\mu_n(A)=2l+1\} = \binom{n}{2l+1}\mu(A)^{2l+1}(1-\mu(A))^{n-(2l+1)}.
$$

If $\mu_n(A) = 2l+1$, then $n\nu_n(A)$ is binomial with parameters $2l+1$ and $\nu(A)/\mu(A)$, thus

$$
\begin{aligned}
&\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=2l+1\right\} \\
&\quad= \mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=l+1\}}\right\} \\
&\quad= \frac{l+1}{n}\mathbf{P}\{n\nu_n(A)=l+1\} \\
&\quad= \frac{l+1}{n}\binom{2l+1}{l+1}\left(\frac{\nu(A)}{\mu(A)}\right)^{l+1}\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^{l} \\
&\quad= \frac{2l+1}{n}\binom{2l}{l}\left(\frac{\nu(A)}{\mu(A)}\right)^{l+1}\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^{l}.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
T_{1,n} &= \sum_{l=0}^{n/2}\frac{2l+1}{n}\binom{2l}{l}\nu(A)^{l+1}(\mu(A)-\nu(A))^{l}\frac{1}{\mu(A)^{2l+1}} \\
&\qquad \binom{n}{2l+1}\mu(A)^{2l+1}(1-\mu(A))^{n-(2l+1)} \\
&= \sum_{l=0}^{n/2}\frac{2l+1}{n}\binom{2l}{l}\nu(A)^{l+1}(\mu(A)-\nu(A))^{l}\binom{n}{2l+1}(1-\mu(A))^{n-(2l+1)} \\
&\le \sum_{l=0}^{n/2}2^{2l}\nu(A)^{l+1}(\mu(A)-\nu(A))^{l}\binom{n}{2l}(1-\mu(A))^{n-(2l+1)} \\
&\qquad \left(\text{since } \tfrac{2l+1}{n}\tbinom{2l}{l}\tbinom{n}{2l+1}=\tbinom{2l}{l}\tbinom{n-1}{2l}\le 2^{2l}\tbinom{n}{2l}\right) \\
&= \frac{\nu(A)}{1-\mu(A)}\sum_{l=0}^{n/2}2^{2l}\nu(A)^{l}(\mu(A)-\nu(A))^{l}\binom{n}{2l}(1-\mu(A))^{n-2l} \\
&= \frac{\nu(A)}{1-\mu(A)}\sum_{l=0}^{n/2}\binom{n}{2l}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{2l}(1-\mu(A))^{n-2l}
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{\nu(A)}{1-\mu(A)}\sum_{k=0}^{n}\binom{n}{k}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^k(1-\mu(A))^{n-k}\\
&= \frac{\nu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n\\
&\le \frac{\mu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n. \qquad (9)
\end{aligned}
$$

Similarly, we can get the same upper bound for the second term in (8):

$$
\begin{aligned}
T_{2,n} &= \mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}\right\}\\
&= \sum_{k=0}^{n}\mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}|n\mu_n(A)=k\right\}\mathbf{P}\{n\mu_n(A)=k\}\\
&= \sum_{l=0}^{n/2}\mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}|n\mu_n(A)=2l\right\}\mathbf{P}\{n\mu_n(A)=2l\}\\
&= \sum_{l=0}^{n/2}\frac{l}{n}\binom{2l}{l}\left(\frac{\nu(A)}{\mu(A)}\right)^l\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^l\binom{n}{2l}\mu(A)^{2l}(1-\mu(A))^{n-2l}\\
&= \sum_{l=0}^{n/2}\frac{l}{n}\binom{2l}{l}\nu(A)^l(\mu(A)-\nu(A))^l\binom{n}{2l}(1-\mu(A))^{n-2l}\\
&\le \frac{1}{n}\sum_{l=0}^{n/2}2l2^{2l}\nu(A)^l(\mu(A)-\nu(A))^l\binom{n}{2l}(1-\mu(A))^{n-2l}\\
&= \frac{1}{n}\sum_{l=0}^{n/2}2l\binom{n}{2l}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{2l}(1-\mu(A))^{n-2l}\\
&\le \frac{1}{n}\sum_{k=0}^{n}k\binom{n}{k}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{k}(1-\mu(A))^{n-k}.
\end{aligned}
$$

The sum in the righthand side is the expected value of the binomial distribution with parameters n and $2\sqrt{\nu(A)(\mu(A)-\nu(A))}/(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A))$ if we normalize with the sum of $2\sqrt{\nu(A)(\mu(A)-\nu(A))}$ and $1-\mu(A)$. Thus

$$
\begin{aligned}
T_{2,n} &\le \frac{1}{n}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n\frac{1}{\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n}\\
&\qquad \sum_{k=0}^{n}k\binom{n}{k}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{k}(1-\mu(A))^{n-k}\\
&= \frac{1}{n}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n\\
&\qquad \mathbf{E}\left\{Binom\left(n,\frac{2\sqrt{\nu(A)(\mu(A)-\nu(A))}}{2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)}\right)\right\}\\
&= \frac{1}{n}\frac{n2\sqrt{\nu(A)(\mu(A)-\nu(A))}}{2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n\\
&\le \frac{\mu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n.
\end{aligned}
$$

From this upper bound and from (8) and (9)

$$
\begin{aligned}
&\sqrt{n}\mathbf{E}(\bar{L}_n - \hat{L}_n) \\
&\leq \sqrt{n}\sum_{j=1}^{m_n} 2\frac{\mu(A_{nj})}{1-\mu(A_{nj})}\left(2\sqrt{\nu(A_{nj})(\mu(A_{nj})-\nu(A_{nj}))}+1-\mu(A_{nj})\right)^n \\
&\leq \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-n\left(\mu(A_{nj})-2\sqrt{\nu(A_{nj})(\mu(A_{nj})-\nu(A_{nj}))}\right)} \\
&= \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-n\mu(A_{nj})\left(1-2\sqrt{\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1-\frac{\nu(A_{nj})}{\mu(A_{nj}}\right)}\right)},
\end{aligned}
$$

where the second inequality comes from the fact that $1 - z \leq e^{-z}$ if $z > 0$, and clearly $\mu(A_{nj}) - 2\sqrt{\nu(A_{nj})(\mu(A_{nj}) - \nu(A_{nj}))} > 0$, and we can assume that $\mu(A_{nj}) < 1/2$.
The condition (1) means that

$$\left|\frac{1}{2} - \frac{\nu(A_{nj})}{\mu(A_{nj})}\right| > \frac{c}{2},$$

and because of this

$$\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1 - \frac{\nu(A_{nj})}{\mu(A_{nj})}\right) \leq \frac{1}{4} - \frac{c^2}{4}.$$

Thus, if we denote $1 - 2\sqrt{\frac{1}{4} - \frac{c^2}{4}} = 1 - \sqrt{1 - c^2}$ by $\delta(c)$ then

$$
\begin{aligned}
\sqrt{n}\mathbf{E}(\bar{L}_n - \hat{L}_n) &\leq \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-n\mu(A_{nj})\left(1-2\sqrt{\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1-\frac{\nu(A_{nj})}{\mu(A_{nj}}\right)}\right)} \\
&\leq \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-\delta(c)n\mu(A_{nj})} \\
&= \sqrt{n}\sum_{j=1}^{m_n} \frac{2}{\delta(c)n}\delta(c)n\mu(A_{nj})e^{-\delta(c)n\mu(A_{nj})} \\
&\leq \sqrt{n}\sum_{j=1}^{m_n} \frac{2}{\delta(c)n}\max_z ze^{-z} \\
&\leq \sqrt{n}m_n\frac{2}{\delta(c)n} = \frac{2m_n}{\delta(c)\sqrt{n}} \to 0
\end{aligned}
$$

because of condition (3). □

Thus, similarly to the resubstitution estimate, the deleted estimate is a good estimate of the true error probability of the partitioning classification rule, and using the result of the theorem a good estimate can be given to the probability that the difference between the deleted estimate and the error probability is greater than $\epsilon/\sqrt{n}$.

A Appendix

Lemma A.1 *(Bernstein (1946)) For independent identically distributed zero mean random variables Z_i with $|Z_i| \le t$, and for all $\epsilon > 0$*

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \epsilon\right) \le 2\exp\left(\frac{-n\epsilon^2}{2(\sigma^2 + t\epsilon)}\right),$$

where $\sigma^2 = \mathbf{E}(Z_i^2)$.

Lemma A.2 *(Györfi and Horváth (1998)) Under the conditions of Theorem 2,*

$$n^{1/2}\left(\hat{L}_n - R_n^*\right)/\sqrt{L^*/2} \xrightarrow{\mathcal{D}} N(0,1).$$

Lemma A.3 *(McDiarmid (1989)) Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set A, and assume that the measurable function $F : A^n \to \mathcal{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,z_i'} |F(z_1,\ldots,z_i,\ldots,z_n) - F(z_1,\ldots,z_i',\ldots,z_n)| \le c_i,\ 1 \le i \le n.$$

Then for all $\epsilon > 0$

$$\mathbf{P}\{|F(Z_1,\ldots,Z_n) - \mathbf{E}F(Z_1,\ldots,Z_n)| > \epsilon\} \le 2e^{-2\epsilon^2/(\sum_{i=1}^n c_i^2)}.$$

Lemma A.4 *(Khinchine's inequality, see e.g. Devroye and Györfi (1985)) Let $Z_1, \ldots, Z_k$ be independent random variables taking the values $+1$ and -1 with equal probability, and let $a_1, \ldots, a_k$ be real numbers. Then there exist positive constants B_p and C_p depending upon $p > 0$ only such that*

$$B_p\sqrt{\sum_{i=1}^{k} a_i^2} \le \mathbf{E}^{1/p}\left(\left|\sum_{i=1}^{k} a_i Z_i\right|^p\right) \le C_p\sqrt{\sum_{i=1}^{k} a_i^2}.$$

The following values are optimal:

$$B_p = \begin{cases} 2^{1/2-1/p}, & \text{if } 0 < p \le p_0 \\ 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, & p_0 \le p < 2 \\ 1, & \text{if } 2 \le p \end{cases}$$

$$C_p = \begin{cases} 1, & \text{if } 0 < p \le 2 \\ 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, & 2 \le p \end{cases}$$

where $\Gamma(\cdot)$ is the classical gamma-function extending the factorial and $p_0 = 1.84742\cdots$ is the solution in $(1,2)$ of $\Gamma((p+1)/2) = \Gamma(3/2)$.

References

[1] Bartlett, M. S. (1938). The characteristic function of a conditional statistic. *J. London Math. Soc.*, 13, pp. 62–67.

[2] Beirlant, J. and Györfi, L. (1998). On the asymptotic normality of L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, 71 (1998), pp. 93–107.

[3] Beirlant, J., Györfi, L. and Lugosi, G.(1994). On the asymptotic normality of L_1- and L_2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22 (1994), no. 3, pp. 309–318.

[4] Beirlant, J. and Mason, D. M. (1995). On the asymptotic normality of L_p-norms of empirical functionals. *Math. Methods Statist.*, 4 (1995), no. 1, pp. 1–19.

[5] Berlinet, A., Devroye, L. and Györfi, L. (1995). Asymptotic normality of L_1-error in histogram density estimation. *Statistics*, 26, pp. 329–343.

[6] Bernstein, S. (1946). *The Theory of Probabilities.* Gostehizdat Publishing House, Moscow.

[7] Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of the nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61, pp. 467–481.

[8] Devroye, L. and Györfi, L. (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, F. Konecny, J. Mogyoródi, W. Wertz (Eds.), pp. 67–76, Akadémiai Kiadó, Budapest, Hungary.

[9] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

[10] Devroye, L., Györfi, L. and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition.* Springer Verlag, New York.

[11] Glick, N. (1973). Sample based multinomial classification. *Biometrics*, 29, pp. 241–256.

[12] Gordon, L. and Olshen, R. (1978). Asymptotically efficient solutions to the calssification problem. *Annals of Statistics*, 6, pp. 515–533.

[13] Gordon, L. and Olshen, R. (1980). Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10, pp. 611–627.

[14] Györfi, L. and Horváth, M. (1998). On the asymptotic normality of the resubstitution error estimate for partitioning classification rule. In *Advances in Data Science and Classification*, A. Rizza, M. Vichi, H. H. Bock (Eds.), Springer, pp. 197–204.

[15] McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989*, pp. 148–188. Cambridge University Press, Cambridge.

[16] Tukey, J. W. (1947). Nonparametric estimation. II. Statistically Equivalent Blocks and Tolerance Regions. *Annals of Mathematical Statistics*, 18, pp. 529–539.