LMD mokslo darbai, III tomas, 361–366

Second-order asymptotics for discriminant analysis in exponential family

Kęstutis DUČINSKAS (KU, MII)
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 $\sum_{l=1}^{2}\pi_i = 1$. Let $X \in \mathbf{X} \subset \mathbf{R}^p$ be a random feature variable which is measured on each individual. Assume that the distribution of X for the individual from Ω_i, $i = 1, 2$ has the probability density function (p. d. f.) $p(x; \Theta_i)$ which belongs to the exponential family of densities $F = \{p(x; \Theta), \Theta \in \mathbf{K} \subset \mathbf{R}^m\}$, (see e.g., Barndorff - Nielsen, 1988).

This family includes many popular distributions (e.g., normal, Weibull, gamma, binomial, etc.). The MLE for Θ is a sufficient statistic in an exponential family, and achieves the Cramer - Rao lower bound if we have chosen the right function of Θ to estimate.

Suppose that there are m_1 elements of all $\{\Theta_i\}$ known a priori to be distinct and let θ_0 be the vector of m_0 elements known a priori to be equal, i.e., $\Theta_i = \left(\theta_0', \theta_i'\right)' = (\theta_0^1, \ldots, \theta_0^{m_0}, \theta_i^1, \ldots, \theta_i^{m_1})$, where $\theta_i^k \neq \theta_j^k$ for $i \neq j$, $(i, j = 1, 2;\ k = 1, \ldots, m_1)$, and $m_0 + m_1 = m$. The prime denotes vector transpose.

The Einstein summation convention will be adopted in this paper.

Denote by α an n-dimensional vector ($n = m_0 + 2m_1$), which consists of θ_0 and (θ_1, θ_2), i.e.,

$$\alpha = \left(\theta_0', \theta_1', \theta_2'\right)' = \left(\alpha^1, \ldots, \alpha^n\right). \tag{1}$$

Let $\mathbf{P} \subset \mathbf{R}^n$ be the set of all possible α, such that $\Theta_i \in \mathbf{K}$ $(i = 1, 2)$.

Further, the dependence of any functions on any distribution parameters will be suppressed in the cases when functions are evaluated 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 which population an individual randomly chosen from $\Omega = \bigcup_{i=1}^{2} \Omega_i$, belongs on the basis of an observed value of X. Let $d(\cdot, \alpha)$ denote a classification rule (CR) formed for this purpose, where $d(x, \alpha) = i$ 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 $\mathbf{X}$ 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 $(i = 1, 2)$. Let $C(i, j)$

denote the cost of allocation when an individual from Ω_i is allocated to Ω_j and let $C(i,j)$ be nonnegative and finite, $i,j = 1,2$.

When prior probabilities $\{\pi_i\}$ and densities $\{p_i(x)\}$ are known, the risk $R(d(\cdot,\alpha))$ associated with rule $d(\cdot,\alpha)$ can be expressed as

$$R(d(\cdot,\alpha)) = \sum_{i=1}^{2} \pi_i \int_{\mathbf{X}} C(i, d(x,\alpha)) p_i(x) \mathrm{d}x. \tag{2}$$

Then Bayes classification rule (BCR) $d_B(x,\alpha^*) = d_B(x)$ minimising the risk $R(d(\cdot,\alpha))$ is defined as

$$d_B(x) = \arg\max_{i=1,2} l_i p_i(x), \tag{3}$$

where

$$l_i = \pi_i\big(C(i,3-i) - C\,(i,i)\big), \quad (i = 1,2). \tag{4}$$

Therefore, Bayes risk $R_B(x,\alpha^*) = R_B$ of $d_B(x)$ is

$$R_B = \sum_{i=1}^{2} \pi_i \int_{\mathbf{X}} C\big(i, d_B(x)\big) p_i(x)\, \mathrm{d}x = \inf_{\{d(\cdot)\in D\}} R\big(d(\cdot)\big), \tag{5}$$

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

The risk becomes the probability of misclassification (PMC) when $C(i,j) = 1 - \delta_{ij}$, where δ_{ij} is the Kronecker's delta.

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 parameters Θ_1, Θ_2 there are M individuals of known origin on which feature vector X has been recorded. That data is referred to in pattern recognition literature as training sample (TS). Only the 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 sample-based classification rule is used. The unknown α is replaced by appropriate estimate $\widehat{\alpha}$ based on TS in the BCR and plug-in rule $d_B\,(x,\widehat{\alpha})$ is obtained. The case when $m_0 = 0$ means that all components of Θ_i are distinct for both populations.

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

$$R_A(\widehat{\alpha}) = \sum_{i=1}^{2} \pi_i \int_{\mathbf{X}} C\big(i, d_B(x, \widehat{\alpha})\big) p_i(x)\, \mathrm{d}x. \tag{6}$$

For $C(i, j) = 1 - \delta_{ij}$, the actual risk becomes the actual error rate (AER), which is usually used for evaluation of performance of a plug-in rule.

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

DEFINITION. Regret risk (RR) for $d_B\ (x, \widehat{\alpha})$ is the difference between the actual risk $R_A(\widehat{\alpha})$ and Bayes risk R_B, and the expected regret risk (ERR) is the expectation of RR, i.e.,

$$ERR(\widehat{\alpha}) = E_T\{R_A(\widehat{\alpha})\} - R_B, \tag{7}$$

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

It is obvious from (4), that RR is nonnegative random variable.

Unfortunately, the exact distributions of RR usually are difficult to obtain. In those cases, large sample approximations to an asymptotic expansions for the distributions and expectations of RR are required.

The purpose of this paper is to find second-order asymptotic expansions of ERR with respect to inverses of training sample sizes, when bias-adjusted MLE of unknown parameters of distributions from exponential family are used. These are used to evaluate the performance of sample-based CR and to find the optimal training sample allocation.

This is an extension of the result of Dučinskas (1995), who presented the first-order asymptotic expansion of expected error regret in the situation when parameter vectors of distributions, being classified, a priori had different all components. Neil (1980) has found the general asymptotic distribution of AER for the classification into one of two populations.

Taniguchi [4] presented sufficient conditions for the plug-in CR to be first- and second-order asymptotically best for the discriminant analysis in exponential families of distributions.

The general asymptotic distribution of RR and the first-order asymptotic expansion of ERR in case of several populations and MLE of unknown parameters of distributions, being classified, are derived in paper of Dučinskas (1997). The general second-order asymptotic expansion of EER in classification of one parametric distributions was obtained by Dučinskas (1999).

2. Notation and assumption

Let ∇_α be the vector partial differential operator given by

$$\nabla'_\alpha = \left(\frac{\partial}{\partial \alpha^1}, \dots, \frac{\partial}{\partial \alpha^n}\right) \quad \text{and} \quad |\nabla_\alpha|^2 = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \alpha^i}\right)^2$$

for any $\alpha = (\alpha^1, \dots, \alpha^n) \in R^n$.

Similarly, ∇_α^2 denotes the matrix second-order differential operator

$$\nabla_\alpha^2 = \left\| \frac{\partial^2}{\partial \alpha^i \partial \alpha^j} \right\|_{i,j=1,\dots,n}.$$

Let $\partial_{x_1} = \frac{\partial}{\partial_{x_1}}, \partial_{x_1}^2 = \frac{\partial^2}{\partial_{x_1^2}}, \dots$, and $G(x) = l_1 p_1(x) - l_2 p_2(x), \Gamma = \{x : x \in R^p, G(x) = 0\}$.

Suppose that $\widetilde{\alpha} = \left(\widetilde{\theta}_0^{\prime}, \widetilde{\theta}_1^{\prime}, \widetilde{\theta}_2^{\prime}\right) \in P$ is a nonrandom point in a neighbourhood of α^* and let L be the loglikelihood based on TS. Set $I = E_T \left\{ \nabla_\alpha L \nabla_\alpha^{\prime} L \right\} / N$ and let I^{ij} be the (i,j) element of I^{-1}.

To develope the asymptotic theory of $R(\widehat{\alpha})$ one needs the following assumption S:

S1. $R(\widetilde{\alpha})$ is five times differentiable with respect to $\widetilde{\alpha}$ in a neighbourhood of α^*;

S2. $M_1 = N$, $M_2 = cN$, where c is a fixed positive constant;

S3. $E(\widehat{\alpha}) - \alpha^* = o(1)$ and $\partial_i = \partial / \partial \alpha^i$ and E are interchangeable;

S4. Let P_N be the probability distribution of $\sqrt{N}(\widehat{\alpha} - \alpha^*)$ $(= v$, say$)$. Then it holds that for every bounded, Borel - measurable function f of v on $\mathbf{R}^n$,

$$\left| \int f \mathrm{d}(P_N - F_N) \right| = o(N^{-2}),$$

where $F_N = F_N(v)$ is the formal Edgeworth expansion of $v = \sqrt{N}(\widehat{\alpha} - \alpha^*)$.

Suppose that $\widehat{\alpha}_M$ is MLE of α based on TS and let

$$\widehat{\alpha}_M^i - \alpha^{*i} = Z^i + Q^i / \sqrt{N} + o_p\left(N^{-1/2}\right),$$
$$E\left(\widehat{\alpha}_M^i - \alpha^{*i}\right) = \mu^i(\alpha^*) / N + o\left(N^{-3/2}\right), \quad i = 1, \dots, n,$$

where $Z^i = I^{ij} \partial_j L / \sqrt{N}$, $Q^i = O_p(1)$, $\mu^i(\alpha^*) = E\{Q^i\}$.

Then bias-adjusted MLE $\widehat{\alpha}_u$ is defined by

$$\widehat{\alpha}_u^i = \widehat{\alpha}_M^i - \mu^i(\widehat{\alpha}) / N, \quad i = 1, \dots, n.$$

Explicit expressions of $\{Q^i\}$ are presented in Taniguchi (1994). In this paper the author also proved, that bias-adjusted MLE $\widehat{\alpha}_u$ minimizes $EER(\widehat{\alpha})$ up to $O\left(N^{-2}\right)$ in the some class of bias-adjusted estimators. The closed-form expressions for the coefficients of the second-order asymptotic expansion of $EER(\widehat{\alpha}_u)$ is derived in the following theorem.

Theorem. *Let the regularity assumption S hold. Then*

$$\begin{aligned} EER(\widehat{\alpha}_u) = {} & \frac{1}{2N}\int_\Gamma \partial_i G \partial_j G \left|\nabla_x G\right|^{-1} \mathrm{d}\gamma \\ & + \frac{1}{2N^2}\Big(\int_\Gamma \partial_i G \partial_j G \left|\nabla_x G\right|^{-1} \mathrm{d}\gamma \operatorname{cov}(Q^i, Q^j) \\ & + R_{ijk}c^{ijk}/3 + R_{ijkl}(I^{ij}I^{kl} + I^{ik}I^{jl} + I^{il}I^{jk})/12 \Big) + o(N^{-2}), \end{aligned}$$

where

$$\begin{aligned} \partial_i\partial_j\partial_k R = {} & \int_\Gamma \Big(\partial_i\partial_j(\partial_k G \cdot G) + 2\big(\partial_i\partial_j\partial_k G \cdot \partial^2_{x_1} G/(\partial_{x_1} G) \\ & - \partial_{x_1}(\partial_i\partial_j\partial_k G)\big)\Big) \left|\nabla_x G\right|^{-1} \mathrm{d}\gamma, \\ \partial_i\partial_j\partial_k\partial_l R = {} & \int_\Gamma \big(\partial_i\partial_j\partial_k(\partial_l G \cdot G) + 2\big(\partial_i\partial_j\partial_k(\partial_l G \cdot G^2) \cdot \partial^2_{x_1} G/(\partial_{x_1} G)^2 \\ & - \partial_{x_1}\big(\partial_i\partial_j\partial_k(\partial_l G \cdot G^2)/\partial_{x_1} G\big) + 3\big(\partial^2_{x_1}(\partial_i G \partial_j G \partial_k G \partial_l G) \\ & -\big(3\partial_{x_1}(\partial_i G \partial_j G \partial_k G \partial_l G) \cdot \partial^2_{x_1} G + (\partial_i G \partial_j G \partial_k G \partial_l G)\partial^3_{x_1} G\big)/\partial_{x_1} G \\ & + 3(\partial_i G \partial_j G \partial_k G \partial_l G)(\partial^2_{x_1} G)^2/(\partial_{x_1} G)^2)\big) \cdot \big(\partial_{x_1} G^2 \left|\nabla_x G\right|\big)^{-1} \mathrm{d}\gamma, \end{aligned}$$

$$c^{ijk} = (I^{il}\partial_l I^{jk} + I^{jl}\partial_l I^{ik} + I^{kl}\partial_l I^{ij} - E_T(Z^i Z^j Z^k))/2.$$

Proof. The proof of Theorem is based on the fifth-order Taylor expansion of $EER(\widehat{\alpha}_u)$ in some neighbourhood of α^*. By taking the expectation according to the distribution of TS of this expansion and using the results of Lemma 2 of [6] and Theorem 2 of [4] the proof of stated Theorem is completed.

References

[1] O. Barndorff - Nielsen, Parametric statistical models and likelihood, it Lecture Notes in Statistics, Springer-Verlag, New-York/Berlin, **50** (1988).

[2] K. Dučinskas, Optimal training sample allocation and asymptotic expansions for error rates in discriminant analysis, *Acta Appl. Math.*, **38**, 3–11 (1995).

[3] T.J. O'Neill, The general distribution of the error rate of a classification procedure with application to logistic regression discriminant, *J. Amer. Statist., Assoc.*, **75**, 154–160 (1980).

[4] M. Taniguchi, Higher order asymptotic theory for discriminant analysis in exponential families of distributions, *J. Multivariate Analysis*, **48**, 169–187 (1994).

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

[6] K. Dučinskas, Second-order expansion for the expected regret risk in classification of one-paranetric distributions, *Liet. Mat. Rink.*, **39**(2), 220–230 (1999).

Skirstinių, priklausančių eksponentinei klasei, antros eilės asimptotika diskriminantinėje analizėje

K. Dučinskas

Straipsnyje nagrinėjamas skirstinių, priklausančių eksponentinei klasei, diskriminantinės analizės uždavinys. Gautas klasifikavimo rizikos padidėjimo antros eilės asimptotinis skleidinys atvejui, kai naudojame maksimalaus tikėtinumo parametrų įverčius pagal stratifikuotas mokymo imtis.