Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 89 (2004) 371–383
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

A lower bound on the performance of the quadratic discriminant function

Tristrom Cooke

Cooperative Research Centre for Sensor Signals and Information Processing, SPRI Building, 1 Mawson Lakes Boulevard, Mawson Lakes, S.A. 5095, Australia

Received 8 March 2001

Abstract

The quadratic discriminant function is often used to separate two classes of points in a multidimensional space. When the two classes are normally distributed, this results in the optimum separation. In some cases however, the assumption of normality is a poor one and the classification error is increased. The current paper derives an upper bound for the classification error due to a quadratic decision surface. The bound is strict when the class means and covariances and the quadratic discriminant surface satisfy certain specified symmetry conditions.

AMS 2000 subject classifications: 62H30

Keywords: Quadratic discriminant; Lower error bound; Non-Gaussian

1. Introduction

The binary classification problem for distinguishing between two classes of points in a multidimensional hyperspace has been well studied in the literature. Many of the recent theoretical papers in this area have dealt with classifiers having complicated decision surfaces such as support vector machines, neural networks and decision trees. While these discriminants have the potential to give much lower classification errors than their simpler counterparts, many application-based papers continue to make use of the simple classifiers such as Fisher's linear discriminant or the

E-mail address: tcooke@cssip.edu.au.

0047-259X/$ - see front matter
doi:10.1016/j.jmva.2003.11.005

quadratic discriminant function (QDF). For this reason, it is useful to have some theoretical understanding of the limitations of these techniques.

Of the simple classifiers, the Fisher linear discriminant and the k-nearest-neighbour method have received the most theoretical attention. The QDF however has been much more resistant to analysis, despite the simplicity of its implementation. Since the QDF is the optimal discriminant between two multi-dimensional Gaussian distributions, many papers have considered its robustness to non-normality. The majority of these papers have been based on computer simulations, and so have produced only qualitative results. For instance, Moore [5] found that in some situations the quadratic discriminant produced a higher classification error than a linear discriminant applied to the same problem. This observation was followed up by a more extensive study by Clarke et al. [2] which found that the skewness of the class distributions played a large role in the accuracy of the QDF. Since the simulations employed in this paper relied on a finite sample size, it was thought that the covariance estimate used to compute the QDF may have been responsible for this result. The hypothesis was later rejected in Broffitt et al. [1], which used a technique called Huberizing to obtain a more accurate covariance estimate for use in the QDF. It was still found that the QDF gave a poor performance for highly skewed distributions.

While the above papers gave qualitative results concerning the effects of non-normality on the QDF, no attempt was made to mathematically quantify the worst possible performance. The current paper derives a mathematical upper bound for the classification error due to a quadratic decision surface for a particular condition on the class covariance matrices. The symmetric problem is defined more rigorously in Section 2, and a strict error bound is obtained for linear, parabolic, hyperbolic and elliptic decision surfaces in Section 3. Section 4 describes how this symmetric result can be used to obtain an error bound for non-symmetric problems, although this bound is no longer guaranteed to be strict. Finally, Section 5 provides some numerical results from this solution.

2. Problem definition

Suppose two classes of N-dimensional distributions $f_1(x_1, x_2, \ldots, x_N)$ and $f_2(x_1, x_2, \ldots, x_N)$ are known to have means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariances $\boldsymbol{C}_1, \boldsymbol{C}_2$. Also suppose that the whole space is split by a quadratic decision surface, so that x is classified as from class 1 when $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} < c$, and class 2 otherwise. The problem is now to find the forms of the distributions $f_1(x_1, x_2, \ldots, x_N)$ and $f_2(x_1, x_2, \ldots, x_N)$ which maximise the classification error rate, and so provide a lower bound on the performance of the quadratic discriminant for given first and second moments of each class.

Without loss of generality, we can rotate and scale the feature space so that the distribution means are $\mathbf{0}$ and $[1, 0, 0, \ldots, 0]$. Due to the difficulty of finding a strict

bound for the general case, the analysis of Section 3 is presented using two symmetry assumptions:

- The covariance matrices are diagonal.
- The quadratic decision surface is symmetric about $x_i = 0$ for $i = 2 \ldots N$.

It is only the first assumption which is particularly restrictive, since practically all methods for generating a quadratic discriminant, given this first assumption, will produce a decision surface satisfying the second automatically.

Since there is no dependence between the exact forms of the distributions $f_1(x_1, x_2, \ldots, x_N)$ and $f_2(x_1, x_2, \ldots, x_N)$, the classification error due to each may be maximised separately. The solution to the problem is thus mathematically equivalent to:

$$\text{Maximise} \quad \int_{\Omega} f(x_1, x_2, \ldots, x_N)\, dx_1\, dx_2 \ldots dx_N \tag{1}$$

$$\text{Subject to} \quad \Omega = \left\{ \boldsymbol{x} : x_1 + A_1(x_1 - d)^2 + \sum_{i=2}^{N} A_i x_i^2 > d \right\}, \tag{2}$$

$$\int f(\boldsymbol{x})\, d\boldsymbol{x} = 1, \tag{3}$$

$$\int \boldsymbol{x} f(\boldsymbol{x})\, d\boldsymbol{x} = \mathbf{0}, \tag{4}$$

$$\int x_i^2 f(\boldsymbol{x})\, d\boldsymbol{x} = C_i, \tag{5}$$

$$f(\boldsymbol{x}) \geqslant 0 \text{ is symmetric about } x_i = 0 \quad \forall i > 2, \tag{6}$$

where A_i and C_i are the diagonal elements of the matrix $\mathbf{A}$ and the covariance matrix $\mathbf{C}$, respectively.

3. The maximum error for a class of quadratic decision surfaces

It is obvious when solving Eqs. (1)–(6) that if $\mathbf{0} \in \Omega$ then f can be made to be

$$f(\boldsymbol{x}) = \lim_{\varepsilon \to 0} \left\{ (1 - \varepsilon^2)\delta(\boldsymbol{x}) + \varepsilon^2 \sum_{i=1}^{N} \frac{1}{N} C_i \delta\left(\boldsymbol{x} - \left[0, 0, \ldots, \frac{\sqrt{N}}{\varepsilon}, 0, \ldots, 0 \right] \right) \right\}, \tag{7}$$

where δ is the Dirac delta function. This results in the maximum classification error of 100 percent. The remaining analysis will concern the case where $\mathbf{0} \in \bar{\Omega}$ (the overline referring to the complement). In this case, f will be expressed as

$$f(\boldsymbol{x}) = \chi_{\Omega} f(\boldsymbol{x}) + (1 - \chi_{\Omega}) f(\boldsymbol{x}),$$

where χ_{Ω} is the indicator function of Ω.

In the case when $A_i = 0$ for some $i \geqslant 2$, the problem's only constraint depending on x_i is that the variance in that direction must be C_i. Since the classification error completely loses its dependence on x_i, the entire problem can be reduced to a lower dimensional form where $A_i \neq 0$.

Now when $A_i < 0$ for some $i \geqslant 2$, the decision surface curves away from the origin in the x_i direction. This means that there is a greater restriction on the way in which the misclassified observations from $\chi_\Omega f(\boldsymbol{x})$ are distributed. For this reason, the maximum possible classification error must be less than or equal to that achievable when $A_i = 0$. Also, there will also be no restriction due to the decision surface on the distribution of $(1 - \chi_\Omega) f(\boldsymbol{x})$. Hence by restricting the function $\chi_\Omega f(\boldsymbol{x})$ to be non-zero only when $x_i = 0$, the same classification error will be achieved as for the case when $A_i = 0$.

It now remains to solve the problem for the case when $A_i > 0 \ \forall i > 2$. For this case, the shape of the decision surface will depend on the sign of A_1. $A_1 > 0$ produces an ellipsoid decision surface, $A_1 < 0$ produces a kind of hyperbolic decision surface, while $A_1 = 0$ yields a paraboloid. After deriving a solution for a single dimension, these three cases will be considered separately.

3.1. Solution to the one-dimensional problem

The single-dimension version of Eqs. (1)–(6) can be written as

$$\begin{aligned} \text{Maximise} \quad & \int_{x_1 > c} f(x_1)\, dx_1 \\ \text{Subject to} \quad & \int f(x_1)\, dx_1 = 1, \\ & \int x_1 f(x_1)\, dx_1 = 0, \\ & \int x_1^2 f(x_1)\, dx_1 = C_1, \\ & f(x_1) \geqslant 0. \end{aligned}$$

The solution to this problem is explained in detail in [3] as the first step in the derivation of a result for a multidimensional hyperplane decision surface. Since the arguments used in the derivation prove useful later in this paper, they are briefly explained here.

The first step in the derivation is to make a guess for the optimal $f(x_1)$ which satisfies all of the constraints. One intuitive guess is

$$f(x_1) = \frac{C_1}{C_1 + c^2}\, \delta(x_1 - c^+) + \frac{c^2}{C_1 + c^2}\, \delta\left(x_1 + \frac{C_1}{c}\right), \tag{8}$$

where c^+ is a number infinitesimally larger than c. It is important to note that the classification error $\int_{x_1 > c} f(x_1)\, dx_1 = C_1/(C_1 + c^2)$ can only be zero when $C_1 = 0$. From this observation, it is now shown that this guess was optimal. This is done by first assuming an optimal solution and showing that another solution exists having

an even larger classification error exists unless the solution satisfies the form of Eq. (8).

Suppose the optimal solution is $f^*(x_1) = \chi_\Omega f^*(x_1) + (1-\chi_\Omega)f^*(x_1)$, where $\Omega = x_1 : x_1 > c$. Then the optimal solution of the form $f(x_1) = \chi_\Omega f^*(x_1) + (1-\alpha)g(x_1)$ for $\alpha = \int_\Omega f^*(x_1)\,dx_1$ must have the same classification error as $f^*(x_1)$. Substitution of this form into the problem gives

$$\begin{aligned}
\text{Maximise} \quad & \int_{x_1>c} g(x_1)\,dx_1 \\
\text{Subject to} \quad & \int g(x_1)\,dx_1 = 1, \\
& \int x_1 g(x_1)\,dx_1 = \mu_g, \\
& \int (x_1-\mu_g)^2 g(x_1)\,dx_1 = C_g, \\
& g(x_1) \geqslant 0
\end{aligned}$$

which is a translated version of the original problem. As noted earlier, the classification error of $g(x_1)$ must greater than zero unless $C_g = 0$. Hence for the new solution,

$$\text{Classification error} = \int_\Omega f^*(x_1)\,dx_1 + \left(1 - \int_\Omega f^*(x_1)\,dx_1\right)\int_\Omega g(x_1)\,dx_1.$$

From the earlier observation, this will be larger than the optimal classification error (and thus lead to contradiction) unless $C_g = 0$. This is equivalent to the constraint that the function $(1-\chi_\Omega)f^*(x_1)$ must be expressible as a constant times a delta function, as in Eq. (8).

For the second part of the proof that Eq. (8) is the optimal, write $f^*(x_1) = (1-\alpha)\delta(x_1 + A) + (1-\chi_\Omega)f^*(x_1)$ where $\Omega = x_1 : x_1 > c$. Now consider the function

$$h(x_1) = \lim_{\varepsilon\to 0} \varepsilon^2 \delta\left(x_1 - \frac{B}{\varepsilon}\right) + (1-\alpha)\delta\left(x + \frac{\alpha}{1-\alpha}c\right) + \alpha\delta(x - c^+),$$

where the constant $B \geqslant 0$ is chosen to satisfy the variance constraint. The classification error is α for both $h(x_1)$ and $f^*(x_1)$, but the left part of the new distribution $\chi_\Omega h(x_1)$ contributes B to the variance. From the first part of the proof, if $B > 0$ then there exists a distribution function giving a higher classification error than $h(x_1)$, which gives a contradiction because it was already assumed that the classification error was a maximum. The only possibility left is that $B = 0$, which means that either $\alpha = 0$ (which is of course the minimum error) or $(1-\chi_\Omega)f^*(x_1) = \alpha\delta(x - c^+)$. This means that the optimal distribution $f^*(x_1)$ must be given by Eq. (8).

3.2. Paraboloid decision surface

In this subsection, a solution to Eqs. (1)–(6) is obtained for the special case when $A_1 = 0$ and $A_i > 0\ \forall i \geqslant 2$. To do this, it is first necessary to show that when the

maximum classification error is less than a 100 percent, the optimal solution will be of the form

$$f^*(\boldsymbol{x}) = (1-\alpha)\delta(\boldsymbol{x}+\boldsymbol{r}) + \chi_{\partial\Omega} f^*(\boldsymbol{x}),$$

where $\partial\Omega$ is the parabolic decision surface, and $\boldsymbol{r} \in \bar{\Omega}$ ($\boldsymbol{r}$ will in fact lie on the x_1 axis due to symmetry, so will have coordinates $[r_1, 0, 0, \ldots, 0]$). This is proved by employing arguments almost identical to those described in the previous subsection for the one-dimensional case. For brevity, these arguments have not been repeated here.

Substitution of the above form of the optimal solution into Eqs. (1)–(6), and using the fact that $x_1 = d_1 - \sum A_i x_i^2$ over $\partial\Omega$ gives

$$\text{Maximise} \quad \alpha = \int_{\partial\Omega} f(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\text{Subject to} \quad (1-\alpha)r_1 + \int_{\partial\Omega} \left(d_1 - \sum_{i\neq 1} A_i x_i^2 \right) f(\boldsymbol{x})\, d\boldsymbol{x} = 0, \tag{9}$$

$$(1-\alpha)r_1^2 + \int_{\partial\Omega} \left(d_1 - \sum_{i\neq 1} A_i x_i^2 \right)^2 f(\boldsymbol{x})\, d\boldsymbol{x} = C_1, \tag{10}$$

$$\int_{\partial\Omega} x_i^2 f(\boldsymbol{x})\, dx_i = C_i \quad \forall\ i \geqslant 2, \tag{11}$$

where the function $f(\boldsymbol{x}) \geqslant 0$ and is symmetric about $x_i \forall i \geqslant 2$. Substituting (11) into (10) gives

$$\int_{\partial\Omega} \left(\sum_{i\neq 1} A_i x_i^2 \right)^2 f(\boldsymbol{x})\, d\boldsymbol{x} = C_1 - (1-\alpha)r_1^2 - \alpha d_1^2 + 2d_1 \sum_{i\neq 1} A_i C_i. \tag{12}$$

Similarly, substitution of (11) into (9) gives

$$(1-\alpha)r_1 = \sum_{i\neq 1} A_i C_i - \alpha d_1$$

which when used in Eq. (12) gives

$$\int_{\partial\Omega} \left(\sum_{i\neq 1} A_i x_i^2 \right)^2 f(\boldsymbol{x})\, d\boldsymbol{x} = C_1 + \frac{1}{1-\alpha} \left[2d_1 \sum_{i\neq 1} A_i C_i - \alpha d_1^2 - \left(\sum_{i\neq 1} A_i C_i \right)^2 \right].$$

Hence all three constraints (9)–(11) may be replaced by the single constraint above. The problem is to maximise α, but as α tends to one in the above expression, the right-hand side tends to negative infinity from a single maximum, while the left-hand side is constrained to be positive. Hence α may be maximised by choosing the function $f(\boldsymbol{x})$ so that the left-hand side is minimised. Since $A_i > 0$,

from Hölder's inequality,

$$\int_{\partial\Omega} \sum_{i\neq 1} A_i x_i^2 f(\boldsymbol{x})\, d\boldsymbol{x} \leqslant \sqrt{\left(\int_{\partial\Omega} f(\boldsymbol{x})\, d\boldsymbol{x}\right) \int_{\partial\Omega} \left(\sum_{i\neq 1} A_i x_i^2\right)^2 f(\boldsymbol{x})\, d\boldsymbol{x}}$$

and so from Eq. (11) and the above constraint,

$$C_1 + \frac{1}{1-\alpha}\left[2d_1 \sum\nolimits_{i\neq 1} A_i C_i - \alpha d_1^2 - \left(\sum_{i\neq 1} A_i C_i\right)^2\right] \geqslant \frac{1}{\alpha}\left(\sum_{i\neq 1} A_i C_i\right)^2.$$

Equality occurs when $\chi_{\partial\Omega} f(\boldsymbol{x})$ is only non-zero for $\sum A_i x_i^2 = c$ where c is some constant (i.e. $\chi_{\partial\Omega} f(\boldsymbol{x})$ is distributed over an ellipse centred somewhere on the x_1-axis). For this case, the classification error is a maximum, and is given by

$$\alpha = \frac{1}{2(C_1 + d_1^2)}\left[C_1 + 2d_1 \sum_{i\neq 1} A_i C_i + S\right],$$

where

$$S = \sqrt{C_1^2 + 4d_1 C_1 \sum\nolimits_{i\neq 1} A_i C_i - 4C_1 \left(\sum_{i\neq 1} A_i C_i\right)^2}.$$

3.3. Ellipsoid and hyperboloid decision surfaces

As with the previous case of a paraboloid decision surface, the distribution giving the largest classification error for an ellipsoid ($A_1 > 0$) or a hyperboloid ($A_1 < 0$) decision surface will, for $A_i > 0\ \forall i \geqslant 2$, also be of the form

$$f^*(\boldsymbol{x}) = (1-\alpha)\delta(\boldsymbol{x} + \boldsymbol{r}) + \chi_{\partial\Omega} f^*(\boldsymbol{x}).$$

Here $\partial\Omega$ is the decision surface, and $\boldsymbol{r} \in \bar{\Omega}$ is a point on the x_1 axis. Substitution of this optimal solution into Eqs. (1)–(6), gives

$$\text{Maximise} \quad \alpha = \int_{\partial\Omega} f(\boldsymbol{x})\, d\boldsymbol{x}$$

$$\text{Subject to} \quad (1-\alpha) r_1 + \int_{\partial\Omega} x_1 f(\boldsymbol{x})\, d\boldsymbol{x} = 0, \tag{13}$$

$$(1-\alpha) r_1^2 + \int_{\partial\Omega} x_1^2 f(\boldsymbol{x})\, d\boldsymbol{x} = C_1, \tag{14}$$

$$\int_{\partial\Omega} x_i^2 f(\boldsymbol{x})\, dx_i = C_i \quad \forall i \geqslant 2. \tag{15}$$

Now the form of the decision surface is $\partial\Omega = \{\boldsymbol{x} : x_1 + A_1(x_1 - d)^2 + \sum_{i\neq 1} A_i x_i^2 = d\}$, which means

$$\int_{\partial\Omega} \left(x_1 + A_1(x_1 - d)^2 + \sum_{i\neq 1} A_i x_i^2 \right) f(\boldsymbol{x})\, d\boldsymbol{x} = \alpha d. \tag{16}$$

Substitution of Eqs. (13)–(15) thus gives

$$(2dA_1 - 1)(1 - \alpha)r_1 + A_1(C_1 - (1 - \alpha)r_1^2) + \sum_{i\neq 1} A_i C_i = \alpha(d - A_1 d^2)$$

which can be rearranged, yielding

$$\alpha = 1 - \frac{(d - A_1 d^2) - \sum A_i C_i}{d - (r_1 + A_1(r_1 - d)^2)}. \tag{17}$$

Since r_1 is the location along the x_1-axis of the correctly classified component of the distribution, then r_1 must be to the left of the decision surface. This means that the denominator must be positive. When the numerator is negative, α will appear to be greater than one, which is of course impossible. This is because for that case it is possible to find a distribution for which the classification error is 100 percent, so the original assumptions no longer hold.

When the numerator of the term for α in Eq. (17) is positive, α is maximised when the denominator is minimised. This occurs at $r_1 = d - 1/(2A_1)$ (when $A_1 > 0$, this will be the centre of the ellipsoid). This maximum however may not be achievable due to constraints (13)–(15).

Now the decision surface was chosen in such a way that every point on it satisfies $x_1 < d$. Hence Hölder's inequality may be applied to the following integral to give

$$\begin{aligned} \int_{\partial\Omega} (d - x_1) f(\boldsymbol{x})\, d\boldsymbol{x} &= \int_{\partial\Omega} \sqrt{(d - x_1)^2 f(\boldsymbol{x})} \sqrt{f(\boldsymbol{x})}\, d\boldsymbol{x} \\ &\leqslant \sqrt{\alpha \int_{\partial\Omega} (d - x_1)^2 f(\boldsymbol{x})\, d\boldsymbol{x}}. \end{aligned}$$

Rearranging the inequality and substituting Eqs. (13), (15) and (16) gives

$$\alpha A_1 \left[\alpha d(1 - A_1 d) + (1 - \alpha) r_1 (1 - 2dA_1) - \sum\nolimits_{i\neq 1} A_i C_i \right] - A_1^2 (1 - \alpha)^2 r_1^2 \geqslant 0.$$

Then using the expression for α in Eq. (17) gives

$$\left(d(A_1 d - 1) + \sum_{i\neq 1} A_i C_i \right) r_1^2 + C_1 (2dA_1 - 1) r_1 + C_1 \sum_i A_i C_i \geqslant 0. \tag{18}$$

It has already been pointed out that when the numerator in Eq. (17) is negative that the maximum classification error will be 100 percent. In the remaining cases, the first term in the above quadratic in r_1 will be negative, and so r_1 must lie somewhere between the two roots. When $r_1 = d - 1/(2A_1)$ lies between the two roots, this will

correspond to the optimal solution, otherwise r_1 will be one of the two quadratic roots. For either case, the maximum classification error will be given by Eq. (17).

4. The non-symmetric problem

The symmetrical problem described in Section 2 may be directly applicable in some specialised applications, but in most situations there is some degree of asymmetry. As with the symmetrical problem, the error may still be maximised with respect to each class distribution separately. Also, it can be assumed without loss of generality that the feature space can be rotated so that the quadratic discriminant surface is symmetric about $x_i = 0$ for $i = 2 \ldots N$, so the matrix $\mathbf{A}$ is still diagonal. The mean, however, will no longer be on the x_1-axis, and the covariance matrix will no longer be diagonal.

Suppose $f^*(\mathbf{x})$ is the unknown distribution, with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, giving the largest classification error for a particular quadratic discriminant. Then all combinations of $N-1$ reflections of the form $R_i : [x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_N] \to [x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_N]$ for $i = 2 \ldots N$ can be used to construct new distributions $G_i(\mathbf{x})$ for $i = 1 \ldots 2^{N-1}$ from $f^*(\mathbf{x})$. Each new distribution will have exactly the same classification error as the original distribution, due to the symmetry of the decision surface. Therefore an equal mixture of these distributions, given by

$$H(\mathbf{x}) = \frac{1}{2^{N-1}} \sum_{i=1}^{2^{N-1}} G_i(\mathbf{x})$$

will also have an identical classification error, but will have mean $\boldsymbol{\mu}' = [\mu_1, 0, 0, \ldots, 0]$ and a diagonal covariance matrix with elements

$$C_i' = \begin{cases} C_1 & \text{for } i = 1, \\ C_i + \mu_i & \text{for } i = 2 \ldots N. \end{cases}$$

An upper bound for the classification error for $H(\mathbf{x})$ can be found from Section 3, so this will also be an upper bound for the classification error for $f^*(\mathbf{x})$. There is, however, no guarantee that this will be a strict bound.

5. Numerical results

In this section, two symmetric problems are discussed. The first of these is a simple two-dimensional example which is used to demonstrate the effect of changing the eccentricity of the quadratic decision surface on the upper bound of the classification error. Classification results obtained using the quadratic discriminant function are also compared to those obtained for a linear discriminant. The second example compares linear and quadratic discriminants for a problem with higher dimensionality.

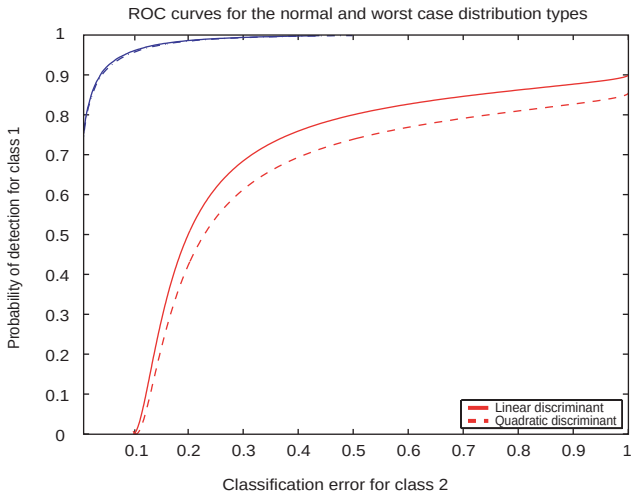

Fig. 1. ROC curves for linear and quadratic discriminants, assuming either normal or the worst possible distribution type.

In the first example, consider two classes in two dimensions having means of $\boldsymbol{\mu}_1 = [0,0]$, $\boldsymbol{\mu}_2 = [3,0]$ and covariances

$$\boldsymbol{C}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

In this case, the second variable has exactly the same mean for both classes, but the variance differs. This sort of scenario is occasionally used to argue for the superiority of non-linear discriminants, since linear discriminants will be unable to make use of the extra class separating information contained in the second variable.

Fig. 1 shows four receiver operating characteristic (ROC) curves. The dashed curves are obtained using a quadratic discriminant decision surface of the form $x_2^2 + 12x_1 = \textit{constant}$, which should be optimal if the classes are normal. The solid lines in the figure were obtained from the family of linear discriminants $x_1 = \textit{constant}$. The two thin almost overlapping lines in the top left corner show the classifier performance for normally distributed classes. In this case, since the classes are reasonably well separated, the linear discriminant performs almost identically to the optimal quadratic discriminant. On the other hand, the remaining thick curves show that for the worst possible type of class distributions, the classifier performance is substantially improved by using a linear discriminant.

Fig. 2 shows the classification error produced by the decision surface $x_1 + A_1(x_1 - 1.625)^2 + x_2^2/12 = 1.625$ for various class distributions. The decision surface is an ellipse for $A_1 > 0$, a hyperbola for $A_1 < 0$ and a parabola for $A_1 = 0$. The continuity of the curves verifies that the solution in Section 3.3 converges to the solution for a parabolic decision surface from Section 3.2 as $A_1 \to 0$.

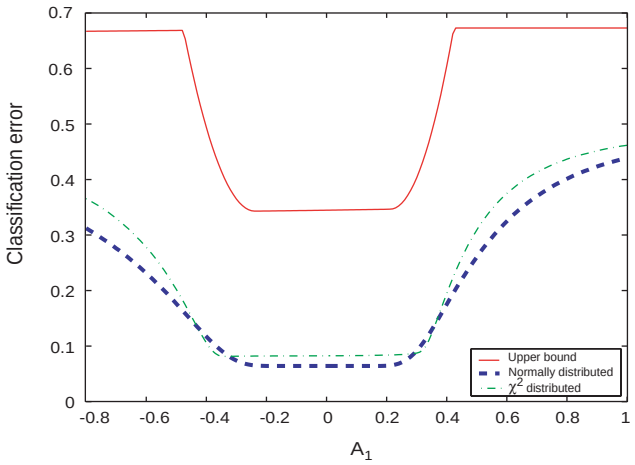

Fig. 2. The change in classifier performance as the decision surface changes.

The solid curve in Fig. 2 is the upper bound on the classification error, while the two lower curves assume that the class distributions are normal or based on the χ^2 distribution (this distribution was simulated by generating two independent χ^2 distributed variables with parameter five and then translating and scaling to give the appropriate means and variances). For this case, the upper bound is extremely loose. This is because the distributions giving the worst possible classification error have an extremely low kurtosis which is not found in the majority of applications. It is not possible to produce a stricter bound without using extra assumptions about the class distributions such as being unimodal or having some lower bound on the kurtosis. This problem, however, is outside the scope of the current paper.

For the second example, suppose that the two classes have means $\boldsymbol{\mu}_1 = [-1,-1,-1,-1]$, $\boldsymbol{\mu}_2 = [1,1,1,1]$ and covariances

$$\boldsymbol{C}_1 = \begin{bmatrix} 2 & -0.5 & -0.5 & 0 \\ -0.5 & 2 & 0 & -0.5 \\ -0.5 & 0 & 2 & -0.5 \\ 0 & -0.5 & -0.5 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}_2 = \begin{bmatrix} 1.75 & 0.75 & 0.25 & 0.25 \\ 0.75 & 1.75 & 0.25 & 0.25 \\ 0.25 & 0.25 & 1.75 & 0.75 \\ 0.25 & 0.25 & 0.75 & 1.75 \end{bmatrix}.$$

These means and covariances have been chosen so that after the appropriate rotation an scaling, they satisfy the constraints described in Section 2.

Fig. 3 shows ROC curves for linear and quadratic discriminants (solid or dashed lines) under the assumptions of normal classes or the worst possible class distributions (thin or thick lines). As with the previous two-dimensional example, there is little improvement in performance obtained by using the maximum likelihood solution instead of a linear discriminant for normal classes. The linear discriminant shows a much more substantial improvement in its lower performance

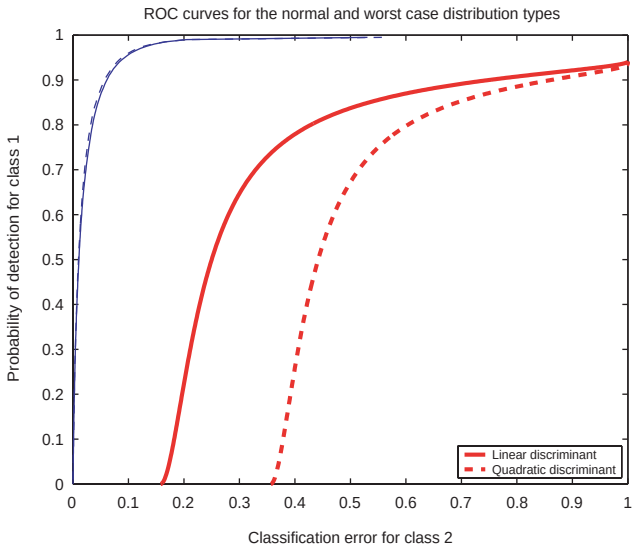

Fig. 3. ROC curves for linear and quadratic discriminants from example two.

bound however. This indicates that the linear classifier is much more robust to non-normality, especially in problems with high dimensionality.

6. Discussion and conclusions

In this paper, a theoretical upper bound has been found for the classification error for a quadratic discriminant surface. The bound is strict when the problem satisfies the symmetry requirements of Section 2, and for this case, the quadratic discriminant which minimises the worst possible classification error is the linear discriminant. Numerical results indicate that for well-separated Gaussian distributions, the use of a maximum likelihood discriminant only gives a small improvement in the error rates when compared with a linear discriminant. On the other hand, the linear discriminant gives a significant improvement to the upper bound on the error in the case when the classes are non-Gaussian, which makes it much more robust.

From this point, there are couple of further points worth considering. First, while it would be worthwhile to find a strict error bound for the non-symmetric problem, a problem of more practical significance is the effect of estimation error. In practice, the means and covariances of each class will not be known exactly, but must be estimated from a set of samples. One method for dealing with the estimation error, as described by Lanckriet et al. [4], is to find a confidence interval for the mean and covariance estimates, and then choose the case which gives the worst error bound from the set of feasible parameters. For this case, the worst scenario occurs for larger variance values that are closer to the decision surface, and the numerical value for the error bound will be as described earlier.

The final point concerns the usefulness of worst case type bounds in general. Similar worst case bounds exist for estimation of computation speed of various heuristic methods for solving combinatorial problems. More recently in this area however, there has been a shift away from measuring performance by the worst possible case toward an average performance. This is because in combinatorics, the worst case scenario happens extremely rarely, and one is unlikely to encounter such an event in practice. Similarly, in this paper, the form of the distribution which gives the worst possible classification error is not one that is likely to appear in practical problems. It would seem that a similar average scenario formulation would also be applicable for measuring the performance of classifiers, but the author is not aware of any useful models for an "average distribution".

References

[1] J. Broffit, W. Clarke, P. Lachenbruch, The effect of Huberizing and trimming on the quadratic discriminant function, Comm. Statist. Theory A9 (1980) 13–25.
[2] W. Clarke, P. Lachenbruch, J. Broffitt, How non-normality affects the quadratic discriminant function, Comm. Statist. Theory A8 (1979) 1285–1301.
[3] T. Cooke, M. Peake, The optimal classification using a linear discriminant for two point classes having known mean and covariance, J. Multivariate Anal. 82 (2) (2002) 379–394.
[4] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, M. Jordan, A robust minimax approach to classification, J. Mach. Learning Res. 3 (2002) 555–582.
[5] D. Moore, Evaluation of five discriminant procedures for binary variables, J. Amer. Statist. Assoc. 68 (1973) 399–404.