Comparison of likelihood ratio rule with plug-in rule in classification

Kęstutis DUČINSKAS (KU)
e-mail: duce@gmf.ku.lt

1. Introduction

Suppose that individuals come from one of two mutually exclusive and exhaustive populations Ω_1, Ω_2 with positive prior probabilities π_1, π_2, respectively, where $\sum_{i=1}^{2}\pi_i = 1$. Let $X \in \mathcal{X} \subset R^p$ be random feature variable which is measured on each individual. Assume that the distribution of X for the individual from Ω_i has the probability density function (p.d.f.) $p_i(x;\Theta_i)$ which belongs to the parametric family of regular densities $F_i = \{p_i(x;\Theta_i), \Theta_i \in K \subset R^m\}, (i = 1,2)$.

Further, the dependence of any functions on any distribution parameters will be suppressed in the cases when functions are evaluted at the true values of these parameters denoted by asterisk $*$, e.g., $p_i(x;\Theta_i^*) = p_i(x)$. A decision is to be made as to witch population an individual randomly chosen from $\Omega = \cup_{i=1}^{2}\Omega_i$, belongs on the basis of an observed value of X. Let $d(\cdot)$ denote a classification rule (CR) formed for this purpose, where $d(x) = i$ implies that an individual with feature vector $X = x$ is to be assigned to the population Ω_i $(i = 1,2)$. In effect, CR divides the feature space **X** into L mutually exclusive and exhaustive assignment regions U_1, U_2, where if X falls in U_i then the individual is allocated to Ω_i $(i = 1,2)$.

When prior probabilities $\{\pi_i\}$ and densities $\{p_i(x)\}$ are known, the probability of misclassification (PM) $P(d(\cdot))$ associated with rule $d(\cdot)$ can be expressed as

$$P(d(\cdot)) = \sum_{i=1}^{2}\pi_i \int_X (1 - \delta(i, d(x))\, p_i(x)\mathrm{d}x, \tag{1}$$

where $\delta(i,j)$ is Kronecker's delta.

Then Bayes classificaion rule (BCR) $d_B(x)$ minimising the PM $P(d(\cdot))$ is defined as

$$d_B(x) = \operatorname*{argmax}_{i=1,2} \pi_i p_i(x). \tag{2}$$

Therefore, Bayes PM P_B is

$$P_B = \sum_{i=1}^{2}\pi_i \int_X (1 - \delta(i, d_B(x))\, p_i(x)\mathrm{d}x = \inf_{\{d(\cdot)\in D\}} P(d(\cdot)), \tag{3}$$

where D is the set of all CR $d(\cdot)$ defined before.

In practical applications, the density functions $\{p_i(x)\}$ are seldom completely known. Often they are only known up to the parameters $\{\theta_i\}$, i.e., we can only assert that $p_i(x)$ is an element of the parametric family of density functions F_i. Under such conditions, it is customary to estimate unknown parameters from given data.

Suppose that in order to estimate unknown parametrs θ_1, θ_2 there are M individuals of known origin on witch feature vector X has been recorded. That data is referred to in pattern recognition literature as training sample (TS). The only case of independent observations in TS will be considered in this paper. Suppose that TS realized under separate sampling (SS) design. This sample often is called stratified sample. Then the feature vectors are observed for a sample of M_i individuals taken separately from each population Ω_i $(i = 1, 2)$.

The so-called estimative approach to the choice of plug-in classification rule $d_s(x)$ is used. The unkown parametrs θ_1, θ_2 are replaced by appropriate estimates $\widehat{\theta}_1$, $\widehat{\theta}_2$ obtained from the training data T in the BCR, i.e., $d_s(x) = d(x, \widehat{\alpha})$, where $\widehat{\alpha}' = (\widehat{\theta}'_1, \widehat{\theta}'_2)'$.

The actual error rate for the rule $d(x, \widehat{\alpha})$ is the error rate of classifying a randomly selected individual with feature X and is designated by

$$P_A(\widehat{\alpha}) = \sum_{i=1}^{2} \pi_i \int_X (1 - \delta(i, d(x, \widehat{\alpha}))\, p_i(x)\mathrm{d}x. \tag{4}$$

It is obvious that $P_A(\alpha^*) = P_B$, where α^* is the true value of α.

Definition 1. Error regret (ER) for $d(x, \widehat{\alpha})$ is the difference between the actual error rate $P_A(\widehat{\alpha})$ and Bayes PM P_B, and the expected error regret (EER) is the expectation of ER, i.e.,

$$EER = E_T\{P_A(\widehat{\alpha})\} - P_B, \tag{5}$$

where $E_T\{P_A(\widehat{\alpha})\}$ denotes the expectation with respect to TS distribution.

Let $T = (T_1, T_2)$ be the training sample realized under SS scheme. Here

$$T'_i = (X'_{i1}, \ldots, X'_{i\mu_i}),$$

where X_{ij} is the j-th observation from Ω_i, $i = 1, 2$ and $M_1 + M_2 = M$.

Another classical CR usually called likelihood ratio (LR) rule is defined in the following way (see T.W. Anderson, 1984, Chapter 6)

$$d_{LR}(X) = \underset{i=1,2}{\operatorname{argmax}} \sup_{\{\theta_1, \theta_2 \in K\}} \{p_i(X; \theta_i) L_1(T_1; \theta_1) L_2(T_2; \theta_2)\}, \tag{6}$$

where

$$L_i(T_1; \theta_i) = \prod_{j=1}^{M_i} p_i(X_{i,j}; \theta_i).$$

The actual error rate and expected error regret for the LR rule is designated by

$$P_{LR}(T) = \sum_{i=1}^{2} \pi_i \int_X (1 - \delta(i, d_{LR}(x)))\,\mathrm{d}x, \tag{7}$$

$$EER_{LR} = E_T\{P_{LR}(T)\} - P_B. \tag{8}$$

Unfortunetely the close form expressions of EER and EER_{LR} are difficult to obtain. In those cases, large sample approximations to and asymptotic expansions are required.

The purpose of this paper is to compare EER in case of pluged ML estimators of θ_1, θ_2 with EER_{LR}. The asymptotic expansion of the difference between EER and EER_{LR} in the case of homoscedastic Normal populations is obtained. This is used to compare the proposed CR.

2. The main result

Suppose that distribution of X for the individual from Ω_i is $N_p(\mu_i, I_p)$, i.e., $\theta_i = \mu_i$, $i = 1, 2$.

Then ML estimators of μ_i based on T_i is

$$\hat{\mu}_i = \sum_{i=1}^{M_i} x_{ij}/M_i, \tag{9}$$

and ML estimators of μ_i based on T_i and X is

$$\hat{\mu}_i^* = (\hat{\mu}_i * M_i + X)/(M_i + 1), \quad i = 1, 2. \tag{10}$$

Assuming without loss of generatity that $\pi_1 = \pi_2$ and using (9), (10) we have

$$d(x; \{\hat{\mu}_i\}) = \underset{i=1,2}{\operatorname{argmin}}(x - \hat{\mu}_i)'(x - \hat{\mu}_i),$$

and

$$d_{LR}(x; \{\hat{\mu}_i\}) = \underset{i=1,2}{\operatorname{argmin}}\{(x - \hat{\mu}_i)'(x - \hat{\mu}_i)M_i/(M_i + 1)\}.$$

Let $e(x)$ be Heaviside function and let

$$G(x; \{\hat{\mu}_i\}) = (x - \hat{\mu}_1)'(x - \hat{\mu}_1) - (x - \hat{\mu}_2)'(x - \hat{\mu}_2),$$

$$G_1(x; \{\hat{\mu}_i\}) = (x-\hat{\mu}_1)'(x-\hat{\mu}_1)M_1/(M_1+1) - (x-\hat{\mu}_2)'(x-\hat{\mu}_2)M_2/(M_2+1),$$

Theorem. *Suppose that CR $d(x;\{\hat{\mu}_i\})$ and $d_{LR}(x;\{\hat{\mu}_i\})$ are used. Then, as $M_i \to \infty$ and $M_i/M \to r_i > 0$, $i = 1, 2$,*

$$EER_{LR} - EER = \phi(-\Delta/2) \times(\Delta^4/16-\Delta^2 p/2-3p^2-\Delta^2/2-1)(M_1-M_2)^2/(16\Delta M_1^2 M_2^2)+o(M^{-2}), \quad (11)$$

where

$$\Delta = ((\mu_1 - \mu_2)'(\mu_1 - \mu_2))^{1/2},$$
$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Proof. It is obvious that (see Ducinskas, 1997)

$$EER_{LR} - EER = E_T\left\{ \int (e(G(x;\{\hat{\mu}_i\})) - e\,(G_1(x;\{\hat{\mu}_i\}))\,G_0(x)\mathrm{d}x \right\}, \quad (12)$$

where

$$G_0(x) = p_1(x;\mu_1) - p_2(x;\mu_2). \quad (13)$$

Expanding integral in (12) at true values of parameters μ_1, μ_2 and taking the expectations we obtain (11).

The sign of the expression in the first bracked of (11) indicate the advantage of one compared rule against another.

References

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

[2] K. Ducinskas, An asymptotic analysis of the regret risk in discriminant analysis under various training schemes, *Lith. Math. J.*, 37(4), 337–351 (1997).

„Plug-in“ ir tikėtinumo santykio taisyklių palyginimas klasifikacijoje

K. Dučinskas

Straipsnyje pateiktas vidutinės klasifikavimo klaidos padidėjimo asimptotinis skleidinys „plug-in“ ir tikėtinumo santykio taisyklėms homoskedastinių normalinių populiacijų atvejui.