On the Asymptotic Normality of a Resubstitution Error Estimate [1]

László Györfi
Dept. of Computer Science and Information Theory
Technical University of Budapest
1521 Stoczek u. 2, Budapest, Hungary. e-mail:gyorfi@inf.bme.hu

Márta Horváth
Dept. of Computer Science and Information Theory
Technical University of Budapest
1521 Stoczek u. 2, Budapest, Hungary. e-mail:marti@inf.bme.hu

Summary: The resubstitution error estimate for the partitioning classification rule from a sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is shown to be asymptotically normal under the condition that X has a density f, if the partition consists of rectangles.

Key Words: error estimation, central limit theorem, partitioning classification rule

1. Introduction

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

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

and the Bayes error probability by

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

For the pattern recognition problem let $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. random variables such that (X_i, Y_i) have the same distribution as (X, Y). Let $\mathcal{P}_n = \{A_{nj}, j = 1, 2, \ldots\}$ be a partition of $\mathcal{R}^d$, and let $A_n(x)$ denote the the subset in the partition that includes x. The resubstitution estimate $\hat{L}_n$ counts the number of errors commited on the training sequence $(X_1, Y_1), (X_2, Y_2), \ldots,\ (X_n, Y_n)$ by the classifier, 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

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

[1] The research was supported by the Computer and Automation Institute of the Hungarian Academy of Sciences (MTA SZTAKI).

can be written in the following form:

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

where

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}} Y_i$$

and

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}}.$$

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

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

where

$$\nu(A) = \mathbf{E}\nu_n(A).$$

2. Asympotic Normality

Concerning the resubstitution error estimate for partitioning rule the following inequalities are known (see Sec. 23.2 in Devroye, Györfi and Lugosi (1996)): for arbitrary partition

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

and

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

For finite partition of size m_n

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

Our aim is to derive the asymptotic normality of the distribution of $\hat{L}_n$.

Theorem 1 *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 for each sphere S centered at the origin

$$\lim_{n\to\infty} \sup_{A_{nj}\cap S \neq \emptyset} diam(A_{nj}) = 0 \tag{2}$$

and

$$\lim_{n\to\infty} \frac{|\{A_{nj} \cap S \neq \emptyset\}|^2}{n} = 0 \tag{3}$$

and

$$\lim_{n\to\infty} \frac{\log n}{n\lambda(A_n(x))} = 0, \tag{4}$$

where λ is the Lebesgue measure. If μ has a density f and

$$\mathbf{P}\{D(X) = 0\} = 0,$$

where $D(x) = P_0(x) - P_1(x)$, then there is a sequence C_n such that

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

Theorem 2 *If, in addition to the conditions of Theorem 1, for*

$$\tau(A) = \mu(A) - 2\nu(A) = \int_A D(x)\mu(dx)$$

there is a constant c such that

$$|\tau(A_{nj})| > c\mu(A_{nj}) \tag{5}$$

for all n and j then

$$n^{1/2}\left(\hat{L}_n - R_n^*\right)/\sqrt{L^*/2} \xrightarrow{\mathcal{D}} N(0,1). \tag{6}$$

Remark.
For cubic partitions with size h_n these conditions mean $h_n \to 0$, $nh_n^{2d} \to \infty$ and $nh_n^d/\log n \to \infty$ as $n \to \infty$.

3. Sketch of the proof of Theorem 1

The main difficulty in proving the theorem is that it states asymptotic normality of infinite sums of dependent random variables. Thus, we need a technique to extend asymptotic normality of the partial sums to the infinite sum. Assume that the $\{A_{nj}\}$ are ordered according to non-decreasing distances of their centers from the origin. For $\alpha \in (0,1)$ choose the integer m_n such that

$$1 - \alpha_n = \sum_{j=1}^{m_n} \mu(A_{nj}) \leq 1 - \alpha < \sum_{j=1}^{m_n+1} \mu(A_{nj}).$$

Roughly speaking, $S_{\alpha_n} = \bigcup_{j=1}^{m_n} A_{nj}$ is approximately a sphere centered at the origin with $1 - \alpha \sim \mu(S_{\alpha_n})$. Because of (2) and the absolute continuity of μ

$$0 \leq \alpha_n - \alpha \leq \mu(A_{n,m_n+1}) \leq \max_j \mu(A_{nj}) \to 0$$

as $n \to \infty$. Also define S_α as the sphere centered at the origin with the property $1-\alpha = \mu(S_\alpha)$.

The basics of our proof is the Poissonization. Introduce the notation N_n for a Poisson (n) random variable independent of $(X_1, Y_1), (X_2, Y_2), \ldots$, moreover put

$$n\tilde{\nu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}} Y_i$$

and

$$n\tilde{\mu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}}.$$

Proposition 1 *(Beirlant, Mason (1995)) Let g_{nj} be real measurable functions and let $M_{n,\alpha} = \sum_{j=1}^{m_n} g_{nj}\left(\tilde{\nu}_n(A_{nj}), \tilde{\mu}_n(A_{nj})\right)$. Assume that*

$$\Phi_{n,\alpha}(t,v) = \mathbf{E}\left(\exp\left(itM_{n,\alpha} + iv\frac{N_n - n}{\sqrt{n}}\right)\right) \to e^{-(t^2\rho_\alpha^2 + 2tv\gamma_\alpha + v^2)/2},$$

with $\rho_\alpha^2 = \int_{S_\alpha} h(x)dx$ and $\gamma_\alpha = \int_{S_\alpha} g(x)dx$, where $h(x)$ and $g(x)$ are some measurable functions such that $\gamma_0 = \int_{\mathcal{R}^d} g(x)dx < \rho_0^2 = \int_{\mathcal{R}^d} h(x)dx < \infty$. Then

$$\sum_{j=1}^{\infty} g_{nj}\left(\nu_n(A_{nj}), \mu_n(A_{nj})\right) \Big/ \sqrt{\rho_0^2 - \gamma_0} \xrightarrow{\mathcal{D}} N(0,1).$$

Proof of Theorem 1. Because of $\min\{a,b\} = \frac{1}{2}(a+b-|a-b|)$

$$\hat{L}_n = \frac{1}{2}(1 - J_n), \tag{7}$$

where $J_n = \sum_{j=1}^{\infty} |\tau_n(A_{nj})|$ with $\tau_n(A) = \mu_n(A) - 2\nu_n(A)$, therefore the first statement of Theorem 1 is equivalent to

$$n^{1/2}\left(J_n - (1 - 2C_n)\right)/\sqrt{2L^*} \xrightarrow{\mathcal{D}} N(0,1). \tag{8}$$

In order to prove (8) let

$$1 - 2C_n = \sum_{j=1}^{\infty} \mathbf{E}|\tilde{\tau}_n(A_{nj})|,$$

where $\tilde{\tau}_n(A) = \tilde{\mu}_n(A) - 2\tilde{\nu}_n(A)$. So we need to prove that

$$n^{1/2}\left(\sum_{j=1}^{\infty}\left(|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)\right)/\sqrt{2L^*} \xrightarrow{\mathcal{D}} N(0,1),$$

for which we will use the Proposition. We choose the functions g_{nj} as

$$g_{nj}(u,v) = \sqrt{n}\left(|v - 2u| - \mathbf{E}\left|\tilde{\mu}_n(A_{nj}) - 2\tilde{\nu}_n(A_{nj})\right|\right) \quad (j = 1, \ldots, m_n)$$

and let $h(x)=f(x)$ and $g(x)=|D(x)|f(x)$, so

$$\rho_0^2-\gamma_0=1-\int_{\mathcal{R}^d}|D(x)|f(x)dx=2L^*.$$

Now, we check the conditions of the Proposition. Introduce

$$S_n=t\sqrt{n}\sum_{j=1}^{m_n}\left(|\tilde{\mu}_n(A_{nj})-2\tilde{\nu}_n(A_{nj})|-\mathbf{E}|\tilde{\mu}_n(A_{nj})-2\tilde{\nu}_n(A_{nj})|\right)+v\frac{N_n-n}{\sqrt{n}}$$

for which a central limit result is to be shown to hold. Remark that

$$Var(S_n)=n\sum_{j=1}^{m_n}\left\{t^2Var(|\tilde{\tau}_n(A_{nj})|)+2tv\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\right)\right\}+v^2,$$

Apply Lemma 1 for $Z_i=I_{\{X_i\in A\}}(1-2Y_i)$ and $p=2$ then

$$|nVar(\tilde{\tau}_n(A))-\mu(A)|\le\frac{A_2}{n^{1/2}}(\mu(A)^{1/2}+\mu(A))\le\frac{2A_2}{n^{1/2}}\mu(A)^{1/2},\qquad(9)$$

therefore, on the one hand we get that

$$\limsup_n n\sum_{j=1}^{m_n}Var(|\tilde{\tau}_n(A_{nj})|)\le\rho_\alpha^2.$$

Obviously

$$\begin{aligned}&\liminf_n n\sum_{j=1}^{m_n}Var(|\tilde{\tau}_n(A_{nj})|)\\&\quad\ge\liminf_n n\sum_{j=1}^{m_n}Var(\tilde{\tau}_n(A_{nj}))-\limsup_n n\sum_{j=1}^{m_n}((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-(\mathbf{E}\tilde{\tau}_n(A_{nj}))^2)\ge\rho_\alpha^2,\end{aligned}$$

if

$$\limsup_n n\sum_{j=1}^{m_n}((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-\tau(A_{nj})^2)=0.$$

It can be easily seen that by Lemma 2 for a set A

$$n((\mathbf{E}|\tilde{\tau}_n(A)|)^2-\tau(A)^2)\le 8n\mu(A)^2e^{-n\tau(A)^2/(3\mu(A))}.$$

Because of (11)

$$n((\mathbf{E}|\tilde{\tau}_n(A)|)^2-\tau(A)^2)\le\frac{1}{n}Var(n\tilde{\tau}_n(A))\le\mu(A)+\frac{2A_2}{n^{1/2}}\mu(A)^{1/2}.$$

From these by the inequality

$$\min(a+b,c)\le\min(a,c)+\min(b,c)\le\min(a,c)+b^{2/3}c^{1/3}$$

we get that

$$
\begin{aligned}
& n\sum_{j=1}^{m_n}((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-\tau(A_{nj})^2)\\
& \quad \le \sum_{j=1}^{m_n}\min\left(8n\mu(A_{nj})^2e^{-n\tau(A_{nj})^2/(3\mu(A_{nj}))},\mu(A_{nj})\right)\\
& \quad +8^{1/3}(2A_2)^{2/3}\sum_{j=1}^{m_n}\mu(A_{nj})e^{-n\tau(A_{nj})^2/(9\mu(A_{nj}))}\to 0 \qquad (10)
\end{aligned}
$$

if $\mathbf{P}(D(X)=0)=0$. To see this limit introduce the functions $\bar{f}_n(x)=\frac{\mu(A_n(x))}{\lambda(A_n(x))}$ and $D_n(x)=\frac{\tau(A_n(x))}{\mu(A_n(x))}$, then

$$
\begin{aligned}
& \sum_{j=1}^{m_n}\min\left(8n\mu(A_{nj})^2e^{-n\tau(A_{nj})^2/(3\mu(A_{nj}))},\mu(A_{nj})\right)\\
& \quad =\int_{S_\alpha}\min\left(8\mu(A(_{nj})e^{-\frac{1}{3}n\lambda(A_n(x))\bar{f}_n(x)D_n(x)^2+\log n},1\right)f(x)dx\\
& \quad \le\int_{S_\alpha}\min\left(8e^{-\frac{1}{3}n\lambda(A_n(x))\bar{f}_n(x)D_n(x)^2+\log n},1\right)f(x)dx
\end{aligned}
$$

where the convergence is dominated, so we need to show that for $f(x)>0$ the integrand tends to 0 almost everywhere w.r.t. λ, for which we can apply Theorem 7.16 in Wheeden, Zygmund (1977). The second term in (10) is similar. To complete the asymptotics for $Var(S_n)$ it remains to show that

$$
n\sum_{j=1}^{m_n}\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\right)\to\gamma_\alpha=\int_{S_\alpha}|D(x)|f(x)dx \text{ as } n\to\infty.
$$

For a set A

$$
n\mathbf{E}\left(|\tilde{\tau}_n(A)|(\tilde{\mu}_n(A)-\mu(A))\right)=n\mathbf{E}(|\tilde{\tau}_n(A)|\tilde{\mu}_n(A))-n\mathbf{E}(|\tilde{\tau}_n(A)|)\mu(A)
$$

For a set A let $Y(A)$ be a random variable with

$$
\mathbf{P}(Y(A)\in B)=\mathbf{P}(Y\in B|X\in A)
$$

then it can be easily seen that

$$
\begin{aligned}
& n\mathbf{E}\left(|\tilde{\tau}_n(A)|(\tilde{\mu}_n(A)-\mu(A))\right)\\
& \quad =\mu(A)\sum_{k=0}^{\infty}\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))|-\mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right)\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)}.
\end{aligned}
$$

Without loss of generality assume that $\tau(A)\ge 0$ then

$$
\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))|-\mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right)
$$

$$
\begin{aligned}
&= \mu(A)\left(\left(\mathbf{E}\sum_{i=1}^{k+1}(1-2Y_i(A)) + 2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+\right)\right.\\
&\left.-\left(\mathbf{E}\sum_{i=1}^{k}(1-2Y_i(A)) + 2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+\right)\right)\\
&= \mu(A)\left(\mathbf{E}(1-2Y_1(A)) + 2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+ - 2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+\right)
\end{aligned}
$$

Because of $|2(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+ - 2(-\sum_{i=1}^{k}(1-2Y_i(A)))^+| \le 2$ apply Hoeffdings inequality as in the proof of Lemma 2

$$
\begin{aligned}
&|\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right) - \mu(A)\mathbf{E}(1-2Y_1(A))|\\
&\quad = |\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right) - |\tau(A)||\\
&\quad = |2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^- 2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+|\\
&\quad \le 2\mu(A)\left(\mathbf{P}\{\sum_{i=1}^{k+1}(1-2Y_i(A))<0\} + \mathbf{P}\{\sum_{i=1}^{k}(1-2Y_i(A))<0\}\right)\\
&\quad \le 2\mu(A)\left(e^{-(k+1)\tau(A)^2/(2\mu(A)^2)} + e^{-k\tau(A)^2/(2\mu(A)^2)}\right)\\
&\quad \le 4\mu(A)e^{-k\tau(A)^2/(2\mu(A)^2)}.
\end{aligned}
$$

Summarizing the consequences of these inequlities we get that

$$
\begin{aligned}
&|n\sum_{j=1}^{m_n}\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\right) - \sum_{j=1}^{m_n}|\tau(A_{nj})||\\
&\quad \le \sum_{j=1}^{m_n}4\mu(A_{nj})\sum_{k=0}^{\infty}e^{-k\tau(A_{nj})^2/(2\mu(A_{nj})^2)}\frac{(n\mu(A_{nj}))^k}{k!}e^{-n\mu(A_{nj})}\\
&\quad \le \sum_{j=1}^{m_n}4\mu(A_{nj})e^{-n\tau(A_{nj})^2/(3\mu(A_{nj}))} \to 0,
\end{aligned}
$$

therefore

$$
n\sum_{j=1}^{m_n}\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\right) = o(1)+\sum_{j=1}^{m_n}|\tau(A_{nj})| \to \int_{S_\alpha}|D(x)|f(x)dx = \gamma_\alpha.
$$

To finish the proof of

$$
S_n \xrightarrow{\mathcal{D}} N\left(0, t^2\rho_\alpha^2 + 2tv\gamma_\alpha + v^2\right) \text{ as } n\to\infty
$$

by Lyapunov's central limit theorem, it suffices to prove that

$$n^{3/2}\left(\sum_{j=1}^{m_n}\mathbf{E}\left(\left|\,t\left(|\tilde{\tau}_n(A_{nj})|-\mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)+v(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\,\right|^3\right)\right.$$
$$\left.+\sum_{j=m_n+1}^{\infty}\mathbf{E}|v(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))|^3\right)\to 0,$$

or, by invoking the c_r inequality, that

$$n^{3/2}\sum_{j=1}^{m_n}\mathbf{E}\left(\left||\tilde{\tau}_n(A_{nj})|-\mathbf{E}|\tilde{\tau}_n(A_{nj})|\right|^3\right)\to 0,$$

and

$$n^{3/2}\sum_{j=1}^{\infty}\mathbf{E}\left(|\tilde{\mu}_n(A_{nj})-\mu(A_{nj})|^3\right)\to 0.$$

The second limit is shown in Beirlant, Györfi and Lugosi (1994). Concerning the first one apply Lemma 3 and Lemma 1 for $p=3$. □

The proof of Theorem 2 is similar, it is left to the reader.

Lemma 1 *(Beirlant, Mason (1995)) Let $\{Z_i\}$ i.i.d. random variables with $\mathbf{E}|Z_1|^p<\infty$, let N_n be a Poisson (n) random variable independent of $\{Z_i\}$, and N a standard normal random variable. Then there is a constant A_p (depending only on p) such that for $1\le p\le 3$*

$$\left|\mathbf{E}\left|\frac{\sum_{i=1}^{N_n}Z_i-n\mathbf{E}Z_1}{\sqrt{n\mathbf{E}Z_1^2}}\right|^p-\mathbf{E}|N|^p\right|\le\frac{A_p}{n^{1/2}}\left(\frac{\mathbf{E}|Z_1|^3}{(\mathbf{E}Z_1^2)^{3/2}}+1\right).$$

Lemma 2 *For a set A*

$$0\le\mathbf{E}|n\tilde{\tau}_n(A)|-|n\tau(A)|\le 2n\mu(A)e^{-\frac{n\tau(A)^2}{3\mu(A)}}.$$

Lemma 3 *If Z is a random variable with $\mathbf{E}|Z|^3<\infty$ then*

$$\mathbf{E}||Z|-\mathbf{E}|Z||^3<27\mathbf{E}|Z-EZ|^3.$$

References

Beirlant, J., Györfi, L. and Lugosi, G. (1994). On the asymptotic normality of the L_1- and L_2- errors in histogram density estimation. *Canadian J. Statistics*, **22**, 309-318.

Beirlant, J. and Mason, D. (1995). On the asymptotic normality of L_p-norms of empirical functionals. *Mathematical Methods of Statistics*, **4**, 1-19.

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

Wheeden, R. L. and Zygmund, A. (1977). *Measure and Integral.* Marcel Dekker Inc., New York, Basel.