On optimal reject rules and ROC curves ⋆

Carla M. Santos-Pereira [a] and Ana M. Pires [b]

[a] *Universidade Portucalense Infante D. Henrique, Oporto, Portugal and Centre for Mathematics and its Applications (CEMAT), IST, Technical University of Lisbon, Portugal.*

[b] *Department of Mathematics and Centre for Mathematics and its Applications (CEMAT), IST, Technical University of Lisbon, Avenida Rovisco Pais - 1049-001, Lisboa, Portugal.*

Abstract

In this paper we make the connection between two approaches for supervised classification with a rejection option. The first approach is due to Tortorella and is based on ROC curves and the second is a generalisation of Chow's optimal rule.

Key words: Chow's rejection rule, Derivative of ROC curves, Supervised classification, Optimal decision-rule, Rejection threshold

⋆ This work was supported by *Programa Operacional "Ciência, Tecnologia, Inovação"* (POCTI) of the *Fundação para a Ciência e a Tecnologia* (FCT), cofinanced by the European Community fund FEDER.

Email address: carla@upt.pt (Carla M. Santos-Pereira).

1 Introduction

In a supervised classification problem the aim is to build a decision rule according to which a new object will be assigned to one of c predefined classes. In order to build this decision rule we need a training set of patterns correctly classified. For a comprehensive treatment of this subject see, e.g., Ripley (1996) or Webb (1999). When there is some uncertainty it may be better to introduce a rejection option to avoid high error rates and/or to reduce the overall costs. Thus, finding a reject rule which achieves the best trade-off between error rate and reject rate is undoubtedly of practical interest in real applications.

The problem of defining an optimal reject option has been tackled by Chow (1957, 1970). In the first article he formulated an optimal decision rule. In the second he derived a general relation between the error and reject probabilities and presented the optimum error-reject curve. More specifically Chow's rule minimises the error rate for a given reject rate, or vice versa, and consists of rejecting an object if the highest a posteriori probability is lower than some threshold $1-t$ (rejection threshold), $t \in [0, 1-1/c]$. Chow's rule is optimal in the sense that for some reject rate specified by the threshold t, no other rule can yield a lower error rate. Chow considered a particular cost function. Let $C(i|j)$ be the cost incurred by classifying in G_i when the true class is G_j. Then

$$C(i|j) = \begin{cases} 0 \text{ if } i = j \\ 1 \text{ if } i \neq j \quad i, j = 1, 2, \ldots, c \\ t \text{ if } i = I \end{cases} , \tag{1}$$

where I is the class of rejection (*Indecision* class). Moreover, it is known (Chow, 1957) that the optimum decision rule is also a minimum-risk rule if the cost function is uniform within each class of decisions. In this case, the rejection threshold is related to the other costs as follows, $t = (C_r - C_c)/(C_e - C_c)$, where C_e, C_r and C_c denote the costs of making an error, rejection and correct recognition, respectively (Chow, 1970).

The rejection option is desirable in those applications where it may be more convenient to withhold making a decision than to make a wrong decision. Furthermore, the costs may not be symmetrical, because the consequences of different errors are usually not equivalent and depend on the nature of the particular application. As an example, in the case of medical diagnosis a false negative outcome can be much more costly than a false positive. Similarly, in the detection of fraud, the cost of missing a particular case of fraud will depend on the amount of money involved. In this paper we consider the generalisation of Chow's rule (optimum error-reject tradeoff) for different class conditional costs of misclassification and correct classification. That is, for the general cost function

$$C(i|j) = \begin{cases} C_{ii} \text{ if } i = j \\ C_{ji} \text{ if } i \neq j \quad i, j = 1, 2, \ldots, c \\ C_r \text{ if } i = I \end{cases} . \tag{2}$$

An optimal decision rule based either on a posteriori probabilities or on conditional densities can then be obtained. This rule, although based on classical Bayes decision theory, is in fact a generalisation of Chow's original rule because of the more general cost structure. As far as we know it has not been

widely exploited or used in practice (there is a brief reference in Webb, 1999, p. 13). For this optimal rule we derive the particular expressions for the two class case.

Tortorella (2000) also presented an optimal reject rule for binary classifiers based on the *Receiver Operating Characteristic curve* (ROC curve). The rule is optimal since it maximizes a classification utility function, defined on the basis of classification and error costs particular for the application at hand.

The aim of our paper is to make the connection between the classification approach presented by Tortorella and the generalisation of Chow's rule for the two class case, and prove that these two approaches are theoretically equivalent.

This paper is organized as follows. Section 2 discusses the generalisation of Chow's optimal reject rule. The connection between the two approaches is presented in Section 3. To exemplify the methods an application is discussed in Section 4. The paper concludes with summary and discussion in Section 5.

2 Generalisation of Chow's rule

Consider a pattern with feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T \in \mathcal{X}$, a set of $c+1$ classes G_1, G_2,..., G_c, I (*Indecision* class) and a training set $(\mathbf{x}_1, G_{i_1})$, $(\mathbf{x}_2, G_{i_2})$, ..., $(\mathbf{x}_n, G_{i_n})$, where $i_k = j$ if and only if the $k-$observation belongs to G_j, $j = 1, \ldots, c$, $k = 1, \ldots, n$. A general classifier is an application $\mathcal{X} \rightarrow \{G_1, G_2, \ldots, G_c, I\}$ and the classification task is to estimate the true class of a new observation characterised by a feature vector $\mathbf{x}_{new}$. Let π_i denote the a priori probability of G_i and $f_i(\mathbf{x})$ the (conditional) density of $\mathbf{x}$ given G_i. The

a posteriori probability of G_i given $\mathbf{x}$ is denoted by $P(G_i|\mathbf{x})$ and is related with the previous quantities by

$$P(G_i|\mathbf{x}) = \frac{\pi_i f_i(\mathbf{x})}{\sum_{j=1}^{c} \pi_j f_j(\mathbf{x})}.$$

Chow's rule is optimal since it is a minimum-risk rule if the cost function is uniform within each class of decisions i.e., if no distinction is made among the errors (Chow, 1957).

Denote by $D_1, D_2, \ldots, D_c, D_I$ the decision regions associated to classes $G_1, G_2, \ldots, G_c, I$. Chow's rule partitions the feature space, $\mathcal{X}$, into two disjoint sets: D_I, a reject region, and $A = \bigcup_{i=1}^{c} D_i$, an acceptance region of the decision rule. The classification rule which minimizes the risk under (1) can be stated as:

$$\begin{array}{ll} \mathbf{x} \in A & \text{if } \max_i P(G_i|\mathbf{x}) \geq 1 - t \\ & \qquad\qquad i = 1, 2, \ldots c. \\ \mathbf{x} \in D_I & \text{if } \max_i P(G_i|\mathbf{x}) < 1 - t \end{array}$$

Consider now the general cost function (2). The conditional risk (or expected loss) of assigning an object $\mathbf{x}$ to class G_i is defined as

$$C^i(\mathbf{x}) = \sum_{j=1}^{c} C(i|j) P(G_j|\mathbf{x}) = \sum_{j=1}^{c} C(i|j) \frac{\pi_j f_j(\mathbf{x})}{f(\mathbf{x})},$$

where $f(\mathbf{x}) = \sum_{k=1}^{c} \pi_k f_k(\mathbf{x})$. The average risk over region D_i is

$$C^i = \int_{D_i} C^i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}.$$

So we can write that

$$A = \{\mathbf{x} : \min_i C^i(\mathbf{x}) \leq C_r\}$$

$$D_I = \{\mathbf{x} : \min_i C^i(\mathbf{x}) > C_r\}$$

and the classification rule that minimizes the Bayes risk under (2) is given by

$$\mathbf{x} \in D_i \quad \text{if} \quad C^i(\mathbf{x}) < C^k(\mathbf{x}), \ \ \forall_{k \neq i} \ \text{ and } \ C^i(\mathbf{x}) \leq C_r, \quad i = 1, 2, \ldots, c,$$

$$\mathbf{x} \in D_I \quad \text{if} \quad \min_i C^i(\mathbf{x}) > C_r, \quad i = 1, 2, \ldots, c, \ .$$

In a two-class problem, it is straightforward to verify that, in practice, the classification rule converts to

$$\begin{aligned}
\text{Classify } \ \mathbf{x} \ \text{ in } \ G_1 \quad &\text{if} \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{C(1|2) - C_r}{C_r - C(1|1)} \times \frac{\pi_2}{\pi_1} = t_1 \\
\text{Classify } \ \mathbf{x} \ \text{ in } \ G_2 \quad &\text{if} \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \leq \frac{C_r - C(2|2)}{C(2|1) - C_r} \times \frac{\pi_2}{\pi_1} = t_2 \\
\text{Reject } \ \mathbf{x} \quad &\text{if} \quad t_2 < \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < t_1 \ .
\end{aligned}$$

The same rule can equivalently be stated in terms of a posteriori probabilities:

$$\begin{aligned}
\text{Classify } \ \mathbf{x} \ \text{ in } \ G_1 \quad &\text{if} \quad \frac{P(G_1|\mathbf{x})}{1 - P(G_1|\mathbf{x})} \geq \frac{C(1|2) - C_r}{C_r - C(1|1)} = \frac{\pi_1}{\pi_2} \times t_1 \\
\text{Classify } \ \mathbf{x} \ \text{ in } \ G_2 \quad &\text{if} \quad \frac{P(G_1|\mathbf{x})}{1 - P(G_1|\mathbf{x})} \leq \frac{C_r - C(2|2)}{C(2|1) - C_r} = \frac{\pi_1}{\pi_2} \times t_2 \\
\text{Reject } \ \mathbf{x} \quad &\text{if} \quad \frac{\pi_1}{\pi_2} \times t_2 < \frac{P(G_1|\mathbf{x})}{1 - P(G_1|\mathbf{x})} < \frac{\pi_1}{\pi_2} \times t_1 \ .
\end{aligned}$$

It is also important to note that, for certain combination of costs, the rejection option may not be activated. It is easy to verify that a necessary, but not sufficient, condition to activate the rejection is that $C_r < C(1|2)$ and $C_r < C(2|1)$. Some of the cost combinations considered in the example discussed in Section 4 further illustrate this point.

3 Relation with the ROC curve

As anticipated in the introduction, the aim of this section is to establish the connection between the rule presented in Section 2 and the rule based on the ROC curve, due to Tortorella (2000).

Tortorella assumes that the classifier provides for each feature vector, $\mathbf{x}$, a value x in the range $[0, 1]$ which is a confidence degree that the pattern belongs to class P (positive class). In this section, as in Section 2, we will adopt G_1 instead of P.

To build the classification rule (without rejection) it is necessary to fix a threshold u (which Tortorella calls "confidence threshold" and denotes by t) such that a new observation is assigned to class G_1 if $x > u$. Obviously, Tortorella is assuming that the classifier provides an estimate of the a posteriori probability of class G_1. Since the aim is to establish a theoretical connection between the two approaches, let us assume that the classifier provides, not an estimate, but the true a posteriori probability of class G_1, that is $x(\mathbf{x}) = P(G_1|\mathbf{x})$.

Tortorella considered not one but two thresholds to cope with a rejection option: $\mathbf{x}$ is assigned to class G_2 if $0 \leq x < u_1$, to class G_1 if $u_2 < x \leq 1$ and it is rejected if $u_1 \leq x \leq u_2$. Tortorella considers earnings and maximizes a classification utility function. Chow takes costs and minimizes the risk. At this point there is no doubt of the correspondence between these two approaches. According to Chow's notation, and ours, in the case of binary classifiers such costs can be specified by $C_{11} = C(1|1)$ and $C_{22} = C(2|2)$ (costs of correct classifications, therefore negative); $C_{21} = C(1|2)$ and $C_{12} = C(2|1)$ (costs of misclassifications, therefore positive) and finally a cost of rejection, $C_r > 0$.

The theoretical ROC curve is defined as a function of the threshold u. For each $u \in [0,1]$, the true positive rate, TPR, and the false positive rate, FPR, are defined as,

$$TPR(u) = \int_u^1 f_P(x)dx \quad \text{and} \quad FPR(u) = \int_u^1 f_N(x)dx, \tag{3}$$

where $f_P(x)$ and $f_N(x)$ are the density functions of the univariate random variable $x(\mathbf{x}) = P(G_1|\mathbf{x})$ (a function of the random vector $\mathbf{x}$) conditional with respect to each class. The ROC curve is the set of points given by the pairs $\{(FPR(u), TPR(u)), u \in [0,1]\}$. Figure 1 shows an example of a theoretical ROC curve (with a tangent line). Note that although, in general, $u \in [0,1]$, there are cases for which the range of $P(G_1|\mathbf{x})$ is a strict subset of this interval, say $[u_{min}, u_{max}]$. In these cases the points of the ROC curve for $u \in [0, u_{min}]$ all collapse at the upper right point of the ROC curve and the points for $u \in [u_{max}, 1]$ are all coincident at the lower left point of the ROC curve.

The rates defined in (3) could be written in a different form, directly, without computing the densities $f_P(x)$ and $f_N(x)$. Note that

$$x > u \Leftrightarrow P(G_1|\mathbf{x}) > u \Leftrightarrow \frac{\pi_1 f_1(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})} > u \Leftrightarrow \mathbf{x} \in R_u = D_1(u),$$

where $R_u \subset \mathbb{R}^p$ symbolize, for each $u \in [0,1]$, the usual classification region (denoted previously as D_1) for class G_1 in the space of the original variables, now represented in this way to emphasise the dependence on u. Note also that $u = 0 \Rightarrow R_u = \mathbb{R}^p$, $u = 1 \Rightarrow R_u = \emptyset$ and $R_{u_1} \supset R_{u_2}$ for every $u_1, u_2 : 0 \leq u_1 < u_2 \leq 1$. As usual, suppose that the set $\{\mathbf{x} : P(G_1|\mathbf{x}) = u\}$ has null probability measure for any of the classes and for every $u : u_{min} < u < u_{max}$.

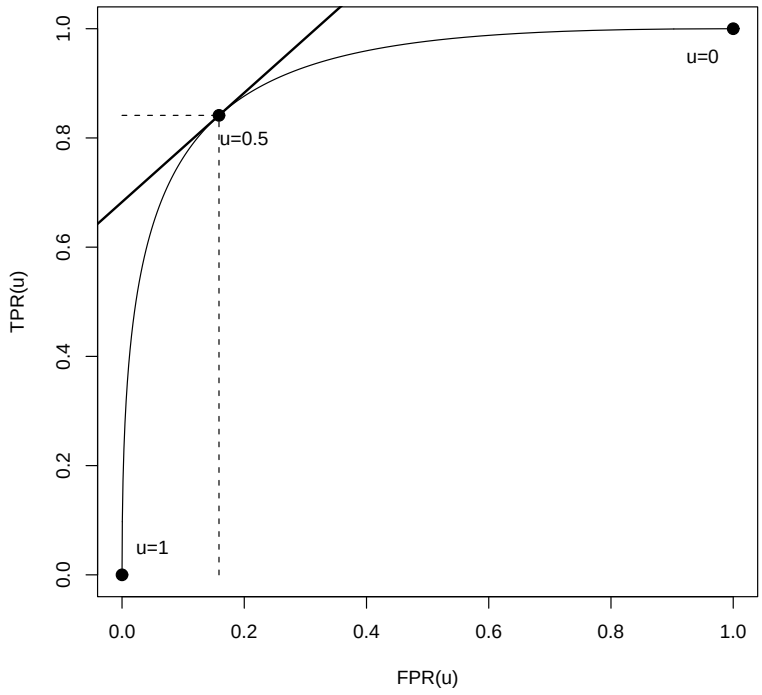

Fig. 1. Example of a theoretical ROC curve ($X|G_1 \sim N(0,1)$, $X|G_2 \sim N(2,1)$, $\pi_1 = \pi_2 = 1/2$), with the tangent line at $u = 1/2$.

Then we have

$$TPR(u) = \int_{R_u} f_1(\mathbf{x})d\mathbf{x} \quad \text{and} \quad FPR(u) = \int_{R_u} f_2(\mathbf{x})d\mathbf{x}. \tag{4}$$

On the other hand, from (3)

$$f_P(u) = -\frac{d}{du}TPR(u) \quad \text{and} \quad f_N(u) = -\frac{d}{du}FPR(u).$$

What Tortorella showed is that the profit is maximized (or equivalently, the risk minimized) if we choose thresholds u_1 and u_2 such that the derivative of the ROC curve at those thresholds is, respectively,

$$m_1 = \frac{\pi_2}{\pi_1} \times \frac{C_r - C_{22}}{C_{12} - C_r} \quad \text{and} \quad m_2 = \frac{\pi_2}{\pi_1} \times \frac{C_{21} - C_r}{C_r - C_{11}}. \tag{5}$$

Afterwards Tortorella proposed a geometrical procedure to compute the thresh-

olds for empirical ROC curves. On the other hand, as we saw in Section 2

$$\mathbf{x} \in D_I \text{ if } \frac{\pi_2}{\pi_1} \times \frac{C_r - C_{22}}{C_{12} - C_r} < \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{\pi_2}{\pi_1} \times \frac{C_{21} - C_r}{C_r - C_{11}},$$

but

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq m \Leftrightarrow P(G_1|\mathbf{x}) \geq u, \text{ iff } m = \frac{\pi_2}{\pi_1} \times \frac{u}{1-u},$$

therefore, we can conclude that the two approaches are equivalent, in a theoretical way, if and only if the derivative of the ROC curve is

$$\frac{\pi_2}{\pi_1} \times \frac{u}{1-u}.$$

This is shown in the following proposition.

Proposition: Consider a random vector $\mathbf{x}$ with distinct probability distributions in classes G_1 and G_2, specified by their densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, with a priori probabilities π_1 and π_2. Then the derivative of the ROC curve, $\{(FPR(u), TPR(u)), u \in [0,1]\}$, at the point specified by u_0, is given by

$$\left.\frac{dTPR(u)}{dFPR(u)}\right|_{u=u_0} = \frac{\pi_2}{\pi_1} \times \frac{u_0}{1-u_0}, \quad u_{min} < u_0 < u_{max} \ ,$$

where $u_{min} = \min_{\mathbf{x}} P(G_1|\mathbf{x})$ and $u_{max} = \max_{\mathbf{x}} P(G_1|\mathbf{x})$.

Proof: First note that

$$\left.\frac{dTPR(u)}{dFPR(u)}\right|_{u=u_0} = \frac{\left.\frac{dTPR(u)}{du}\right|_{u=u_0}}{\left.\frac{dTPR(u)}{du}\right|_{u=u_0}} = \frac{\lim\limits_{h\to 0} \frac{TPR(u_0+h) - TPR(u_0)}{h}}{\lim\limits_{h\to 0} \frac{FPR(u_0+h) - FPR(u_0)}{h}} =$$

$$= \frac{\lim\limits_{h\to 0} \frac{\int_{R_{u_0+h}} f_1(\mathbf{x})d\mathbf{x} - \int_{R_{u_0}} f_1(\mathbf{x})d\mathbf{x}}{h}}{\lim\limits_{h\to 0} \frac{\int_{R_{u_0+h}} f_2(\mathbf{x})d\mathbf{x} - \int_{R_{u_0}} f_2(\mathbf{x})d\mathbf{x}}{h}} = \lim_{h\to 0} \frac{\int\limits_{R_{u_0+h}} f_1(\mathbf{x})d\mathbf{x} - \int\limits_{R_{u_0}} f_1(\mathbf{x})d\mathbf{x}}{\int\limits_{R_{u_0+h}} f_2(\mathbf{x})d\mathbf{x} - \int\limits_{R_{u_0}} f_2(\mathbf{x})d\mathbf{x}}.$$

Consider now separately the cases $h > 0$ and $h < 0$. If $h > 0$, $R_{u_0+h} \subset R_{u_0}$, so we can write

$$\int\limits_{R_{u_0+h}} f_i(\mathbf{x})d\mathbf{x} - \int\limits_{R_{u_0}} f_i(\mathbf{x})d\mathbf{x} = -\int\limits_{R_{u_0}\backslash R_{u_0+h}} f_i(\mathbf{x})d\mathbf{x}\,,\ i = 1, 2.$$

On the other hand if $h < 0$, $R_{u_0+h} \supset R_{u_0}$,

$$\int\limits_{R_{u_0+h}} f_i(\mathbf{x})d\mathbf{x} - \int\limits_{R_{u_0}} f_i(\mathbf{x})d\mathbf{x} = \int\limits_{R_{u_0+h}\backslash R_{u_0}} f_i(\mathbf{x})d\mathbf{x}\,,\ i = 1, 2.$$

Therefore, the left and right derivatives are given by

$$\left.\frac{dTPR(u)}{dFPR(u)}\right|_{u=u_0^+} = \lim_{h\to 0^+} \frac{\int_{R_{u_0}\backslash R_{u_0+h}} f_1(\mathbf{x})d\mathbf{x}}{\int_{R_{u_0}\backslash R_{u_0+h}} f_2(\mathbf{x})d\mathbf{x}}\,,$$

$$\left.\frac{dTPR(u)}{dFPR(u)}\right|_{u=u_0^-} = \lim_{h\to 0^-} \frac{\int_{R_{u_0+h}\backslash R_{u_0}} f_1(\mathbf{x})d\mathbf{x}}{\int_{R_{u_0+h}\backslash R_{u_0}} f_2(\mathbf{x})d\mathbf{x}}\,.$$

Moreover,

$$\text{if } \mathbf{x} \in R_{u_0}\backslash R_{u_0+h} \ \ (h > 0),\ \ u_0 \le P(G_1|\mathbf{x}) \le u_0 + h\,,$$

$$\text{if } \mathbf{x} \in R_{u_0+h}\backslash R_{u_0} \ \ (h < 0),\ \ u_0 + h \le P(G_1|\mathbf{x}) \le u_0\,,$$

and as $h \to 0^+$ ($h \to 0^-$), the region $R_{u_0}\backslash R_{u_0+h}$ ($R_{u_0+h}\backslash R_{u_0}$) "shrinks" and is such that $P(G_1|\mathbf{x})$ tends to u_0 and $P(G_2|\mathbf{x}) = 1 - P(G_1|\mathbf{x})$ tends to $1 - u_0$. Thus, if $h \to 0^+$,

$$\lim_{h\to 0^+} \frac{\int_{R_{u_0}\backslash R_{u_0+h}} f_1(\mathbf{x})d\mathbf{x}}{\int_{R_{u_0}\backslash R_{u_0+h}} f_2(\mathbf{x})d\mathbf{x}} = \lim_{h\to 0^+} \frac{\int_{R_{u_0}\backslash R_{u_0+h}} \frac{f(\mathbf{x})P(G_1|\mathbf{x})}{\pi_1}d\mathbf{x}}{\int_{R_{u_0}\backslash R_{u_0+h}} \frac{f(\mathbf{x})P(G_2|\mathbf{x})}{\pi_2}d\mathbf{x}} = \frac{\pi_2}{\pi_1} \times \frac{u_0}{1-u_0},$$

and similarly if $h \to 0^-$.

Remark 1: This proposition can be related to a well known result in ROC analysis (see, e.g., van Trees, 1968, Vol. 1, p. 44, Property 3).

Remark 2: Although we have showed the theoretical equivalence of the two approaches, the results may differ in a real application due to different estimation procedures.

Remark 3: Other proofs could be given for particular cases (for example, any univariate distributions, normal distributions with different means and equal covariances, normal distributions with equal covariances and different means, etc.). To fix ideas we consider a simple example.

Example: Consider two bivariate normal distributions with means $\boldsymbol{\mu}_1 = (0,0)^T$ and $\boldsymbol{\mu}_2 = (2,0)^T$, respectively, and equal covariance matrices $\boldsymbol{\Sigma} = \mathbf{I}$. Then we have

$$P(G_1|\mathbf{x}) \geq t \Leftrightarrow \frac{\pi_1}{\pi_1 + \pi_2 \exp(2x_1 - 2)} \geq t \Leftrightarrow x_1 \leq \frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] + 1\,.$$

So we can write

$$TPR(t) = \int_{R_t} f_1(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right)+\ln\left(\frac{1-t}{t}\right)\right]+1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx =$$

$$= \Phi\left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] + 1\right\}\,.$$

Analogously,

$$FPR(t) = \int_{R_t} f_2(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right)+\ln\left(\frac{1-t}{t}\right)\right]+1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-2)^2}{2}\right) dx =$$

$$= \Phi\left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] - 1\right\}.$$

In this case it is straightforward to verify the proposition:

$$\frac{dTPR(t)}{dFPR(t)} = \frac{\frac{dTPR(t)}{dt}}{\frac{dFPR(t)}{dt}} =$$

$$= \frac{\varphi\left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] + 1\right\} \times \left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] + 1\right\}'}{\varphi\left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] - 1\right\} \times \left\{\frac{1}{2}\left[\ln\left(\frac{\pi_1}{\pi_2}\right) + \ln\left(\frac{1-t}{t}\right)\right] - 1\right\}'} = \frac{\pi_2}{\pi_1} \times \frac{t}{1-t},$$

where $\Phi(x)$ is the cumulative distribution function of the standard univariate normal distribution and $\varphi(x) = \exp(-x^2/2)$ is its derivative.

4 Application

In view of Remark 2 of the previous section it is important to compare the results obtained with the application of the methods just discussed to real data. For that purpose we have chosen the well known *Pima Indian Diabetes* dataset, also used by Tortorella (2000). The dataset is available on the web from the UCI Machine Learning Repository and contains 768 labelled cases (500 healthy, G_2 or negative class, N, and 268 diabetes, G_1 or positive class, P) with 8 features. The a priori probability estimates are therefore $\hat{\pi}_1 = 0.35$ and $\hat{\pi}_2 = 0.65$.

We have chosen three classifiers: linear discriminant analysis (LDA, that is, the Bayes classifier assuming multivariate normal distributions with a common covariance matrix and parameters estimated by the sample means and pooled sample covariance matrix of the cases in the training set), logistic dis-

Table 1

The combinations of costs.

	CFN	CFP	CTN	CTP	CR
Case	$C(2\|1)$	$C(1\|2)$	$C(2\|2)$	$C(1\|1)$	C_r
a	50	25	-200	-400	12.5
b	50	25	-100	-200	12.5
c	50	25	-50	-100	12.5
d	50	25	-25	-50	12.5
e	100	50	-25	-50	12.5
f	200	100	-25	-50	12.5
g	400	200	-25	-50	12.5

crimination (LD, see e.g. Anderson, 1972, 1982) and a Multi Layer Perceptron (MLP) with 8 input units, 4 hidden units and 1 output unit (implemented with the `nnet` library from S-Plus 2000, with `rang`=0.1, `decay`=0.01 and 20000 iterations, for details on this implementation see Venables and Ripley, 2002). Each of these classifiers provide estimates of the posterior probability of class membership, $\hat{P}(G_1|\mathbf{x})$ and $\hat{P}(G_2|\mathbf{x}) = 1 - \hat{P}(G_1|\mathbf{x})$.

As in Tortorella (2000) we have used 614 ($\simeq$ 80%) randomly selected observations for training, 77 ($\simeq$ 10%) for evaluating the ROC curves and the remaining 77 as test set. We have also adopted the seven combinations of costs chosen by Tortorella (2000) and reproduced in Table 1.

When the rejection option is not considered all the cost combinations lead to

the following rule: classify in G_1 (or P) if

$$\hat{P}(G_1|\mathbf{x}) \geq \frac{CFP - CTN}{CFP - CTN + CFN - CTP} = \frac{1}{3}\,,$$

and in G_2 (or N) otherwise. For each classifier and case we then estimated, using the test set, the error rates (false positive rate, FPR, and false negative rate, FNR), as well as the rates of correct classification (true positive rate, $TPR = 1 - FNR$, and true negative rate, $TNR = 1 - FPR$).

When the rejection option is considered we have to compute, for each of the two methods, two thresholds, u_1 and u_2, such that an observation is rejected if $u_1 < \hat{P}(G_1|\mathbf{x}) < u_2$.

For the method based on Chow's rule these are, independently of the classifier, given by

$$u_1 = \frac{CR - CTN}{CFN - CTN}, \quad u_2 = \frac{CFP - CR}{CFP - CTP}\,. \tag{6}$$

For the method based on the ROC curve we obtained, for each classifier, an independent estimate of the ROC curve (using the second set of observations and eleven equidistant candidate thresholds $u = 0, 0.1, \ldots, 1$). Then its convex hull was obtained and from it the final thresholds, u_1 and u_2, which are such that the corresponding slopes are as close as possible to the values given by m_1 and m_2 defined in (5). Figure 2 shows the three ROC curves with their convex hulls.

After obtaining the thresholds, u_1 and u_2, by the two methods we again classified the observations in the test set, estimating FPR, FNR, TPR and TNR, as before, and also the probabilities of rejecting a negative case, RN, and a

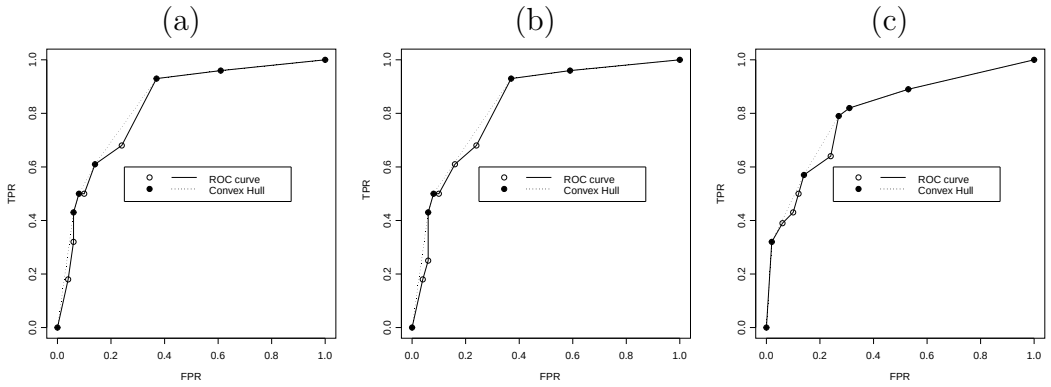

Fig. 2. Estimated ROC curves for the *Pima Indian Diabetes* dataset and three classifiers. (a) LDA (b) LD (c) MLP.

positive case, RP. Table 2 summarises the results. Note that for the first four cost combinations (cases **a**, **b**, **c** and **d**) the rejection option is not activated. This happens not because of the failure of the necessary condition mentioned at the end of Section 2, but because $u_2 < u_1$.

We finally estimated, again for each cost case and classifier, the total associated risk (that is, the symmetric of the utility in Tortorella, 2000). Without rejection this risk is given by

$$\widehat{TR} = \hat{\pi}_1 \left(CTP \cdot TPR + CFN \cdot FNR\right) + \hat{\pi}_2 \left(CTN \cdot TNR + CFP \cdot FPR\right),$$

whereas with rejection it is given by

$$\begin{aligned}\widehat{TR}_{rej} = &\hat{\pi}_1 \left(CTP \cdot TPR + CFN \cdot FNR + CR \cdot RP\right) + \\ &+\hat{\pi}_2 \left(CTN \cdot TNR + CFP \cdot FPR + CR \cdot RN\right).\end{aligned}$$

These results are given in Table 3. Since the aim is to minimize the risk the lowest value in each line is shown in boldface. From the three classifiers LDA gives the best results. When comparing the results with and without rejection we can see that, for the cost combinations for which the rejection

Table 2

Classification results and rejection thresholds obtained for the *Pima Indians Diabetes* test set.

Classifier Method	Case	FPR	TPR	FNR	TNR	RP	RN	u_1	u_2
	a b c d	0.34	0.83	0.17	0.66	–	–	–	–
LDA	**e**	0.23	0.80	0.17	0.54	0.03	0.23	0.30	0.38
Chow	**f**	0.02	0.67	0.03	0.34	0.30	0.64	0.17	0.58
	g	0.02	0.47	0.00	0.15	0.53	0.83	0.09	0.75
	a b c d	0.34	0.83	0.17	0.66	–	–	–	–
LDA	**e**	0.62	0.90	0.10	0.38	–	–	0.20	0.20
ROC	**f**	0.02	0.67	0.10	0.38	0.23	0.60	0.20	0.60
	g	0.02	0.63	0.10	0.38	0.27	0.60	0.20	0.70
	a b c d	0.28	0.77	0.23	0.72	–	–	–	–
LD	**e**	0.15	0.77	0.20	0.68	0.03	0.17	0.30	0.38
Chow	**f**	0.02	0.67	0.13	0.40	0.20	0.58	0.17	0.58
	g	0.02	0.40	0.00	0.19	0.60	0.79	0.09	0.75
	a b c d	0.28	0.77	0.23	0.72	–	–	–	–
LD	**e**	0.57	0.83	0.17	0.43	–	–	0.20	0.20
ROC	**f**	0.02	0.63	0.14	0.45	0.23	0.53	0.20	0.60
	g	0.02	0.50	0.13	0.45	0.37	0.53	0.20	0.70
	a b c d	0.28	0.70	0.30	0.72	–	–	–	–
MLP	**e**	0.21	0.70	0.30	0.68	0.00	0.11	0.30	0.38
Chow	**f**	0.11	0.53	0.14	0.55	0.33	0.34	0.17	0.58
	g	0.00	0.37	0.07	0.36	0.56	0.64	0.09	0.75
	a b c d	0.28	0.70	0.30	0.72	–	–	–	–
MLP	**e**	0.32	0.70	0.30	0.68	–	–	0.30	0.30
ROC	**f**	0.07	0.00	0.57	0.20	0.73	0.43	0.20	0.90
	g	0.07	0.00	0.00	0.00	0.93	1.00	0.00	0.90

Table 3

Estimates of the total risk, with and without rejection, obtained for the *Pima Indians Diabetes* test set.

	LDA			LD			MLP		
Case	No rej.	Chow	ROC	No rej.	Chow	ROC	No rej.	Chow	ROC
a	**−193.5**	–	–	−192.8	–	–	−181.8	–	–
b	**−92.5**	–	–	−92.1	–	–	−86.0	–	–
c	**−42.0**	–	–	−41.8	–	–	−38.1	–	–
d	**−16.8**	–	–	−16.6	–	–	−14.2	–	–
e	−8.3	−7.4	–	−8.0	**−11.1**	–	−4.4	−5.1	–
f	8.8	**−7.3**	−3.7	9.1	−2.2	−1.9	15.3	2.9	47.9
g	42.8	**1.0**	5.5	43.4	1.6	10.7	54.5	5.1	21.3

option is activated, the risk with rejection is usually smaller than the risk without rejection (the exceptions are LDA/Chow at case **e** and MLP/ROC at case **f**). Finally we conclude that for this example the rejection method based on Chow's rule leads to smaller risks than the rejection method based on the ROC curve.

5 Summary and discussion

We have reviewed Tortorella's rejection approach for binary classifiers and compared it with a generalisation of Chow's rule, showing the theoretical equivalence of the two approaches, through an important invariance property

of the derivative of the ROC curve.

Both methods include a rejection class and can cope with a rather general cost structure, however, Tortorella's approach is restricted to binary classifiers whereas Chow's rule is not. Furthermore, the first approach relies on a geometrical procedure which consists of finding the lines with slopes given in (5), intercepting the ROC curve and having largest TPR. Therefore the threshold values are chosen from a discrete set (associated to the ROC convex hull vertices). In the second approach there is not this limitation, and the threshold values are optimally derived from the cost structure. In our opinion this explains why in the application considered in Section 4 the method based on the ROC curve gave poorer results.

To finish we want to emphasise that, for the second approach, it is not necessary to rely on the exact knowledge of the a posteriori probabilities or the class conditional densities. It is sufficient to have, for a given observation, an estimate of any of those quantities. Many classifiers provide directly the estimates of the a posteriori probabilities (e.g. neural networks, logistic discrimination, k-NN). In cases where it may be necessary to estimate the class conditional densities we suggest, for example, a mixture of multivariate normal densities (as in Fraley and Raftery, 2003, Sec. 10.2).

In conclusion, we believe that the procedure discussed in this work can be useful for several applications. As far as our future work is concerned great effort will be devoted to the development of a new method that could take into account not only the rejection by indecision but also a new class of rejection for atypical observations.

References

Anderson, J.A., 1972. Separate sample logistic discrimination. Biometrika 59, 19-35.

Anderson, J.A., 1982. Logistic discrimination. In: P.R. Krishnaiah, L.N. Kanal, Eds., *Handbook of Statistics (Classification pattern recognition and reduction of dimensionality)* (Vol. 2). North-Holland, Amsterdam. pp. 169-191.

Chow, C.K., 1957. An optimum character recognition using decision functions. IRE Transactions on Electronic Computers 6, 247-254.

Chow, C.K., 1970. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16, 41-46.

Fraley, C., Raftery, A.E., 2003. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification 20, 263-286.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Tortorella, F., 2000. An optimal reject rule for binary classifiers. In: F.J. Ferri et al. Eds., *Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000. Lecture Notes in Computer Science, Vol. 1876.* Springer-Verlag, Heidelberg. pp. 611-620.

van Trees, H.L., 1968. Detection, Estimation, and Modulation Theory. Wiley, New York.

Venables, W.N., Ripley, B. D., 2002. *Modern Applied Statistics with S.* Springer-Verlag, New York.

Webb, A., 1999. Statistical Pattern Recognition. Arnold Publishers, London.