Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition 38 (2005) 1197 – 1207

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

A one-dimensional analysis for the probability of error of linear classifiers for normally distributed classes

Luis Rueda*

School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, Ont., N9B 3P4, Canada

Received 7 October 2004; accepted 7 December 2004

Abstract

Computing the probability of error is an important problem in evaluating classifiers. When dealing with normally distributed classes, this problem becomes intricate due to the fact that there is no closed-form expression for integrating the probability density function. In this paper, we derive lower and upper bounds for the probability of error for a linear classifier, where the random vectors representing the underlying classes obey the *multivariate* normal distribution. The expression of the error is derived in the *one-dimensional space*, *independently of the dimensionality of the original problem*. Based on the two bounds, we propose an approximating expression for the error of a generic linear classifier. In particular, we derive the corresponding bounds and the expression for approximating the error of Fisher's classifier. Our empirical results on synthetic data, including up to two-hundred-dimensional featured samples, show that the computations for the error are extremely fast and quite accurate; it differs from the *actual* error in at most $\varepsilon = 0.0184340683$. The scheme has also been successfully tested on real-life data sets drawn from the UCI machine learning repository.

Keywords: Linear discriminant analysis; Fisher's classifier; Error rate evaluation; Gaussian distributions; Curse of dimensionality

1. Introduction

Assessing the performance of classifiers is a fundamental problem is pattern recognition, for which various approaches have been proposed in the literature. The main idea is to measure the discriminability of the classifier by means of its *misclassification rate* or *error rate*. The error rate or classification error, in general, measured as the *probability of error*, provides a quite useful insight about the quality of a classifier. We consider the classical problem of deriving the *true* error rate for a linear classifier, which we presently refer to as the *classification error* or *probability of error*. We deal with two classes, ω_1 and ω_2, whose a priori probabilities are $P(\omega_1)$ and $P(\omega_2)$, respectively, and which are represented by two normally distributed *d*-dimensional random vectors,[1] $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively.

A more realistic scenario involves two data sets containing labeled samples, $\mathscr{D}_1 = \{\mathbf{x}_{1_1}, \mathbf{x}_{1_2}, \ldots, \mathbf{x}_{1_{n_1}}\}$ and $\mathscr{D}_2 = \{\mathbf{x}_{2_1}, \mathbf{x}_{2_2}, \ldots, \mathbf{x}_{2_{n_2}}\}$, where $\mathbf{x}_{1_j}$ and $\mathbf{x}_{2_j}$ are drawn independently from their respective classes, ω_1 and ω_2, respectively. In order to derive a linear classification scheme, the aim is to find a linear function of the form:

$$g(\mathbf{x}) = \mathbf{w}^{\mathrm{t}}\mathbf{x} + w_0 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0, \tag{1}$$

* Tel.: +1 519 253 3000x3780; fax: +1 519 973 7093.
E-mail address: lrueda@uwindsor.ca.

[1] In this paper, we use the notation $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to refer to a normally distributed random vector, $\mathbf{x}$, where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix.

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2004.12.002

that classifies an unknown object, represented by a real-valued feature vector, $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{t}}$, into the respective class, where $\mathbf{w}$ is a d-dimensional *weight* vector, such that[2] $\mathbf{w} \neq \mathbf{0}_d$, and w_0 is a *threshold* weight. We assume that the underlying classes are normally distributed, and that the parameters of the distribution are also known, i.e. they are estimated by the *maximum likelihood estimate* (MLE) or the *Bayesian approach*, for example. The linear classifier can be derived from the respective data sets, or once the parameters have been estimated, the linear classifier can be obtained from these parameters. In our notation, a vector $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{t}}$ is composed of d real-valued variables $x_1, \ldots, x_d$, which also refer to the axes of the system in $\mathbb{R}^d$, in that order. Similarly, we use the same notation for a d-dimensional random vector, $\mathbf{x} = [x_1, \ldots, x_d]^{\mathrm{t}}$, where $x_1, \ldots, x_d$ are random variables. It is clear that the assumption of normally distributed classes is restricted to particular cases in real-life scenarios. However, it is important to highlight that this assumption has strong theoretical justifications (see Refs. [1,2]). Also, normal distributions are typically used in unsupervised learning schemes, in which the distribution and the number of classes is unknown.

The problem of estimating the classification error has been studied for various cases, including an asymptotic formula for the expected error of the pseudo-Fisher classifier for the case in which the dimensionality of the problem is relatively larger than the size of the training data set [3]. Bounds for more generic scenarios have been derived for linear classifiers that use kernel classifiers [4,5]. In the case of the Bayesian (quadratic) classifier for normal distributions, it is well-known that bounds on the classification error exist, namely Chernoff's and Battacharyya's bounds [1,2,6], and the approximation method introduced by Lee and Choi [7]. These bounds are applicable to the *optimal* classifier, which is linear only for equal covariance matrices, and are not tight enough for the majority of the cases.

Since we are dealing with normally distributed classes, an algebraic analysis of the probability of error is not possible as it involves integrating the normal distribution probability density function, which has no closed-form algebraic expression. Although bounds on the probability of error exist, finding the exact (or eventually an approximate) value for integrating the normal distribution probability density function still remains a well-known, open problem, which has many applications in statistics, engineering, and computer science. In this paper, we derive lower and upper bounds for the classification error. These bounds are obtained from a *one-dimensional* algebraic expression for the probability of error of a linear classifier, where the random vectors representing the underlying classes obey the *multivariate* normal distribution. Using these bounds, an approximating expression is derived, which has been proved to be quite accurate in estimating the *actual* probability of error, differing from the latter in at most $\varepsilon = 0.0184340683$. By instantiating the generic case to a specific scenario, we derive the bounds and approximation for the well-known Fisher's classifier.

2. Bounds and approximations for the error

We assume that we are dealing with the case in which $\boldsymbol{\mu}_1$ is on the "negative side" of the classifier, i.e. $g(\boldsymbol{\mu}_1) < 0$, which implies that $g(\boldsymbol{\mu}_2) > 0$. It is thus, easy to see that to evaluate the opposite case, i.e. when $g(\boldsymbol{\mu}_1) > 0$ and $g(\boldsymbol{\mu}_2) < 0$, it suffices to rename the classes in such a way that the new class ω_1 satisfies $g(\boldsymbol{\mu}_1) < 0$. Additionally, although this approach is valid for a fairly efficient linear classifier that at least separates the means, a similar analysis can be done to consider even the case of a *quite inefficient* classifier that *does not* separate the means.

Given a linear classifier of the form of Eq. (1), the intent of the exercise is to compute (or eventually estimate) the classification error. The probability of error measures the likelihood that an unknown sample, $\mathbf{x}$, which belongs to ω_1, is assigned to ω_2, or $\mathbf{x}$ is assigned to ω_1 when it belongs to ω_2. Thus, if the space is divided into two regions, $\mathscr{R}_1$ and $\mathscr{R}_2$, which represent the areas in which an object is assigned to ω_1 and ω_2, respectively, the probability of error, Pr[error], is calculated as follows [1]:

$$\Pr[\text{error}] = \int_{\mathscr{R}_2} p_{x_1}(\mathbf{x}|\omega_1) P(\omega_1)\,\mathrm{d}\mathbf{x} + \int_{\mathscr{R}_1} p_{x_2}(\mathbf{x}|\omega_2) P(\omega_2)\,\mathrm{d}\mathbf{x}, \tag{2}$$

where $p_{x_i}(\mathbf{x}|\omega_i)$ is the probability of $\mathbf{x}$ given ω_i, $\mathscr{R}_1$ is the region determined by $g(\mathbf{x}) < 0$, and $\mathscr{R}_2$ is the region determined by $g(\mathbf{x}) > 0$. Note that $g(\mathbf{x}) = 0$ is excluded from (2) since $\int_{g(\mathbf{x})=0} p_{x_i}(\mathbf{x}|\omega_i) P(\omega_i) = 0$.

The expression given in Eq. (2) is quite involved due to the fact that it invokes integrating multivariate normal distributions. Fortunately, when dealing with normal distributions, simpler expressions that involve $N(0,1)$ random variables can be used (see Ref. [8]). An elegant way of writing the integrals of Eq. (2) in terms of a $N(0,1)$ random variable, x, is as follows [2]:

$$\Pr[\text{error}] = P(\omega_1) \int_{-\infty}^{a_1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x + P(\omega_2) \int_{-\infty}^{a_2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x, \tag{3}$$

where x is a $N(0,1)$ random variable, and

$$a_1 = (\mathbf{w}^{\mathrm{t}} \boldsymbol{\Sigma}_1 \mathbf{w})^{-1/2} (w_0 + \mathbf{w}^{\mathrm{t}} \boldsymbol{\mu}_1) \tag{4}$$

and

$$a_2 = -(\mathbf{w}^{\mathrm{t}} \boldsymbol{\Sigma}_2 \mathbf{w})^{-1/2} (w_0 + \mathbf{w}^{\mathrm{t}} \boldsymbol{\mu}_2). \tag{5}$$

[2] Note that due to the nature of the problem, $\mathbf{w}$ is a non-null vector, otherwise the hyperplane classifier would not have a defined orientation. To state this assumption, we use the notation $\mathbf{w} \neq \mathbf{0}_d$, where $\mathbf{0}_d$ is the null vector in the Euclidean d-dimensional space.

Although the multi-dimensional problem is reduced to its equivalent in the one-dimensional space, the analytical form of the integral for the univariate normal distribution density function is still not possible, and thus an algebraic analysis is only possible by means of bounds. To derive the expressions for the lower and upper bounds, we resort on the following inequality for the cumulative normal distribution function [9]. Let x be a $N(0,1)$ random variable. Then, for all $a > 0$:

$$\frac{2\mathrm{e}^{-a^2/2}}{(a+\sqrt{a^2+4})\sqrt{2\pi}} \leqslant 1-\int_{-\infty}^{a}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \leqslant \frac{2\mathrm{e}^{-a^2/2}}{(a+\sqrt{a^2+2})\sqrt{2\pi}}. \qquad (6)$$

Using this inequality, we now obtain the lower and upper bounds for the probability of error. Since the normal probability density function is symmetric around the mean (*zero* in this case), we can re-write the inequality in Eq. (6) for all $-a < 0$ as follows:

$$\frac{2\mathrm{e}^{-a^2/2}}{(a+\sqrt{a^2+4})\sqrt{2\pi}} \leqslant \int_{-\infty}^{-a}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \leqslant \frac{2\mathrm{e}^{-a^2/2}}{(a+\sqrt{a^2+2})\sqrt{2\pi}}. \qquad (7)$$

On the other hand, we know that $a_1=(\mathbf{w}^t\mathbf{\Sigma}_1\mathbf{w})^{-1/2}(w_0+\mathbf{w}^t\boldsymbol{\mu}_1)$. Since $g(\boldsymbol{\mu}_1)<0$, it then follows that $\mathbf{w}^t\boldsymbol{\mu}_1+w_0<0$. Also, $\mathbf{\Sigma}_1$ is positive definite, which implies that $(\mathbf{w}^t\mathbf{\Sigma}_1\mathbf{w})^{-1/2}>0$. Thus, it follows that $a_1<0$.

Additionally, we know that $a_2=-(\mathbf{w}^t\mathbf{\Sigma}_2\mathbf{w})^{-1/2}(w_0+\mathbf{w}^t\boldsymbol{\mu}_2)$. Since $g(\boldsymbol{\mu}_2)=\mathbf{w}^t\boldsymbol{\mu}_2+w_0>0$, and $\mathbf{\Sigma}_1$ is positive definite, which implies that $(\mathbf{w}^t\mathbf{\Sigma}_2\mathbf{w})^{-1/2}>0$, it follows that $a_2<0$.

Since we know that $a_i<0$ for $i=1,2$, we can then express Eq. (7) in terms of a_i as follows:

$$\frac{2\mathrm{e}^{-a_i^2/2}}{(-a_i+\sqrt{a_i^2+4})\sqrt{2\pi}} \leqslant \int_{-\infty}^{a_i}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x \leqslant \frac{2\mathrm{e}^{-a_i^2/2}}{(-a_i+\sqrt{a_i^2+2})\sqrt{2\pi}}. \qquad (8)$$

Pre-multiplying by $P(\omega_i)$, and adding the corresponding terms of Eq. (8) for $i=1,2$, we obtain the following inequality:

$$\sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-a_1^2/2}}{-a_1+\sqrt{a_1^2+4}}+\frac{P(\omega_2)\mathrm{e}^{-a_2^2/2}}{-a_2+\sqrt{a_2^2+4}}\right) \leqslant \Pr[\text{error}] \leqslant \sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-a_1^2/2}}{-a_1+\sqrt{a_1^2+2}}+\frac{P(\omega_2)\mathrm{e}^{-a_2^2/2}}{-a_2+\sqrt{a_2^2+2}}\right), \qquad (9)$$

where a_1 and a_2 are obtained as in Eqs. (4) and (5), respectively.

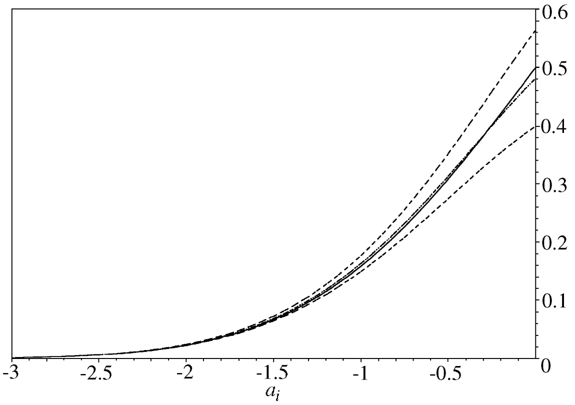

Fig. 1. Lower and upper bounds, actual and approximate values for the cumulative normal distribution for values of a_i in $[-3,0]$. The dashed lines represent the lower and upper bounds, and the solid line represents the actual value of the cumulative distribution function, as computed in Eq. (8). The dotted line near the solid one corresponds to the approximation of Eq. (10).

The bounds for the classification error are important to obtain a fair assessment about the classifier, without computing the error using numeric integration methods. One should note, however, that these bounds are not tight enough for values of a_i close to 0, while being asymptotically accurate for $a_i < -1$. This relationship is shown in Fig. 1, where the lower and upper bounds, as well as the actual values for the cumulative distribution functions for values of a_i are plotted. An alternative for this is to use the following approximating function for the probability of error:

$$\Pr[\text{error}] \cong \frac{1}{\sqrt{2\pi}}\left[P(\omega_1)\mathrm{e}^{-a_1^2/2}\left(\frac{1}{-a_1+\sqrt{a_1^2+4}}+\frac{1}{-a_1+\sqrt{a_1^2+2}}\right)+P(\omega_2)\mathrm{e}^{-a_2^2/2}\times\left(\frac{1}{-a_2+\sqrt{a_2^2+4}}+\frac{1}{-a_2+\sqrt{a_2^2+2}}\right)\right]. \qquad (10)$$

As can observed in Fig. 1, taking the average between the lower and the upper bounds as in Eq. (10) appears to be a very good approximation for the actual probability of error. Based on this observation, we show that a good approximation for the probability of error can be achieved by averaging the lower and upper bounds. To prove this result, we, first of all, analyze the relationship between the integral and the average of the two bounds in Eq. (8). We are tempted to call the result given below *lemma*, but unfortunately, one of the steps cannot be proved algebraically. Thus, we merely provide a sketch of proof for such a result.

Result 1. For all $a_i < 0$

$$|g(a_i)| = \left| \int_{-\infty}^{a_i} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x - \frac{\mathrm{e}^{-a_i^2/2}}{\sqrt{2\pi}} \times \left(\frac{1}{-a_i + \sqrt{a_i^2+4}} + \frac{1}{-a_i + \sqrt{a_i^2+2}} \right) \right| \leqslant \varepsilon, \qquad (11)$$

where $\varepsilon = 0.0184340683$.

Sketch of Proof. By taking the first derivative of $g(a_i)$ with respect to a_i we obtain the necessary condition for a maximum or a minimum:

$$\frac{\partial g}{\partial a_i} = \frac{\mathrm{e}^{-a_i^2/2}}{\sqrt{2\pi}} \frac{a_i^4 + \left(4 - \sqrt{a_i^2+4}\sqrt{a_i^2+2}\right) a_i^2 + \left(\sqrt{a_i^2+4} + \sqrt{a_i^2+2}\right) a_i + 2}{\sqrt{a_i^2+4}\sqrt{a_i^2+2}\left(\sqrt{a_i^2+4} - a_i\right)\left(\sqrt{a_i^2+2} - a_i\right)} = 0. \qquad (12)$$

It can be shown that for any value of $a_i < 0$, $\sqrt{a_i^2+4}\sqrt{a_i^2+2}\left(\sqrt{a_i^2+4} - a_i\right)\left(\sqrt{a_i^2+2} - a_i\right) > 0$. Thus, to satisfy the equality of Eq. (12), either of the following equalities must be satisfied:

$$\mathrm{e}^{-a_i^2/2} = 0 \qquad (13)$$

or

$$a_i^4 + \left(4 - \sqrt{a_i^2+4}\sqrt{a_i^2+2}\right) a_i^2 + \left(\sqrt{a_i^2+4} + \sqrt{a_i^2+2}\right) a_i + 2 = 0. \qquad (14)$$

It is true that $\mathrm{e}^{-a_i^2/2}$ will never reach the value *zero* exactly but asymptotically, i.e. $\lim_{a_i \to -\infty} \mathrm{e}^{-a_i^2/2} = 0$. Since we are interested in all the values of $a_i < 0$, we should then find the values of a_i that satisfy Eq. (14). Unfortunately, this equation does not have a closed-form solution for a_i, and hence we are obliged to find a numerical solution instead.

In Fig. 2, we plot the function $g(a_i)$ for values of a_i between -4 and 0. Noting that $g(a_i)$ reaches a minimum for a value of a_i that is between 0 and -1, we numerically found the roots of Eq. (14), being the one of our interest $a_i' = -0.690181517$, where $|g(a_i')| \approx 0.005882668 < \varepsilon$.

On the other hand, we observe that for all $a_i' < a_i < 0$, $g(a_i)$ is monotonically increasing, and hence it attains a maximum at $a_i'' = 0$, where $|g(a_i'')| \approx \varepsilon = 0.0184340683$. To verify that this is a maximum in $(-\infty, 0]$, we should still analyze the behavior of $g(a_i)$ for $a_i \to -\infty$. In other words, we should find the following limit:

$$\lim_{a_i \to -\infty} g(a_i) = \lim_{a_i \to -\infty} \left[\int_{-\infty}^{a_i} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x - \frac{\mathrm{e}^{-a_i^2/2}}{\sqrt{2\pi}} \times \left(\frac{1}{-a_i + \sqrt{a_i^2+4}} + \frac{1}{-a_i + \sqrt{a_i^2+2}} \right) \right]. \qquad (15)$$

It is clear that $\lim_{a_i \to -\infty} \int_{-\infty}^{a_i} (1/\sqrt{2\pi}) \mathrm{e}^{-x^2/2}\,\mathrm{d}x = 0$, since it leads to an integral with coincident boundaries. Also, as observed earlier, $\lim_{a_i \to -\infty} \mathrm{e}^{-a_i^2/2} = 0$. Additionally, as $a_i \to -\infty$, it implies that $a_i^2 \to \infty$ and $-a_i \to \infty$. This implies that $\sqrt{a_i^2+4} \to \infty$ and $\sqrt{a_i^2+2} \to \infty$, yielding $\lim_{a_i \to -\infty} 1/\left(-a_i + \sqrt{a_i^2+4}\right) + 1/\left(-a_i + \sqrt{a_i^2+2}\right) \to 0$, and hence $\lim_{a_i \to -\infty} g(a_i) = 0$.

Consequently, for all $a_i < 0$, $|g(a_i)| \leqslant \varepsilon = 0.0184340683$. The result follows. □

Using the result shown above, we now state and prove the relationship between the approximation function in Eq. (10) and the actual probability of error, which is computed as in Eq. (3).

Theorem 1. *The approximation given in Eq.* (10) *differs from the actual error in Eq.* (3) *in at most* ε=0.0184340683.

Proof. Since $P(\omega_1) \geqslant 0$ and $P(\omega_2) \geqslant 0$, using the inequality given in Eq. (11), we can write:

$$|P(\omega_1) g(a_1)| \leqslant P(\omega_1)\varepsilon \qquad (16)$$

and

$$|P(\omega_2) g(a_2)| \leqslant P(\omega_2)\varepsilon = [1 - P(\omega_1)]\varepsilon. \qquad (17)$$

Adding Eqs. (16) and (17), and using properties of absolute value, we have:

$$\begin{aligned} &|P(\omega_1) g(a_1) + P(\omega_2) g(a_2)| \\ &\quad \leqslant |P(\omega_1) g(a_1)| + |P(\omega_2) g(a_2)| \\ &\quad \leqslant P(\omega_1)\varepsilon + [1 - P(\omega_1)]\varepsilon. \end{aligned} \qquad (18)$$

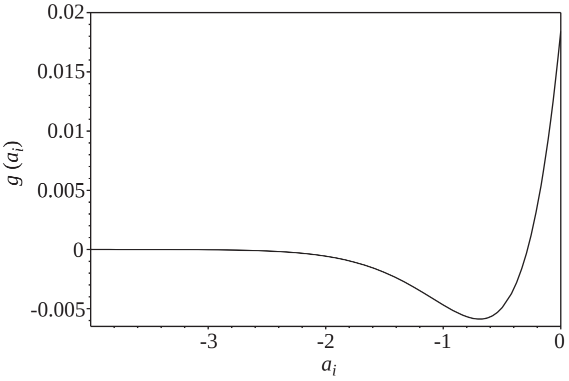

Fig. 2. Plot of the function $g(a_i)$, which is obtained as in Eq. (11), for values of a_i between -4 and 0.

Substituting $g(a_i)$ for its equivalent expression in Eq. (11), we obtain:

$$\left| P(\omega_1)\int_{-\infty}^{a_1}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x+P(\omega_2)\int_{-\infty}^{a_2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x - \frac{1}{\sqrt{2\pi}}\left[P(\omega_1)\mathrm{e}^{-a_1^2/2}\left(\frac{1}{-a_1+\sqrt{a_1^2+4}} + \frac{1}{-a_1+\sqrt{a_1^2+2}}\right) + P(\omega_2)\mathrm{e}^{-a_2^2/2} \times \left(\frac{1}{-a_2+\sqrt{a_2^2+4}}+\frac{1}{-a_2+\sqrt{a_2^2+2}}\right)\right]\right| \leqslant P(\omega_1)\varepsilon+[1-P(\omega_1)]\varepsilon=\varepsilon, \tag{19}$$

where a_1 and a_2 are obtained as in Eqs. (4) and (5), respectively.

The theorem is thus proved. □

The result of Theorem 1 is quite important in our analysis, since it shows that the approximation for the error is very accurate, e.g. it differs from the actual error in at most two digits. This relationship is corroborated in the empirical results discussed in Section 5.

3. Error analysis for Fisher's classifier

It is well known that Fisher's classifier can be derived from the training samples or from the parameters of the distributions, when available. We assume that the parameters of the distributions are known or they can be obtained from the training samples by using the MLE or Bayesian estimation. While we consider the two-class case, the general Fisher's classifier design that involves more than two classes can be found in Refs. [1,10].

Suppose then that we deal with two classes, ω_1 and ω_2, which are represented by two normally distributed d-dimensional random vectors, $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, respectively, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and whose a priori probabilities are $P(\omega_1)$ and $P(\omega_2)$, respectively. The aim is to find a vector, $\mathbf{w}$, which leads to the *maximum* class separability in the projected, one-dimensional space. The solution for the vector $\mathbf{w}$ that maximizes the aforementioned criterion is given by:

$$\mathbf{w}=\mathbf{S}_W^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1), \tag{20}$$

where $\mathbf{S}_W=\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)$. Note that only the direction of $\mathbf{w}$ is important here, and thus, the scaling factor $\frac{1}{2}$ can be omitted at this point. However, to avoid inconsistencies in our derivations, we will continue using the scaling factor.

To complete the linear classifier, the threshold w_0 has to be obtained. A simple approach (suggested in Ref. [6]) is to assume that the distributions in the original space have identical covariance matrices, and take the independent term of the optimal quadratic or Bayesian classifier, which results in:

$$w_0=-\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}\mathbf{S}_W^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)-\log\frac{P(\omega_1)}{P(\omega_2)}. \tag{21}$$

Once we have derived the corresponding linear classifier by means of vector $\mathbf{w}$ and the corresponding threshold, we obtain the boundaries for the integrals in Eq. (3). The algebraic expression for the error is stated in the following theorem. We use Pr[error(F)] to refer to the probability of error of Fisher's classifier so that it is distinguished from that of the *generic* linear classifier.

Theorem 2. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors representing two classes,* ω_1 *and* ω_2*, whose a priori probabilities are* $P(\omega_1)$ *and* $P(\omega_2)$*, respectively, and* $g(\mathbf{x})=\mathbf{w}^{\mathrm{t}}\mathbf{x}+w_0$ *be Fisher's classifier, where* $\mathbf{w}$ *and* w_0 *are obtained as in Eqs.* (20) *and* (21)*, respectively. If* $r^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) > \max\{\log(P(\omega_1)/P(\omega_2)), \log(P(\omega_2)/P(\omega_1))\}$, *then*:

$$\Pr[\mathrm{error}(F)]=P(\omega_1)\int_{-\infty}^{b_1}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x + P(\omega_2)\int_{-\infty}^{b_2}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x, \tag{22}$$

where x is a $N(0,1)$ *random variable, and*

$$b_1=\frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right]^{-1/2} \left[-(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)-\log\frac{P(\omega_1)}{P(\omega_2)}\right] \tag{23}$$

and

$$b_2=-\frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right]^{-1/2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - \log\frac{P(\omega_1)}{P(\omega_2)}\right]. \tag{24}$$

Proof. We prove, first, that Fisher's classifier separates the means, i.e. $g(\boldsymbol{\mu}_1)<0$ and $g(\boldsymbol{\mu}_2)>0$. Substituting $\mathbf{x}$ for $\boldsymbol{\mu}_1$ in Fisher's classifier we have:

$$g(\boldsymbol{\mu}_1)=\left\{\left[\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\right]^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right\}^{\mathrm{t}}\boldsymbol{\mu}_1-\frac{1}{2}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}} \times\left[\frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)\right]^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)-\log\frac{P(\omega_1)}{P(\omega_2)}, \tag{25}$$

$$g(\boldsymbol{\mu}_1) = 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\mu}_1 - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)}, \tag{26}$$

$$g(\boldsymbol{\mu}_1) = -(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)}, \tag{27}$$

$$g(\boldsymbol{\mu}_1) = -r^2 - \log\frac{P(\omega_1)}{P(\omega_2)}, \tag{28}$$

where $r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Since $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite, it follows that $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ and $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}$ are also positive definite, implying that r^2 is a positive real number. Thus, it is true that $g(\boldsymbol{\mu}_1) = -r^2 - \log(P(\omega_1)/P(\omega_2)) < 0$, if

$$r^2 > -\log\frac{P(\omega_1)}{P(\omega_2)} = \log\frac{P(\omega_2)}{P(\omega_1)}. \tag{29}$$

Similarly, replacing $\mathbf{x}$ by $\boldsymbol{\mu}_2$ in Fisher's classifier, and replicating the steps from Eqs. (25) to (28), we obtain:

$$g(\boldsymbol{\mu}_2) = 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\mu}_2 - (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}} \times (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)}, \tag{30}$$

$$g(\boldsymbol{\mu}_2) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)}, \tag{31}$$

$$g(\boldsymbol{\mu}_2) = r^2 - \log\frac{P(\omega_1)}{P(\omega_2)}. \tag{32}$$

Again, since r^2 is a positive real number, it follows that $g(\boldsymbol{\mu}_2) > 0$, if

$$r^2 > \log\frac{P(\omega_1)}{P(\omega_2)}. \tag{33}$$

From the inequalities in Eqs. (29) and (33), we conclude that Fisher's classifier separates the means, i.e. $g(\boldsymbol{\mu}_1) < 0$ and $g(\boldsymbol{\mu}_2) > 0$, if

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \max\left\{\log\frac{P(\omega_1)}{P(\omega_2)}, \log\frac{P(\omega_2)}{P(\omega_1)}\right\}. \tag{34}$$

We now use the result of the generic linear classifier. Substituting $\mathbf{w}$ and w_0 for their corresponding values obtained from Eqs. (20) and (21), respectively, the boundary for the first integral of Eq. (3) results in:

$$b_1 = (\mathbf{w}^{\mathrm{t}}\boldsymbol{\Sigma}_1\mathbf{w})^{-1/2}(w_0 + \mathbf{w}^{\mathrm{t}}\boldsymbol{\mu}_1), \tag{35}$$

$$b_1 = \left\{\left[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{\mathrm{t}}\boldsymbol{\Sigma}_1\left[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]\right\}^{-1/2}\left\{-\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)} + [2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^{\mathrm{t}}\boldsymbol{\mu}_1\right\}, \tag{36}$$

$$b_1 = \frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{-1/2}\left[-(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)} + 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\mu}_1\right], \tag{37}$$

$$b_1 = \frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{-1/2}\left[-(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log\frac{P(\omega_1)}{P(\omega_2)}\right], \tag{38}$$

where Eq. (37) is obtained from Eq. (36) after expanding $[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^{\mathrm{t}}$ and some algebraic manipulations; and Eq. (38) follows from Eq. (37) because $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are symmetric, which implies that $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ and $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}$ are both symmetric, and other minor algebraic manipulations.

Similarly, the boundary for the second integral of Eq. (3) can be expressed as follows:

$$b_2 = -(\mathbf{w}^{\mathrm{t}}\boldsymbol{\Sigma}_2\mathbf{w})^{-1/2}(w_0 + \mathbf{w}^{\mathrm{t}}\boldsymbol{\mu}_2), \tag{39}$$

$$b_2 = -\left\{\left[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{\mathrm{t}}\boldsymbol{\Sigma}_2\left[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]\right\}^{-1/2} \times \left\{-\frac{1}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)} + \left[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{\mathrm{t}}\boldsymbol{\mu}_2\right\}, \tag{40}$$

$$b_2 = -\frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2 (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{-1/2} \times \left[-(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \log\frac{P(\omega_1)}{P(\omega_2)} + 2(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\mu}_2\right], \tag{41}$$

$$b_2 = -\frac{1}{2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} \times (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)\right]^{-1/2}\left[(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \log\frac{P(\omega_1)}{P(\omega_2)}\right], \tag{42}$$

where, again, we expanded $[2(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)]^{\mathrm{t}}$, used the fact that $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}$ is symmetric, and performed other minor algebraic manipulations.

The theorem is thus proved. □

Corollary 1. *Under the conditions of Theorem* 2, *and assuming that* $P(\omega_1) = P(\omega_2) = 0.5$, *the probability of error of Fisher's classifier can* always *be computed as in Eq.* (22).

Proof. The proof of this corollary is straightforward. From the proof of Theorem 2, we know that the probability of error can be computed as in (22), if $r^2 > \max\{\log(P(\omega_1)/P(\omega_2)), \log(P(\omega_2)/P(\omega_1))\}$. Since, $P(\omega_1)=P(\omega_2)=0.5$, it implies that $\log(P(\omega_1)/P(\omega_2))=\log(P(\omega_2)/P(\omega_1))=0$. Also, r^2 is a positive real number, and thus, the result follows. □

The algebraic expression for the error obtained in Theorem 2 shows that the classification error for Fisher's classifier can be derived directly from the parameters of the distributions, i.e. without finding the corresponding classifier. It is important to note, however, that the threshold in Fisher's classifier, and in general, for any classifier, can be obtained in many different ways. In Ref. [11], it has been shown experimentally that the probability of error can be reduced if the threshold is computed by invoking a Bayes classifier in the transformed, one-dimensional space. It is then easy to see that the expressions for $b_1$and b_2 given in Eqs. (23) and (24), respectively, can also be expressed in terms of the aforementioned thresholding method. The derivation of these algebraic expressions is quite involved, and left for future research work.

4. Bounds on the error of Fisher's classifier

Using the algebraic analysis of the probability of error discussed in the previous subsection, we obtain bounds on the error for Fisher's classifier. To achieve this, we use the inequality of Eq. (9).

Theorem 3. *Let* $\mathbf{x}_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ *and* $\mathbf{x}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *be two normally distributed random vectors representing two classes*, ω_1 *and* ω_2, *whose a priori probabilities are* $P(\omega_1)$ *and* $P(\omega_2)$, *respectively, and* $g(\mathbf{x}) = \mathbf{w}^{\mathrm{t}}\mathbf{x} + w_0$ *be Fisher's classifier, where* $\mathbf{w}$ *and* w_0 *are obtained as in Eqs.* (20) *and* (21), *respectively. If* $r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \max\{\log(P(\omega_1)/P(\omega_2)), \log(P(\omega_2)/P(\omega_1))\}$, *then*:

$$\sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-b_1^2/2}}{-b_1+\sqrt{b_1^2+4}}+\frac{P(\omega_2)\mathrm{e}^{-b_2^2/2}}{-b_2+\sqrt{b_2^2+4}}\right) \leqslant \Pr[\mathrm{error}(F)]$$
$$\leqslant \sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-b_2^2/2}}{-b_1+\sqrt{b_1^2+2}}+\frac{P(\omega_2)\mathrm{e}^{-b_2^2/2}}{-b_2+\sqrt{b_2^2+2}}\right), \tag{43}$$

where b_1 *and* b_2 *are obtained as in Eqs.* (23) *and* (24), *respectively*.

Proof. Since $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite, it then follows that $\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$, its inverse, $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}$, and $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}$ are positive definite, implying that $(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ is a positive real number for all $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Also, it is true that $r^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) > 0$, and hence $b_1 < 0$ if $r^2 > \log(P(\omega_1)/P(\omega_2))$.

Similarly, since $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are positive definite, it follows that $(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}$ is also positive definite. As a result, $(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^{\mathrm{t}}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\boldsymbol{\Sigma}_2(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) > 0$ for all $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Consequently, $b_2 < 0$ if $r^2 > \log(P(\omega_2)/P(\omega_1))$, implying that $b_1 < 0$ and $b_2 < 0$, if $r^2 > \max\{\log(P(\omega_1)/P(\omega_2)), \log(P(\omega_2)/P(\omega_1))\}$.

Substituting a_1 and a_2 for b_1 and b_2, respectively, Eq. (9) can be written as follows:

$$\sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-b_1^2/2}}{-b_1+\sqrt{b_1^2+4}}+\frac{P(\omega_2)\mathrm{e}^{-b_2^2/2}}{-b_2+\sqrt{b_2^2+4}}\right)$$
$$\leqslant P(\omega_1)\int_{-\infty}^{b_1}\frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-x^2/2}\,\mathrm{d}x$$
$$+\,P(\omega_2)\int_{-\infty}^{b_2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x$$
$$\leqslant\sqrt{\frac{2}{\pi}}\left(\frac{P(\omega_1)\mathrm{e}^{-b_1^2/2}}{-b_1+\sqrt{b_1^2+2}}+\frac{P(\omega_2)\mathrm{e}^{-b_2^2/2}}{-b_2+\sqrt{b_2^2+2}}\right), \tag{44}$$

where b_1 and b_2 are obtained as in Eqs. (23) and (24), respectively.

The theorem is thus proved. □

The lower and upper bounds can also be obtained by taking Eq. (3), and substituting $\mathbf{w}$ and w_0 for their corresponding values obtained from Eqs. (20) and (21), respectively. The algebraic steps involved in the derivation are straightforward and omitted to avoid repetition.

As in the general case, we can use the expression given in Eq. (10) to yield an approximation for the probability of error of Fisher's classifier. This expression can be derived by taking the average of the two bounds in Eq. (43) as follows:

$$\Pr[\mathrm{error(F)}] \cong \frac{1}{\sqrt{2\pi}}\left[P(\omega_1)\mathrm{e}^{-b_1^2/2}\left(\frac{1}{-b_1+\sqrt{b_1^2+4}}\right.\right.$$
$$\left.+\frac{1}{-b_1+\sqrt{b_1^2+2}}\right)+P(\omega_2)\mathrm{e}^{-b_2^2/2}$$
$$\left.\times\left(\frac{1}{-b_2+\sqrt{b_2^2+4}}+\frac{1}{-b_2+\sqrt{b_2^2+2}}\right)\right]. \tag{45}$$

As shown in the general case, the expression given in Eq. (45) provides a good approximation for the probability of error of Fisher's classifier, differing from the actual value in

Table 1
Comparison of the probability of error obtained using traditional methods and the approximation method discussed in this paper. The lower and upper bounds, as well as the approximation given in Eq. (10) are shown

Dim.	$P(\omega_1)$	a_1	a_2	Pr[error]	Lower bnd.	Upper bnd.	Approx.	Difference
10	0.42	−0.451764	−1.048493	0.2219389	0.2015077	0.2510731	0.2262904	0.0043516
20	0.21	−0.523061	−1.793484	0.0929374	0.0855113	0.1041107	0.0948110	0.0018735
30	0.29	−0.670047	−1.443671	0.1260964	0.1179252	0.1407908	0.1293580	0.0032616
40	0.13	−0.865691	−2.102256	0.0414202	0.0394080	0.0456212	0.0425146	0.0010943
50	0.54	−1.500102	−1.590889	0.0617688	0.0599570	0.0671311	0.0635440	0.0017752
60	0.74	−2.120503	−1.914417	0.0197510	0.0194277	0.0210349	0.0202313	0.0004804
70	0.47	−1.970613	−1.977531	0.0241757	0.0237525	0.0258121	0.0247823	0.0006066
80	0.15	−2.144580	−2.923591	0.0039152	0.0038698	0.0041255	0.0039976	0.0000824
90	0.77	−2.496978	−2.074405	0.0092195	0.0091016	0.0097450	0.0094233	0.0002037
100	0.73	−2.434394	−2.235407	0.0088897	0.0087858	0.0093767	0.0090812	0.0001915

at most $\varepsilon = 0.0184340683$. This is empirically demonstrated in the experiments discussed in Section 5.

5. Experimental results

To test the accuracy and computational efficiency of the error analysis discussed in this paper, we performed a few simulations on synthetic data and standard real-life data sets. The experiments involve two normally distributed classes and Fisher's linear classifier, which is obtained as in Eqs. (20) and (21).

5.1. Experiments on synthetic data

For this set of experiments, we randomly generated normally distributed random vectors for two classes. These classes, ω_1 and ω_2, are then fully specified by their parameters, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. In order to make the experiments more realistic, we randomly generated $P(\omega_1)$ from a uniform distribution in [0, 1], and set $P(\omega_2) = 1 - P(\omega_1)$.

One of the tests involves the analysis of the *actual* classification error,[3] as well as the bounds and the approximation introduced in this paper. To conduct the test, we generated random parameters for *d*-dimensional classes, where $d = 10, 20, \ldots, 100$. The linear classifier used to test our method is the traditional Fisher's classifier, where the threshold is obtained as in Eq. (21). The mean vectors for the two classes were generated randomly from a uniform distribution specified by the intervals [0, 0.4] for $\boldsymbol{\mu}_1$, and [0.4, 0.8] for $\boldsymbol{\mu}_2$. The means generated are very close so that for lower dimensions, the classification task is more demanding. The covariance matrices were generated by invoking the random correlation method for generating positive semidefinite matrices [12]. The empirical results obtained from our simulations are shown in Table 1. The first column corresponds to the dimension of the feature space, and the second column contains the a priori probability of ω_1. The third and fourth columns contain the boundaries for the two integrals, computed as in Eqs. (4) and (5), respectively. The fifth column corresponds to the probability of error for Fisher's classifier, obtained as in Eq. (3), where the integrals were computed numerically by invoking the near-minimax Chebyschev approximations for the error function [13]. The sixth and seventh columns contain the lower and upper bounds for the classification error, which were computed as in Eq. (9). The eighth column represents the approximation of the classification error, obtained as in Eq. (10), and the last column corresponds to the difference (in absolute value) between the *actual* probability of error, the fifth column, and the *approximation* of the error, the eighth column.

The results from the table show that the lower and upper bounds for the error are quite *loose* for large values of the classification error. Conversely, they are very *tight* for small values of the classification error. This is observed in the last row for dimension 100, in which the bounds are found to be very close to each other. A similar behavior is observed when analyzing the difference between the actual error and the approximation. The approximation differs from the actual error in nearly 10^{-4} for dimensions 90 and 100. Observe also, that in all cases, even in the case of dimension 10 in which the error is large, the approximation of the error differs from the actual value in less than $\varepsilon = 0.0184340683$, and hence achieving at least *two digits* of precision.

To experimentally analyze the computational efficiency for computing the classification error, we conducted simulations on test suites involving dimensions $d=20, 40, \ldots, 200$. The parameters used in our simulations were obtained as explained above. To assess the running time of the methods, we ran the experiments for each dimension 100 times. The results obtained are depicted in Table 2. The second and third columns contain the average for the probability of

[3] It should be noted that the *actual* classification error cannot be obtained, due to the impossibility of obtaining the integrals for the normal distribution density function. We, indeed, use the term "*actual*" to refer to values obtained using numeric integration methods.

Table 2
Results for the running times and probabilities of error for different simulations on normally distributed classes whose dimensions range from 20 to 200

Dim.	Avg. error	Avg. diff.	Time
20	0.29636033	0.00505519	0.01743
40	0.22418577	0.00571323	0.03264
60	0.17657586	0.00504054	0.05367
80	0.14238103	0.00423317	0.09574
100	0.11416514	0.00343556	0.15212
120	0.09429007	0.00282129	0.26718
140	0.07453343	0.00218855	0.43082
160	0.06238508	0.00179567	0.74598
180	0.05257992	0.00148198	0.91441
200	0.04395622	0.00120941	1.46771

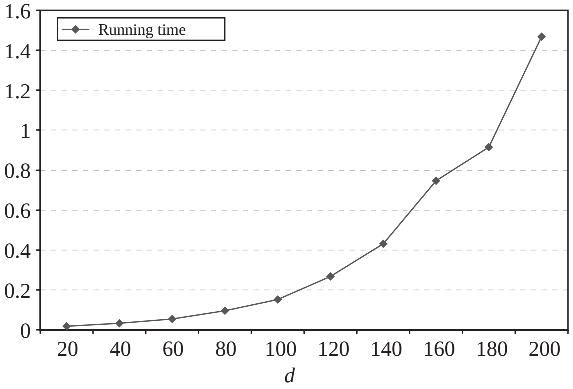

Fig. 3. Plot of the running time vs. the dimension of the feature space for simulations involving normally distributed classes and Fisher's classification error.

error, and the average of the difference between the actual value and the approximation. The fourth column contains the CPU time (in seconds) for computing the actual probability of error, the bounds, and the approximation for each of the experiments. The methods were tested using Matlab on an Intel 2.0 GHz workstation running Windows XP.

The results from the table show that our computational method is extremely fast for dimensions less than 100, performing the computations in less than *a tenth* of a second. As the dimension of the feature space increases, we observe that the running times also increase. We also observe that the average differences between the actual error and the approximation are very small, below $\varepsilon = 0.0184340683$.

On the other hand, as can be observed in Eqs. (4) and (5), the running time for computing the boundaries of the integrals is proportional to the square of the dimension of the feature space, i.e. its time complexity is $\Theta(d^2)$. This can be observed in Fig. 3, in which the plot of the dimension against the running time is depicted. The shape of the curve in the figure corroborates the aforementioned complexity analysis for our approximation method to compute the classification error. This behavior indicates that there are still open problems in this direction, such as devising more efficient algorithms to implement our approximation methods for computing the classification error.

5.2. Results on real-life data

For the experiments on real-life data we have selected various data sets drawn from the UCI machine learning repository.[4] From each data set, we have selected pairs of classes, in order to conduct the experiments for two classes. The parameters of the distributions have been estimated using the MLE method, and the threshold for Fisher's classifier has been computed as in Eq. (21). The a priori probabilities for the two classes has also been estimated by using the MLE method, i.e. by dividing the number of samples that belong to the respective class by the total number of samples. The results obtained are displayed in Table 3. We observe that the results shown in the table corroborate our theoretical analysis for the classification error, regarding the accuracy in approximating the error, i.e. the error is approximated by a factor of at most $\varepsilon = 0.0184340683$. In some data sets, and specifically, in some pairs of classes, the resulting error, approximation and bounds are arbitrarily small, due to the fact that the classes are well-separated (linearly). This is the case of versicolor and setosa, from the Iris data set, and type3 and type1 from the Wine data set.

6. The multi-class case

When dealing with more than two classes the problem of estimating the error, and in general, deriving a linear classifier is quite intricate. A generic classification scheme

[4] Available electronically at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Table 3
Empirical results for the lower and upper bounds, as well as the approximation for the probability of error on data sets from the UCI machine learning repository

Data set	ω_1	ω_2	$P(\omega_1)$	Pr[error]	Lwr bnd.	Upp bnd.	Approx.	Diff.
Yeast	ystCYT	ystMIT	0.655	0.187076	0.165495	0.211298	0.188396	0.001321
Iris	versicolor	virginica	0.500	0.028485	0.027937	0.030503	0.029220	0.000735
Iris	versicolor	setosa	0.500	0.000002	0.000002	0.000002	0.000002	0.000000
Balance	left	right	0.500	0.073745	0.071336	0.080478	0.075907	0.002162
Balance	right	balanced	0.855	0.117241	0.110187	0.130346	0.120266	0.003026
WDBC	nonrecur	recur	0.763	0.128400	0.120771	0.143026	0.131899	0.003499
WDBC	benign	malignant	0.627	0.031091	0.030312	0.033562	0.031937	0.000846
CPU-Perf.	nas	ncr	0.594	0.032361	0.031642	0.034805	0.033224	0.000863
Letter	A	B	0.507	0.001107	0.001100	0.001151	0.001126	0.000019
Spect	class0	class1	0.500	0.051545	0.050115	0.055884	0.052999	0.001455
Vehicle	bus	van	0.500	0.000503	0.000500	0.000521	0.000510	0.000008
Wine	type1	type2	0.454	0.007422	0.007329	0.007847	0.007588	0.000165
Wine	type3	type1	0.449	0.000001	0.000001	0.000001	0.000001	0.000000
Wine	type3	type2	0.403	0.004361	0.004319	0.004578	0.004448	0.000088

(suggested in Ref. [1]) that avoids ambiguous regions is as follows. Consider c classes, $\{\omega_1, \ldots, \omega_c\}$. The aim is to find c linear functions $g_i(\mathbf{x})$, and given an unknown object, $\mathbf{x}$, assign it to class ω_i for which $g_i(\mathbf{x})$ is maximum. As observed in Ref. [1], finding the probability of error in this scheme is easier if considering the probability of being correct. That is:

$$\Pr[\text{error}] = 1 - \Pr[\text{correct}], \tag{46}$$

where

$$\Pr[\text{correct}] = \sum_{i=1}^{c} \Pr[\mathbf{x} \text{ assigned to } \omega_i \mid \mathbf{x} \in \omega_i]. \tag{47}$$

Given a set W of linear functions of the form $\mathbf{w}_j^{\mathrm{t}}\mathbf{x} + w_{0j}$, the above-mentioned scheme leads to c convex regions, where the region for ω_i is bounded by a subset of W containing k_i functions $\mathbf{w}_k^{\mathrm{t}}\mathbf{x} + w_{0k}$, for $k = 1, \ldots, k_i$ as follows:

$$\mathbf{x} \in \omega_i \text{ if } \bigwedge_{k=1}^{k_i} \mathbf{w}_k^{\mathrm{t}}\mathbf{x} + w_{0k} > 0. \tag{48}$$

Pr[correct] is then computed by adding the probabilities of the events in Eq. (48), for $i = 1, \ldots, c$, where each term is multiplied by the corresponding a priori probability, $P(\omega_i)$. This is always possible, since events that $\mathbf{x} \in \omega_i$ and $\mathbf{x} \in \omega_j$ are mutually exclusive, for all $i \neq j$. Computing such a probability is not an easy task, and many approaches have been proposed [14–17]. Unfortunately, all of these approaches provide a numerical solution, and thus, deriving closed-form expressions for the bounds or approximation of the error is not possible. This problem which, to date, remains open, is currently being investigated.

7. Conclusions

In this paper, we derive lower and upper bounds, and an expression that approximates the probability of error, which can be obtained directly from the parameters of the distributions. This result can be used for *any* linear classifier, even though the underlying distributions are not normal. We have shown that the approximating expression differs from the *actual* value for the error in at most $\varepsilon = 0.0184340683$. By instantiating the expression of the generic linear classifier to a particular case, we derive the lower and upper bounds for the probability of error of the traditional Fisher's classifier. For this classifier, we also derive the corresponding expression that approximates the classification error.

Our empirical results on synthetic, *higher*-dimensional data show that the bounds are very tight for small values of the classification error. Also, the approximation expression has been empirically shown to be very precise in the estimation of the error—the error is approximated by a factor at most $\varepsilon = 0.0184340683$. The method has been shown to work efficiently for classification problems involving up to *two hundred* features. Experiments on real-life data from the UCI machine learning repository have also been conducted, which demonstrate that our scheme is quite accurate in approximating the classification error for real-life scenarios.

Many directions for future work exist, including the generalization of this model for more than two classes. As pointed out in Section 6, this problem, which we are currently investigating, is far from trivial. Another problem that is worth investigating is the generalization of the model for piecewise linear classifiers including more than one hyperplane. This problem is quite difficult since the linear transformations have to be applied simultaneously for all hyperplanes, leading to a multivariate integration problem.

Acknowledgements

The author is very grateful to the anonymous referees who pointed out important issues about this paper, such as the multi-class problem and experiments on real-life data, which helped in substantially enhancing the quality of the paper. The author would also like to thank Ataul Bari, Semsettin Ilyas, Guorui Li, and Ahmed Ullah, for their help in the experiments on real-life data. This research has been partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Foundation for Innovation (CFI), and the Ontario Innovation Trust (OIT).

References

[1] R. Duda, P. Hart, D. Stork, Pattern Classification, second ed., Wiley, New York, NY, 2000.

[2] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.

[3] S. Raudys, R. Duin, Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix, Pattern Recogn. Lett. 19 (1999) 385–392.

[4] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, MIT Press, Cambridge, MA, 2001.

[5] R. Herbrich, T. Graepel, A PAC-Bayesian margin bound for linear classifiers, IEEE Trans. Inf. Theory 48 (12) (2002) 3140–3150.

[6] A. Webb, Statistical Pattern Recognition, second ed., Wiley, New York, 2002.

[7] C. Lee, E. Choi, Bayes error evaluation of the Gaussian ML classifier, IEEE Trans. Geosci. Remote Sensing 38 (2000) 1471–1475.

[8] N. Vaswani, A linear classifier for Gaussian class conditional distributions with unequal covariance matrices, in: Proceedings of the 16th International Conference on Pattern Recognition, Quebec, Canada, vol. 2, 2002, 60–63.

[9] M. Kendall, A. Stuart, Kendall's Advanced Theory of Statistics, vol. I: Distribution Theory, sixth ed., Edward Arnold, Paris, 1998.

[10] Y. Xu, Y. Yang, Z. Jin, A novel method for Fisher discriminant analysis, Pattern Recognition 37 (2004) 381–384.

[11] L. Rueda, An efficient approach to compute the threshold for multi-dimensional linear classifiers, Pattern Recognition 37 (2004) 811–826.

[12] P. Davies, N. Higham, Numerically stable generation of correlation matrices and their factors, Technical Report, 354, Manchester, England, 1999.

[13] W. Cody, A portable FORTRAN package of special function routines and test drivers, ACM Trans. Math. Software 19 (1993) 22–32.

[14] A. Genz, Numerical computation of multivariate normal probabilities, J. Comput. Graphical Stat. 1 (1992) 141–149.

[15] H. Joe, Approximations to multivariate normal rectangle probabilities based on conditional expectations, J. Am. Stat. Assoc. 90 (431) (1995) 957–964.

[16] T. Miwa, A. Hayter, S. Kuriki, The evaluation of general non-centred orthant probabilities, J. R. Stat. Soc. Ser. B 65 (Part 1) (2003) 223–234.

[17] P. Somerville, Numerical computation of multivariate normal and multivariate-*t* probabilities over convex regions, J. Comput. Graphical Stat. 7 (4) (1998) 529–544.

About the Author—LUIS RUEDA received the degree of "Licenciado" in computer science from the National University of San Juan, Argentina, in 1993, and his Master's and Ph.D. degrees in computer science from Carleton University, Canada, in 1998 and 2002, respectively. He is currently working as Assistant Professor in the School of Computer Science at the University of Windsor, Canada. His research interests include bioinformatics, pattern recognition and microarray data analysis. He holds one patent and more than 25 publications in prestigious journals and refereed conferences. He is a member of the IEEE and the IAPR.