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 $\delta$ 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 $X = (x_1, x_2, \ldots, x_p)'$ be $p \times 1$ observation vector of an individual from one or another of two $p$-variate classes (populations) $\pi_1$ and $\pi_2$. Suppose we have two learning sets, of sizes $N_1$ and $N_2$, of $p$-dimensional data from populations $\pi_1$ and $\pi_2$ respectively. The problem is to utilize an information contained in the learning sets and to design a classification rule assigning vector $X$ to one of the classes.

In statistic-theoretical approach it is assumed vector $X$ to be random one with class conditional probability density function $f_i(X|\pi_i)$. Let $q_i$ be a priori probability of class $\pi_i\ (q_1 + q_2 = 1)$. Then an optimal (Bayes) classification
The expected probability of misclassification rule (which minimizes the probability of incorrect classification) will assign vector \( X \) to one of classes \( \pi_i \) according a sign of the following discriminant function (DF):

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

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

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

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 approach 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 \( f_i(X|\pi_i) \)) 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 \( E P_N \) over all possible random learning sets of size \( N_1 \) and \( N_2 \) is called an expected PMC. A theoretical limit

\[
\lim_{N_1 \to \infty, N_2 \to \infty} E P_N = P_\infty
\]

is called an asymptotic PMC.

The expected PMC was studied in a number of research papers beginning from pioneering work of John (1961) who obtained first exact and approximate formulae for the standard linear DF for the Gaussian classes for case when \( \Sigma \), the covariance matrix, is known. Best known asymptotic expansion for case when \( \Sigma \) is known due to Okamoto (1963). Principal results 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 referred 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 \( f(X|\pi_i) \) are multivariate spherically Gaussian.

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

\[
g(X) = \Lambda^T X + a = \sum_{i=1}^{p} a_i x_i + a.
\]

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 \( f_{\text{prior}}(a, \Lambda) \) of weight vector \( (a, \Lambda) \) one generates a set of random weights \( a, a_1, a_2, \ldots, a_p \). We will say that training is successful if conditions \( S \) are satisfied, where

\[
S : \begin{cases} 
\text{for all training pattern vectors from } \pi_1: & g(X|a, \Lambda) > 0, \\
\text{for all training pattern vectors from } \pi_2: & g(X|a, \Lambda) \leq 0.
\end{cases}
\]

We shall compute the expected PMC \( E_{\text{PN}} \) of successfully trained linear discriminant function.

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

- two multivariate spherically Gaussian classes \( \pi_1, \pi_2 \) with densities \( N(X, C_1, I) \) and \( N(X, C_2, I) \) 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 \( X^{(1)}_1, X^{(1)}_2, \ldots, X^{(1)}_N, X^{(2)}_1, X^{(2)}_2, \ldots, X^{(2)}_N \) are statistically independent and identically distributed in their own classes;
- we assume that \( C_1 + C_2 = 0 \) (this assumption enables us further to simplify the calculations);
- the components of the vector \( (a, \Lambda) \) are chosen random from Gaussian distribution with zero mean and variance 1: \( a_i \sim N(0, 1) \).
We shall analyze a limit case when $N \to \infty$, $p \to \infty$ and $p/N \to \text{const.}$

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 $(a, A)$, 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 $(a, A)$ (Raudys, 1993):

$$f_{\text{post}}(a, A | S) = \frac{\Pr(S = \text{true} | a, A) f_{\text{prior}}(a, A)}{\Pr(S = \text{true})} = \frac{\Pr(S = \text{true} | a, A) f_{\text{prior}}(a, A)}{\iiint \Pr(S = \text{true} | a, A) f_{\text{prior}}(a, A) \, da \, dA},$$  \hspace{1cm} (4)

$$EP_N = \Pr(MC | S) = \iiint \Pr(MC | a, A) f_{\text{post}}(a, A | S) \, da \, dA,$$  \hspace{1cm} (5)

where $\Pr(MC | a, A)$ is a conditional probability of misclassification given the set of weights $(a, A)$ and $f_{\text{post}}(a, A | S = \text{true})$ 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 $g(X)$ will be Gaussian and the conditional probability of misclassification

$$\Pr(MC | A, a) = \frac{1}{2} \Pr(A'X + a \leq 0 \mid X \in \pi_1) + \frac{1}{2} \Pr(A'X + a > 0 \mid X \in \pi_2)$$
$$= \frac{1}{2} \Phi \left( \frac{-A'C_1 + a}{\sqrt{A'A}} \right) + \frac{1}{2} \Phi \left( \frac{A'C_2 + a}{\sqrt{A'A}} \right),$$  \hspace{1cm} (6)

where

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

Conditional PMC (6) depends on $(p+1)$-variate vector $(a, A)'$. For spherical case we can show this PMC depends only on two independent scalar variables.
Let us perform a transformation

\[ V = TA = \begin{bmatrix}
    v_1 \\
    v_2 \\
    \vdots \\
    v_p
\end{bmatrix} \quad \text{and} \quad T(C_1 - C_2) = \begin{bmatrix}
    \delta \\
    l_2 \\
    \vdots \\
    l_p
\end{bmatrix}, \]

where \( T \) is \( p \times p \) orthogonal matrix with a first row vector

\[ t_1 = \frac{(C_1 - C_2)}{(C_1 - C_2)'(C_1 - C_2)}, \quad (7) \]

and \( \delta^2 = (C_1 - C_2)'(C_1 - C_2) \) is a squared Mahalanobis distance. Then

\[
\frac{\Lambda'C_1 + a}{\sqrt{\Lambda'A}} = \frac{(TA)'(T(C_1 - C_2) + T(C_1 + C_2)) + 2a}{2\sqrt{(TA)'(TA)}} = \frac{v_1\delta + w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p}v_i^2}} = u\frac{\delta}{2} + w, \quad (8)
\]

where

\[ w_0 = (TA)'(T(C_1 + C_2)) + 2a = \Lambda'(C_1 + C_2) + 2a = 2a, \quad (9) \]

as \( C_1 + C_2 = 0 \) by our assumption, and

\[ u = \frac{v_1}{\sqrt{v_1^2 + \sum_{i=2}^{p}v_i^2}}, \quad w = \frac{w_0}{2\sqrt{v_1^2 + \sum_{i=2}^{p}v_i^2}}. \]

Analogously

\[ \frac{\Lambda'C_2 + a}{\sqrt{\Lambda'A}} = -u\frac{\delta}{2} + w. \quad (10) \]

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

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

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

Student \( S_t(p) \) distributions accordingly. Their density functions \( h(u) \) and \( p(w) \) accordingly are

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

and

\[
p(w) = K_2(1 + w^2)^{-(p+1)/2}, \quad w \in \mathbb{R}.
\]  
(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

\[
\Pr(S = \text{true} \mid a, A) = \prod_{j=1}^{N} \Pr\{A'X_j^{(1)} + a > 0\} \prod_{j=1}^{N} \Pr\{A'X_j^{(2)} + a \leq 0\}
\]

\[
= \left[ \Pr\{A'X + a > 0 \mid X \in \pi_1\} \right]^N \left[ \Pr\{A'X + a \leq 0 \mid X \in \pi_2\} \right]^N
\]

\[
= \left[ 1 - \Pr\{A'X + a \leq 0 \mid X \in \pi_1\} \right]^N \times \left[ 1 - \Pr\{A'X + a > 0 \mid X \in \pi_2\} \right]^N
\]

\[
= \left[ 1 - \Phi \left( -\frac{A'C_1 + a}{\sqrt{A'A}} \right) \right]^N \times \left[ 1 - \Phi \left( \frac{A'C_2 + a}{\sqrt{A'A}} \right) \right]^N
\]

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

\[
\Pr(S = \text{true} \mid a, A) = \Pr(S = \text{true} \mid u, w)
\]

\[
= \left[ 1 - \Phi(-u\delta/2 - w) \right]^N \left[ 1 - \Phi(-u\delta/2 + w) \right]^N
\]

\[
= \left[ \Phi(u\delta/2 + w) \Phi(u\delta/2 - w) \right]^N.
\]  
(14)

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

\[
EP_N = \frac{I_1 + I_2}{J_1 + J_2},
\]  
(15)

where

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

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

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

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

\[ F(u, w) = \frac{1}{2} (\Phi(-u \delta/2 + w) + \Phi(-u \delta/2 - w)). \]

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 \( \lambda_1 = \frac{p + 1}{2N} \),

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

where \( \lambda_2 = \frac{p - 3}{2N} \).

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

\[ I_1 = \int_{0}^{1} I_1(u)(1 - u^2)^{-N \lambda_2} du, \quad J_1 = \int_{0}^{1} J_1(u)(1 - u^2)^{-N \lambda_2} du, \]

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

where

\[ I_1(u) = \int_{-\infty}^{\infty} F(u, w)e^{NS_1(u, w)} dw, \quad J_1(u) = \int_{-\infty}^{\infty} e^{NS_1(u, w)} dw, \]

\[ I_2(w) = \int_{0}^{1} F(-u, w)e^{NS_2(u, w)} du, \quad J_2(w) = \int_{0}^{1} e^{NS_2(u, w)} du. \]

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 \( N \) increasing to infinity
The expected probability of misclassification

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 $w \in \mathbb{R}$. As

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

then

$$S_1(u, w) = \ln \left( \Phi(u \delta/2 + w)\Phi(u \delta/2 - w)(1 + w^2)^{-1/2} \right)$$

$$\leq \ln \left( \Phi(u \delta/2 - w)(2\Phi(u \delta/2) - \Phi(u \delta/2 - w)) \right)$$

$$= \ln \left( \Phi^2(u \delta/2) - (\Phi(u \delta/2 - w) - \Phi(u \delta/2))^2 \right)$$

$$\leq \ln \Phi^2(u \delta/2) = S_1(u, 0). \quad (21)$$

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 $w \in \mathbb{R}$.

Moreover,

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

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

$$I_1(u) \sim e^{N S_1(u, 0)} \sqrt{\frac{\pi}{N}} \left( a_0(u) + \frac{a_1(u)}{N} + \cdots \right), \quad (22)$$

$$J_1(u) \sim e^{N S_1(u, 0)} \sqrt{\frac{\pi}{N}} \left( b_0(u) + \frac{b_1(u)}{N} + \cdots \right), \quad (23)$$

where

$$a_0(u) = F(u, 0) \beta_1^{-1/2}(u), \quad b_0(u) = \beta_1^{-1/2}(u),$$

$$a_1(u) = -\frac{1}{4} \left( \beta_2(u) \Phi(-u \delta/2) \beta_1^{-5/2}(u) + \frac{1}{3} \beta_1^{-3/2}(u) \Phi'(x)|_{x=u \delta/2} \right)$$

and

$$b_1(u) = -\frac{1}{4} \beta_2(u) \beta_1^{-5/2}(u).$$

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

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

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

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

Let us denote

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

Then

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

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

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 derivatives of $Z(u)$:

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

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

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 its 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

$$\frac{\delta}{\sqrt{2\pi}} e^{-u^2 j/8} (1 - u^2) = 2\lambda_2 u \Phi \left( \frac{u^j}{2} \right). \quad (26)$$
The expected probability of misclassification

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

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

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

where

\[
a_{i0} = a_i(u_0) \sqrt{-\frac{2}{Z''(u_0)}}, \quad b_{i0} = b_i(u_0) \sqrt{-\frac{2}{Z''(u_0)}}, \quad (29)
\]

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

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

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

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

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

Analogously obtain

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

\[
= \frac{\pi}{N} \sqrt{\frac{1}{2}} e^{2Nz(u_0)} \left[ b_{00} + \frac{1}{N} \left( b_{01} + b_{10} \right) + \cdots \right]. \quad (33)
\]
Let us now deal with integrals $I_2(w)$, $J_2(w)$:

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

It is easy to see that

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

Since

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

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

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

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

where

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

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

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

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

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

$$\max_{w \in \mathbb{R}} \Phi(w)\Phi(-w) = \Phi^2(0) = \frac{1}{4} \quad \text{and} \quad S_2(0, w) = \ln(\Phi(w)\Phi(-w)) - \lambda_2 \ln(1 + w^2) \leq S_2(0, 0) = -\ln 4.$$
As
\[
\frac{dS_2(0, w)}{dw} \bigg|_{w=0} = 0 \quad \text{and} \quad \frac{d^2S_2(0, w)}{dw^2} \bigg|_{w=0} \neq 0,
\]
then by Theorem 1.3 from Fedorchuk (1987, p.66) we have
\[
\int_{-\infty}^{\infty} e^{NS_2(0, w)} c_0(w) dw \sim e^{NS_2(0, 0)} \sqrt{\frac{\pi}{N}} \left[ c_{00} + \frac{c_{01}}{N} + \cdots \right], \quad (36)
\]
\[
\int_{-\infty}^{\infty} e^{NS_2(0, w)} d_0(w) dw \sim e^{NS_2(0, 0)} \sqrt{\frac{\pi}{N}} \left[ d_{00} + \frac{d_{01}}{N} + \cdots \right], \quad (37)
\]
where
\[
c_{00} = \frac{1}{4\delta} \sqrt{\frac{2\pi}{\lambda_2 + 4/\pi}}, \quad d_{00} = \frac{1}{2\delta} \sqrt{\frac{2\pi}{\lambda_2 + 4/\pi}}.
\]
Inserting (36), (37) into (34), (35) and observing that \( S_2(0, 0) = 2Z(0) \) we obtain
\[
I_2 \sim \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}} \left[ h_0 + \frac{h_1}{N} + \cdots \right], \quad (38)
\]
\[
J_2 \sim \frac{e^{2NZ(0)}\sqrt{\pi}}{N\sqrt{N}} \left[ g_0 + \frac{g_1}{N} + \cdots \right], \quad (39)
\]
where \( h_0 = c_{00}, \ g_0 = d_{00}. \) Now (32), (33), (38) and (39) yield that
\[
EP_N \sim \frac{\pi}{N} \sqrt{\frac{1}{2} e^{2NZ(u_0)} \left[ a_{00} + \frac{1}{N} \left( \frac{a_{01}}{2} + a_{10} \right) + \cdots \right]} \]
\[
+ \frac{\pi}{N} \sqrt{\frac{1}{2} e^{2NZ(u_2)} H_1 + \frac{e^{2NZ(u_0)}\sqrt{\pi}}{N\sqrt{N}} H_2} \]
\[
= \frac{a_{00} + \frac{1}{N} \left( \frac{a_{01}}{2} + a_{10} \right) + \cdots + \frac{e^{2NZ(0) - Z(u_0)}\sqrt{\pi}}{\sqrt{N}} [h_0 + \frac{h_1}{N} + \cdots] \]
\[
+ \frac{1}{b_{00} + \frac{1}{N} \left( \frac{b_{01}}{2} + b_{10} \right) + \cdots + \frac{e^{2NZ(0) - Z(u_0)}\sqrt{\pi}}{\sqrt{N}} [g_0 + \frac{g_1}{N} + \cdots] \]
\[
\sim \frac{a_{00}}{b_{00}} + \frac{1}{N} \left( \frac{a_{01}}{2b_{00}} + a_{10} \frac{b_{01}}{b_{00}} + \frac{a_{00} \left( b_{01} + b_{10} \right)}{b_{00} \left( 2b_{00} + b_{00} \right)} \right) + \cdots \quad (40)
\]
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

\[
\frac{a_{00}}{b_{00}} = \frac{a_{00}(u_0)}{b_{00}(u_0)} = F(u_0, 0) = \Phi \left( -\frac{\delta u_0}{2} \right), \tag{41}
\]
\[
\frac{a_{01}}{2b_{00}} - \frac{a_{00}}{2b_{00}} = \frac{\delta e^{-u_0^2 \delta^2/8}}{4\sqrt{2\pi} Z''(u_0)} \left( \frac{b_0(u_0)}{b_0(u_0)} + \frac{u_0 \delta^2}{8} - \frac{Z'''(u_0)}{2Z''(u_0)} \right), \tag{42}
\]
\[
\frac{a_{10}}{b_{00}} - \frac{a_{00}}{b_{00}} = \frac{u_0 \delta e^{-u_0^2 \delta^2/8}}{8\sqrt{2\pi} \beta_1(u_0)}. \tag{43}
\]

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

\[
\begin{align*}
EP_N & \sim \Phi \left( -\frac{\delta u_0}{2} \right) + \frac{\delta e^{-u_0^2 \delta^2/8}}{8N\sqrt{2\pi}} \left( \frac{u_0}{\beta_1(u_0)} + \frac{u_0 \delta^2}{4Z''(u_0)} \right) \\
& \quad - \left( \frac{Z'''(u_0)}{(Z''(u_0))^3} - \frac{\beta'_1(u_0)}{Z''(u_0)\beta_1(u_0)} \right), \tag{44}
\end{align*}
\]

where

\[
\beta_1(u_0) = \lambda_1 + m(u_0)m_1(u_0),
\]
\[
\beta'_1(u_0) = \frac{\delta}{2} m(u_0)(1 - m_1(u_0)(m_1(u_0) + m(u_0))),
\]
\[
Z''(u_0) = -\frac{\lambda_2(1 + u_0^2)}{1 - u_0^2} - \frac{\delta^2}{4} m(u_0)m_1(u_0),
\]
\[
Z'''(u_0) = -\frac{2u_0 \lambda_2(2 + u_0^2)}{(1 - u_0^2)^3} - \frac{\delta^2}{4} \beta'_1(u_0),
\]
\[
m(u_0) = \frac{e^{-u_0^2 \delta^2/8}}{\sqrt{2\pi} \Phi(u_0 \delta/2)}, \quad m_1(u_0) = \frac{u_0 \delta}{2} + m(u_0),
\]
\[
\lambda_1 = \frac{p + 1}{2N}, \quad \lambda_2 = \frac{p - 3}{2N},
\]
The expected probability of misclassification

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

<table>
<thead>
<tr>
<th>$N$</th>
<th>6</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>0.418703</td>
<td>0.397796</td>
<td>0.364124</td>
<td>0.332233</td>
<td>0.322517</td>
</tr>
<tr>
<td>Formula</td>
<td>0.429074</td>
<td>0.401674</td>
<td>0.367643</td>
<td>0.336833</td>
<td>0.323717</td>
</tr>
<tr>
<td>Main term</td>
<td>0.410828</td>
<td>0.385036</td>
<td>0.355329</td>
<td>0.330168</td>
<td>0.319955</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.46502</td>
<td>0.42058</td>
<td>0.37967</td>
<td>0.34321</td>
<td>0.32723</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>30</th>
<th>50</th>
<th>100</th>
<th>250</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>0.426664</td>
<td>0.401198</td>
<td>0.368473</td>
<td>0.335025</td>
<td>0.322911</td>
</tr>
<tr>
<td>Formula</td>
<td>0.428475</td>
<td>0.40247</td>
<td>0.369079</td>
<td>0.337830</td>
<td>0.324307</td>
</tr>
<tr>
<td>Main term</td>
<td>0.425718</td>
<td>0.399755</td>
<td>0.366908</td>
<td>0.336580</td>
<td>0.323581</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.46186</td>
<td>0.41944</td>
<td>0.37972</td>
<td>0.34405</td>
<td>0.32802</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>120</th>
<th>200</th>
<th>400</th>
<th>1000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integral</td>
<td>0.426589</td>
<td>0.402039</td>
<td>0.369032</td>
<td>0.337458</td>
<td>0.324232</td>
</tr>
<tr>
<td>Formula</td>
<td>0.428667</td>
<td>0.402800</td>
<td>0.369429</td>
<td>0.338043</td>
<td>0.324426</td>
</tr>
<tr>
<td>Main term</td>
<td>0.428011</td>
<td>0.402144</td>
<td>0.368898</td>
<td>0.337734</td>
<td>0.324246</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.460943</td>
<td>0.419128</td>
<td>0.379731</td>
<td>0.344250</td>
<td>0.328204</td>
</tr>
</tbody>
</table>

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 useful 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 Diciunas (1996)). Looking at formula (44) of expected error $E_{PN}$ 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 $E_{PN}$ as a function of learning sample size $N$ and
dimensionality $p$ for $\delta = 4$ ($P_{\infty} = 0.022850$)

\begin{center}
\begin{tabular}{|c|c|c|c|c|c|}
\hline
\textbf{$p = 10$} &  &  &  &  \\
\hline
\textbf{$N$} & 6 & 10 & 20 & 50 & 100 \\
\hline
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 \\
\hline
\textbf{$N$} & 30 & 50 & 100 & 250 & 500 \\
\hline
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 \\
\hline
\textbf{$N$} & 120 & 200 & 400 & 1000 & 2000 \\
\hline
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 \\
\hline
\end{tabular}
\end{center}

term. Therefore, for large $N$ we obtain

$$E_{PN} \approx \Phi \left( -\frac{\delta u_0}{2} \right).$$

(45)

We will call (45) the main term.

Pikelis (1976) gives the Table of exact values of the expected error $E_{PN}$
for the Fisher linear DF with different values of parameters $p$, $N$ and $\delta$. We
carried out numerical calculation of $E_{PN}$ for the zero empirical error classifier
with the same values of parameters as in Pikelis (1976). Moreover, we used
three different formulæ for $E_{PN}$:

1) numerically calculated integral (15),
2) asymptotic formula (44) and
3) main term (45).

Numerical results are presented in Tables 1 and 2.
The expected probability of misclassification

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 \( N \) \((p \ll 10, \ N \ll 20)\), 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 \( N \gg p \), 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


Received April 1996
The expected probability of misclassification

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.