INFORMATICA, 1996, Vol. 7, No. 2, 137–154

THE EXPECTED PROBABILITY OF MISCLASSIFICATION OF LINEAR ZERO EMPIRICAL ERROR CLASSIFIER

Alfredas BASALYKAS

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St. 4, Lithuania

Abstract. There exist two principally different approaches to design the classification rule. In classical (parametric) approach one parametrizes conditional density functions of the pattern classes. In a second (nonparametric) approach one parametrizes a type of the discriminant function and minimizes an empirical classification error to find unknown coefficients of the discriminant function. There is a number of asymptotic expansions for an expected probability of misclassification of parametric classifiers. Error bounds exist for nonparametric classifiers so far. In this paper an exact analytical expression for the expected error EP_N of nonparametric linear zero empirical error classifier is derived for a case when the distributions of pattern classes are spherically Gaussian. The asymptotic expansion of EP_N is obtained for a case when both the number of learning patterns N and their dimensionality p increase infinitely. The tables for exact and approximate expected errors as functions of N, dimensionality p and the distance δ between pattern classes are presented and compared with the expected error of the Fisher's linear classifier and indicate that the minimum empirical error classifier can be used even in cases where dimensionality exceeds the number of learning examples.

Key words: expected error, Fisher's discriminant function, zero empirical error classifier, dimensionality, learning set's size.

1. Introduction. Let $\mathbf{X} = (x_1, x_2, \ldots, x_p)'$ be $p \times 1$ observation vector of an individual from one or another of two p-variate classes (populations) π_1 and π_2. Suppose we have two learning sets, of sizes N_1 and N_2, of p-dimensional data from populations π_1 and π_2 respectively. The problem is to utilize an information contained in the learning sets and to design a classification rule assigning vector $\mathbf{X}$ to one of the classes.

In statistic-theoretical approach it is assumed vector $\mathbf{X}$ to be random one with class conditional probability density function $f_i(\mathbf{X}|\pi_i)$. Let q_i be a priori probability of class π_i $(q_1 + q_2 = 1)$. Then an optimal (Bayes) classification

rule (which minimizes the probability of incorrect classification) will assign vector $\mathbf{X}$ to one of classes π_i according a sign of the following discriminant function (DF):

$$g(\mathbf{X}) = \ln \frac{q_1 f_1(\mathbf{X}|\pi_1)}{q_2 f_2(\mathbf{X}|\pi_2)}. \tag{1}$$

Principally different approach is if one instead of parametrization of the probability density functions (p.d.f.) will parametrize the discriminant function itself. For example one can assume the DF has a linear form

$$g(\mathbf{X}) = \sum_{i=1}^{p} a_i x_i + a. \tag{2}$$

To find unknown coefficients (weights of the DF) $a, a_1, a_2, \ldots, a_p$ one introduces a certain loss function (empirical classification error, sum of squares error etc.) and minimizes it. Latter approch became very popular in recent years in an analysis and development of Artificial Neural Networks.

In both classifier design approaches resulting DF depends on the learning set data. In finite learning set case the data does not represent the populations (probability density functions $f_i(\mathbf{X}|\pi_i)$) exactly. Therefore the resulting classification rule is not optimal. Its classification performance will differ from Bayes error, i.e., probability of misclassification P_B of optimal Bayes classifier (1). A probability of misclassification P_N of sample based classification rule will depend of particular learning sets. Therefore it is called a conditional probability of misclassification (PMC). Its expectation EP_N over all possible random learning sets of size N_1 and N_2 is called an expected PMC. A theoretical limit

$$\lim_{\substack{N_1 \to \infty \\ N_2 \to \infty}} EP_N = P_\infty$$

is called an asymptotic PMC.

The expected PMC was studied in a number of research papers beginning from pionering work of John (1961) who obtained first exact and approximate formulae for the standard linear DF for the Gaussian classes for case when Σ, the covariance matrix, is known. Best known asymptotic expansion for case when Σ is known due to Okamoto (1963). Principal resuls were obtained by Deev (1970, 1972), Raudys (1967, 1972). Most of results on the subject are summarized in McLachlan's monograph (1992). Results of Soviet investigators

are reffered in Aivazian *et al.* monograph (1989), Raudys and Jain review (1991) and also in Wyman *et al.* (1990) experimental comparison of several asymptotic expansions for expected error of the standard Fisher linear DF.

An objective of this paper is to obtain a formula for expected PMC of the linear zero empirical error classifier (a special case of minimal empirical error classifier) for a case when true densities $f(\mathbf{X}|\pi_i)$ are multivariate spherically Gaussian.

2. Main assumptions. Let us assume we have the linear classifier with discriminant function $g(\mathbf{X})$:

$$g(\mathbf{X}) = \mathbf{A'X} + a = \sum_{i=1}^{p} a_i x_i + a.$$

To find weights $a, a_1, a_2, \ldots, a_p$ we'll use following hypothetical procedure (Raudys, 1993).

According to some chosen prior density $f_{\text{prior}}(a, \mathbf{A})$ of weight vector $(a, \mathbf{A})$ one generates a set of random weights $a, a_1, a_2, \ldots, a_p$. We will say that training is sucessful if conditions S are satisfied, where

$$S: \begin{cases} \text{for all training pattern vectors from } \pi_1 & g(\mathbf{X}|a, \mathbf{A}) > 0, \\ \text{for all training pattern vectors from } \pi_2 & g(\mathbf{X}|a, \mathbf{A}) \leqslant 0. \end{cases} \tag{3}$$

We shall compute the expected PMC EP_N of successfully trained linear discriminant function.

In order to obtain an analytical expression for the expected PMC suitable for numerical evalution of the error rate we need to specify prior density $f_{\text{prior}}(a, \mathbf{A})$ and true probability density functions of the pattern classes $f(\mathbf{X}|\pi_1), f(\mathbf{X}|\pi_2)$. Thus we shall analyze a case of simple distributions:

– two multivariate spherically Gaussian classes π_1, π_2 with densities $\mathrm{N}(\mathbf{X}, \mathbf{C}_1, \mathbf{I})$ and $\mathrm{N}(\mathbf{X}, \mathbf{C}_2, \mathbf{I})$ accordingly, equal prior probabilities $q_1 = q_2 = 1/2$ and equal number of training vectors from each class: $N_2 = N_1 = N$;

– the training vectors $\mathbf{X}_1^{(1)}, \mathbf{X}_2^{(1)}, \ldots, \mathbf{X}_N^{(1)}, \mathbf{X}_1^{(2)}, \mathbf{X}_2^{(2)}, \ldots, \mathbf{X}_N^{(2)}$ are statistically independent and identically distributed in their own classes;

– we assume that $\mathbf{C}_1 + \mathbf{C}_2 = \mathbf{0}$ (this assumption enables us further to simplify the calculations);

– the components of the vector $(a, \mathbf{A})$ are chosen random from Gaussian distribution with zero mean and variance 1: $a_i \sim \mathrm{N}(0, 1)$.

We shall analyze a limit case when $N \to \infty$, $p \to \infty$ and $p/N \to$ const.

3. Integral representation of the expected error. A derivation of the mean expected error EP_N is based on calculation of the conditional probability of misclassification of the linear classifier conditioned on the set of weights $(a, \mathbf{A})$, on representations of conditional error rates in terms of two independent scalar random variables and subsequent averaging of these error rates over aposteriori distribution of weights $(a, \mathbf{A})$ (Raudys, 1993):

$$f_{\text{apost}}(a, \mathbf{A}|S) = \frac{\Pr(S = \text{true} \mid a, \mathbf{A})\, f_{\text{prior}}(a, \mathbf{A})}{\Pr(S = \text{true})} = \frac{\Pr(S = \text{true} \mid a, \mathbf{A})\, f_{\text{prior}}(a, \mathbf{A})}{\iint \Pr(S = \text{true} \mid a, \mathbf{A})\, f_{\text{prior}}(a, \mathbf{A})\, da\, d\mathbf{A}'}, \tag{4}$$

$$EP_N = \Pr(MC|S) = \int\int \Pr(MC|a, \mathbf{A}) f_{\text{apost}}(a, \mathbf{A}|S)\, da\, d\mathbf{A}, \tag{5}$$

where $\Pr(MC|a, \mathbf{A})$ is a conditional probability of misclassification given the set of weights $(a, \mathbf{A})$ and $f_{\text{apost}}(a, \mathbf{A} \mid S = \text{true})$ is aposteriori density function of the weights if the training was successful, i.e., conditions S were satisfied.

Due to our assumptions the distribution of discriminant function $g(\mathbf{X})$ will be Gaussian and the conditional probability of misclassification

$$\begin{aligned}\Pr(MC|\mathbf{A}, a) &= \frac{1}{2} \Pr\left(\mathbf{A'X} + a \leqslant 0 \mid \mathbf{X} \in \pi_1\right) \\ &\quad + \frac{1}{2} \Pr\left(\mathbf{A'X} + a > 0 \mid \mathbf{X} \in \pi_2\right) \\ &= \frac{1}{2} \Phi\left(-\frac{\mathbf{A'C}_1 + a}{\sqrt{\mathbf{A'A}}}\right) + \frac{1}{2} \Phi\left(\frac{\mathbf{A'C}_2 + a}{\sqrt{\mathbf{A'A}}}\right),\end{aligned} \tag{6}$$

where

$$\Phi(u) = \int_{-\infty}^{u} \varphi(t)\, dt \qquad \text{and} \qquad \varphi(t) = \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}.$$

Conditional PMC (6) depends on $(p+1)$-variate vector $(a, \mathbf{A})'$. For spherical case we can show this PMC depends only on two independent scalar variables.

Let us perform a transformation

$$\mathbf{V} = \mathbf{TA} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix} \quad \text{and} \quad \mathbf{T}(\mathbf{C}_1 - \mathbf{C}_2) = \begin{bmatrix} \delta \\ l_2 \\ \vdots \\ l_p \end{bmatrix},$$

where $\mathbf{T}$ is $p \times p$ orthogonal matrix with a first row vector

$$\mathbf{t}_1 = \frac{(\mathbf{C}_1 - \mathbf{C}_2)}{\sqrt{(\mathbf{C}_1 - \mathbf{C}_2)'(\mathbf{C}_1 - \mathbf{C}_2)}}, \tag{7}$$

and $\delta^2 = (\mathbf{C}_1 - \mathbf{C}_2)'(\mathbf{C}_1 - \mathbf{C}_2)$ is a squared Mahalanobis distance. Then

$$\begin{aligned} \frac{\mathbf{A'C}_1 + a}{\sqrt{\mathbf{A'A}}} &= \frac{(\mathbf{TA})'(\mathbf{T}(\mathbf{C}_1 - \mathbf{C}_2) + \mathbf{T}(\mathbf{C}_1 + \mathbf{C}_2)) + 2a}{2\sqrt{(\mathbf{TA})'(\mathbf{TA})}} \\ &= \frac{v_1\delta + w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}} = u\frac{\delta}{2} + w, \end{aligned} \tag{8}$$

where

$$w_0 = (\mathbf{TA})'(\mathbf{T}(\mathbf{C}_1 + \mathbf{C}_2)) + 2a = \mathbf{A}'(\mathbf{C}_1 + \mathbf{C}_2) + 2a = 2a, \tag{9}$$

as $\mathbf{C}_1 + \mathbf{C}_2 = \mathbf{0}$ by our assumption, and

$$u = \frac{v_1}{\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}}, \qquad w = \frac{w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p} v_i^2}}.$$

Analogously

$$\frac{\mathbf{A'C}_2 + a}{\sqrt{\mathbf{A'A}}} = -u\frac{\delta}{2} + w. \tag{10}$$

Therefore, conditional error rate can be represented in terms of two independent scalar variables, u and w:

$$\Pr(MC|a, \mathbf{A}) = \Pr(MC|u, w) = \frac{1}{2}\Phi\left(-u\frac{\delta}{2} - w\right) + \frac{1}{2}\Phi\left(-u\frac{\delta}{2} + w\right). \tag{11}$$

As $a_i \sim \mathrm{N}(0, 1)$ are independent then it is not difficult to show that random variables u and w are independent and have Beta $\mathrm{Be}((p-1)/2,\ (p-1)/2)$ and

Student St(p) distributions accordingly. Their density functions $h(u)$ and $p(w)$ accordingly are

$$h(u) = \begin{cases} K_1(1-u^2)^{(p-3)/2}, & \text{if } |u| \leqslant 1, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

and

$$p(w) = K_2(1+w^2)^{-(p+1)/2}, \quad w \in \mathbf{R}. \tag{13}$$

Here K_1 and K_2 are positive terms which depend only on p.

For independent identically distributed training pattern vectors the conditional probability

$$\begin{aligned} \Pr(S = \text{true} \mid a, \mathbf{A}) &= \prod_{j=1}^{N} \Pr\{\mathbf{A}'\mathbf{X}_j^{(1)} + a > 0\} \prod_{j=1}^{N} \Pr\{\mathbf{A}'\mathbf{X}_j^{(2)} + a \leqslant 0\} \\ &= \left[\Pr\{\mathbf{A}'\mathbf{X} + a > 0 \mid \mathbf{X} \in \pi_1\}\right]^N \left[\Pr\{\mathbf{A}'\mathbf{X} + a \leqslant 0 \mid \mathbf{X} \in \pi_2\}\right]^N \\ &= \left[1 - \Pr\{\mathbf{A}'\mathbf{X} + a \leqslant 0 \mid \mathbf{X} \in \pi_1\}\right]^N \\ &\quad \times \left[1 - \Pr\{\mathbf{A}'\mathbf{X} + a > 0 \mid \mathbf{X} \in \pi_2\}\right]^N \\ &= \left[1 - \Phi\left(-\frac{\mathbf{A}'\mathbf{C}_1 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\right)\right]^N \times \left[1 - \Phi\left(\frac{\mathbf{A}'\mathbf{C}_2 + a}{\sqrt{\mathbf{A}'\mathbf{A}}}\right)\right]^N. \end{aligned}$$

Taking into account (8) and (10) the above equation can be rewriten in a form

$$\begin{aligned} \Pr(S = \text{true} \mid a, \mathbf{A}) &= \Pr(S = \text{true} \mid u, w) \\ &= [1 - \Phi(-u\delta/2 - w)]^N [1 - \Phi(-u\delta/2 + w)]^N \\ &= [\Phi(u\delta/2 + w)\Phi(u\delta/2 - w)]^N. \end{aligned} \tag{14}$$

Noticing that $f_{\text{prior}}(u, w) = h(u)p(w)$ and inserting (11) – (14) into (4), (5) we obtain

$$EP_N = \frac{I_1 + I_2}{J_1 + J_2}, \tag{15}$$

where

$$I_1 = \int_{-\infty}^{\infty} \int_{0}^{1} F(u, w) \left[\Phi\left(u\frac{\delta}{2} + w\right)\Phi\left(u\frac{\delta}{2} - w\right)\right]^N h(u)p(w)\,du\,dw, \tag{16}$$

$$I_2 = \int_{-\infty}^{\infty}\int_{0}^{1} F(-u,w)\left[\Phi\left(-u\frac{\delta}{2}+w\right)\Phi\left(-u\frac{\delta}{2}-w\right)\right]^N h(u)p(w)dudw, \quad (17)$$

$$J_1 = \int_{-\infty}^{\infty}\int_{0}^{1} \left[\Phi\left(u\frac{\delta}{2}+w\right)\Phi\left(u\frac{\delta}{2}-w\right)\right]^N h(u)p(w)\,du\,dw, \quad (18)$$

$$J_2 = \int_{-\infty}^{\infty}\int_{0}^{1} \left[\Phi\left(-u\frac{\delta}{2}+w\right)\Phi\left(-u\frac{\delta}{2}-w\right)\right]^N h(u)p(w)\,du\,dw, \quad (19)$$

$$F(u,w) = \frac{1}{2}\left(\Phi(-u\delta/2+w)+\Phi(-u\delta/2-w)\right).$$

4. Asymptotic expansion for the expected error. Let us denote

$$S_1(u,w) = \ln\Phi(u\delta/2+w) + \ln\Phi(u\delta/2-w) - \lambda_1\ln(1+w^2),$$

where $\lambda_1 = \dfrac{p+1}{2N}$,

$$S_2(u,w) = \ln\Phi(-u\delta/2+w) + \ln\Phi(-u\delta/2-w) - \lambda_2\ln(1-u^2),$$

where $\lambda_2 = \dfrac{p-3}{2N}$.

Then the integrals (16)–(19) we can write in the following form:

$$I_1 = \int_0^1 I_1(u)(1-u^2)^{N\lambda_2}\,du, \qquad J_1 = \int_0^1 J_1(u)(1-u^2)^{N\lambda_2}\,du,$$

$$I_2 = \int_{-\infty}^{\infty} I_2(w)(1+w^2)^{-N\lambda_1}\,dw, \qquad J_2 = \int_{-\infty}^{\infty} J_2(w)(1+w^2)^{-N\lambda_1}\,dw,$$

where

$$I_1(u) = \int_{-\infty}^{\infty} F(u,w)e^{NS_1(u,w)}\,dw, \qquad J_1(u) = \int_{-\infty}^{\infty} e^{NS_1(u,w)}\,dw,$$

$$I_2(w) = \int_0^1 F(-u,w)e^{NS_2(u,w)}du, \qquad J_2(w) = \int_0^1 e^{NS_2(u,w)}\,du.$$

First at all let us deal with integrals $I_1(u)$ and $J_1(u)$. It is not difficult to see that these integrals are Laplace integrals with parameter N increacing to infinity

and phase function $S_1(u, w)$. In order to compute $I_1(u)$ and $J_1(u)$ we shall use the methods contained in Fedorchuk monograph (1987). Therefore we have to find the maximal point of phase function $S_1(u, w)$ by $w \in \mathbf{R}$. As

$$\Phi(u\delta/2 + w) + \Phi(u\delta/2 - w) \leqslant 2\Phi(u\delta/2), \quad w \in \mathbf{R},\ u \in [0, 1], \tag{20}$$

then

$$\begin{aligned} S_1(u, w) &= \ln\left(\Phi(u\delta/2 + w)\Phi(u\delta/2 - w)(1 + w^2)^{-\lambda_1}\right) \\ &\leqslant \ln\left(\Phi(u\delta/2 - w)\left(2\Phi(u\delta/2) - \Phi(u\delta/2 - w)\right)\right) \\ &= \ln\left(\Phi^2(u\delta/2) - \left(\Phi(u\delta/2 - w) - \Phi(u\delta/2)\right)^2\right) \\ &\leqslant \ln \Phi^2(u\delta/2) = S_1(u, 0). \end{aligned} \tag{21}$$

Inequality (20) holds because point $w = 0$ for function $R(w) = \Phi(u\delta/2 + w) + \Phi(u\delta/2 - w)$ is its stationary point and $R'(w) > 0$ as $w < 0$ and $R'(w) < 0$ as $w > 0$.

Therefore (21) yields that $w = 0$ is maximal point of $S_1(u, w)$ by $w \in \mathbf{R}$. Moreover,

$$\left.\frac{d^2 S_1(u, w)}{dw^2}\right|_{w=0} \neq 0.$$

Now by Th.1.3 from Fedorchuk (1987, p.66) we obtain that as $N \to \infty$

$$I_1(u) \sim e^{NS_1(u,0)}\sqrt{\frac{\pi}{N}}\left(a_0(u) + \frac{a_1(u)}{N} + \cdots\right), \tag{22}$$

$$J_1(u) \sim e^{NS_1(u,0)}\sqrt{\frac{\pi}{N}}\left(b_0(u) + \frac{b_1(u)}{N} + \cdots\right), \tag{23}$$

where

$$\begin{aligned} a_0(u) &= F(u, 0)\beta_1^{-1/2}(u), \qquad b_0(u) = \beta_1^{-1/2}(u), \\ a_1(u) &= -\frac{1}{4}\left(\beta_2(u)\Phi(-u\delta/2)\beta_1^{-5/2}(u) + \frac{1}{3}\beta_1^{-3/2}(u)\Phi''(x)|_{x=u\delta/2}\right) \quad \text{and} \\ b_1(u) &= -\frac{1}{4}\beta_2(u)\beta_1^{-5/2}(u). \end{aligned}$$

In formulae (22), (23) and further symbol $V_N \sim U_N$ as $N \to \infty$ means that

$$\lim_{N\to\infty} V_N/U_N = 1.$$

Also the terms following after $a_1(u)/N$ and $b_1(u)/N$ as $N \to \infty$ are of order $1/N^2$ and

$$\beta_j(u) = -\frac{1}{(2j)!}\frac{d^{2j}S_1(u,w)}{dw^{2j}}\bigg|_{w=0}$$

$$= (-1)^{j-1}\frac{\lambda_1}{j} - \frac{2}{(2j)!}\ln^{(2j)}\Phi(x)|_{x=u\delta/2}, \qquad j = 1,2.$$

Let us denote

$$Z(u) = \ln\Phi\left(u\frac{\delta}{2}\right) + \frac{\lambda_2}{2}\ln(1-u^2), \quad u \in [0,1).$$

Then

$$I_1 \sim \frac{\sqrt{\pi}}{\sqrt{N}}\int_0^1 e^{2NZ(u)}\left(a_0(u) + \frac{a_1(u)}{N} + \cdots\right)du, \tag{24}$$

$$J_1 \sim \frac{\sqrt{\pi}}{\sqrt{N}}\int_0^1 e^{2NZ(u)}\left(b_0(u) + \frac{b_1(u)}{N} + \cdots\right)du. \tag{25}$$

In order to calculate the integrals (24), (25) we have to explore the behaviour of the function $Z(u)$ in interval [0,1). For this purpose we calculate two first derivaties of $Z(u)$:

$$Z'(u) = \frac{\delta}{2\sqrt{2\pi}}\frac{e^{-u^2\delta^2/8}}{\Phi(u\delta/2)} - \frac{\lambda_2 u}{(1-u^2)},$$

$$Z''(u) = -\frac{\delta^2}{4\sqrt{2\pi}}\frac{e^{-u^2\delta^2/8}}{\Phi(u\delta/2)}\left(\frac{u\delta}{2} + \frac{1}{\sqrt{2\pi}}\frac{e^{-u^2\delta^2/8}}{\Phi(u\delta/2)}\right) - \frac{\lambda_2(1+u^2)}{(1-u^2)^2}.$$

It is not difficult to see that $Z''(u)$ in the interval [0,1) is negative and the first derivative $Z'(u)$ in the same interval is changing it's sign from $+$ to $-$. This means that in this interval there is point u_0 in which function $Z(u)$ has its maximal value $Z(u_0)$ and $Z(u_0) > Z(0) = -\ln 2$. We may find this point by solving equation $Z'(u) = 0$ or

$$\frac{\delta}{\sqrt{2\pi}}e^{-u^2\delta^2/8}(1-u^2) = 2\lambda_2 u\Phi\left(u\frac{\delta}{2}\right). \tag{26}$$

Again, as $Z'(u_0) = 0$ and $Z''(u_0) \neq 0$, applying the same theorem from Fedorchuk monograph we obtain that

$$\int_0^1 e^{2NZ(u)} a_i(u)du \sim e^{2NZ(u_0)} \sqrt{\frac{\pi}{2N}} \left(a_{i0} + \frac{a_{i1}}{2N} + \cdots \right), \tag{27}$$

$$\int_0^1 e^{2NZ(u)} b_i(u)du \sim e^{2NZ(u_0)} \sqrt{\frac{\pi}{2N}} \left(b_{i0} + \frac{b_{i1}}{2N} + \cdots \right), \quad i = 0, 1, \tag{28}$$

where

$$a_{i0} = a_i(u_0) \sqrt{-\frac{2}{Z''(u_0)}}, \qquad b_{i0} = b_i(u_0) \sqrt{-\frac{2}{Z''(u_0)}}, \tag{29}$$

$$a_{i1} = \frac{1}{4} \left(-\frac{2}{Z''(u_0)} \right)^{3/2} \left(a_i''(u_0) - a_i'(u_0) \frac{Z'''(u_0)}{Z''(u_0)} + a_i(u_0) \left(\frac{5}{12} \left(\frac{Z'''(u_0)}{Z''(u_0)} \right)^2 - \frac{Z^{(IV)}(u_0)}{4Z''(u_0)} \right) \right), \tag{30}$$

$$b_{i1} = \frac{1}{4} \left(-\frac{2}{Z''(u_0)} \right)^{3/2} \left(b_i''(u_0) - b_i'(u_0) \frac{Z'''(u_0)}{Z''(u_0)} + b_i(u_0) \left(\frac{5}{12} \left(\frac{Z'''(u_0)}{Z''(u_0)} \right)^2 - \frac{Z^{(IV)}(u_0)}{4Z''(u_0)} \right) \right). \tag{31}$$

Inserting (27), (28) into (24) and (25) we obtain

$$I_1 \sim \frac{\pi}{N} \sqrt{\frac{1}{2}} e^{2NZ(u_0)} \left[\left(a_{00} + \frac{a_{01}}{2N} + \cdots \right) + \frac{1}{N} \left(a_{10} + \frac{a_{11}}{2N} + \cdots \right) + \cdots \right]$$

$$= \frac{\pi}{N} \sqrt{\frac{1}{2}} e^{2NZ(u_0)} \left[a_{00} + \frac{1}{N} \left(\frac{a_{01}}{2} + a_{10} \right) + \cdots \right]. \tag{32}$$

Analogously obtain

$$J_1 \sim \frac{\pi}{N} \sqrt{\frac{1}{2}} e^{2NZ(u_0)} \left[\left(b_{00} + \frac{b_{01}}{2N} + \cdots \right) + \frac{1}{N} \left(b_{10} + \frac{b_{11}}{2N} + \cdots \right) + \cdots \right]$$

$$= \frac{\pi}{N} \sqrt{\frac{1}{2}} e^{2NZ(u_0)} \left[b_{00} + \frac{1}{N} \left(\frac{b_{01}}{2} + b_{10} \right) + \cdots \right]. \tag{33}$$

Let us now deal with integrals $I_2(w)$, $J_2(w)$:

$$I_2(w) = \int_0^1 F(-u, w)e^{NS_2(u,w)}du, \qquad J_2(w) = \int_0^1 e^{NS_2(u,w)}du.$$

It is easy to see that

$$\max_{u\in[0,1)} S_2(u, w) = S_2(0, w) = \ln\Phi(w) + \ln\Phi(-w).$$

Since

$$\left.\frac{dS_2(u, w)}{du}\right|_{u=0} \neq 0,$$

then by Theorem 1.1 from Fedorchuk (1987, p.62) monograph

$$I_2(w) \sim \frac{1}{N}e^{NS_2(0,w)}\left[c_0(w) + \frac{c_1(w)}{N} + \cdots\right], \tag{34}$$

$$J_2(w) \sim \frac{1}{N}e^{NS_2(0,w)}\left[d_0(w) + \frac{d_1(w)}{N} + \cdots\right], \tag{35}$$

where

$$c_0(w) = -\left(\left.\frac{dS_2(u, w)}{du}\right|_{u=0}\right)^{-1} F(0, w) = e^{w^2}\delta^{-1}\sqrt{2\pi}\Phi(w)\Phi(-w),$$

$$d_0(w) = -\left(\left.\frac{dS_2(u, w)}{du}\right|_{u=0}\right)^{-1} = 2e^{w^2}\delta^{-1}\sqrt{2\pi}\Phi(w)\Phi(-w),$$

since $F(0, w) = \frac{1}{2}(\Phi(w) + \Phi(-w)) = \frac{1}{2}$. The expresions of the terms $c_1(w), d_1(w)$ will be not usefull for us and we omit them. It is easy to see that the first derivative of function

$$\big(\Phi(w)\Phi(-w)\big)' = \big(1/\sqrt{2\pi}\big)e^{-w^2/2}(1 - 2\Phi(w))$$

is positive for $w < 0$, negative for $w > 0$ and equals zero when $w = 0$. Therefore,

$$\max_{w\in R}\Phi(w)\Phi(-w) = \Phi^2(0) = \frac{1}{4} \qquad \text{and}$$

$$S_2(0, w) = \ln(\Phi(w)\Phi(-w)) - \lambda_2\ln(1 + w^2) \leqslant S_2(0, 0) = -\ln 4.$$

As

$$\left.\frac{dS_2(0,w)}{dw}\right|_{w=0} = 0 \qquad \text{and} \qquad \left.\frac{d^2S_2(0,w)}{dw^2}\right|_{w=0} \neq 0,$$

then by Theorem 1.3 from Fedorchuk (1987, p.66) we have

$$\int_{-\infty}^{\infty} e^{NS_2(0,w)}c_0(w)dw \sim e^{NS_2(0,0)}\sqrt{\frac{\pi}{N}}\left[c_{00} + \frac{c_{01}}{N} + \cdots\right], \tag{36}$$

$$\int_{-\infty}^{\infty} e^{NS_2(0,w)}d_0(w)dw \sim e^{NS_2(0,0)}\sqrt{\frac{\pi}{N}}\left[d_{00} + \frac{d_{01}}{N} + \cdots\right], \tag{37}$$

where

$$c_{00} = \frac{1}{4\delta}\sqrt{\frac{2\pi}{\lambda_2 + 4/\pi}}, \qquad d_{00} = \frac{1}{2\delta}\sqrt{\frac{2\pi}{\lambda_2 + 4/\pi}}.$$

Inserting (36), (37) into (34), (35) and observing that $S_2(0,0) = 2Z(0)$ we obtain

$$I_2 \sim \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}}\left[h_0 + \frac{h_1}{N} + \cdots\right], \tag{38}$$

$$J_2 \sim \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}}\left[g_0 + \frac{g_1}{N} + \cdots\right], \tag{39}$$

where $h_0 = c_{00}$, $g_0 = d_{00}$. Now (32), (33), (38) and (39) yield that

$$\begin{aligned}
EP_N \sim & \frac{\frac{\pi}{N}\sqrt{\frac{1}{2}}e^{2NZ(u_0)}[a_{00} + \frac{1}{N}(\frac{a_{01}}{2} + a_{10}) + \cdots]}{\frac{\pi}{N}\sqrt{\frac{1}{2}}e^{2NZ(u_0)}H_1 + \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}}H_2} \\
& + \frac{\frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}}[h_0 + \frac{h_1}{N} + \cdots]}{\frac{\pi}{N}\sqrt{\frac{1}{2}}e^{2NZ(u_0)}H_1 + \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}}H_2} \\
= & \frac{a_{00} + \frac{1}{N}(\frac{a_{01}}{2} + a_{10}) + \cdots + \frac{e^{2N(Z(0)-Z(u_0))\sqrt{2}}}{\sqrt{N\pi}}[h_0 + \frac{h_1}{N} + \cdots]}{b_{00} + \frac{1}{N}(\frac{b_{01}}{2} + b_{10}) + \cdots + \frac{e^{2N(Z(0)-Z(u_0))\sqrt{2}}}{\sqrt{N\pi}}[g_0 + \frac{g_1}{N} + \cdots]} \\
\sim & \frac{a_{00} + \frac{1}{N}(\frac{a_{01}}{2} + a_{10}) + \cdots}{b_{00} + \frac{1}{N}(\frac{b_{01}}{2} + b_{10}) + \cdots} \\
\sim & \frac{a_{00}}{b_{00}} + \frac{1}{N}\left(\frac{a_{01}}{2b_{00}} + \frac{a_{10}}{b_{00}} - \frac{a_{00}}{b_{00}}\left(\frac{b_{01}}{2b_{00}} + \frac{b_{10}}{b_{00}}\right)\right) + \cdots
\end{aligned} \tag{40}$$

since $Z(u_0) - Z(0) > \text{const} > 0$ and $e^{2N(Z(0)-Z(u_0))} \sim e^{-\text{const}N}$.
Here

$$H_1 = b_{00} + \frac{1}{N}\left(\frac{b_{01}}{2} + b_{10}\right) + \cdots$$
$$H_2 = g_0 + \frac{g_1}{N} + \cdots$$

From (29)–(31) we obtain

$$\frac{a_{00}}{b_{00}} = \frac{a_0(u_0)}{b_0(u_0)} = F(u_0, 0) = \Phi\left(-\frac{\delta u_0}{2}\right), \tag{41}$$

$$\frac{a_{01}}{2b_{00}} - \frac{a_{00}}{2b_{00}}\frac{b_{01}}{b_{00}} = \frac{\delta e^{-u_0^2\delta^2/8}}{4\sqrt{2}\pi Z''(u_0)}\left(\frac{b_0'(u_0)}{b_0(u_0)} + \frac{u_0\delta^2}{8} - \frac{Z'''(u_0)}{2Z''(u_0)}\right), \tag{42}$$

$$\frac{a_{10}}{b_{00}} - \frac{a_{00}}{b_{00}}\frac{b_{10}}{b_{00}} = \frac{u_0\delta e^{-u_0^2\delta^2/8}}{8\sqrt{2}\pi\beta_1(u_0)}. \tag{43}$$

Inserting into (42), (43) the expressions of the $b_0(u_0)$, $b_0'(u_0)$, $Z''(u_0)$, $Z'''(u_0)$ and $\beta_1(u_0)$ we finally obtain

$$EP_N \sim \Phi\left(-\frac{\delta u_0}{2}\right) + \frac{\delta e^{-u_0^2\delta^2/8}}{8N\sqrt{2}\pi}\left(\frac{u_0}{\beta_1(u_0)} + \frac{u_0\delta^2}{4Z''(u_0)} - \frac{Z'''(u_0)}{(Z''(u_0))^2} - \frac{\beta_1'(u_0)}{Z''(u_0)\beta_1(u_0)}\right), \tag{44}$$

where

$$\beta_1(u_0) = \lambda_1 + m(u_0)m_1(u_0),$$
$$\beta_1'(u_0) = \frac{\delta}{2}m(u_0)(1 - m_1(u_0)(m_1(u_0) + m(u_0))),$$
$$Z''(u_0) = -\frac{\lambda_2(1+u_0^2)}{(1-u_0^2)^2} - \frac{\delta^2}{4}m(u_0)m_1(u_0),$$
$$Z'''(u_0) = -\frac{2u_0\lambda_2(3+u_0^2)}{(1-u_0^2)^3} - \frac{\delta^2}{4}\beta_1'(u_0),$$

$$m(u_0) = \frac{e^{-u_0^2\delta^2/8}}{\sqrt{2}\pi\Phi(u_0\delta/2)}, \quad m_1(u_0) = \frac{u_0\delta}{2} + m(u_0),$$
$$\lambda_1 = \frac{p+1}{2N}, \quad \lambda_2 = \frac{p-3}{2N},$$

Table 1. The values of EP_N as a function of learning sample size N and dimensionality p for $\delta = 1$ ($P_\infty = 0.308538$)

$p = 10$					
N	6	10	20	50	100
Integral	0.418703	0.397796	0.364124	0.332233	0.322517
Formula	0.429074	0.401674	0.367643	0.336833	0.323717
Main term	0.410828	0.385036	0.355329	0.330168	0.319955
Fisher	0.46502	0.42058	0.37967	0.34321	0.32723
$p = 50$					
N	30	50	100	250	500
Integral	0.426664	0.401198	0.368473	0.335025	0.322911
Formula	0.428475	0.40247	0.369079	0.337830	0.324307
Main term	0.425718	0.399755	0.366908	0.336580	0.323581
Fisher	0.46186	0.41944	0.37972	0.34405	0.32802
$p = 200$					
N	120	200	400	1000	2000
Integral	0.426589	0.402039	0.369032	0.337458	0.324232
Formula	0.428667	0.402800	0.369429	0.338043	0.324426
Main term	0.428011	0.402144	0.368898	0.337734	0.324246
Fisher	0.460943	0.419128	0.379731	0.344250	0.328204

and u_0 is found by solving Eq. (26).

For u_0 we propose the following analytic formula :

$$u_0 \approx \left(1 + 2\lambda_2\delta^{-1}(B + \sqrt{B^2 + \lambda_2^2/4})^{-1}\right)^{-1/2}$$

where

$$B = \frac{e^{-\delta^2/8}}{\sqrt{2\pi}\Phi(\delta/2)} + \frac{\lambda_2}{2}\left(\frac{\delta}{4} - \frac{1}{\delta}\right).$$

Numerical calculations show this formula is usefull for small $\delta(\delta = 1)$ or when $p/N \to 0$. In other cases it works bad.

5. Numerical results. Let us now compare the zero empirical error and Fisher's linear classifiers (this comparison was done by Dičiūnas (1996)). Looking at formula (44) of expected error EP_N for the zero empirical error classifier we see that for $N \to \infty$ the main contribution to sum is determined by the first

Table 2. The values of EP_N as a function of learning sample size N and dimensionality p for $\delta = 4$ ($P_\infty = 0.022850$)

$p = 10$					
N	6	10	20	50	100
Integral	0.155296	0.117827	0.081907	0.052822	0.040051
Formula	0.130178	0.101495	0.073414	0.049963	0.039132
Main term	0.129269	0.098716	0.069374	0.046084	0.036065
Fisher	0.26349	0.099509	0.047685	0.030520	0.026336
$p = 50$					
N	30	50	100	250	500
Integral	0.154316	0.118291	0.082491	0.053347	0.039701
Formula	0.151101	0.115802	0.080259	0.052741	0.040262
Main term	0.150904	0.115312	0.080351	0.052076	0.039723
Fisher	0.23684	0.093015	0.046723	0.030479	0.026356
$p = 200$					
N	120	200	400	1000	2000
Integral	0.154087	0.118378	0.082721	0.053379	0.040376
Formula	0.154667	0.118316	0.082450	0.053306	0.040511
Main term	0.154614	0.118201	0.082286	0.053143	0.040379
Fisher	0.228807	0.091211	0.046446	0.030459	0.026355

term. Therefore, for large N we obtain

$$EP_N \approx \Phi\left(-\frac{\delta u_0}{2}\right). \tag{45}$$

We will call (45) the *main term.*

Pikelis (1976) gives the Table of exact values of the expected error EP_N for the Fisher linear DF with different values of parameters p, N and δ. We carried out numerical calculation of EP_N for the zero empirical error classifier with the same values of parameters as in Pikelis (1976). Moreover, we used three different formulae for EP_N:

1) numerically calculated *integral* (15),

2) asymptotic *formula* (44) and

3) main term (45).

Numerical results are presented in Tables 1 and 2.

Inspection of these tables leads to the following two conclusions:

1. In almost all observed cases both formulae, (44) and even (45), are very accurate (matches with (15)). Only for very small values of p and N ($p \leqslant 10,\ N \leqslant 20$), integral (15) is more preferable.
2. The Fisher's classifier outperforms the zero empirical error classifier for a big distance between the classes ($\delta = 4$) when $N \geqslant p$, while for a small distance ($\delta = 1$) and for cases when $N < p$ the zero empirical error classifier is preferable. It means, the linear zero empirical error classifier can be used in cases when the number of dimensions is higher than number of learning samples.

REFERENCES

Aivazyan, S. A., V.M. Buchstaber, I.S. Yenyukov, and L.D. Meshalkin (1989). *Applied Statistics.* Finansy i statistika (in Russian).

Deev, A. D. (1970). Representation of statistics of discriminant analysis and asymptotic expansions in dimensionalities comparable with sample size. *Reports of Academy of Sciences of the USSR,* **195**(4), 756–762 (in Russian).

Deev, A. D. (1972). Asymptotic expansions of statistic distributions of discriminant analysis W, M, W'. In *Stat. Metody Klassif.* MGU, Moscow, 6–51 (in Russian).

Dičiūnas, V. (1996). Numerical comparisation of linear zero empirical error classifier. (unpublished).

Fedorchuk, M. V. (1987). *Asymptotics, Integrals and Series.* Nauka, Moscow.

John, S. (1961). Errors in discrimination. *Annals of Mathematical Statistics,* **32**, 1125–1144.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Annals of Mathematical Statistics,* **34**, 1286–1301.

Pikelis, V. (1976). Comparison of methods of computing the expected classification errors. *Automation and Remote Control,* **5**, 59–63 (in Russian, English translation).

Raudys, Š. (1967). On determining training sample size of linear classifier. *Computing systems, Proc. of Institute of Mathematics, Academy of Sciences of USSR,* **28**, 79–87 (in Russian).

Raudys, Š. (1972). On the amount of prior information in designing the classification algorithm. *Proceedings of Academy of Sciences of the USSR. Technical Cybernetics,* **4**, 168–174 (in Russian).

Raudys, Š. (1991). Methods of overcome dimensionality problems in statistical pattern recognition. A review. *Zavodskaja Laboratorija,* **3**. Nauka, Moscow (in Russian).

Raudys, Š. (1993). On shape of pattern error function, initializations and intrinsic dimensionality in ANN classifier design. *Informatica,* **4**(3–4), 360–383.

Raudys, Š., and A.K. Jain (1991). Small sample size effects in statistical pattern recognition. Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **13**, 252–264.

Raudys, Š., and V. Pikelis (1980). On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **2**(3), 242–252.

Wyman, F., D. Young, and D. Turner (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition,* **23**(7), 775–783 .

Received April 1996

A. Basalykas is a Doctor of Mathematical Sciences, a senior researcher of the Probability Theory Department at the Institute of Mathematics and Informatics. His scientific interests include asymptotic analysis of nonlinear statistics and classification.

NULINĖS EMPIRINĖS KLAIDOS TIESINIO KLASIFIKATORIAUS PIRMOS RŪŠIES KLAIDOS TIKIMYBĖ

Alfredas BASALYKAS

Šiame straipsnyje pateikiama tiksli analizinė nulinės empirinės klaidos tiesinio klasifikatoriaus pirmos rūšies klaidos tikimybės išraiška. Asimptotinis šios klaidos skleidinys gautas tuo atveju kai abi duomenų klasės turi sferinį Gauso pasiskirstymą ir kai mokymo imties tūris N ir matavimų skaičius p auga į begalybę taip, kad $p/N \to \text{const}$. Pateikiamos lentelės kur lyginamos tikslios ir apytikslės (asimptotinio skleidinio pagrindinio nario) šios klaidos reikšmės priklausomai nuo p, N ir atstumo tarp klasių δ. Be to, šio klasifikatoriaus klaidos tikimybė yra palyginama su Fišerio tiesiniu klasifikatoriumi ir nustatoma, kad mūsų nagrinėjamas klasifikatorius gali būti sėkmingai naudojamas ir tada kai matavimų skaičius viršija mokymo imties tūrį.