Journal of Machine Learning Research 13 (2012) 2813-2869 Submitted 1/12; Revised 7/12; Published 10/12

A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss

José Hernández-Orallo JORALLO@DSIC.UPV.ES
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46020, València, Spain

Peter Flach PETER.FLACH@BRISTOL.AC.UK
Intelligent Systems Laboratory
University of Bristol, United Kingdom
Merchant Venturers Building, Woodland Road
Bristol, BS8 1UB, United Kingdom

Cèsar Ferri CFERRI@DSIC.UPV.ES
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46020, València, Spain

Editor: Charles Elkan

Abstract

Many performance metrics have been introduced in the literature for the evaluation of classification performance, each of them with different origins and areas of application. These metrics include accuracy, unweighted accuracy, the area under the ROC curve or the ROC convex hull, the mean absolute error and the Brier score or mean squared error (with its decomposition into refinement and calibration). One way of understanding the relations among these metrics is by means of variable operating conditions (in the form of misclassification costs and/or class distributions). Thus, a metric may correspond to some expected loss over different operating conditions. One dimension for the analysis has been the distribution for this range of operating conditions, leading to some important connections in the area of proper scoring rules. We demonstrate in this paper that there is an equally important dimension which has so far received much less attention in the analysis of performance metrics. This dimension is given by the decision rule, which is typically implemented as a *threshold choice method* when using scoring models. In this paper, we explore many old and new threshold choice methods: fixed, score-uniform, score-driven, rate-driven and optimal, among others. By calculating the expected loss obtained with these threshold choice methods for a uniform range of operating conditions we give clear interpretations of the 0-1 loss, the absolute error, the Brier score, the *AUC* and the refinement loss respectively. Our analysis provides a comprehensive view of performance metrics as well as a systematic approach to loss minimisation which can be summarised as follows: given a model, apply the threshold choice methods that correspond with the available information about the operating condition, and compare their expected losses. In order to assist in this procedure we also derive several connections between the aforementioned performance metrics, and we highlight the role of calibration in choosing the threshold choice method.

Keywords: classification performance metrics, cost-sensitive evaluation, operating condition, Brier score, area under the ROC curve (*AUC*), calibration loss, refinement loss

1. Introduction

The choice of a proper performance metric for evaluating classification (Hand, 1997) is an old but still lively debate which has incorporated many different performance metrics along the way. Besides accuracy (*Acc*, or, equivalently, the error rate or 0-1 loss), many other performance metrics have been studied. The most prominent and well-known metrics are the Brier Score (*BS*, also known as Mean Squared Error) (Brier, 1950) and its decomposition in terms of refinement and calibration (Murphy, 1973), the absolute error (*MAE*), the log(arithmic) loss (or cross-entropy) (Good, 1952) and the area under the ROC curve (*AUC*, also known as the Wilcoxon-Mann-Whitney statistic, linearly related to the Gini coefficient and to the Kendall's tau distance to a perfect model) (Swets et al., 2000; Fawcett, 2006). There are also many graphical representations and tools for model evaluation, such as ROC curves (Swets et al., 2000; Fawcett, 2006), ROC isometrics (Flach, 2003), cost curves (Drummond and Holte, 2000, 2006), DET curves (Martin et al., 1997), lift charts (Piatetsky-Shapiro and Masand, 1999), and calibration maps (Cohen and Goldszmidt, 2004). A survey of graphical methods for classification predictive performance evaluation can be found in the work of Prati et al. (2011).

Many classification models can be regarded as functions which output a score for each example and class. This score represents a probability estimate of each example to be in one of the classes (or may just represent an unscaled magnitude which is monotonically related with a probability estimate). A score can then be converted into a class label using a decision rule. One of the reasons for evaluation being so multi-faceted is that models may be learnt in one context (misclassification costs, class distribution, etc.) but *deployed* in a different context. A context is usually described by a set of parameters, known as *operating condition*. When we have a clear operating condition at deployment time, there are effective tools such as ROC analysis (Swets et al., 2000; Fawcett, 2006) to establish which model is best and what its expected loss will be. However, the question is more difficult in the general case when we do not have information about the operating condition where the model will be applied. In this case, we want our models to perform well in a wide range of operating conditions. In this context, the notion of 'proper scoring rule', see, for example, the work of Murphy and Winkler (1970), sheds some light on some performance metrics. Some proper scoring rules, such as the Brier Score (MSE loss), the logloss, boosting loss and error rate (0-1 loss) have been shown by Buja et al. (2005) to be special cases of an integral over a Beta density of costs, see, for example, the works of Gneiting and Raftery (2007), Reid and Williamson (2010, 2011) and Brümmer (2010). Each performance metric is derived as a special case of the Beta distribution. However, this analysis focusses on scoring rules which are 'proper', that is, metrics that are minimised for well-calibrated probability assessments or, in other words, get the best (lowest) score by forecasting the true beliefs. Much less is known (in terms of expected loss for varying distributions) about other performance metrics which are non-proper scoring rules, such as *AUC*. Moreover, even its role as a classification performance metric has been put into question (Hand, 2009, 2010; Hand and Anagnostopoulos, 2011).

All these approaches make some (generally implicit and poorly understood) assumptions on how the model will work for each operating condition. In particular, it is generally assumed that the threshold which is used to discriminate between the classes will be set according to the operating condition. In addition, it is assumed that the threshold will be set in such a way that the estimated probability where the threshold is set is made equal to the operating condition. This is natural if we focus on proper scoring rules. Once all this is settled and fixed, different performance metrics

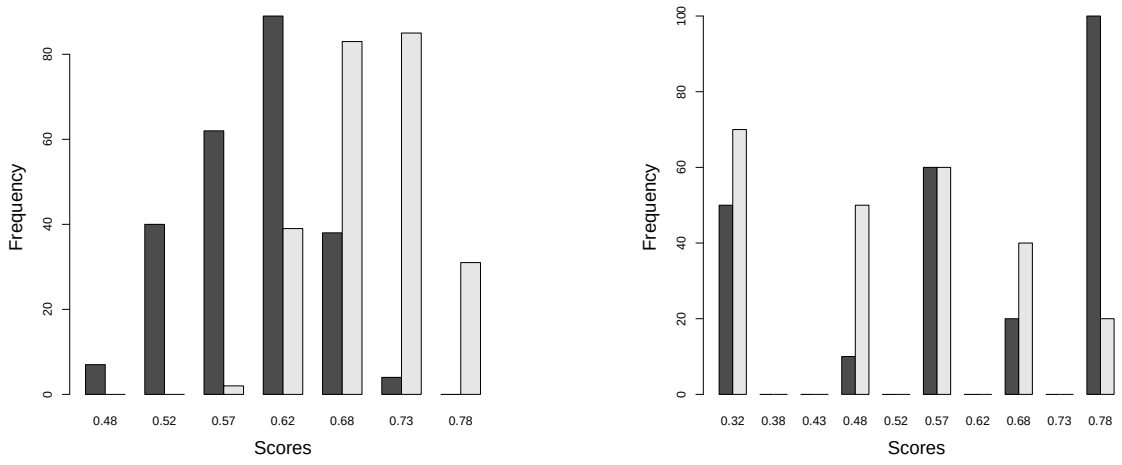

Figure 1: Histograms of the score distribution for model *A* (left) and model *B* (right).

represent different expected losses by using the distribution over the operating condition as a parameter. However, this *threshold choice* is only one of the many possibilities, as some other works have explored or mentioned in a more or less explicit way (Wieand et al., 1989; Drummond and Holte, 2006; Hand, 2009, 2010).

In our work we make these assumptions explicit through the concept of a *threshold choice method*, which systematically links performance metrics and expected loss. A threshold choice method sets a single threshold on the scores of a model in order to arrive at classifications, possibly taking circumstances in the deployment context into account, such as the operating condition (the class or cost distribution) or the intended proportion of positive predictions (the predicted positive rate). Building on this notion of threshold choice method, we are able to systematically explore how known performance metrics are linked to expected loss, resulting in a range of results that are not only theoretically well-founded but also practically relevant.

The basic insight is the realisation that there are many ways of converting a model (understood throughout this paper as a function assigning scores to instances) into a classifier that maps instances to classes (we assume binary classification throughout). Put differently, there are many ways of setting the threshold given a model and an operating condition. We illustrate this with an example concerning a very common scenario in machine learning research. Consider two models *A* and *B*, a naive Bayes model and a decision tree respectively (induced from a training data set), which are evaluated against a test data set, producing a score distribution for the positive and negative classes as shown in Figure 1. We see that scores are in the $[0,1]$ interval and in this example are interpreted as probability estimates for the negative class. ROC curves of both models are shown in Figure 2. We will assume that at this *evaluation time* we do not have information about the operating condition, but we expect that this information will be available at *deployment time*.

If we ask the question of which model is best we may rush to calculate its *AUC* and *BS* (and perhaps other metrics), as given by Table 1. However, we cannot give an answer because the question is *underspecified*. First, we need to know the range of operating conditions the model will work with. Second, we need to know how we will make the classifications, or in other words, we need a *deci-*

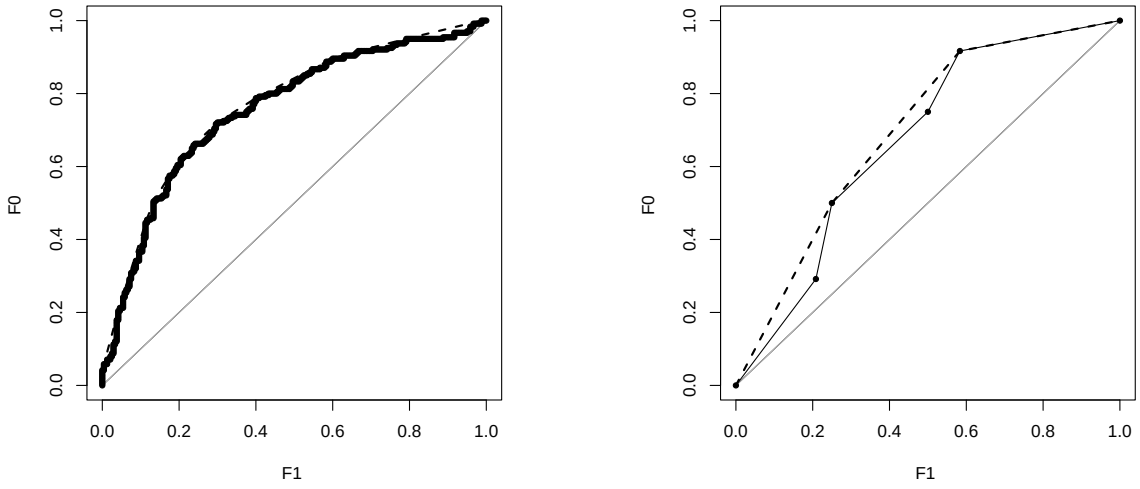

Figure 2: ROC Curves for model *A* (left) and model *B* (right).

sion rule, which can be implemented as a *threshold choice method* when the model outputs scores. For the first dimension (already considered by the work on proper scoring rules), if we have no precise knowledge about the operating condition, we can assume any of many distributions, depending on whether we have some information about how the cost may be distributed or no information at all. For instance, we can use a symmetric Beta distribution (Hand, 2009), an asymmetric Beta distribution (Hand and Anagnostopoulos, 2011) or, a partial or truncated distribution only considering a range of costs, or a simple uniform distribution (Drummond and Holte, 2006), as we also do here, which considers all operating conditions equally likely. For the second (new) dimension, we *also* have many options.

performance metric	model *A*	model *B*
AUC	0.791	0.671
Brier score	0.328	0.231

Table 1: Results from two models on a data set.

For instance, we can just set a fixed threshold at 0.5. This is what naive Bayes and decision trees do by default. This decision rule works as follows: if the score is greater than 0.5 then predict negative (1), otherwise predict positive (0). With this precise decision rule, we can now ask the question about the expected misclassification loss for a range of different misclassification costs (and/or class distributions), that is, for a distribution of operating conditions. Assuming a uniform distribution for operating conditions (cost proportions), we can effectively calculate the answer on the data set: 0.51.

But we can use other decision rules. We can use decision rules which adapt to the operating condition. One of these decision rules is the score-driven threshold choice method, which sets the threshold equal to the operating condition or, more precisely, to a cost proportion c. Another decision rule is the rate-driven threshold choice method, which sets the threshold in such a way that

the proportion of predicted positives (or predicted positive rate), simply known as 'rate' and denoted by r, equals the operating condition. Using these three different threshold choice methods for the models A and B, and assuming cost proportions are uniformly distributed, we get the expected losses shown in Table 2.

threshold choice method	expected loss model A	expected loss model B
Fixed ($T = 0.5$)	0.510	0.375
Score-driven ($T = c$)	0.328	0.231
Rate-driven (T s.t. $r = c$)	0.188	0.248

Table 2: Extension of Table 1 where two models are applied with three different threshold choice methods each, leading to six different classifiers and corresponding expected losses. In all cases, the expected loss is calculated over a range of cost proportions (operating conditions), which is assumed to be uniformly distributed. We denote the threshold by T, the cost proportion by c and the predicted positive rate by r).

In other words, only when we specify or assume a threshold choice method can we convert a model into a classifier for which it makes sense to consider its expected loss (for a range or distribution of costs). In fact, as we can see in Table 2, very different expected losses are obtained for the same model with different threshold choice methods. And this is the case even assuming the same uniform cost distribution for all of them.

Once we have made this (new) dimension explicit, we are ready to ask new questions. How many threshold choice methods are there? Table 3 shows six of the threshold choice methods we will analyse in this work, along with their notation. Only the score-fixed and the score-driven methods have been analysed in previous works in the area of proper scoring rules. The use of rates, instead of scores, is assumed in *screening* applications where an inspection, pre-diagnosis or coverage rate is intended (Murphy et al., 1987; Wieand et al., 1989), and the idea which underlies the distinction between rate-uniform and rate-driven is suggested by Hand (2010). In addition, a seventh threshold choice method, known as optimal threshold choice method, denoted by T^o, has been (implicitly) used in a few works (Drummond and Holte, 2000, 2006; Hand, 2009).

Threshold choice method	Fixed	Chosen uniformly	Driven by o.c.
Using scores	score-fixed (T^{sf})	score-uniform (T^{su})	score-driven (T^{sd})
Using rates	rate-fixed (T^{rf})	rate-uniform (T^{ru})	rate-driven (T^{rd})

Table 3: Possible threshold choice methods. The first family uses scores (as they were probabilities) and the second family uses rates (using scores as rank indicators). For both families we can fix a threshold or assume them ranging uniformly, which makes the threshold choice method independent from the operating condition. Only the last column takes the operating condition (o.c.) into account, and hence are the most interesting threshold choice methods.

We will see that each threshold choice method is linked to a specific performance metric. This means that if we decide (or are forced) to use a threshold choice method then there is a recommended performance metric for it. The results in this paper show that accuracy is the appropriate performance metric for the score-fixed method, *MAE* fits the score-uniform method, *BS* is the appropriate performance metric for the score-driven method, and *AUC* fits both the rate-uniform and the rate-driven methods. The latter two results assume a uniform cost distribution. It is important to make this explicit since the uniform cost distribution may be unrealistic in many particular situations and it is only one of many choices for a reference standard in the general case. As we will mention at the end of the paper, this suggests that new metrics can be derived by changing this distribution, as Hand (2009) has already done for the optimal threshold choice method with a Beta distribution.

The good news is that inter-comparisons are still possible: given a threshold choice method we can calculate expected loss from the relevant performance metric. The results in Table 2 allow us to conclude that model *A* achieves the lowest expected loss for uniformly sampled cost proportions, *if* we are wise enough to choose the appropriate threshold choice method (in this case the rate-driven method) to turn model *A* into a successful classifier. Notice that this cannot be said by just looking at Table 1 because the metrics in this table are not comparable to each other. In fact, there is no single performance metric that ranks the models in the correct order, because, as already said, expected loss cannot be calculated for models, only for classifiers.

1.1 Contributions and Structure of the Paper

The contributions of this paper to the subject of model evaluation for classification can be summarised as follows.

1. The expected loss of a model can only be determined if we select a distribution of operating conditions and a threshold choice method. We need to set a point in this two-dimensional space. Along the second (usually neglected) dimension, several new threshold choice methods are introduced in this paper.

2. We answer the question: "if one is choosing thresholds in a particular way, which performance metric is appropriate?" by giving an explicit expression for the expected loss for each threshold choice method. We derive linear relationships between expected loss and many common performance metrics.

3. Our results reinvigorate AUC as a well-founded measure of expected classification loss for both the rate-uniform and rate-driven methods. While Hand (2009, 2010) raised objections against AUC for the optimal threshold choice method only, noting that AUC can be consistent with other threshold choice methods, we encountered a widespread misunderstanding in the machine learning community that the AUC is fundamentally flawed as a performance metric—a clear misinterpretation of Hand's papers that we hope that this paper helps to further rectify.

4. One fundamental and novel result shows that the refinement loss of the convex hull of a ROC curve is equal to expected *optimal* loss as measured by the area under the optimal cost curve. This sets an optimistic (but also unrealistic) bound for the expected loss.

5. Conversely, from the usual calculation of several well-known performance metrics we can derive expected loss. Thus, classifiers and performance metrics become easily comparable.

With this we do not choose the best model but rather the best classifier (a model with a particular threshold choice method).

6. By cleverly manipulating scores we can connect several of these performance metrics, either by the notion of evenly-spaced scores or perfectly calibrated scores. This provides an additional way of analysing the relation between performance metrics and, of course, threshold choice methods.

7. We use all these connections to better understand which threshold choice method should be used, and in which cases some are better than others. The analysis of calibration plays a central role in this understanding, and also shows that non-proper scoring rules do have their role and can lead to lower expected loss than proper scoring rules, which are, as expected, more appropriate when the model is well-calibrated.

This set of contributions provides an integrated perspective on performance metrics for classification around the systematic exploration of the notion of threshold choice method that we develop in this paper.

The remainder of the paper is structured as follows. Section 2 introduces some notation, the basic definitions for operating condition, threshold, expected loss, and particularly the notion of threshold choice method, which we will use throughout the paper. Section 3 investigates expected loss for fixed threshold choice methods (score-fixed and rate-fixed), which are the base for the rest. We show that, not surprisingly, the expected loss for these threshold choice method are the 0-1 loss (weighted or unweighted accuracy depending on whether we use cost proportions or skews). Section 4 presents the results that the score-uniform threshold choice method has *MAE* as associate performance metric and the score-driven threshold choice method leads to the Brier score. We also show that one dominates over the other. Section 5 analyses the non-fixed methods based on rates. Somewhat surprisingly, both the rate-uniform threshold choice method and the rate-driven threshold choice method lead to linear functions of *AUC*, with the latter always been better than the former. All this vindicates the rate-driven threshold choice method but also *AUC* as a performance metric for classification. Section 6 uses the optimal threshold choice method, connects the expected loss in this case with the area under the optimal cost curve, and derives its corresponding metric, which is refinement loss, one of the components of the Brier score decomposition. Section 7 analyses the connections between the previous threshold choice methods and metrics by considering several properties of the scores: evenly-spaced scores and perfectly calibrated scores. This also helps to understand which threshold choice method should be used depending on how good scores are. Finally, Section 8 closes the paper with a thorough discussion of results, related work, and an overall conclusion with future work and open questions. Two appendices include a derivation of univariate operating conditions for costs and skews and some technical results for the optimal threshold choice method.

2. Background

In this section we introduce some basic notation and definitions we will need throughout the paper. Further definitions will be introduced when needed. The most important definitions we will need are introduced below: the notion of threshold choice method and the expression of expected loss.

2.1 Notation and Basic Definitions

A *classifier* is a function that maps instances x from an instance space X to classes y from an output space Y. For this paper we will assume binary classifiers, that is, $Y = \{0,1\}$. A *model* is a function $m : X \to \mathbb{R}$ that maps examples to real numbers (scores) on an unspecified scale. We use the convention that higher scores express a stronger belief that the instance is of class 1. A *probabilistic model* is a function $m : X \to [0,1]$ that maps examples to estimates $\hat{p}(1|x)$ of the probability of example x to be of class 1. Throughout the paper we will use the term *score* (usually denoted by s) both for unscaled values (in an unbounded interval) and probability estimates (in the interval [0,1]). Nonetheless, we will make the interpretation explicit whenever we use them in one way or the other. We will do similarly for thresholds. In order to make predictions in the Y domain, a model can be converted to a classifier by fixing a decision threshold t on the scores. Given a predicted score $s = m(x)$, the instance x is classified in class 1 if $s > t$, and in class 0 otherwise.

For a given, unspecified model and population from which data are drawn, we denote the score density for class k by f_k and the cumulative distribution function by F_k. Thus, $F_0(t) = \int_{-\infty}^{t} f_0(s)ds = P(s \leq t|0)$ is the proportion of class 0 points correctly classified if the decision threshold is t, which is the sensitivity or true positive rate at t. Similarly, $F_1(t) = \int_{-\infty}^{t} f_1(s)ds = P(s \leq t|1)$ is the proportion of class 1 points incorrectly classified as 0 or the false positive rate at threshold t; $1 - F_1(t)$ is the true negative rate or specificity. Note that we use 0 for the positive class and 1 for the negative class, but scores increase with $\hat{p}(1|x)$. That is, $F_0(t)$ and $F_1(t)$ are monotonically non-decreasing with t. This has some notational advantages and is the same convention as used by, for example, Hand (2009).

Given a data set $D \subset \langle X, Y \rangle$ of size $n = |D|$, we denote by D_k the subset of examples in class $k \in \{0,1\}$, and set $n_k = |D_k|$ and $\pi_k = n_k/n$. Clearly $\pi_0 + \pi_1 = 1$. We will use the term *class proportion* for π_0 (other terms such as 'class ratio' or 'class prior' have been used in the literature). Given a model and a threshold t, we denote by $R(t)$ the predicted positive rate, that is, the proportion of examples that will be predicted positive (class 0) if the threshold is set at t. This can also be defined as $R(t) = \pi_0 F_0(t) + \pi_1 F_1(t)$. The average score of actual class k is $\bar{s}_k = \int_0^1 s f_k(s)ds$. Given any strict order for a data set of n examples we will use the index i on that order to refer to the i-th example. Thus, s_i denotes the score of the i-th example and y_i its true class.

We define partial class accuracies as $Acc_0(t) = F_0(t)$ and $Acc_1(t) = 1 - F_1(t)$. From here, (weighted or micro-average) accuracy is defined as $Acc(t) = \pi_0 Acc_0(t) + \pi_1 Acc_1(t)$ and (unweighted or macro-average) accuracy as $uAcc(t) = (Acc_0(t) + Acc_1(t))/2$ (also known as 'average recall', Flach, 2012), which computes accuracy while assuming balanced classes.

We denote by $U_S(x)$ the continuous uniform distribution of variable x over an interval $S \subset \mathbb{R}$. If this interval S is $[0,1]$ then S can be omitted. The family of continuous distributions Beta is denoted by $\beta_{\alpha,\beta}$. The Beta distributions are always defined in the interval $[0,1]$. Note that the uniform distribution is a special case of the Beta family, that is, $\beta_{1,1} = U$.

2.2 Operating Conditions and Expected Loss

When a model is deployed for classification, the conditions might be different to those during training. In fact, a model can be used in several deployment contexts, with different results. A context can entail different class distributions, different classification-related costs (either for the attributes, for the class or any other kind of cost), or some other details about the effects that the application of a model might entail and the severity of its errors. In practice, a deployment context or *operating*

condition is usually defined by a misclassification cost function and a class distribution. Clearly, there is a difference between operating when the cost of misclassifying 0 into 1 is equal to the cost of misclassifying 1 into 0 and doing so when the former is ten times the latter. Similarly, operating when classes are balanced is different from when there is an overwhelming majority of instances of one class.

One general approach to cost-sensitive learning assumes that the cost does not depend on the example but only on its class. In this way, misclassification costs are usually simplified by means of cost matrices, where we can express that some misclassification costs are higher than others (Elkan, 2001). Typically, the costs of correct classifications are assumed to be 0. This means that for binary models we can describe the cost matrix by two values $c_k \geq 0$ with at least one of both being strictly greater than 0, representing the misclassification cost of an example of class k. Additionally, we can normalise the costs by setting $b = c_0 + c_1$, which will be referred to as the *cost magnitude* (which is clearly strictly greater than 0), and $c = c_0/b$; we will refer to c as the *cost proportion* Since this can also be expressed as $c = (1 + c_1/c_0)^{-1}$, it is often called 'cost ratio' even though, technically, it is a proportion ranging between 0 and 1.

Under these assumptions, an operating condition can be defined as a tuple $\theta = \langle b, c, \pi_0 \rangle$. The space of operating conditions is denoted by Θ. These three parameters are not necessarily independent, as we will discuss in more detail below. The loss for an operating condition is defined as follows:

$$\begin{aligned} Q(t;\theta) = Q(t;\langle b,c,\pi_0\rangle) &\triangleq b\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \\ &= c_0\pi_0(1-F_0(t)) + c_1\pi_1 F_1(t). \end{aligned} \tag{1}$$

It is important to distinguish the information we may have available at each stage of the process. At evaluation time we may not have access to some information that is available later, at deployment time. In many real-world problems, when we have to evaluate or compare models, we do not know the operating condition that will apply during deployment. One general approach is to evaluate the model on a range of possible operating points. In order to do this, we have to set a weight or distribution for operating conditions.

A key issue when applying a model under different operating conditions is how the threshold is chosen in each of them. If we work with a classifier, this question vanishes, since the threshold is already settled. However, in the general case when we work with a model, we have to decide how to establish the threshold. The key idea proposed in this paper is the notion of a threshold choice method, a function which converts an operating condition into an appropriate threshold for the classifier.

Definition 1 Threshold choice method. *A threshold choice method*[1] *is a (possibly non-deterministic) function $T : \Theta \to \mathbb{R}$ such that given an operating condition it returns a decision threshold.*

When we say that T may be non-deterministic, it means that the result may depend on a random variable and hence may itself be a random variable according to some distribution. We introduce

1. The notion of threshold choice method could be further generalised to cover situations where we have some information about the operating condition which cannot be expressed in terms of a specific value of Θ, such a distribution on Θ or information about $\mathbb{E}\{b\}$, $\mathbb{E}\{bc\}$, etc. This generalisation could be explored, but it is not necessary for the cases discussed in this paper.

the threshold choice method as an abstract concept since there are several reasonable options for the function T, essentially because there may be different degrees of information about the model and the operating conditions at evaluation time. We can set a fixed threshold ignoring the operating condition; we can set the threshold by looking at the ROC curve (or its convex hull) and using the cost proportion to intersect the ROC curve (as ROC analysis does); we can set a threshold looking at the estimated scores; or we can set a threshold independently from the rank or the scores. The way in which we set the threshold may dramatically affect performance. But, not less importantly, the performance metric used for evaluation must be in accordance with the threshold choice method.

Given a threshold choice function T, the loss for a particular operating condition θ is given by $Q(T(\theta);\theta)$. However, if we do not know the operating condition precisely, we can define a distribution for operating conditions as a multivariate distribution, $w(\theta)$. From here, we can now calculate expected loss as a weighted average over operating conditions (Adams and Hand, 1999):

$$L \triangleq \int_{\Theta} Q(T(\theta);\theta)w(\theta)d\theta. \tag{2}$$

Calculating this integral for a particular case depends on the threshold choice method and the kind of model, but particularly on the space of operating conditions Θ and its associated distribution $w(\theta)$. Typically, the representation of operating conditions is simplified from a three-parameter tuple $\langle b, c, \pi_0 \rangle$ to a single parameter. This reduces w to a univariate distribution. However, this reduction must carry some assumptions. For instance, the cost magnitude b is not always independent of c and π_0, since costs in very imbalanced cases tend to have higher magnitude. For instance, we may have two different operating conditions, one with $c_0 = 10$ and $c_1 = 1$ and another with $c_0 = 5$ and $c_1 = 50$. While the cost ratios are symmetric (10:1 with $c = 10/11$ for the first case, 1:10 with $c = 1/11$ for the second), the second operating condition will clearly have more impact on the expected loss, because its magnitude is five times higher. Moreover, c is usually closely linked to π_0, since the higher the imbalance (class proportion), the higher the cost proportion. For instance, if positives are rare, we usually want them to be detected (especially in diagnosis and fault detection applications), and false negatives (i.e., a positive which has not been detected but misclassified as negative) will have higher cost.

Despite these dependencies, one common option for this simplified operating condition is to consider that costs are normalised (the cost matrix always sums up to a constant, that is, the cost magnitude b is constant), or less strongly, that b and c are independent. Another option which does not require independence of b and c relies on noticing that b is a multiplicative factor in Equation (1). From here, we just need to assume that the threshold choice method is independent of b. This is not a strong assumption, since all the threshold choice methods that have been used systematically in the literature (e.g., the optimal threshold choice method and the score-driven method) are independent of b and so are the rest of methods we work with in this paper. With this, as deployed in appendix A, we can incorporate b in a bivariate distribution $v_H(\langle c, \pi_0 \rangle)$ for cost and class proportions. This does not mean that we ignore the magnitude b or assume it constant, but that we can embed its variability in v_H. From here, we just derive two univariate cost functions and the corresponding expected losses.

The first one assumes π_0 constant, leading to a loss expression which only depends on cost proportions c:

$$Q_c(t;c) \triangleq \mathbb{E}\{b\}\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}. \tag{3}$$

Using this expression, expected loss will be derived as an integral using the univariate distribution w_c, which incorporates the variability of b jointly with c (see appendix A for details).

A different approach to reducing the operating condition to a single parameter is the notion of *skew*, which is a normalisation of the product between cost proportion and class proportion:

$$z \triangleq \frac{c\pi_0}{c\pi_0 + (1-c)(1-\pi_0)}.$$

This means that π_0 is no longer fixed, but neither is it independent of c. What z does is to combine both parameters. This is a different way of reducing the operating condition to one single parameter. We thus (re-)define loss as depending solely on z.

$$Q_z(t;z) \triangleq z(1-F_0(t)) + (1-z)F_1(t).$$

Similarly, we also define a weight $w_z(z)$ which also incorporates the variability of b and π_0 (see appendix A for details), which will be used in the integral for calculating the expected loss below.

As a result, in what follows, we will just work with operating conditions which are either defined by the cost proportion c (assuming a fixed class distribution π_0) or by the skew z (which combines c and π_0). For convenience, as justified in appendix A, we will assume $\mathbb{E}\{b\} = 2$. Interestingly, we can relate both approaches (using costs and skews) with the following lemma (proven in appendix A):

Lemma 2 *Assuming $\mathbb{E}\{b\} = 2$, if $\pi_0 = \pi_1$ then $z = c$ and $Q_z(t;z) = Q_c(t;c)$.*

This will allow us to translate the results for cost proportions to skews.

From now on, since the operating condition can be either a cost proportion c or a skew z we will use the subscript c or z to differentiate them. In fact, threshold choice methods will be represented by T_c and T_z and they will be defined as $T_c : [0,1] \to \mathbb{R}$ and $T_z : [0,1] \to \mathbb{R}$ respectively. Superscripts will be used to identify particular threshold choice methods. Some threshold choice methods we consider in this paper take additional information into account, such as a default threshold or a target predicted positive rate; such information is indicated by square brackets. So, for example, the score-fixed threshold choice method for cost proportions considered in the next section is indicated thus: $T_c^{sf}[t](c)$. In the rest of this paper, we explore a range of different methods to choose the threshold (some deterministic and some non-deterministic). We will give proper definitions of all these threshold choice methods in its due section.

The expected loss for costs and skews is then adapted from Equation (2) as follows:

Definition 3 *Given a threshold choice method for cost proportions T_c and a probability density function over cost proportions w_c,* expected loss *L_c is defined as*

$$L_c \triangleq \int_0^1 Q_c(T_c(c);c)w_c(c)dc. \tag{4}$$

Incorporating the class distribution into the operating condition as skews and defining a distribution over skews w_z, we obtain expected loss over a distribution of skews:

$$L_z \triangleq \int_0^1 Q_z(T_z(z);z)w_z(z)dz. \tag{5}$$

It is worth noting that if we plot Q_c or Q_z against c and z, respectively, we obtain *cost curves* as defined by Drummond and Holte (2000, 2006). Cost curves are also known as risk curves (e.g., Reid and Williamson, 2011, where the plot can also be shown in terms of *priors*, that is, class proportions).

Equations (4) and (5) illustrate the space we explore in this paper. Two parameters determine the expected loss: $w_c(c)$ and $T_c(c)$ (respectively $w_z(z)$ and $T_z(z)$). While much work has been done on a first dimension, by changing $w_c(c)$ or $w_z(z)$, particularly in the area of proper scoring rules, no work has *systematically* analysed what happens when changing the second dimension, $T_c(c)$ or $T_z(z)$.

This means that in this paper we focus on this second dimension, and just make some simple choices for the first dimension. Except for cases where the threshold choice is independent of the operating condition, we will assume a uniform distribution for $w_c(c)$ and $w_z(z)$. This is of course just one possible choice, but not an arbitrary choice for a number of reasons:

- The uniform distribution is arguably the simplest distribution for a value between 0 and 1 and requires no parameters.
- This distribution makes the representation of the loss straightforward, since we can plot Q on the y-axis versus c (or z) on the x-axis, where the x-axis can be shown linearly from 0 to 1, without any distribution warping. This makes metrics correspond exactly with the areas under many cost curves, such as the optimal cost curves (Drummond and Holte, 2006), the Brier curves (Hernández-Orallo et al., 2011) or the rate-driven/Kendall curves (Hernández-Orallo et al., 2012).
- The uniform distribution is a reasonable choice if we want a model to behave well in a wide range of situations, from high costs for false positives to the other extreme. In this sense, it gives more relevance to models which perform well when the cost matrices or the class proportions are highly imbalanced.
- Most of the connections with the existing metrics are obtained with this distribution and not with others, which is informative about what the metrics implicitly assume (if understood as measures of expected loss).

Many expressions in this paper can be fine-tuned with other distributions, such as the Beta distribution $\beta(2,2)$, as suggested by Hand (2009), or using imbalance (Hand, 2010). However, it is the uniform distribution which leads us to many well-known evaluation metrics.

3. Expected Loss for Fixed-Threshold Classifiers

The easiest way to choose the threshold is to set it to a pre-defined value t_{fixed}, independently from the model and also from the operating condition. This is, in fact, what many classifiers do (e.g., Naive Bayes chooses $t_{\text{fixed}} = 0.5$ independently from the model and independently from the operating condition). We will see the straightforward result that this threshold choice method corresponds to 0-1 loss. Part of these results will be useful to better understand some other threshold choice methods.

Definition 4 *The* score-fixed threshold choice method *is defined as follows:*

$$T_c^{sf}[t](c) \triangleq T_z^{sf}[t](z) \triangleq t. \tag{6}$$

This choice has been criticised in two ways, but is still frequently used. Firstly, choosing 0.5 as a threshold is not generally the best choice even for balanced data sets or for applications where the test distribution is equal to the training distribution (see, for example, the work of Lachiche and Flach, 2003 on how to get much more from a Bayes classifier by simply changing the threshold). Secondly, even if we are able to find a better value than 0.5, this does not mean that this value is best for every skew or cost proportion—this is precisely one of the reasons why ROC analysis is used (Provost and Fawcett, 2001). Only when we know the deployment operating condition at evaluation time is it reasonable to fix the threshold according to this information. So either by common choice or because we have this latter case, consider then that we are going to use the same threshold t independently of skews or cost proportions. Given this threshold choice method, then the question is: *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?* This is what we answer below.

If we plug T_c^{sf} (Equation 6) into the general formula of the expected loss for a range of cost proportions (Equation 4) we have:

$$L_c^{sf}(t) \triangleq \int_0^1 Q_c(T_c^{sf}[t](c);c)w_c(c)dc.$$

We obtain the following straightforward result.

Theorem 5 *If a classifier sets the decision threshold at a fixed value t irrespective of the operating condition or the model, then expected loss for any cost distribution w_c is given by:*

$$L_c^{sf}(t) = 2\mathbb{E}_{w_c}\{c\}\,(1-Acc(t)) + 4\pi_1 F_1(t)\left(\frac{1}{2} - \mathbb{E}_{w_c}\{c\}\right).$$

Proof

$$\begin{aligned}
L_c^{sf}(t) &= \int_0^1 Q_c(T_c^{sf}[t](c);c)w_c(c)dc = \int_0^1 Q_c(t;c)w_c(c)dc \\
&= \int_0^1 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}w_c(c)dc \\
&= 2\pi_0(1-F_0(t))\int_0^1 cw_c(c)dc + 2\pi_1 F_1(t)\int_0^1 (1-c)w_c(c)dc \\
&= 2\pi_0(1-F_0(t))\mathbb{E}_{w_c}\{c\} + 2\pi_1 F_1(t)(1-\mathbb{E}_{w_c}\{c\}) \\
&= 2\pi_0(1-F_0(t))\mathbb{E}_{w_c}\{c\} + 2\pi_1 F_1(t)\left(\mathbb{E}_{w_c}\{c\} + 2\left(\frac{1}{2} - \mathbb{E}_{w_c}\{c\}\right)\right) \\
&= 2\mathbb{E}_{w_c}\{c\}\,(\pi_0(1-F_0(t)) + \pi_1 F_1(t)) + 4\pi_1 F_1(t)\left(\frac{1}{2} - \mathbb{E}_{w_c}\{c\}\right) \\
&= 2\mathbb{E}_{w_c}\{c\}\,(1-Acc(t)) + 4\pi_1 F_1(t)\left(\frac{1}{2} - \mathbb{E}_{w_c}\{c\}\right).
\end{aligned}$$

■

This gives an expression of expected loss which depends on error rate and false positive rate at t and the expected value for the distribution of costs.[2] Similarly, if we plug T_z^{sf} (Equation 6) into the general formula of the expected loss for a range of skews (Equation 5) we have:

$$L_z^{sf}(t) \triangleq \int_0^1 Q_z(T_z^{sf}[t](z);z)w_z(z)dz.$$

Using Lemma 2 we obtain the equivalent result for skews:

$$L_z^{sf}(t) = 2\mathbb{E}_{w_z}\{z\}\,(1-uAcc(t)) + 2F_1(t)\left(\frac{1}{2} - \mathbb{E}_{w_z}\{z\}\right).$$

Corollary 6 *If a classifier sets the decision threshold at a fixed value irrespective of the operating condition or the model, then expected loss under a distribution of cost proportions w_c with expected value $\mathbb{E}_{w_c}\{c\} = 1/2$ is equal to the error rate at that decision threshold.*

$$L^{sf}_{\mathbb{E}\{c\}=1/2}(t) = \pi_0(1-F_0(t)) + \pi_1 F_1(t) = 1 - Acc(t).$$

Using Lemma 2 we obtain the equivalent result for skews:

$$L^{sf}_{\mathbb{E}\{z\}=1/2}(t) = (1-F_0(t))/2 + F_1(t)/2 = 1 - uAcc(t).$$

So the expected loss under a distribution of cost proportions with mean 1/2 for the *score-fixed threshold choice method* is the error rate of the classifier at that threshold. Clearly, a uniform distribution is a special case, but the result also applies to, for instance, a symmetric Beta distribution centered at 1/2. That means that accuracy can be seen as a measure of classification performance in a range of cost proportions when we choose a fixed threshold. This interpretation is reasonable, since accuracy is a performance metric which is typically applied to classifiers (where the threshold is fixed) and not to models outputting scores. This is exactly what we did in Table 2. We calculated the expected loss for the fixed threshold at 0.5 for a uniform distribution of cost proportions, and we obtained $1 - Acc = 0.51$ and 0.375 for models A and B respectively.

The previous results show that 0-1 losses are appropriate to evaluate models in a range of operating conditions if the threshold is fixed for all of them and we do not have any information about a possible asymmetry in the cost matrix at deployment time. In other words, accuracy and unweighted accuracy can be the right performance metrics for classifiers even in a cost-sensitive learning scenario. The situation occurs when one assumes a particular operating condition at evaluation time while the classifier has to deal with a range of operating conditions in deployment time.

In order to prepare for later results we also define a particular way of setting a fixed classification threshold, namely to achieve a particular predicted positive rate. One could say that such a method *quantifies* the proportion of positive predictions made by the classifier. For example, we could say that our threshold is fixed to achieve a rate of 30% positive predictions and the rest negatives. This

2. As mentioned above, the value of t is usually calculated disregarding the information (if any) about the operating condition, and frequently set to 0.5. In fact, this threshold choice method is called 'fixed' because of this. However, we can estimate and fix the value of t by taking the expected value for the operating condition $\mathbb{E}_{w_c}\{c\}$ into account, if we have some information about the *distribution* w_c. For instance, we may choose $t = \mathbb{E}_{w_c}\{c\}$ or we may choose the value of t which minimises the expression of expected loss in Theorem 5.

of course involves ranking the examples by their scores and setting a cutting point at the appropriate position, something which is frequent in 'screening' applications (Murphy et al., 1987; Wieand et al., 1989).

Let us denote the predicted positive rate at threshold t as $R(t) = \pi_0 F_0(t) + \pi_1 F_1(t)$. Then,

Definition 7 *If R is invertible, then we define the* rate-fixed threshold choice method *for rate r as:*

$$T_c^{rf}[r](c) \triangleq R^{-1}(r).$$

Similarly to the cost case, the rate-fixed threshold choice method for skews, assuming R is invertible, is defined as:

$$T_z^{rf}[r](z) \triangleq R_z^{-1}(r).$$

where $R_z(t) = F_0(t)/2 + F_1(t)/2$.

If R is not invertible, it has plateaus and so does R. This can be handled by deriving t from the centroid of a plateau. Nonetheless, in what follows, we will explicitly state when the invertibility of R is necessary. The corresponding expected loss for cost proportions is

$$L_c^{rf} \triangleq \int_0^1 Q_c(T_c^{rf}[r](c); c) w_c(c) dc = \int_0^1 Q_c(R^{-1}(r); c) w_c(c) dc.$$

As already mentioned, the notion of setting a threshold based on a rate is typically seen in screening applications but it also closely related to the task of class prevalence estimation (Neyman, 1938; Tenenbein, 1970; Alonzo et al., 2003), which is also known as quantification in machine learning and data mining (Forman, 2008; Bella et al., 2010). The goal of this task is to correctly estimate the proportion for each of the classes. This threshold choice method allows the user to set the quantity of positives, which might be known (from a sample of the test) or can be estimated using a quantification method. In fact, some quantification methods can be seen as methods to determine an absolute fixed threshold t that ensures a correct proportion for the test set. Fortunately, it is immediate to get the threshold which produces a rate; it can just be derived by sorting the examples by their scores and placing the cutpoint where the rate equals the rank divided by the number of examples (e.g., if we have n examples, the cutpoint i makes $r = i/n$).

4. Threshold Choice Methods Using Scores

In the previous section we looked at accuracy and error rate as performance metrics for classifiers and gave their interpretation as expected losses. In this and the following sections we consider performance metrics for models that do not require fixing a threshold choice method in advance. Such metrics include *AUC* which evaluates ranking performance and the Brier score or mean squared error which evaluates the quality of probability estimates. We will deal with the latter in this section. For the rest of this section, we will therefore assume that scores range between 0 and 1 and represent posterior probabilities for class 1, unless otherwise stated. This means that we can sample thresholds uniformly or derive them from the operating condition. We first introduce two performance metrics that are applicable to probabilistic scores.

The Brier score is a well-known performance metric for probabilistic models. It is an alternative name for the Mean Squared Error or MSE loss (Brier, 1950), especially for binary classification.

Definition 8 *The Brier score, BS, is defined as follows:*

$$BS \triangleq \pi_0 BS_0 + \pi_1 BS_1.$$

where the partial expressions for the positive and negative class are given by:

$$BS_0 \triangleq \int_0^1 s^2 f_0(s) ds.$$
$$BS_1 \triangleq \int_0^1 (1-s)^2 f_1(s) ds.$$

From here, we can define a prior-independent version of the Brier score (or an unweighted Brier score) as follows:

$$uBS \triangleq \frac{BS_0 + BS_1}{2}.$$

The Mean Absolute Error (*MAE*) is another simple performance metric which has been rediscovered many times under different names.

Definition 9 *The Mean Absolute Error, MAE, is defined as follows:*

$$MAE \triangleq \pi_0 MAE_0 + \pi_1 MAE_1.$$

where the partial expressions for the positive and negative class are given by:

$$MAE_0 \triangleq \int_0^1 s f_0(s) ds = \bar{s}_0.$$
$$MAE_1 \triangleq \int_0^1 (1-s) f_1(s) ds = 1 - \bar{s}_1.$$

We can define an unweighted MAE as follows:

$$uMAE \triangleq \frac{MAE_0 + MAE_1}{2} = \frac{\bar{s}_0 + (1-\bar{s}_1)}{2}.$$

It can be shown that *MAE* is equivalent to the Mean Probability Rate (MPR) (Lebanon and Lafferty, 2002) for discrete classification (Ferri et al., 2009).

4.1 The Score-Uniform Threshold Choice Method Leads to MAE

We now demonstrate how varying a model's threshold leads to an expected loss that is different from accuracy. First, we explore a threshold choice method which considers that we have no information at all about the operating condition, neither at evaluation time nor at deployment time. We just employ the interval between the maximum and minimum value of the scores, and we randomly select the threshold using a uniform distribution over this interval. It can be argued that this threshold choice method is unrealistic, because we almost always have some information about the operating condition, especially at deployment time. A possible interpretation is that this threshold choice method is useful to make a *worst-case* evaluation. In other words, expected loss using this method gives a robust assessment for situations where the information about the operating condition is not only unavailable, but maybe unrealiable or even malicious. So what we show next is that there are evaluation metrics which can be expressed as an expected loss under these assumptions, adding support to the idea that the metrics related to this threshold choice method are blind to (or unaware of) any cost information.

Definition 10 *Assuming a model's scores are expressed on a bounded scale* $[l,u]$*, the* score-uniform threshold choice method *is defined as follows:*

$$T_c^{su}(c) \triangleq T_z^{su}(z) \triangleq T_c^{sf}[U_{l,u}](c).$$

Given this threshold choice method, then the question is: *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?*

Theorem 11 *Assuming probabilistic scores and the score-uniform threshold choice method, expected loss under a distribution of cost proportions* w_c *is equal to:*

$$L_c^{su} = 2\{\mathbb{E}_{w_c}\{c\}\pi_0(\bar{s}_0) + (1-\mathbb{E}_{w_c}\{c\})\pi_1(1-\bar{s}_1)\}.$$

Proof First we derive Q_c:

$$\begin{aligned} Q_c(T_c^{su}(c);c) = Q_c(T_c^{sf}[U_{l,u}](c);c) &= \int_l^u Q_c(T_c^{sf}[t](c);c)\frac{1}{u-l}dt \\ &= \frac{1}{u-l}\int_l^u Q_c(t;c)dt = \frac{1}{u-l}\int_l^u 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}dt \\ &= 2\frac{c\pi_0(\bar{s}_0 - l) + (1-c)\pi_1(u-\bar{s}_1)}{(u-l)}. \end{aligned}$$

The last step makes use of the following useful property.

$$\int_l^u F_k(t)dt = [tF_k(t)]_l^u - \int_l^u t\,f_k(t)dt = uF_k(u) - lF_k(l) - \bar{s}_k = u - \bar{s}_k.$$

Setting $l=0$ and $u=1$ for probabilistic scores, we obtain the final result:

$$Q_c(T_c^{su}(c);c) \quad = 2\{c\pi_0(\bar{s}_0) + (1-c)\pi_1(1-\bar{s}_1)\}.$$

And now, we calculate the expected loss for the distribution $w_c(c)$.

$$\begin{aligned} L_c^{su} &= \int_0^1 Q_c(T_c^{su}(c);c)w_c(c)dc \\ &= \int_0^1 2\{c\pi_0(\bar{s}_0) + (1-c)\pi_1(1-\bar{s}_1)\}w_c(c)dc \\ &= 2\{\mathbb{E}_{w_c}\{c\}\pi_0(\bar{s}_0) + (1-\mathbb{E}_{w_c}\{c\})\pi_1(1-\bar{s}_1)\}. \end{aligned}$$

■

Corollary 12 *Assuming probabilistic scores and the score-uniform threshold choice method, expected loss under a distribution of cost proportions* w_c *with expected value* $\mathbb{E}_{w_c}\{c\} = 1/2$ *is equal to the model's mean absolute error.*

$$L^{su}_{\mathbb{E}\{c\}=1/2} = \pi_0\bar{s}_0 + \pi_1(1-\bar{s}_1) = MAE.$$

This gives a baseline loss if we choose thresholds randomly and independently of the model. Using Lemma 2 we obtain the equivalent result for skews:

$$L^{su}_{\mathbb{E}\{z\}=1/2} \quad = \quad \frac{\bar{s}_0 + (1-\bar{s}_1)}{2} = uMAE.$$

4.2 The Score-Driven Threshold Choice Method Leads to the Brier Score

We will now consider the first threshold choice method to take the operating condition into account. Since we are dealing with probabilistic scores, this method simply sets the threshold equal to the operating condition (cost proportion or skew). This is a natural criterion as it has been used especially when the model is a probability estimator and we expect to have perfect information about the operating condition at deployment time. In fact, this is a direct choice when working with proper scoring rules, since when rules are proper, scores are assumed to be a probabilistic assessment. The use of this threshold choice method can be traced back to Murphy (1966) and, perhaps, implicitly, much earlier. More recently, and in a different context from proper scoring rules, Drummond and Holte (2006) say "the performance independent criterion, in this case, is to set the threshold to correspond to the operating conditions. For example, if $PC(+) = 0.2$ the Naive Bayes threshold is set to 0.2". The term $PC(+)$ is equivalent to our 'skew'.

Definition 13 *Assuming the model's scores are expressed on a probability scale* $[0,1]$*, the* score-driven threshold choice method *is defined for cost proportions as follows:*

$$T_c^{sd}(c) \triangleq c \tag{7}$$

and for skews as

$$T_z^{sd}(z) \triangleq z.$$

Given this threshold choice method, then the question is: *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?* This is what we answer below.

Theorem 14 (Hernández-Orallo et al., 2011) *Assuming probabilistic scores and the score-driven threshold choice method, expected loss under a uniform distribution of cost proportions is equal to the model's Brier score.*

Proof If we plug T_c^{sd} (Equation 7) into the general formula of the expected loss (Equation 4) we have the expected score-driven loss:

$$L_c^{sd} \triangleq \int_0^1 Q_c(T_c^{sd}(c);c)w_c(c)dc = \int_0^1 Q_c(c;c)w_c(c)dc. \tag{8}$$

And if we use the uniform distribution and the definition of Q_c (Equation 3):

$$L_{U(c)}^{sd} = \int_0^1 Q_c(c;c)U(c)dc = \int_0^1 2\{c\pi_0(1-F_0(c)) + (1-c)\pi_1 F_1(c)\}dc. \tag{9}$$

In order to show this is equal to the Brier score, we expand the definition of BS_0 and BS_1 using integration by parts:

$$\begin{aligned} BS_0 &= \int_0^1 s^2 f_0(s)ds = \left[s^2 F_0(s)\right]_{s=0}^1 - \int_0^1 2sF_0(s)ds = 1 - \int_0^1 2sF_0(s)ds \\ &= \int_0^1 2sds - \int_0^1 2sF_0(s)ds = \int_0^1 2s(1-F_0(s))ds. \\ BS_1 &= \int_0^1 (1-s)^2 f_1(s)ds = \left[(1-s)^2 F_1(s)\right]_{s=0}^1 + \int_0^1 2(1-s)F_1(s)ds = \int_0^1 2(1-s)F_1(s)ds. \end{aligned}$$

Taking their weighted average, we obtain

$$BS = \pi_0 BS_0 + \pi_1 BS_1 = \int_0^1 \{\pi_0 2s(1-F_0(s)) + \pi_1 2(1-s)F_1(s)\}ds. \quad (10)$$

which, after reordering of terms and change of variable, is the same expression as Equation (9). ■

It is now clear why we just put the Brier score from Table 1 as the expected loss in Table 2. We calculated the expected loss for the score-driven threshold choice method for a uniform distribution of cost proportions as its Brier score.

Theorem 14 was obtained by Hernández-Orallo et al. (2011) (the threshold choice method there was called 'probabilistic') but it is not completely new in itself. Murphy (1966) found a similar relation to expected utility (in our notation, $-(1/4)PS + (1/2)(1+\pi_0)$, where the so-called probability score $PS = 2BS$). Apart from the sign (which is explained because Murphy works with utilities and we work with costs), the difference in the second constant term is explained because Murphy's utility (cost) model is based on a cost matrix where we have a cost for one of the classes (in meteorology the class 'protect') independently of whether we have a right or wrong prediction ('adverse' or 'good' weather). The only case in the matrix with a 0 cost is when we have 'good' weather and 'no protect'. It is interesting to see that the result only differs by a constant term, which supports the idea that whenever we can express the operating condition with a cost proportion or skew, the results will be portable to each situation with the inclusion of some constant terms (which are the same for all classifiers). In addition to this result, it is also worth mentioning another work by Murphy (1969) where he makes a general derivation for the Beta distribution.

After Murphy, in the last four decades, there has been extensive work on the so-called proper scoring rules, where several utility (cost) models have been used and several distributions for the cost have been used. This has led to relating Brier score (square loss), logarithmic loss, 0-1 loss and other losses which take the scores into account. For instance, Buja et al. (2005) give a comprehensive account of how all these losses can be obtained as special cases of the Beta distribution. The result given in Theorem 14 would be a particular case for the uniform distribution (which is a special case of the Beta distribution) and a variant of Murphy's results. In fact, the *BS* decomposition can also be connected to more general decompositions of Bregman divergences (Reid and Williamson, 2011). Nonetheless, it is important to remark that the results we have just obtained in Section 4.1 (and those we will get in Section 5) are new because they are not obtained by changing the cost distribution but rather by changing the threshold choice method. The threshold choice method used (the score-driven one) is not put into question in the area of proper scoring rules. But Theorem 14 can now be seen as a result which connects these two different dimensions: cost distribution and threshold choice method, so placing the Brier score at an even more predominant role.

Hernández-Orallo et al. (2011) derive an equivalent result using empirical distributions. In that paper we show how the loss can be plotted in cost space, leading to the *Brier curve* whose area underneath is the Brier score.

Finally, using skews we arrive at the prior-independent version of the Brier score.

Corollary 15 $L^{sd}_{U(z)} = uBS = (BS_0 + BS_1)/2$.

It is interesting to analyse the relation between $L^{su}_{U(c)}$ and $L^{sd}_{U(c)}$ (similarly between $L^{su}_{U(z)}$ and $L^{sd}_{U(z)}$). Since the former gives the *MAE* and the second gives the Brier score (which is the MSE),

from the definitions of *MAE* and Brier score, we get that, assuming scores are between 0 and 1:

$$MAE = L^{su}_{U(c)} \geq L^{sd}_{U(c)} = BS.$$

$$uMAE = L^{su}_{U(z)} \geq L^{sd}_{U(z)} = uBS.$$

Since *MAE* and *BS* have the same terms but the second squares them, and all the values which are squared are between 0 and 1, then the *BS* must be lower or equal. This is natural, since the expected loss is lower if we get reliable information about the operating condition at deployment time. So, the difference between the Brier score and *MAE* is precisely the gain we can get by having (and using) the information about the operating condition at deployment time. Notice that all this holds regardless of the quality of the probability estimates.

Finally, the difference between the results of Section 3 (Corollary 6) and these results fits well with a phenomenon which is observed when trying to optimise classification models: good probability estimation does not imply good classification and vice versa (see, for example, the work of Friedman, 1997). In the context of these results, we can re-interpret this phenomenon from a new perspective. The Brier score is seen as expected loss for the score-driven threshold choice method, while accuracy assumes a fixed threshold. The expected losses shown in Table 2 are a clear example of this.

5. Threshold Choice Methods Using Rates

We show in this section that *AUC* can be translated into expected loss for varying operating conditions in more than one way, depending on the threshold choice method used. We consider two threshold choice methods, where each of them sets the threshold to achieve a particular predicted positive rate: the rate-uniform method, which sets the rate in a uniform way; and the rate-driven method, which sets the rate equal to the operating condition. Some of these approaches have been used or mentioned in the literature, but choosing or ranging over sensitivity (or, complementary, specificity) instead of ranging over the *rate* (which is a weighted sum of sensitivity, that is, F_0, and $1-$ specificity, that is, F_1). For instance, Wieand et al. (1989) take a uniform distribution on a restricted range of sensitivities (or, similarly, specificities, Wieand et al., 1989). Also, Hand (2010) mentions that *AUC* can be seen as 'the mean specificity value, assuming a uniform distribution for the sensitivity'.

We recall the definition of a ROC curve and its area first.

Definition 16 *The ROC curve (Swets et al., 2000; Fawcett, 2006) is defined as a plot of $F_1(t)$ (i.e., false positive rate at decision threshold t) on the x-axis against $F_0(t)$ (true positive rate at t) on the y-axis, with both quantities monotonically non-decreasing with increasing t (remember that scores increase with $\hat{p}(1|x)$ and 1 stands for the negative class). The Area Under the ROC curve (AUC) is defined as:*

$$\begin{aligned} AUC &\triangleq \int_0^1 F_0(s)dF_1(s) = \int_{-\infty}^{+\infty} F_0(s)\,f_1(s)ds = \int_{-\infty}^{+\infty}\int_{-\infty}^{s} f_0(t)\,f_1(s)dtds \\ &= \int_0^1 (1-F_1(s))dF_0(s) = \int_{-\infty}^{+\infty} (1-F_1(s))\,f_0(s)ds = \int_{-\infty}^{+\infty}\int_{s}^{+\infty} f_1(t)\,f_0(s)dtds. \end{aligned}$$

Note that in this section scores are not necessarily assumed to be probability estimates and so s ranges from $-\infty$ to ∞.

5.1 The Rate Uniform Threshold Choice Method Leads to AUC

The rate-fixed threshold choice method places the threshold in such a way that a given predictive positive rate is achieved. However, this proportion may change, or we might not have reliable information about the operating condition *at deployment time*. An option in this case is to fix a predictive positive rate equal to 0.5 (predict exactly half of the examples as positive), which boils down to a special case of Theorem 5, but another option is to consider a non-deterministic choice or a distribution for this quantity. One natural choice can be a uniform distribution. This complete absence of information will hardly ever be the case, as we discussed for the score-uniform threshold choice method, but it is still instructive to explore what the outcome would be with this choice.

Definition 17 *The rate-uniform threshold choice method non-deterministically sets the threshold to achieve a uniformly randomly selected rate:*

$$T_c^{ru}(c) \triangleq T_c^{rf}[U_{0,1}](c).$$
$$T_z^{ru}(z) \triangleq T_z^{rf}[U_{0,1}](z).$$

In other words, it sets a relative quantity (from 0% positives to 100% positives) in a uniform way, and obtains the threshold from this uniform distribution over rates. Note that for a large number of examples, this is the same as defining a uniform distribution over examples or, alternatively, over cutpoints (between examples), as explored by Flach et al. (2011).

This threshold choice method is a generalisation of the rate-fixed threshold choice method which considers all the imbalances (class proportions) equally likely whenever we make a classification. It is important to clearly state that this makes the strong assumption that we will not have any information about the operating condition at deployment time.

As done before for other threshold choice methods, we analyse the question: given this threshold choice method, *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?*

The corresponding expected loss for cost proportions is (assuming R is invertible)

$$L_c^{ru} \triangleq \int_0^1 Q_c(T_c^{ru}(c);c)w_c(c)dc = \int_0^1 \int_0^1 Q_c(R^{-1}(r);c)U(r)w_c(c)dr\,dc.$$

We then have the following result.

Theorem 18 *Assuming the rate-uniform threshold choice method and invertible R, expected loss under a distribution of cost proportions w_c decreases linearly with AUC as follows:*

$$L_c^{ru} = \pi_0\pi_1(1-2AUC) + \pi_0\mathbb{E}_{w_c}\{c\} + \pi_1(1-\mathbb{E}_{w_c}\{c\}).$$

Proof First of all we note that $r = R(t)$ and hence $U(r)dr = R'(t)dt = \{\pi_0 f_0(t) + \pi_1 f_1(t)\}dt$. Under the same change of variable, $Q_c(R^{-1}(r);c) = Q_c(t;c) = 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}$. Hence:

$$\begin{aligned}
L_c^{ru} &= \int_0^1 \int_{-\infty}^{\infty} \{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} w_c(c) \{\pi_0 f_0(t) + \pi_1 f_1(t)\} dt\, dc \\
&= \int_{-\infty}^{\infty} \int_0^1 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \{\pi_0 f_0(t) + \pi_1 f_1(t)\} w_c(c) dc\, dt \\
&= \int_{-\infty}^{\infty} 2\{\mathbb{E}_{w_c}\{c\}\pi_0(1-F_0(t)) + (1-\mathbb{E}_{w_c}\{c\})\pi_1 F_1(t)\} \{\pi_0 f_0(t) + \pi_1 f_1(t)\} dt \\
&= 2\pi_0\pi_1 \mathbb{E}_{w_c}\{c\} \int_{-\infty}^{\infty} (1-F_0(t)) f_1(t) dt + 2\pi_0\pi_1 (1-\mathbb{E}_{w_c}\{c\}) \int_{-\infty}^{\infty} F_1(t) f_0(t) dt \\
&+ 2\pi_0^2 \mathbb{E}_{w_c}\{c\} \int_{-\infty}^{\infty} (1-F_0(t)) f_0(t)\, dt + 2\pi_1^2 (1-\mathbb{E}_{w_c}\{c\}) \int_{-\infty}^{\infty} F_1(t) f_1(t)\, dt.
\end{aligned}$$

The first two integrals in this last expression are both equal to $1-AUC$. The remaining two integrals reduce to a constant:

$$\int_{-\infty}^{\infty} (1-F_0(t)) f_0(t)\, dt = -\int_1^0 (1-F_0(t))\, d(1-F_0(t)) = 1/2.$$

$$\int_{-\infty}^{\infty} F_1(t) f_1(t)\, dt = \int_0^1 F_1(t)\, dF_1(t) = 1/2.$$

Putting everything together we obtain

$$\begin{aligned}
L_c^{ru} &= 2\pi_0\pi_1 \mathbb{E}_{w_c}\{c\}(1-AUC) + 2\pi_0\pi_1(1-\mathbb{E}_{w_c}\{c\})(1-AUC) + \pi_0^2 \mathbb{E}_{w_c}\{c\} + \pi_1^2(1-\mathbb{E}_{w_c}\{c\}) \\
&= 2\pi_0\pi_1(1-AUC) + \pi_0^2 \mathbb{E}_{w_c}\{c\} + \pi_1^2(1-\mathbb{E}_{w_c}\{c\}) \\
&= \pi_0\pi_1(1-2AUC) + \pi_0\pi_1 + \pi_0^2 \mathbb{E}_{w_c}\{c\} + \pi_1^2(1-\mathbb{E}_{w_c}\{c\}) \\
&= \pi_0\pi_1(1-2AUC) + \pi_0\pi_1 \mathbb{E}_{w_c}\{c\} + \pi_0^2 \mathbb{E}_{w_c}\{c\} + \pi_0\pi_1(1-\mathbb{E}_{w_c}\{c\}) + \pi_1^2(1-\mathbb{E}_{w_c}\{c\})
\end{aligned}$$

and the result follows. ■

The following two results were originally obtained by Flach, Hernández-Orallo, and Ferri (2011) for the special case of uniform cost and skew distributions. We are grateful to David Hand for suggesting that the earlier results might be generalised.

Corollary 19 *Assuming the rate-uniform threshold choice method, invertible R, and a distribution of cost proportions w_c with expected value $\mathbb{E}_{w_c}\{c\} = 1/2$, expected loss decreases linearly with AUC as follows:*

$$L^{ru}_{\mathbb{E}\{c\}=1/2} = \pi_0\pi_1(1-2AUC) + 1/2.$$

Corollary 20 *For any distribution of skews w_z, assuming the rate-uniform threshold choice method and invertible R, expected loss decreases linearly with AUC as follows:*

$$L_z^{ru} = (1-2AUC)/4 + 1/2.$$

Proof By assuming a uniform class distribution in Theorem 18 we obtain:

$$L_c^{ru} = \frac{1}{2}\frac{1}{2}(1-2AUC) + \frac{1}{2}\mathbb{E}_{w_c}\{c\} + \frac{1}{2}(1-\mathbb{E}_{w_c}\{c\}) = (1-2AUC)/4 + 1/2.$$

By Lemma 2 this is equal to L_z^{ru}. ■

Notice that Corollary 20 does not make any assumption about the expected value of w_z, and in that sense is more general than Corollary 19 for cost proportions. We see that expected loss for uniform skew ranges from 1/4 for a perfect ranker that is harmed by sub-optimal threshold choices, to 3/4 for the worst possible ranker that puts positives and negatives the wrong way round, yet gains some performance by putting the threshold at or close to one of the extremes.

Intuitively, a result like Corollary 20 can be understood as follows. Setting a randomly sampled rate is equivalent to setting the decision threshold to the score of a randomly sampled example. With probability π_0 we select a positive and with probability π_1 we select a negative. If we select a positive, then the expected true positive rate is $1/2$ (as on average we select the middle one); and the expected false positive rate is $1-AUC$ (as one interpretation of *AUC* is the expected proportion of negatives ranked correctly wrt. a random positive). Similarly, if we select a negative then the expected true positive rate is *AUC* and the expected false positive rate is $1/2$. Put together, the expected true positive rate is $\pi_0/2+\pi_1 AUC$ and the expected false positive rate is $\pi_1/2+\pi_0(1-AUC)$. The proportion of true positives among all examples is thus

$$\pi_0\,(\pi_0/2+\pi_1 AUC) = \frac{\pi_0^2}{2} + \pi_0\pi_1 AUC$$

and the proportion of false positives is

$$\pi_1\,(\pi_1/2+\pi_0(1-AUC)) = \frac{\pi_1^2}{2} + \pi_0\pi_1(1-AUC).$$

We can summarise these expectations in the following contingency table (all numbers are proportions relative to the total number of examples):

	Predicted +	Predicted −	
Actual +	$\pi_0^2/2+\pi_0\pi_1 AUC$	$\pi_0^2/2+\pi_0\pi_1(1-AUC)$	π_0
Actual −	$\pi_1^2/2+\pi_0\pi_1(1-AUC)$	$\pi_1^2/2+\pi_0\pi_1 AUC$	π_1
	$1/2$	$1/2$	1

The column totals are, of course, as expected: if we randomly select an example to split on, then the expected split is in the middle.

While in this paper we concentrate on the case where we have access to population densities $f_k(s)$ and distribution functions $F_k(t)$, in practice we have to work with empirical estimates. Flach et al. (2011) provides an alternative formulation of the main results in this section, relating empirical loss to the *AUC* of the empirical ROC curve. For instance, the expected loss for uniform skew and uniform instance selection is calculated by Flach et al. (2011) to be $\left(\frac{n}{n+1}\right)\frac{1-2AUC}{4}+\frac{1}{2}$, showing that for smaller samples the reduction in loss due to *AUC* is somewhat smaller.

5.2 The Rate-Driven Threshold Choice Method Leads to AUC

Naturally, if we can have precise information of the operating condition at deployment time, we can use the information about the skew or cost to adjust the rate of positives and negatives to that proportion. This leads to a new threshold selection method: if we are given skew (or cost proportion) z (or c), we choose the threshold t in such a way that we get a proportion of z (or c) positives. This is an elaboration of the rate-fixed threshold choice method which *does* take the operating condition into account.

Definition 21 *The* rate-driven threshold choice method *for cost proportions is defined as*

$$T_c^{rd}(c) \triangleq T_c^{rf}[c](c) = R^{-1}(c). \tag{11}$$

The rate-driven threshold choice method for skews is defined as

$$T_z^{rd}(z) \triangleq T_z^{rf}[z](z) = R_z^{-1}(z).$$

Given this threshold choice method, the question is again: *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?* This is what we answer below.

If we plug T_c^{rd} (Equation 11) into the general formula of the expected loss for a range of cost proportions (Equation 4) we have:

$$L_c^{rd} \triangleq \int_0^1 Q_c(T_c^{rd}(c);c)w_c(c)dc.$$

And now, from this definition, if we use the uniform distribution for $w_c(c)$, we obtain this new result.

Theorem 22 *Expected loss for uniform cost proportions using the rate-driven threshold choice method is linearly related to AUC as follows:*

$$L_{U(c)}^{rd} = \pi_1\pi_0(1-2AUC) + 1/3.$$

Proof

$$\begin{aligned} L_{U(c)}^{rd} &= \int_0^1 Q_c(T_c^{rd}(c);c)U(c)dc = \int_0^1 Q_c(R^{-1}(c);c)dc \\ &= 2\int_0^1 \{c\pi_0(1-F_0(R^{-1}(c))) + (1-c)\pi_1F_1(R^{-1}(c))\}dc \\ &= 2\int_0^1 \{c\pi_0 - c\pi_0F_0(R^{-1}(c))) + \pi_1F_1(R^{-}1(c)) - c\pi_1F_1(R^{-1}(c))\}dc. \end{aligned}$$

Since $\pi_0F_0(R^{-}1(c))) + \pi_1F_1(R^{-1}(c)) = R(R^{-1}(c)) = c$,

$$\begin{aligned} L_{U(c)}^{rd} &= 2\int_0^1 \{c\pi_0 - c^2 + \pi_1F_1(R^{-1}(c))\}dc \\ &= \pi_0 - \frac{2}{3} + 2\pi_1\int_0^1 F_1(R^{-1}(c))dc. \end{aligned}$$

Taking the rightmost term and using the change of variable $R^{-1}(c) = t$ we have $c = R(t)$ and hence $dc = R'(t)dt = \{\pi_0 f_0(t) + \pi_1 f_1(t)\}dt = R'(t)dt$, and thus this term is rewritten as

$$\begin{aligned} 2\pi_1 \int_0^1 F_1(R^{-1}(c))dc &= 2\pi_1 \int_\infty^\infty F_1(t)\{\pi_0 f_0(t) + \pi_1 f_1(t)\}dt \\ &= 2\pi_1\pi_0 \int_0^1 F_1(t)dF_0(t) + 2\pi_1^2 \int_0^1 F_1(t)dF_1(t) \\ &= 2\pi_1\pi_0(1-AUC) + 2\pi_1^2\frac{1}{2} = 2\pi_1\pi_0(1-AUC) + \pi_1(1-\pi_0). \end{aligned}$$

Putting everything together we have:

$$\begin{aligned} L^{rd}_{U(c)} &= \pi_0 - \frac{2}{3} + 2\pi_1\pi_0(1-AUC) + \pi_1(1-\pi_0) \\ &= \frac{1}{3} + \pi_1\pi_0(1-2AUC). \end{aligned}$$

■

Now we can unveil and understand how we obtained the results for the expected loss in Table 2 for the rate-driven method. We just took the *AUC* of the models and applied the previous formula: $\pi_1\pi_0(1-2AUC) + \frac{1}{3}$.

Corollary 23 *Expected loss for uniform skews using the rate-driven threshold choice method is linearly related to AUC as follows:*

$$L^{rd}_{U(z)} = (1-2AUC)/4 + 1/3.$$

If we compare Corollary 20 with Corollary 23, we see that $L^{ru}_{U(z)} > L^{rd}_{U(z)}$, more precisely:

$$L^{ru}_{U(z)} = (1-2AUC)/4 + 1/2 = L^{rd}_{U(z)} + 1/6.$$

So we see that taking the operating condition into account when choosing thresholds based on rates reduces the expected loss with $1/6$, *regardless of the quality of the model* as measured by *AUC*. This term is clearly not negligible and demonstrates that the rate-driven threshold choice method is superior to the rate-uniform method. Figure 3 illustrates this. Logically, $L^{rd}_{U(c)}$ and $L^{rd}_{U(z)}$ work upon information about the operating condition at deployment time, while $L^{ru}_{U(c)}$ and $L^{ru}_{U(z)}$ may be suited when this information is unavailable or unreliable.

6. The Optimal Threshold Choice Method

The last threshold choice method we investigate is based on the optimistic assumption that (1) we are having complete information about the operating condition (class proportions and costs) at deployment time and (2) we are able to use that information (also at deployment time) to choose the threshold that will minimise the loss using the current model. ROC analysis is precisely based on these two points since we can calculate the threshold which gives the smallest loss by using the skew and the convex hull.

This threshold choice method, denoted by T^o_c, is defined as follows:

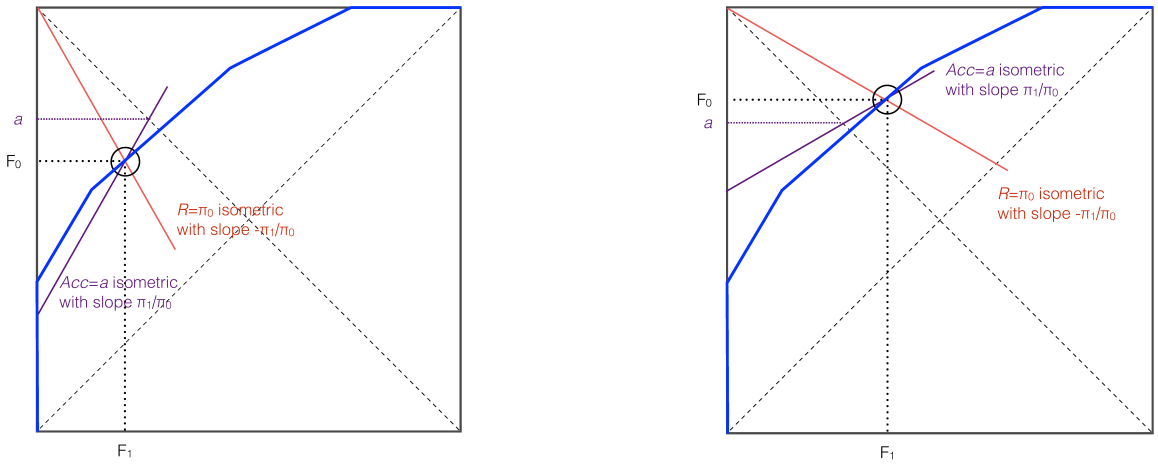

Figure 3: Illustration of the rate-driven threshold choice method. We assume uniform misclassification costs ($c_0 = c_1 = 1$), and hence skew is equal to the proportion of positives ($z = \pi_0$). The majority class is class 1 on the left and class 0 on the right. Unlike the rate-uniform method, the rate-driven method is able to take advantage of knowing the majority class, leading to a lower expected loss.

Definition 24 *The optimal threshold choice method is defined as:*

$$T_c^o(c) \triangleq \arg\min_t \{Q_c(t;c)\} = \arg\min_t 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \tag{12}$$

and similarly for skews:

$$T_z^o(z) \triangleq \arg\min_t \{Q_z(t;z)\}.$$

Note that in both cases, the arg min will typically give a range (interval) of values which give the same optimal value. So these methods can be considered non-deterministic. This threshold choice method is analysed by Fawcett and Provost (1997), and used by Drummond and Holte (2000, 2006) for defining their cost curves and by Hand (2009) to define a new performance metric.

If we plug Equations (12) and (3) into Equation (4) using a uniform distribution for cost proportions, we get:

$$\begin{aligned} L^o_{U(c)} &= \int_0^1 Q_c(\arg\min_t\{Q_c(t,c)\};c)dc = \int_0^1 \min_t\{Q_c(t;c)\}dc \\ &= \int_0^1 \min_t\{2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\}dc. \end{aligned} \tag{13}$$

The connection with the convex hull of a ROC curve (ROCCH) is straightforward. The convex hull is a construction over the ROC curve in such a way that all the points on the convex hull have minimum loss for some choice of c or z. This means that we restrict attention to the *optimal* threshold for a given cost proportion c, as derived from Equation (12).

6.1 Convexification

We can give a corresponding, and more formal, definition of the convex hull as derived from the score distributions. First, we need a more precise definition of a convex model. For that, we rely on the ROC curve, and we use the slope of the curve, defined as usual:

$$slope(T) = \frac{f_0(T)}{f_1(T)}.$$

A related expression we will also use is:

$$c(T) = \frac{\pi_1 f_1(T)}{\pi_0 f_0(T) + \pi_1 f_1(T)}.$$

Sometimes we will use subindices for $c(T)$ depending on the model we are using. In this way, we have, $\frac{\pi_0}{\pi_1} slope(T) = \frac{\pi_0 f_0(T)}{\pi_1 f_1(T)} = \frac{1}{c(T)} - 1$.

Definition 25 (Convex model) *A model m is convex, if for every threshold T, we have that $c(T)$ is non-decreasing (or, equivalently, slope(T) is non-increasing).*

In order to make any model convex, it is not sufficient to repair local concavities, we need to calculate the convex hull. This is clear if we categorise the types of segments. Some threshold values t will never minimise $Q_c(t;c) = 2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\}$ for any value of c. These values will be in one or more intervals of which only the end points will minimise $Q_c(t;c)$ for some value of c. We will call these intervals *non-hull intervals*, and all the rest will be referred to as *hull intervals*. It clearly holds that hull intervals are convex. Non-hull intervals may contain convex and concave subintervals.

From here, a definition of convex hull for continuous distributions is given as follows:

Definition 26 (Convexification) *Let m be any model with score distributions $f_0(T)$ and $f_1(T)$. Define convexified score distributions $e_0(T)$ and $e_1(T)$ as follows.*

1. *For every hull interval $t_{i-1} \leq s \leq t_i$: $e_0(T) = f_0(T)$ and $e_1(T) = f_1(T)$.*
2. *For every non-hull interval $t_{j-1} \leq s \leq t_j$:*

$$e_0(T) = e_{0,j} = \frac{1}{t_j - t_{j-1}} \int_{t_{j-1}}^{t_j} f_0(T) dT.$$
$$e_1(T) = e_{1,j} = \frac{1}{t_j - t_{j-1}} \int_{t_{j-1}}^{t_j} f_1(T) dT.$$

The function Conv *returns the model* Conv(m) *defined by the score distributions $e_0(T)$ and $e_1(T)$.*

We can also define the cumulative distributions $E_x(t) = \int_0^t e_x(T)dT$, where x represents either 0 or 1. By construction we have that for every interval $[t_{j-1}, t_j]$ identified above:

$$[E_x(t)]_{t_{j-1}}^{t_j} = \int_{t_{j-1}}^{t_j} e_x(T)dT = (t_j - t_{j-1}) e_{x,j} = \int_{t_{j-1}}^{t_j} f_x(T)dT = [F_x(t)]_{t_{j-1}}^{t_j} \quad (14)$$

and so the convexified score distributions are proper distributions. Furthermore, since the new score distributions are constant in the convexified intervals—and hence monotonically non-decreasing for the new $c(T)$, denoted by $c_{\text{Conv}(m)}(T)$—so is

$$c_{\text{Conv}(m)}(T) = c_j = \frac{\pi_1 e_{1,j}}{\pi_0 e_{0,j} + \pi_1 e_{1,j}}.$$

It follows that $\text{Conv}(m)$ is everywhere convex. In addition,

Theorem 27 *Optimal loss is invariant under* Conv, *that is:* $L^o_{U(c)}(\text{Conv}(m)) = L^o_{U(c)}(m)$ *for every* m.

Proof By Equation (13) we have that optimal loss is:

$$L^o_{U(c)}(m) \quad = \quad \int_0^1 \min_t \{2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\} dc.$$

By definition, the hull intervals have not been modified by $\text{Conv}(m)$. Only the non-hull intervals have been modified. A non-hull interval was defined as those where there is no t which minimises $Q_c(t;c) = 2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\}$ for any value of c, and only the endpoints attained the minimum. Consequently, we only need to show that the new $e_0(T)$ and $e_1(T)$ do not introduce any new minima.

We now focus on each non-hull segment (t_{j-1}, t_j) using the definition of Conv. We only need to check the expression for the minimum:

$$\min_{t_j \le t \le t_{j-1}} \{2c\pi_0(1-E_0(t)) + 2(1-c)\pi_1 E_1(t)\}.$$

From Equation (14) we derive that $E_x(t) = E_x(t_{j-1}) + (t_j - t_{j-1})e_{x,j}$ inside the interval (they are straight lines in the ROC curve), and we can see that the expression to be minimised is constant (it does not depend on t). Since the end points were the old minima and were equal, we see that this expression cannot find new minima. ■

It is not difficult to see that if we plot $\text{Conv}(m)$ in the cost space defined by Drummond and Holte (2006) with $Q_z(t;z)$ on the y-axis against skew z on the x-axis, we have a cost curve. Its area is then the expected loss for the optimal threshold choice method. In other words, this is the area under the (optimal) cost curve. Similarly, the new performance metric H introduced by Hand (2009) is simply a rescaled version of the area under the optimal cost curve using the $\beta_{2,2}$ distribution instead of the $\beta_{1,1}$ (i.e., uniform) distribution, and using cost proportions instead of skews (so being dependent to class priors). This is further discussed by Flach et al. (2011). While all of these distributions are symmetric, the Beta distribution can be non-symmetric if required by a specific application. In fact, Hand and Anagnostopoulos (2012) suggest that the parameters of the distribution should be linked to class proportion π_0.

6.2 The Optimal Threshold Choice Method Leads to Refinement Loss

Once again, the question now must be stated clearly. Assume that the optimal threshold choice method is set as the method we will use for every application of our model. Furthermore, assume

that each and every application of the model is going to find the perfect threshold. Then, *if we must evaluate a model before application for a wide range of skews and cost proportions, which performance metric should be used?* In what follows, we will find the answer by relating this expected loss with a genuine performance metric: refinement loss. We will now introduce this performance metric.

The Brier score, being a sum of squared probabilistic residuals, can be decomposed in various ways. The most common decomposition of the Brier score is due to Murphy (1973) and decomposes the Brier score into Reliability, Resolution and Uncertainty. Frequently, the two latter components are joined together and the decomposition gives two terms: calibration loss and refinement loss.

This decomposition is usually applied to empirical distributions, requiring a binning of the scores. Scores are assumed to be probability estimates in the interval $[0,1]$. The decomposition is based on a partition $\mathcal{P}_D = \{b_j\}_{j=1..B}$ where D is the data set, B the number of bins, and each bin is denoted by $b_j \subset D$. Since it is a partition $\bigcup_{j=1}^{B} b_j = D$. With this partition the decompoistion is:

$$BS \approx CL^{\mathcal{P}_D} + RL^{\mathcal{P}_D} = \frac{1}{n}\sum_{j=1}^{B} |b_j| \left(s_{b_j} - y_{b_j}\right)^2 + \frac{1}{n}\sum_{j=1}^{B} |b_j| y_{b_j} \left(1 - y_{b_j}\right).$$

Here we use the notation $s_{b_j} = \frac{1}{|b_j|}\sum_{i \in b_j} s_i$ and $y_{b_j} = \frac{1}{|b_j|}\sum_{i \in b_j} y_i$ for the average predicted scores and the average actual classes respectively for bin b_j.

For many partitions the empirical decomposition is not exact. It is only exact for partitions which are coarser than the partition induced by the ROC curve (i.e., ties cannot be spread over different partitions), as shown by Flach and Matsubara (2007). We denote by CL^{ROC} and RL^{ROC} the calibration loss and the refinement loss, respectively, using the segments of the empirical ROC curve as bins. In this case, $BS = CL^{ROC} + RL^{ROC}$.

In this paper we will use a variant of the above decomposition based on the ROC convex hull of a model. In this decomposition, we take each bin as each segment in the convex hull. Naturally, the number of bins in this decomposition is lower or equal than the number of bins in the ROC decomposition. In fact, we may find different values of s_i in the same bin. In some way, we can think about this decomposition as an optimistic/optimal version of the ROC decomposition, as Flach and Matsubara (2007, Th. 3) show. We denote by CL^{ROCCH} and RL^{ROCCH} the calibration loss and the refinement loss, respectively, using the segments of the convex hull of the empirical ROC curve as bins (Flach et al., 2011).

We can define the same decomposition in continuous terms considering Definition 8. We can see that in the continuous case, the partition is irrelevant. Any partition will give the same result, since the composition of consecutive integrals is the same as the whole integral.

Theorem 28 *The continuous decomposition of the Brier Score, $BS = CL + RL$, is exact and gives CL and RL as follows.*

$$CL = \int_0^1 \frac{(s(\pi_0 f_0(s) + \pi_1 f_1(s)) - \pi_1 f_1(s))^2}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds.$$
$$RL = \int_0^1 \frac{\pi_1 f_1(s) \pi_0 f_0(s)}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds.$$

Proof

$$\begin{aligned}
BS &= \int_0^1 \left[s^2\pi_0 f_0(s) + (1-s)^2\pi_1 f_1(s)\right] ds \\
&= \int_0^1 \left[s^2(\pi_0 f_0(s) + \pi_1 f_1(s)) - 2s\pi_1 f_1(s) + \pi_1 f_1(s)\right] ds \\
&= \int_0^1 \frac{s^2(\pi_0 f_0(s) + \pi_1 f_1(s))^2 - 2s(\pi_0 f_0(s) + \pi_1 f_1(s))\pi_1 f_1(s) + \pi_1 f_1(s)(\pi_1 f_1(s) + \pi_0 f_0(s))}{(\pi_0 f_0(s) + \pi_1 f_1(s))} ds \\
&= \int_0^1 \frac{(s(\pi_0 f_0(s) + \pi_1 f_1(s)) - \pi_1 f_1(s))^2 + \pi_1 f_1(s)\pi_0 f_0(s)}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds \\
&= \int_0^1 \frac{(s(\pi_0 f_0(s) + \pi_1 f_1(s)) - \pi_1 f_1(s))^2}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds + \int_0^1 \frac{\pi_1 f_1(s)\pi_0 f_0(s)}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds.
\end{aligned}$$

■

This proof keeps the integral from start to end. That means that the decomposition is not only true for the integral as a whole, but also pointwise for every single score s. Note that y_{b_j} in the empirical case (see Definition 15) corresponds to $c(s) = \frac{\pi_1 f_1(s)}{\pi_0 f_0(s) + \pi_1 f_1(s)}$ (as given by Equation 14) in the continuous case above, and also note that s_{b_j} corresponds to the cardinality $\pi_0 f_0(s) + \pi_1 f_1(s)$. The decomposition for empirical distributions as introduced by Murphy (1973) is still predominant for any reference to the decomposition. To our knowledge this is the first explicit derivation of a continuous version of the decomposition.

And now we are ready for relating the optimal threshold choice method with a performance metric as follows:

Theorem 29 *For every convex model m, we have that:*

$$L^o_{U(c)}(m) = RL(m).$$

The proof of this theorem is found in the appendix as Theorem 48.

Corollary 30 *For every model m the expected loss for the optimal threshold choice method $L^o_{U(c)}$ is equal to the refinement loss using the convex hull.*

$$L^o_{U(c)}(m) = RL(\mathrm{Conv}(m)) \triangleq RL_{\mathrm{Conv}}(m).$$

Proof We have $L^o_{U(c)}(m) = L^o_{U(c)}(\mathrm{Conv}(m))$ by Theorem 27, and $L^o_{U(c)}(\mathrm{Conv}(m)) = RL(\mathrm{Conv}(m))$ by Theorem 29 and the convexity of $\mathrm{Conv}(m)$. ■

It is possible to obtain a version of this theorem for empirical distributions which states that $L^o_{U(c)} = RL^{ROCCH}$ where RL^{ROCCH} is the refinement loss of the empirical distribution using the segments of the convex hull for the decomposition.

Before analysing what the meaning of this threshold choice method is and how it relates to the rest, we have to consider whether this threshold choice method is realistic or not. In the beginning of this section we said that the optimal method assumes that (1) we are having complete information about the operating condition at deployment time and (2) we are able to use that information to choose the threshold that will minimise the loss at deployment time.

While (1) is not always true, there are many occasions where we know the costs and distributions at application time. This is the base of the score-driven and rate-driven methods. However, having this information does not mean that the optimal threshold for a data set (e.g., the training or validation data set) ensures an optimal choice for a test set (2). Drummond and Holte (2006) are conscious of this problem and they reluctantly rely on a threshold choice method which is based on "the ROC convex hull [...] only if this selection criterion happens to make cost-minimizing selections, which in general it will not do". But even if these cost-minimising selections are done, as mentioned above, it is not clear how reliable they are for a test data set. As Drummond and Holte (2006, p. 122) recognise: "there are few examples of the practical application of this technique. One example is given by Fawcett and Provost (1997), in which the decision threshold parameter was tuned to be optimal, empirically, for the test distribution".

In the example shown in Table 2 in Section 1, the evaluation technique was training and test. However, with cross-validation, the convex hull cannot be estimated reliably in general, and the thresholds derived from each fold might be inconsistent. Even with a large validation data set, the decision threshold may be suboptimal. This is one of the reasons why the area under the convex hull has not been used as a performance metric. In any case, we can calculate the values as an optimistic limit, leading to $L^o_{U(c)} = RL^{ROCCH} = 0.0953$ for model A and 0.2094 for model B.

7. Relating Performance Metrics

So far, we have repeatedly answered the following question: "If threshold choice method X is used, which is the corresponding performance metric?" The answers are summarised in Table 4. The seven threshold choice methods are shown in the first column (the two fixed methods are grouped in the same row). The integrated view of performance metrics for classification is given by the next two columns. The expected loss of a model for a uniform distribution of cost proportions or skews for each of these seven threshold choice methods produces most of the common performance metrics in classification: 0-1 loss (either weighted or unweighted accuracy), the Mean Absolute Error (equivalent to Mean Probability Rate), the Brier score, *AUC* (which equals the Wilcoxon-Mann-Whitney statistic and the Kendall tau distance of the model to the perfect model, and is linearly related to the Gini coefficient) and, finally, the refinement loss using the bins given by the convex hull.

All the threshold choice methods seen in this paper consider model scores in different ways. Some of them disregard the score, since the threshold is fixed, some others consider the 'magnitude' of the score as an (accurate) estimated probability, leading to the score-based methods, and others consider the 'rank', 'rate' or 'proportion' given by the scores, leading to the rate-based methods. Since the optimal threshold choice is also based on the convex hull, it is apparently more related to the rate-based methods. This is consistent with the taxonomy proposed by Ferri et al. (2009) based on correlations over more than a dozen performance metrics, where three families of metrics were recognised: performance metrics which account for the quality of classification (such as accuracy), performance metrics which account for a ranking quality (such as *AUC*), and performance metrics which evaluate the quality of scores or how well the model does in terms of probability estimation (such as the Brier score or logloss).

This suggests that the way scores are distributed is crucial in understanding the differences and connections between these metrics. In addition, this may shed light on which threshold choice method is best. We have already seen some relations, such as $L^{su}_{U(c)} \geq L^{sd}_{U(c)}$, and $L^{ru}_{U(c)} > L^{rd}_{U(c)}$, but

Threshold choice method	Cost proportions	Skews	Equivalent (or related) performance metrics
fixed	$L^{sf}_{U(c)} = 1 - Acc$	$L^{sf}_{U(z)} = 1 - uAcc$	0-1 loss: Weighted and unweighted accuracy.
score-uniform	$L^{su}_{U(c)} = MAE$	$L^{su}_{U(z)} = uMAE$	Absolute error, Average score, *pAUC* (Ferri et al., 2005) , Probability Rate (Ferri et al., 2009).
score-driven	$L^{sd}_{U(c)} = BS$	$L^{sd}_{U(z)} = uBS$	Brier score (Brier, 1950), Mean Squared Error (*MSE*).
rate-uniform	$L^{ru}_{U(c)} = \pi_0\pi_1(1-2AUC)+\frac{1}{2}$	$L^{ru}_{U(z)} = \frac{1-2AUC}{4}+\frac{1}{2}$	*AUC* (Swets et al., 2000) and variants (*wAUC*) (Fawcett, 2001; Ferri et al., 2009), Kendall tau, WMW statistic, Gini coefficient.
rate-driven	$L^{rd}_{U(c)} = \pi_0\pi_1(1-2AUC)+\frac{1}{3}$	$L^{rd}_{U(z)} = \frac{1-2AUC}{4}+\frac{1}{3}$	*AUC* (Swets et al., 2000) and variants (*wAUC*) (Fawcett, 2001; Ferri et al., 2009), Kendall tau, WMW statistic, Gini coefficient.
optimal	$L^{o}_{U(c)} = RL_{\text{Conv}}$	$L^{o}_{U(z)} = uRL_{\text{Conv}}$	ROCCH Refinement loss (Flach and Matsubara, 2007), Refinement Loss (Murphy, 1973), Area under the Cost Curve ('Total Expected Cost') (Drummond and Holte, 2006), Hand's H (Hand, 2009).

Table 4: Threshold choice methods and their expected loss for cost proportions and skews. The *u* in *uAcc*, *uMAE*, *uBS* and *uRL* mean that these metrics are unweighted, that is, calculated as if $\pi_0 = \pi_1$, while the *w* in *wAUC* refers to a weighted version of the AUC, and the *x*-axis and *y*-axis are proportional to π_0 and π_1.

what about $L^{sd}_{U(c)}$ and $L^{rd}_{U(c)}$? Are they comparable? And what about $L^{o}_{U(c)}$? It gives the minimum expected loss by definition over the training (or validation) data set, but when does it become a good estimation of the expected loss for the test data set?

In order to answer these questions we need to analyse transformations on the scores and see how these affect the expected loss given by each threshold choice method. For the rest of the section we assume that scores are in the interval $[0,1]$. Given a model, its scores establish a total order over the examples: $\sigma = (s_1, s_2, ..., s_n)$ where $s_i \leq s_{i+1}$. Since there might be ties in the scores, this total order is not necessarily strict. A monotonic transformation is any alteration of the scores, such that the order is kept. We will consider two transformations: the evenly-spaced transformation and PAV calibration.

7.1 Evenly-Spaced Scores. Relating Brier Score, MAE and AUC

If we are given a ranking or order, or we are given a set of scores but its reliability is low, a quite simple way to assign (or re-assign) the scores is to set them evenly-spaced (in the $[0,1]$ interval).

Definition 31 *A discrete evenly-spaced transformation is a procedure* $\text{EST}(\sigma) \to \sigma'$ *which converts any sequence of scores* $\sigma = (s_1, s_2, ..., s_n)$ *where* $s_i < s_{i+1}$ *into scores* $\sigma' = (s'_1, s'_2, ..., s'_n)$ *where* $s'_i = \frac{i-1}{n-1}$.

Notice that such a transformation does not affect the ranking and hence does not alter the *AUC*.

The previous definition can be applied to continuous score distribution as follows:

Definition 32 *A continuous evenly-spaced transformation is a any strictly monotonic transformation function on the score distribution, denoted by* Even, *such that for the new scores* s' *it holds that* $P(s' \leq t) = t$.

It is easy to see that EST is idempotent, that is, $\text{EST}(\text{EST}(\sigma)) = \text{EST}(\sigma)$. So we say a set of scores σ is evenly-spaced if $\text{EST}(\sigma) = \sigma$.

Lemma 33 *Given a model and data set with set of scores σ, such that they evenly-spaced, when $n \to \infty$ then we have $R(t) = t$.*

Proof Remember that by definition the true positive rate $F_0(t) = P(s \leq t|0)$ and the false positive rate $F_1(t) = P(s \leq t|1)$. Consequently, from the definition of rate we have $R(t) = \pi_0 F_0(t) + \pi_1 F_1(t) = \pi_0 P(s \leq t|0) + \pi_1 P(s \leq t|1) = P(s \leq t)$. But, since the scores are evenly-spaced, the number of scores such that $s \leq t$ is $\sum_{i=1}^{n} I(s_i \leq t) = \sum_{i=1}^{n} I(\frac{i-1}{n-1} \leq t)$ with I being the indicator function (1 when true, 0 otherwise). This number of scores is $\sum_{i=1}^{tn} 1$ when $n \to \infty$, which clearly gives tn. So the probability $P(s \leq t)$ is $tn/n = t$. Consequently $R(t) = t$. ■

The following results connect the score-driven threshold choice method with the rate-driven threshold choice method:

Theorem 34 *Given a model and data set with set of scores σ, such that they are evenly-spaced, when $n \to \infty$:*

$$BS = L^{sd}_{U(c)} = L^{rd}_{U(c)} = \pi_0 \pi_1 (1 - 2AUC) + \frac{1}{3}.$$

Proof By Lemma 33 we have $R(t) = t$, and so the rate-driven and score-driven threshold choice methods select the same thresholds. ■

Corollary 35 *Given a model and data set with set of scores σ such that they are evenly-spaced, when $n \to \infty$:*

$$uBS = L^{sd}_{U(z)} = L^{rd}_{U(z)} = \frac{1 - 2AUC}{4} + \frac{1}{3}.$$

These straightforward results connect *AUC* and Brier score for evenly-spaced scores. This connection is enlightening because it says that *AUC* and *BS* are equivalent performance metrics (linearly related) when we set the scores in an evenly-spaced way. In other words, it says that *AUC* is like a Brier score which considers all the scores evenly-spaced. Although the condition is strong, this is the first linear connection which, to our knowledge, has been established so far between *AUC* and the Brier score.

Similarly, we get the same results for the score-uniform threshold choice method and the rate-uniform threshold choice method.

Theorem 36 *Given a model and data set with set of scores σ such that they are evenly-spaced, when $n \to \infty$:*

$$MAE = L^{su}_{U(c)} = L^{ru}_{U(c)} = \pi_0 \pi_1 (1 - 2AUC) + \frac{1}{2}$$

with similar results for skews. This also connects *MAE* with *AUC* and clarifies when they are linearly related.

7.2 Perfectly-Calibrated Scores. Relating BS, CL and RL

In this section we will work with a different condition on the scores. We will study what interesting connections can be established if we assume the scores to be perfectly calibrated.

The informal definition of perfect calibration usually says that a model is calibrated when the estimated probabilities are close to the true probabilities. From this informal definition, we would derive that a model is perfectly calibrated if the estimated probability given by the scores (i.e., $\hat{p}(1|x)$) equals the true probability. However, if this definition is applied to single instances, it implies not only perfect calibration but a perfect model. In order to give a more meaningful definition, the notion of calibration is then usually defined in terms of groups or bins of examples, as we did, for instance, with the Brier score decomposition. So, we need to apply this correspondence between estimated and true (actual) probabilities over bins. We say a bin partition is invariant on the scores if for any two examples with the same score they are in the same bin. In other words, two equal scores cannot be in different bins (equivalence classes cannot be broken). From here, we can give a definition of perfect calibration:

Definition 37 (Perfectly-calibrated for empirical distribution models) *A model is* perfectly calibrated *if for any invariant bin partition $\mathcal{P}$ we have that $y_{b_j} = s_{b_j}$ for all its bins: that is, the average actual probability equals the average estimated probability, thus making $CL^{\mathcal{P}} = 0$.*

Note that it is not sufficient to have $CL = 0$ for one partition, but for all the invariant partitions. Also notice that the bins which are generated by a ROC curve are the minimal invariant partition on the scores (i.e., the quotient set). So, we can give an alternative definition of perfectly calibrated model: a model is perfectly calibrated if and only if $CL^{ROC} = 0$. For the continuous case, the partition is irrelevant and the definition is as follows:

Definition 38 (Perfectly-calibrated for continuous distribution models) *A continuous model is* perfectly calibrated *if and only if $s = \frac{\pi_1 f_1(s)}{\pi_0 f_0(s)+\pi_1 f_1(s)}$ which is exactly $c(s)$ given by Equation (14).*

Note that the previous definition is equivalent to saying that $CL = 0$, as in the empirical case, since CL can be rewritten as follows, following the decomposition of Theorem 28:

$$\begin{aligned} CL &= \int_0^1 \frac{(s(\pi_0 f_0(s) + \pi_1 f_1(s)) - \pi_1 f_1(s))^2}{\pi_0 f_0(s) + \pi_1 f_1(s)} ds \\ &= \int_0^1 (\pi_0 f_0(s) + \pi_1 f_1(s)) \left(s - \frac{\pi_1 f_1(s)}{\pi_0 f_0(s) + \pi_1 f_1(s)} \right)^2 ds. \end{aligned}$$

Lemma 39 *For a perfectly calibrated classifier m:*

$$\frac{1-s}{s} = \frac{f_0(s)}{f_1(s)} \frac{\pi_0}{\pi_1}$$

and m is convex.

Proof The expression is a direct transformation of Definition 38 and convexity just follows from Definition 25. ■

Now that we have two proper and operational definitions of perfect calibration, we define a calibration transformation as follows.

Definition 40 Cal *is a monotonic function over the scores which converts any model m into another calibrated model* m^* *such that* $CL = 0$ *and RL is not modified.*

Cal always produces a convex model, so $\text{Conv}(\text{Cal}(m)) = \text{Cal}(m)$, but a convex model is not always perfectly calibrated (e.g., a binormal model with same variances is always convex but it can be uncalibrated), so $\text{Cal}(\text{Conv}(m) \neq \text{Conv}(m)$. This is summarised in Table 5. If the model is strictly convex, then Cal is strictly monotonic. An instance of the function Cal is the transformation $T \mapsto s = c(T)$ where $c(T) = \frac{\pi_1 f_1(T)}{\pi_0 f_0(T) + \pi_1 f_1(T)}$ as given by Equation (14). This transformation is shown to keep *RL* unchanged in the appendix and makes $CL = 0$.

The previous function is defined for continuous score distributions. The corresponding function for empirical distributions is known as the Pool Adjacent Algorithm (PAV) (Ayer et al., 1955). Following Fawcett and Niculescu-Mizil (2007), the *PAV* function converts any model *m* into another calibrated model m^* such that the following property $s_{b_j} = y_{b_j}$ holds for every segment in its convex hull.

	Evenly-spaced	Convexification	Perfect Calibration
Continuous distributions	Even	Conv	Cal
Empirical distributions	EST	ROCCH	PAV

Table 5: Transformations on scores. Perfect calibration implies a convex model but not vice versa.

Fawcett and Niculescu-Mizil (2007) have shown that isotonic-based calibration (Robertson et al., 1988) is equivalent to the PAV algorithm, and closely related to ROCCH, since, for every *m* and data set, we have:

$$\begin{aligned} BS(PAV(m)) &= CL^{ROC}(PAV(m)) + RL^{ROC}(PAV(m)) = CL^{ROCCH}(PAV(m)) + RL^{ROCCH}(PAV(m)) \\ &= RL^{ROC}(PAV(m)) = RL^{ROCCH}(PAV(m)). \end{aligned}$$

It is also insightful to see that isotonic regression (calibration) is the monotonic function defined as $\arg\min_f \sum (y_i - f(s_i))^2$, that is, the monotonic function over the scores which minimises the Brier score. This leads to the same function if we use any other proper scoring function (such as logloss).

The similar expression for the continuous case is

$$BS(\text{Cal}(m)) = CL(\text{Cal}(m)) + RL(\text{Cal}(m)) = RL(\text{Cal}(m)).$$

Now we analyse what happens with perfectly calibrated models for the score-driven threshold choice and the score-uniform threshold choice methods. This will help us understand the similarities and differences between the threshold choices and their relation with the optimal method. Along the way, we will obtain some straightforward, but interesting, results.

Theorem 41 *If a model is perfectly calibrated then we have:*

$$\pi_0 \bar{s_0} = \pi_1 (1 - \bar{s_1}) \tag{15}$$

or equivalently,

$$\pi_0 MAE_0 = \pi_1 MAE_1.$$

Proof For perfectly calibrated models, we have that for every bin in an invariant partition on the scores we have that $y_{b_j} = s_{b_j}$. Just taking a partition consisting of one single bin (which is an invariant partition), we have that this is the same as saying that $\pi_1 = \pi_1\bar{s_1} + \pi_0\bar{s_0}$. This leads to $\pi_1(1-\bar{s_1}) = \pi_0\bar{s_0}$. ■

Equation (15) is an interesting formula in its own right. It gives a necessary condition for calibration: the extent to which the average score over all examples (which is the weighted mean of per-class averages $\pi_0\bar{s}_0 + \pi_1\bar{s}_1$) deviates from π_1.

We now give a first result which connects two performance metrics:

Theorem 42 *If a model is perfectly calibrated then we have:*

$$BS = \pi_0\bar{s_0} = \pi_1(1-\bar{s_1}) = MAE/2.$$

Proof We use the continuous decomposition (Theorem 28):

$$BS = CL + RL.$$

Since it is perfectly calibrated, $CL = 0$. Then we have:

$$\begin{aligned} BS &= RL = \int_0^1 \frac{\pi_1 f_1(s)\pi_0 f_0(s)}{\pi_0 f_0(s)+\pi_1 f_1(s)} ds = \int_0^1 (\pi_1 f_1(s)) \left(1 - \frac{\pi_1 f_1(s)}{\pi_0 f_0(s)+\pi_1 f_1(s)}\right) ds \\ &= \int_0^1 \left(\pi_1 f_1(s) - \frac{[\pi_1 f_1(s)]^2}{\pi_0 f_0(s)+\pi_1 f_1(s)}\right) ds = \int_0^1 \pi_1 f_1(s) ds - \int_0^1 \frac{[\pi_1 f_1(s)]^2}{\pi_0 f_0(s)+\pi_1 f_1(s)} ds \\ &= \pi_1 - \int_0^1 \frac{\pi_1 f_1(s)}{\frac{\pi_0 f_0(s)}{\pi_1 f_1(s)}+1} ds. \end{aligned}$$

Since it is perfectly calibrated, we have, by Lemma 39:

$$\frac{f_0(s)}{f_1(s)} = \frac{1-s}{s}\frac{\pi_1}{\pi_0}.$$

So:

$$\begin{aligned} BS &= \pi_1 - \int_0^1 \frac{\pi_1 f_1(s)}{\frac{\pi_0}{\pi_1}\frac{1-s}{s}\frac{\pi_1}{\pi_0}+1} ds = \pi_1 - \int_0^1 \frac{s\pi_1 f_1(s)}{(1-s)+s} ds \\ &= \pi_1 - \pi_1 \int_0^1 s f_1(s) ds = \pi_1(1 - \int_0^1 s f_1(s)) ds = \pi_1(1-\bar{s_1}). \end{aligned}$$

■

We will now use the expressions for expected loss to analyse where this result comes from exactly. In the following result, we see that for a calibrated model the optimal threshold T for a given cost proportion c is $T = c$, which is exactly the score-driven threshold choice method. In other words:

Theorem 43 *For a perfectly calibrated model:*
$T_c^o(c) = T_c^{sd}(c) = c$.

Proof We first take Equation (12):

$$T_c^o(c) \quad = \quad \arg\min_t \; 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}.$$

We calculate the derivative and equal it to 0 to get t, but we isolate c:

$$2\{c\pi_0(1-f_0(t)) + (1-c)\pi_1 f_1(t)\} = 0$$
$$c = \frac{\pi_1 f_1(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)}.$$

From Definition 38 (perfect calibration) we have that the right expression above equals t, so we have $t = c$. The proof is identical for T_c^{sd}. ■

And now we can express and relate many of the expressions for the expected loss seen so far. Starting with the expected loss for the optimal threshold choice method, that is, L_c^o (which uses T_c^o), we have, from Theorem 43, that $T_c^o(c) = T_c^{sd}(c) = c$ when the model is perfectly calibrated. Consequently, we have the same as Equation (8), and since we know that $BS = \pi_0\bar{s}_0$ for perfectly calibrated models, we have:

$$L^o_{U(c)} = BS = \pi_0\bar{s}_0 = MAE/2.$$

The following theorem summarises all the previous results.

Theorem 44 *For perfectly calibrated models:*

$$L^{sd}_{U(c)} = L^o_{U(c)} = RL = \frac{L^{su}_{U(c)}}{2} = \frac{MAE}{2} = BS = \pi_0\bar{s}_0 = \pi_1(1-\bar{s}_1).$$

Proof Since $L^{sd}_{U(c)} = BS$ it is clear that $L^{sd}_{U(c)} = \pi_0\bar{s}_0$, as seen above for $L^o_{U(c)}$ as well. Additionally, from Theorem 12, we have that $L^{su}_{U(c)} = \pi_0\bar{s}_0 + \pi_1(1-\bar{s}_1)$, which reduces to $2L^{su}_{U(c)} = 2BS = 2\pi_0\bar{s}_0$. We also use the result of Theorem 29 which states that, in general (not just for perfectly calibrated models), $L^o_{U(c)}(m) = RL(\text{Conv}(m))$. ■

All this gives an interpretation of the optimal threshold choice method as a method which calculates expected loss by assuming perfect calibration. Note that this is clearly seen by the relation $L^{sd}_{U(c)} = L^o_{U(c)} = \frac{L^{su}_{U(c)}}{2}$, since the loss drops to the half if we use scores to adjust to the operating condition. In this situation, we get the best possible result.

7.3 Choosing a Threshold Choice Method

It is enlightening to see that many of the most popular classification performance metrics are just expected losses by changing the threshold choice method and the use of cost proportions or skews. However, it is even more revealing to see how (and under which conditions) these performance

General relations:

$$L^{ru}_{U(c)} = \pi_0\pi_1(1-2AUC) + \frac{1}{2} > L^{rd}_{U(c)} = \pi_0\pi_1(1-2AUC) + \frac{1}{3} \geq L^{o}_{U(c)} = RL_{\text{Conv}}$$

$$L^{su}_{U(c)} = MAE \geq L^{sd}_{U(c)} = BS \geq L^{o}_{U(c)} = RL_{\text{Conv}}$$

If scores are evenly-spaced:

$$L^{ru}_{U(c)} = \pi_0\pi_1(1-2AUC) + \frac{1}{2} = L^{su}_{U(c)} = MAE = \pi_0\bar{s}_0 + \pi_1(1-\bar{s}_1)$$

$$L^{rd}_{U(c)} = \pi_0\pi_1(1-2AUC) + \frac{1}{3} = L^{sd}_{U(c)} = BS$$

If scores are perfectly calibrated:

$$L^{sd}_{U(c)} = L^{o}_{U(c)} = RL = \frac{L^{su}_{U(c)}}{2} = \frac{MAE}{2} = BS = \pi_0\bar{s}_0 = \pi_1(1-\bar{s}_1)$$

If the model has perfect ranking:

$$L^{ru}_{U(c)} = \frac{1}{4} > L^{rd}_{U(c)} = \frac{1}{12} > L^{o}_{U(c)} = 0$$

If the model is random (and $\pi_0 = \pi_1$):

$$L^{su}_{U(c)} = L^{sd}_{U(c)} = L^{ru}_{U(c)} = \frac{1}{2} > L^{rd}_{U(c)} = \frac{1}{3} > L^{o}_{U(c)} = \frac{1}{4}$$

Figure 4: Comparison of losses and performance metrics, in general and under several score conditions.

metrics can be related (in some cases with inequalities and in some other cases with equalities). The notion of score transformation is the key idea for these connections, and is more important that it might seem at first sight. Some threshold choice methods can be seen as a score transformation followed by the score-driven threshold choice method. Even the fixed threshold choice method can be seen as a crisp transformation where scores are set to 1 if $s_i > t$ and 0 otherwise. Another interesting point of view is to see the values of extreme models, such as a model with perfect ranking ($AUC = 1$, $RL^{ROCCH} = 0$) and a random model ($AUC = 0.5$, $RL^{ROCCH} = 0.25$ when $\pi_0 = \pi_1$). Figure 4 summarises all the relations found so far and these extreme cases.

The first apparent observation is that $L^{o}_{U(c)}$ seems the best loss, since it derives from the optimal threshold choice method. We already argued in Section 6 that this is unrealistic. The result given by Theorem 29 is a clear indication of this, since this makes expected loss equal to RL_{Conv}. Hence, this threshold choice method assumes that the calibration which is performed with the convex hull over the training (or a validation data set) is going to be perfect and hold for the test set. Figure 4 also gives the impression that $L^{su}_{U(c)}$ and $L^{ru}_{U(c)}$ are so bad that their corresponding threshold choice methods and metrics are useless. In order to refute this simplistic view, we must realise (again) that not every threshold choice method can be applied in every situation. Some require more information

or more assumptions than others. Table 6 completes Table 3 to illustrate the point. If we know the deployment operating condition at evaluation time, then we can fix the threshold and get the expected loss. If we do not know this information at evaluation time, but we expect to be able to have it and use it at deployment time, then the score-driven, rate-driven and optimal threshold choice methods seem the appropriate ones. Finally, if no information about the operating condition is going to be available at any time then the score-uniform and the rate-uniform may be alternative options, which could account for a worst-case estimation.

Threshold choice method	Fixed	Driven by o.c.	Chosen uniformly
Using scores	score-fixed (T^{sf})	score-driven (T^{sd})	score-uniform (T^{su})
Using rates	rate-fixed (T^{rf})	rate-driven (T^{rd})	rate-uniform (T^{ru})
Using optimal thresholds		optimal (T^{o})	
Required information	o.c. at evaluation time	o.c. at deployment time	no information

Table 6: Information which is required (and when) for the seven threshold choice methods so that they become reasonable (or just not totally unreasonable). Operating condition is denoted by o.c.

From the cases shown in Table 6, the methods driven by the operating condition require further discussion. The relations shown in Figure 4 illustrate that, in addition to the optimal threshold choice method, the other two methods that seem more competitive are the score-driven and the rate-driven. One can argue that the rate-driven threshold choice has an expected loss which is always greater than $1/12$ (if $AUC = 1$, we get $-1/4+1/3$), while the others can be 0. But things are not so clear-cut.

- The score-driven threshold choice method considers that the scores are estimated probabilities and that they are reliable, in the tradition of proper scoring rules. So it just uses these probabilities to set the thresholds.

- The rate-driven threshold choice method completely ignores the scores and only considers their order. It assumes that the ranking is reliable while the scores are not accurate probabilities. It derives the thresholds using the predictive positive rate. It can be seen as the score-driven threshold choice method where the scores have been set evenly-spaced by a transformation.

- The optimal threshold choice method also ignores the scores completely and only considers their order. It assumes that the ranking is reliable while the scores are not accurate probabilities. However, this method derives the thresholds by keeping the order and using the slopes of the segments of the convex hull (typically constructed over the training data set or a validation data set). It can be seen as the score-driven threshold choice method where the scores have been calibrated by the PAV method.

Now that we better understand the meaning of the threshold choice methods we may state the difficult question more clearly: given a model, which threshold choice method should we use to make classifications? The answer is closely related to the calibration problem. Some theoretical and experimental results (Robertson et al., 1988; Ayer et al., 1955; Platt, 1999; Zadrozny and Elkan, 2001,

2002; Niculescu-Mizil and Caruana, 2005; Niculescu-Mizil and Caruana, 2005; Bella et al., 2009; Gebel, 2009) have shown that the PAV method (also known as isotonic regression) is not frequently the best calibration method. Some other calibration methods could do better, such as Platt's calibration or binning averaging. In particular, it has been shown that "isotonic regression is more prone to overfitting, and thus performs worse than Platt scaling, when data is scarce" (Niculescu-Mizil and Caruana, 2005). Even with a large validation data set which allows the construction of an accurate ROC curve and an accurate convex hull, the resulting choices are not necessarily optimal for the test set, since there might be problems with outliers (Rüping, 2006). In fact, if the validation data set is much smaller (or biased) than the training set, the resulting probabilities can be even worse than the original probabilities, as it may happen with cross-validation. So, we have to feel free to use other (possibly better) calibration methods instead and do not stick to the PAV method just because it is linked to the optimal threshold choice method.

So the question of whether we keep the scores or not (and how we replace them in case) depends on our expectations on how well-calibrated the model is, and whether we have tools (calibration methods and validation data sets) to calibrate the scores.

But we can turn the previous question into a much more intelligent procedure. Calculating the three expected losses discussed above (and perhaps the other threshold choice methods as well) provides a rich source of information about how our models behave. This is what performance metrics are all about. It is only after the comparison of all the results and the availability of (validation) data sets when we can make a decision about which threshold choice method to use.

This is what we did with the example shown in Table 2 in Section 1. We evaluated the model for several threshold choice methods and from there we clearly saw which models were better calibrated and we finally made a decision about which model to use and with which threshold choice methods.

In any case, note that the results and comparisons shown in Figure 4 are for expected loss; the actual loss does not necessarily follow these inequalities. In fact, the expected loss calculated over a validation data set may not hold over the test data set, and even some threshold choice methods we have discarded from the discussion above (the fixed ones or the score-uniform and rate-uniform, if probabilities or rankings are very bad respectively) could be better in some particular situations.

8. Discussion

This paper builds upon the notion of threshold choice method and the expected loss we can obtain for a range of cost proportions (or skews) for each of the threshold choice methods we have investigated. The links between threshold choice methods, between performance metrics, in general and for specific score arrangements, have provided us with a much broader (and more elaborate) view of classification performance metrics and the way thresholds can be chosen. In this last section we link our results to the extensive bulk of work on classification evaluation and analyse the most important contributions and open questions which are derived from this paper.

8.1 Related Work

One decade ago there was a scattered view of classification evaluation. Many performance metrics existed and it was not clear what their relationships were. One first step in understanding some of these performance metrics in terms of costs was the notion of cost isometrics (Flach, 2003). With cost isometrics, many classification metrics (and decision tree splitting criteria) are characterised by its skew landscape, that is, the slope of its isometric at any point in the ROC space. Another com-

prehensive view was the empirical evaluation made by Ferri et al. (2009). The analysis of Pearson and Spearman correlations between 18 different performance metrics shows the pairs of metrics for which the differences are significant. However, this work does not elaborate, at least theoretically, on what exactly each metric measures, but rather on whether they give different choices in general.

In addition to these, there have been three lines of research in this area which provide further pieces to understand the whole picture.

- First, the notion of 'proper scoring rules' (which was introduced in the sixties, see for example, the work by Murphy and Winkler, 1970), has been developed to a degree (Buja et al., 2005) in which it has been shown that the Brier score (MSE loss), logloss, boosting loss and error rate (0-1 loss) are all special cases of an integral over a Beta density, and that all these performance metrics can be understood as averages (or integrals), at least theoretically, over a range of cost proportions (see, e.g., the works of Gneiting and Raftery, 2007; Reid and Williamson, 2010 and Brümmer, 2010), so generalising the early works by Murphy on probabilistic predictions when cost-loss ratio is unknown (Murphy, 1966, 1969). Additionally, further connections have been found between proper scoring rules and distribution divergences (f-divergences and Bregman divergences) (Reid and Williamson, 2011).

- Second, the translation of the Brier decomposition using ROC curves (Flach and Matsubara, 2007) suggests a connection between the Brier score and ROC curves, and particularly between refinement loss and *AUC*, since both are performance metrics which do not require the magnitude of the scores of the model.

- Third, an important coup d'effet has been given by Hand (2009), stating that the *AUC* cannot be used as a performance metric for evaluating models (for a range of cost proportions), *assuming the optimal threshold choice method*, because the distribution for these cost proportions depends on the model. This seemed to suggest a definitive rupture between ranking quality and classification performance over a range of cost proportions.

Each of the three lines mentioned above provides a partial view of the problem of classifier evaluation, and suggests that some important connections between performance metrics were waiting to be unveiled. The starting point of this unifying view is that all the previous works above worked with only two threshold choice methods, which we have called the score-driven threshold choice method and the optimal threshold choice method. Only a few works mention these two threshold choice methods together. For instance, Drummond and Holte (2006) talk about 'selection criteria' (instead of 'threshold choice methods') and they distinguish between 'performance-independent' selection criteria and 'cost-minimizing' selection criteria. Hand (personal communication) says that 'Hand (2009) (top of page 122) points out that there are situations where one might choose thresholds independently of cost, and go into more detail in Hand (2010)'. This is related to the fixed threshold choice method, or the rate-uniform and score-uniform threshold choice methods used here. Finally, Flach et al. (2011) explore the rate-uniform threshold choice method while Hernández-Orallo et al. (2011) explore the score-driven threshold choice method.

The notion of proper scoring rule works with the score-driven threshold choice method. This implies that this notion cannot be applied to *AUC*—Reid and Williamson (2011) connects the area under the convex hull (*AUCH*) with other proper scoring rules but not *AUC*—and to RL. As a consequence, the Brier score, log-loss, boosting loss and error rate would only be minor choices depending on the information about the distribution of costs.

Hand (2009) takes a similar view of the cost distribution, as a choice that depends on the information we may have about the problem, but makes an important change over the tradition in proper scoring rules tradition. He considers 'optimal thresholds' (see Equation 12) instead of the score-driven choice. With this threshold choice method, Hand is able to derive *AUC* (or yet again *AUCH*) as a measure of aggregated classification performance, but the distribution he uses (and criticises) depends on the model itself. Then he defines a new performance metric which is proportional to the area under the optimal cost curve. Hand (2010) and Hand and Anagnostopoulos (2011) elaborate on this by the consideration of asymmetries in the cost distribution.

8.2 A Plethora of Evaluation Metrics

The unifying view under the systematic exploration of threshold choice methods in this paper has established a set of connections which portray a much more comprehensive view of the landscape of evaluation metrics for classification. However, it has to be emphasised that each connection between a metric and a kind of expected loss is associated to a particular set of assumptions. The most important assumption is the cost model. For the whole paper, we have assumed that the operating condition θ is simplified to a single parameter, c or z, from a three-dimensional vector $\theta = \langle b, c, \pi_0 \rangle$. In order to make this reduction, we have assumed that the threshold choice method ignores the magnitude b. In addition, we have either assumed π_0 fixed or have linked it to c through the notion of skew (see appendix A). However, in general, we could consider trivariate distributions for the parameters in θ. We could also consider threshold choice methods which are sensitive to the magnitude b or other combinations of the three parameters. For instance, we could consider a threshold choice method which is more 'conservative' when b is high and more 'risky' when b is low. Moreover, in some applications, the operating condition can have even more parameters, since it may be instance-dependent (Turney, 2000) or can change depending on previous errors. Certainly, for a specific application one must consider the distribution which better fits the expectation or knowledge about the possible operating conditions. This first *dimension*, the distribution of operating conditions, has been varied in many different ways by Wieand et al. (1989), Gneiting and Raftery (2007), Reid and Williamson (2010), Reid and Williamson (2011), Brümmer (2010), Hand (2009), Hand (2010) and Hand and Anagnostopoulos (2011), as mentioned above. Here we have considered the simplest option, a uniform distribution (except for the fixed threshold choice methods where the results are more general), but many other options can be explored, including partial, asymmetric or multimodal distributions.

As said in the introduction, this paper works and varies on a different dimension, by changing the threshold choice method systematically. The choice of a particular threshold choice method reflects many things: the information one may have at evaluation time or deployment time, the reliability or calibration degree one expect from a model or the very character of the model, which may be a crisp classifier, a ranker or a probability estimator. Also, the choice can also be just a matter of practice, since some threshold choice methods are simpler than others and develop into simpler decision rules. In fact, we have explored seven possibilities here. Some of them may look more reasonable than others, and some may correspond to frequent practical situations, while others have just been set in order to derive the relation between expected loss and a relevant evaluation metric. And there might be other possibilites. For instance, a particular case of the fixed threshold choice method can be defined by choosing the threshold at $\mathbb{E}\{c\}$, which can be generalised, for example,

in an online setting, if this expected value evolves (or is refined) after the information we get from the actual costs example after example.

All this suggests that many other combinations could be explored by using different options for the two dimensions, and possibly relaxing some other assumptions, such as Hand did with his measure H (Hand, 2009), when using the $\beta_{2,2}$ distribution for the optimal threshold choice method instead of the uniform ($\beta_{1,1}$) distribution. We think that the same thing could be done with the rate-driven threshold choice method, possibly leading to new variants of the *AUC*. This is related to the extensive work where several distributions are presented for calculating an average of sensitivities over a restricted range of specificities (Wieand et al., 1989), leading to other variants of *AUC* (such as partial *AUC*). And, of course, this has also been done with proper scoring rules for the score-driven threshold choice method with many loss functions.

It is, however, also worthwhile to underline the limitations of aggregated metrics for comparing classification models. Graphical plots, where performance is shown in a range of operating conditions, are more powerful than aggregated metrics, since we can keep several methods provided they are not completely dominated by others. Many threshold choice methods give rise to particular kinds of curves that provide at each operating point, rather than just an aggregate.

8.3 Conclusions and Future Work

As a conclusion, if we want to evaluate a model for a wide range of operating conditions (i.e., cost proportion or skews), we have to determine first which threshold choice method is to be used. If it is fixed because we have a non-probabilistic classifier or we are given the actual operating condition at evaluation time, then we get accuracy (and unweighted accuracy) as a good performance metric. If we have no access to the operating condition at evaluation time but neither do we at deployment time, then the score-uniform and the rate-uniform may be considered, with *MAE* and *AUC* as corresponding performance metrics. Finally, in the common situation when we do not know the operating condition at evaluation time but we expect that it will be known and used at deployment time, then we have more options. If a model has no reliable scores or probability estimations, we recommend the refinement loss (RL_{Conv}, which is equivalent to area under the optimal cost curve) if thresholds are being chosen using the convex hull of a reliable ROC curve, or, alternatively, we recommend the area under the ROC curve (*AUC*) if the estimation of this convex hull is not reliable enough to choose thresholds confidently. More readily, if a model has reliable scores because it is a good probability estimator or it has been processed by a calibration method, then we recommend to choose the thresholds according to scores. In this case, the corresponding performance metric is the Brier score.

From this paper, now we have a much better understanding on the relation between the Brier score, the *AUC* and refinement loss. We also know much better what is happening when models are not convex and/or not calibrated. In addition, we find that using evenly-spaced scores, we get that the Brier score and the *AUC* are linearly related. Furthermore, we see that if the model is perfectly calibrated, the expected loss using the score-driven threshold choice method equals the optimal threshold choice method.

The collection of new findings introduced in this paper leads to many other avenues to follow and some questions ahead. For instance, the duality between cost proportions and skews suggests that we could work with loglikelihood ratios as well. Also, there is always the problem of multi-class evaluation. This is as challenging as interesting, since there are many more threshold choice

methods in the multiclass case and the corresponding expected losses could be connected to some multiclass extensions of the binary performance metrics. Finally, more work is needed on the relation between the ROC space and the cost space, and the representation of all these expected losses in the latter space. The notion of Brier curve (Hernández-Orallo et al., 2011) is a first step in this direction, but all the other threshold choice methods also lead to other curves.

Acknowledgments

We thank the anonymous reviewers for their comments, which have helped to improve this paper significantly. We are also grateful to David Hand for insightful discussion. This work was supported by the MEC/MINECO projects CONSOLIDER-INGENIO CSD2007-00022 and TIN 2010-21062-C02-02, GVA project PROMETEO/2008/051, and the *REFRAME* project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net (CHIST-ERA), and funded by the Engineering and Physical Sciences Research Council in the UK and the Ministerio de Economía y Competitividad in Spain.

Appendix A. Univariate Operating Conditions Using Cost Proportions and Skews

This appendix develops the expressions for univariate operating conditions from the general notion of operating condition introduced in Section 2.2. One possibility of reducing an operating condition with three parameters $\theta = \langle b, c, \pi_0 \rangle$ into one single parameter, without assuming independence of b and c, relies on noticing that b is a multiplicative factor in Equation (1). So, we can express loss as follows:

$$Q(t;\theta) = b\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} = bQ_\eta(t;\eta). \tag{16}$$

where $\eta = \langle c, \pi_0 \rangle$. The set of these normalised operating conditions is denoted by H. In other words, loss is just a product of the cost magnitude b and normalised loss. When calculating expected loss for a particular threshold choice method, we can write

$$Q(T(\theta);\theta) = bQ_\eta(T_\eta(\eta),\eta). \tag{17}$$

Note that this *assumes that the threshold choice method is defined in terms of* η*, and hence it is independent of* b*.*

From here, we can just work with Equation (17) and derive expected loss from Equation (2) as follows:

$$\begin{aligned} L &= \int_\Theta Q(T(\theta);\theta)w(\theta)d\theta = \int_\Theta bQ_\eta(T_\eta(\eta),\eta)w(\theta)d\theta \\ &= \int_H \left\{ \int_0^\infty bQ_\eta(T_\eta(\eta),\eta)w_{B|H}(b|\eta)db \right\} w_H(\eta)d\eta \\ &= \int_H Q_\eta(T_\eta(\eta),\eta) \left\{ \int_0^\infty bw_{B|H}(b|\eta)db \right\} w_H(\eta)d\eta \\ &= \int_H Q_\eta(T_\eta(\eta),\eta)\mathbb{E}\{b|\eta\}w_H(\eta)d\eta. \end{aligned}$$

with $w_H(\eta)$ being the marginal distribution density for η, that is, $w_H(\eta) = \int_0^\infty w(\theta)db$, and $w_{B|H}(b|\eta)$ the conditional density for b given η, that is, $w_{B|H}(b|\eta) = w(\theta)/w_H(\eta)$. And now, let us define

$v_H(\eta) = w_H(\eta)\mathbb{E}\{b|\eta\}/\mathbb{E}\{b\}$ with $\mathbb{E}\{b\} = \int_H \mathbb{E}\{b|\eta\}w_H(\eta)d\eta$, leading to the following expression:

$$L \quad = \quad \mathbb{E}\{b\}\int_H Q_\eta(T_\eta(\eta),\eta)v_H(\eta)d\eta. \tag{18}$$

So now we have an expression which seems to isolate b (more precisely, $\mathbb{E}\{b\}$) as a constant factor. Obviously, this is possible since we have constructed $v_H(\eta)$ in terms of the conditional expected value b, incorporating the variability of b, while isolating its average magnitude. This is interesting and very useful, because if we have knowledge about the dependency between b and η, we can incorporate that information into v, without affecting Q_η at all. For instance, if π_0 is fixed and we assume (or know) $w_H(\eta)$ to be a Beta distribution $\beta_{2,2}$ and we also assume (or know) that the values for b are higher for extreme values of c (closer to 1 or 0), then the resulting distribution v could be assumed to account for these two things. This would make increase the probability for extreme values of c, making the resulting distribution flatter and closer, for example, to a uniform distribution. Consequently, in this paper we will frequently assume $v_H(\eta)$ to be uniform (either fixing π_0 or combining π_0 and c into a single parameter z). This really makes it explicit that it is $\frac{w_H(\eta)\mathbb{E}\{b|\eta\}}{\mathbb{E}\{b\}}$ what we are assuming to be uniform.

Now, we will derive the two approaches for univariate operating conditions (costs and skews) that we use in this paper. In one of them, we assume that the class proportion (i.e., π_0) is fixed, leading to the marginal distribution $w_c(c) = v_H(\langle c, \pi_0\rangle)$.

Since we now only have a relevant free parameter c in Q_η, we can now express the normalised loss as a function of c. However, for a mere convenience that will become clear below, we include the factor $\mathbb{E}\{b\}$ in the loss produced at a decision threshold t and a cost proportion c, adapting Equation (16):

$$Q_c(t;c) \triangleq \mathbb{E}\{b\}Q_\eta(t;\langle c,\pi_0\rangle) = \mathbb{E}\{b\}\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}. \tag{19}$$

With this inclusion of $\mathbb{E}\{b\}$ we just have the following simple expression for the calculation of expected loss, derived from Equation (18):

$$L_c \quad = \quad \int_0^1 Q_c(T(c);c)w_c(c)dc. \tag{20}$$

Recall that w_c incorporates the variability of b jointly with c.

A different approach to reducing the operating condition to a single parameter is the notion of *skew*, which is a normalisation of the product between cost proportion and class proportion:

$$z \triangleq \frac{c_0\pi_0}{c_0\pi_0 + c_1\pi_1} = \frac{c\pi_0}{c\pi_0 + (1-c)(1-\pi_0)}.$$

This means that π_0 is no longer fixed, but neither is it independent of c. What z does is to combine both parameters. This is a different way of reducing the operating condition to one single parameter. We thus define loss as depending solely on z. From Equation (19) we obtain

$$\frac{Q_c(t;c)}{\mathbb{E}\{b\}[c\pi_0 + (1-c)(1-\pi_0)]} = z(1-F_0(t)) + (1-z)F_1(t) \triangleq Q_z(t;z). \tag{21}$$

This gives an expression for *standardised*] loss at a threshold t and a skew z.

We then have the following simple but useful result.

Lemma 45 *If $\pi_0 = \pi_1$ then $z = c$ and $Q_z(t;z) = \frac{2}{\mathbb{E}\{b\}} Q_c(t;c)$.*

Proof If classes are balanced we have $c\pi_0 + (1-c)(1-\pi_0) = 1/2$, and the result follows from Equation (21). ■

This justifies taking the expected value of the cost magnitude $\mathbb{E}\{b\} = 2$, which means that Q_z and Q_c are expressed on the same 0-1 scale, and are also commensurate with error rate which assumes $c = 1/2$. The upshot of Lemma 45 is that we can transfer any expression for loss in terms of cost proportion to an equivalent expression in terms of skew by just setting $\pi_0 = \pi_1 = 1/2$ and $z = c$. Notice that if $c = 1/2$ then $z = \pi_0$, so in that case skew denotes the class distribution as operating condition.

In fact, we can also define $w_z(z)$ by incorporating the variability of b (and also π_0 and c). We could choose $w_z(z)dz = \frac{1}{\mathbb{E}\{b\}[c\pi_0+(1-c)(1-\pi_0)]} w_c(c)dc$, but we can use any other distribution. In the paper we will use the uniform distribution for $w_z(z)$. In any case, this leads to the corresponding expression of standardised expected loss (as Equation 20):

$$L_z \quad = \quad \int_0^1 Q_z(T(z);z) w_z(z) dz.$$

So, with this isolation of the average magnitude of b, and the incorporation of its variability into the univariate distribution, in the paper we will just work with operating conditions which are either defined by the cost proportion c (assuming a fixed class distribution π_0) or by the skew z (which combines c and π_0).

Appendix B. Proof of Theorem 29

In this appendix, we give the proof for Theorem 29 in the paper. The theorem works with convex models as given by Definition 25.

In this appendix, we will use:

$$c(T) = \frac{\pi_1 f_1(T)}{\pi_0 f_0(T) + \pi_1 f_1(T)}.$$

throughout, as introduced by Equation (14). Sometimes we will use subindices for $c(T)$ depending on the model we are using. We will also use $slope(T) = \pi_1 f_1(T) = \frac{\pi_1}{\pi_0}\left(\frac{1}{c(T)} - 1\right)$. A convex model is the same as saying that $c(T)$ is non-decreasing or that $slope(T)$ is non-increasing.

We use $c^{-1}(s)$ for the inverse of $c(T)$ (wherever it is well defined). We will use the following transformation $T \mapsto s = c(T)$ and the resulting model will be denoted by $m^{(c)}$. We will use s, c or σ for elements in the codomain of this transformation (cost proportions or scores between 0 and 1) and we will use T or τ for elements in the domain.

For continuous and strictly convex models for which $c(0) = 0$ and $c(1) = 1$, the proof is significantly simpler. In general, for any convex model, including discontinuities and straight segments, things become a little bit more elaborate, as we see below.

B.1 Intervals

Since the model is convex, we know that $c(T)$ is monotone, more precisely, non-decreasing. We can split the codomain and domain of this function into intervals. Intervals in the codomain of

thresholds will be represented with the letter τ and intervals in the domain of cost proportions or scores between 0 and 1 will be denoted by letter σ. The series of intervals are denoted as follows:

$$\begin{aligned} I_\sigma &= (\sigma_0,\sigma_1),(\sigma_1,\sigma_2)\ldots(\sigma_i,\sigma_{i+1})\ldots(\sigma_{n-1},\sigma_n) \\ &\qquad\qquad \Uparrow c(\tau) \qquad\quad \Downarrow c^{-1}(\sigma) \\ I_\tau &= (\tau_0,\tau_1),(\tau_1,\tau_2)\ldots(\tau_i,\tau_{i+1})\ldots(\tau_{n-1},\tau_n) \end{aligned}$$

where $\sigma_0 = 0$, $\sigma_n = 1$, $\tau_0 = -\infty$ and $\tau_n = \infty$. Even though we cannot make a bijective mapping for every point, we can construct a bijective mapping between I_σ and I_τ. Because of this bijection, we may occasionally drop the subindex for I_σ and I_τ.

We need to distinguish three kinds of intervals:

- Intervals where $c(T)$ is strictly increasing, denoted by $\acute{I}$. We call these intervals *bijective*, since $c(T)$ is invertible. These correspond to non-straight parts of the ROC curve. Each point inside these segments is optimal for one specific cost proportion.
- Intervals where $c(T)$ is constant, denoted by $\bar{I}$. We call these non-injective intervals *constant*. These correspond to straight parts of the ROC curve. All the points inside these segments are optimal for just one cost proportion, and we only need to consider any of them (e.g., the extremes).
- Intervals in the codomain where no value T for $c(T)$ has an image, denoted by $\dot{I}$. We call these 'intervals' *singular*, and address non-surjectiveness. In the codomain they may usually correspond to one single point, but also can correspond to an actual interval when the density functions are 0 for some intervals in the codomain. In the end, these correspond to discontinuous points of the ROC curve. The points at $(0,0)$ and $(1,1)$ are generally (but not always) discontinuous. These points are optimal for many cost proportions.

Table 7 shows how these three kinds of intervals work.

bijective $\acute{I}$	constant $\bar{I}$	singular $\dot{I}$
$]\sigma_i,\sigma_{i+1}[$	$[\sigma_i,\sigma_{i+1}]$	$]\sigma_i,\sigma_{i+1}[$
↑↑↑↑↑	↗↖	↖↗
$]\tau_i,\tau_{i+1}[$	$]\tau_i,\tau_{i+1}[$	$[\tau_i,\tau_{i+1}]$

Table 7: Illustration for the three types of intervals.

Now we are ready to get some results:

Lemma 46 *If the model m is convex, we have that minimal expected loss can be expressed as:*

$$L^o_{U(c)}(m) = \acute{\Lambda}(m) + \dot{\Lambda}(m)$$

where:

$$\acute{\Lambda}(m) = \sum_{]\tau_i,\tau_{i+1}[\in\acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} 2c(T)\pi_0(1-F_0(T)) + 2(1-c(T))\pi_1 F_1(T)\}c'(T)dT \tag{22}$$

where $c'(T)$ is the derivative of $c(T)$ and:

$$\dot{\Lambda}(m) = \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2c\pi_0(1-F_0(\tau_i)) + 2(1-c)\pi_1 F_1(\tau_i)\}dc \quad (23)$$

$$= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \left\{\pi_0(1-F_0(\tau_i))({\sigma_{i+1}}^2 - {\sigma_i}^2) + \pi_1 F_1(\tau_i)(2\sigma_{i+1} - {\sigma_{i+1}}^2 - 2\sigma_i + {\sigma_i}^2)\right\}. \quad (24)$$

Note that the constant intervals in $\bar{I}_\sigma$ are not considered (their loss is 0).

Proof We take the expression for optimal loss from Equation (13):

$$L^o_{U(c)} = \int_0^1 \min_t \{2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\}dc. \quad (25)$$

In order to calculate the minimum, we make the derivative of the min expression equal to 0:

$$\begin{aligned} 2c\pi_0(0 - f_0(t)) + 2(1-c)\pi_1 f_1(t) &= 0 \\ -2c \cdot \frac{\pi_0}{\pi_1} slope(T) + 2(1-c) &= 0 \\ \frac{\pi_0}{\pi_1} slope(T) &= \frac{1-c}{c} \\ \frac{1}{c(T)} - 1 &= \frac{1-c}{c} \\ c(T) &= c. \end{aligned}$$

We now check the sign of the second derivative, which is:

$$-2c \cdot \frac{\pi_0}{\pi_1} slope'(t) = -2c \times (\frac{1}{c(T)} - 1)' = -2c\frac{-c'(T)}{{c(T)}^2} = 2c\frac{c'(T)}{{c(T)}^2}.$$

For the bijective intervals $\acute{I}_\sigma$, where the model is strictly convex and $c(T)$ is strictly decreasing, its derivative is > 0. Also, c is always between 0 and 1, so the above expression is positive, and it is a minimum. And this cannot be a 'local' minimum, since the model is convex.

For the constant intervals $\bar{I}_\sigma$ where the model is convex (but not strictly), this means that $c(T)$ is constant, and its derivative is 0. That means that the minimum can be found at any point T in these intervals $]\tau_i, \tau_{i+1}[$ for the same $[\sigma_i = \sigma_{i+1}]$. But their contribution to the loss will be 0, as can be seen since $c'(T)$ equals 0.

For the singular intervals $\dot{I}_\sigma$, on the contrary, all the values in each interval $]\sigma_i, \sigma_{i+1}[$ will give a minimum for the same $[\tau_i = \tau_{i+1}]$.

So we decompose the loss with the bijective and singular intervals only:

$$L^o_{U(c)}(m) = \acute{\Lambda}(m) + \dot{\Lambda}(m).$$

For the strictly convex (bijective) intervals, we now know that the minimum is at $c(T) = c$, and $c(T)$ is invertible. We can use exactly this change of variable over Equation (25) and express this for the series of intervals $]\tau_i, \tau_{i+1}[$.

$$\acute{\Lambda}(m) \quad = \quad \sum_{]\tau_i,\tau_{i+1}[\in \acute{I}_\sigma} \int_{\tau_i}^{\tau_{i+1}} 2c(T)\pi_0(1-F_0(T)) + 2(1-c(T))\pi_1 F_1(T)\}c'(T)dT.$$

which corresponds to Equation (22). Note that when there is only one bijective interval (the model is continuous and strictly convex), we have that there is only one integral in the sum and its limits go from $c^{-1}(0)$ to $c^{-1}(1)$, which in some cases can go from $-\infty$ to ∞, if the scores are not understood as probabilities.

For the singular intervals, we can work from Equation (25):

$$\dot{\Lambda}(m) \quad = \quad \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} \min\{2c\pi_0(1-F_0(t)) + 2(1-c)\pi_1 F_1(t)\}dc.$$

As said, all the values in each interval $]\sigma_i,\sigma_{i+1}[$ will give a minimum for the same $[\tau_i = \tau_{i+1}]$, so this reduces to:

$$\begin{aligned}
\dot{\Lambda}(m) &= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} \{2c\pi_0(1-F_0(\tau_i)) + 2(1-c)\pi_1 F_1(\tau_i)\}dc \\
&= 2 \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \left\{ \pi_0(1-F_0(\tau_i)) \int_{\sigma_i}^{\sigma_{i+1}} cdc + \pi_1 F_1(\tau_i) \int_{\sigma_i}^{\sigma_{i+1}} (1-c)dc \right\} \\
&= 2 \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \left\{ \pi_0(1-F_0(\tau_i)) \left[\frac{c^2}{2}\right]_{\sigma_i}^{\sigma_{i+1}} + \pi_1 F_1(\tau_i) \left[c - \frac{c^2}{2}\right]_{\sigma_i}^{\sigma_{i+1}} \right\} \\
&= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \left\{ \pi_0(1-F_0(\tau_i))({\sigma_{i+1}}^2 - {\sigma_i}^2) + \pi_1 F_1(\tau_i)(2\sigma_{i+1} - {\sigma_{i+1}}^2 - 2\sigma_i + {\sigma_i}^2) \right\}.
\end{aligned}$$

which corresponds to Equation (24). ■

B.2 $c(T)$ is Idempotent

Now we work with the transformation $T \mapsto s = c(T)$. The resulting model using this transformation will be denoted by $m^{(c)}$. We will use $H_0(s)$ and $H_1(s)$ for the cumulative distributions, which are defined as follows. Since $s = c(T)$ by definition we have that $F_0(T) = H_0(c(T)) = H_0(s)$ and similarly $F_1(T) = H_1(c(T)) = H_1(s)$.

For the intervals $]\tau_i,\tau_{i+1}[$ in $\acute{I}_\tau$, we have $c(T)$ is strictly convex we just use $c^{-1}(s)$ to derive H_0 and H_1. This may imply discontinuities at τ_i or τ_{i+1} for those values of s for which constant intervals have been mapped, namely σ_i and σ_{i+1}. So, we need to define the density functions as follows. For the bijective intervals we just use $h_0(s)ds = f_0(T)dT$ and $h_1(s)ds = f_1(T)dT$ as a shorthand for a change of variable, and we can clear h_0 and h_1 using $c^{-1}(s)$. We do that using open intervals $]\tau_i,\tau_{i+1}[$ in T. These correspond to $]c(\tau_i),c(\tau_{i+1})[=]\sigma_i,\sigma_{i+1}[$.

The constant intervals are $[\tau_i, \tau_{i+1}]$ in $\bar{I_\tau}$. There is probability mass for every constant interval $[\tau_i, \tau_{i+1}]$ mapping to a point $s_i = c(\tau_i) = c(\tau_{i+1}) = \sigma_i = \sigma_{i+1}$, as follows:

$$[H_0(T)]_{\sigma_i}^{\sigma_{i+1}} = \int_{\tau_i}^{\tau_{i+1}} f_0(T)dt = [F_0(T)]_{\tau_i}^{\tau_{i+1}} = F_0(\tau_{i+1}) - F_0(\tau_i). \quad (26)$$

$$[H_1(T)]_{\sigma_i}^{\sigma_{i+1}} = \int_{\tau_i}^{\tau_{i+1}} f_1(T)dt = [F_1(T)]_{\tau_i}^{\tau_{i+1}} = F_1(\tau_{i+1}) - F_1(\tau_i). \quad (27)$$

Finally, we just define $h_0(s) = h_1(s) = 0$ for those $s \in [\sigma_i, \sigma_{i+1}] \in \dot{I}_\sigma$, since for the singular intervals there is only one point τ_i and the mass to share is 0.

This makes $m^{(c)}$ well-defined for convex models (not necessarily continuous and strictly convex).

Lemma 47 *For model $m^{(c)}$ we have that, for the non-singular intervals, $c_{m^{(c)}}(s) = \frac{\pi_1 h_1(s)}{\pi_0 h_0(s) + \pi_1 h_1(s)}$ is idempotent, that is:*

$$c_{m^{(c)}}(s) = s.$$

Proof For the bijective (strictly convex) intervals $]\tau_i, \tau_{i+1}[$ mapped into $]c(\tau_i), c(\tau_{i+1})[$, that is, $]\sigma_i, \sigma_{i+1}[$:

$$\begin{aligned} c_{m^{(c)}}(s) &= \frac{\pi_1 h_1(s)}{\pi_0 h_0(s) + \pi_1 h_1(s)} = \frac{\pi_1 h_1(s)ds}{\pi_0 h_0(s)ds + \pi_1 h_1(s)ds} \\ &= \frac{\pi_1 f_1(T)dT}{\pi_0 f_0(T)dT + \pi_1 f_1(T)dT} = \frac{\pi_1 f_1(T)}{\pi_0 f_0(T) + \pi_1 f_1(T)} = c(T) = s. \end{aligned}$$

For the points $s_i = c(\tau_i) = c(\tau_{i+1})$ corresponding to constant intervals, we have that using Equation (26) and (27):

$$c_{m^{(c)}}(s_i) = \frac{\pi_1 h_1(s_i)}{\pi_0 h_0(s_i) + \pi_1 h_1(s_i)} = \frac{\pi_1 [F_1(T)]_{\tau_i}^{\tau_{i+1}}}{\pi_0 [F_0(T)]_{\tau_{i+1}}^{a_{i+1}} + \pi_1 [F_1(T)]_{\tau_i}^{\tau_{i+1}}}.$$

Since $c(T)$ is constant in the interval $]\tau_i, \tau_{i+1}[$, we have:

$$c_{m^{(c)}}(s_i) = \frac{\pi_1 f_1(T)}{\pi_0 f_0(T) + \pi_1 f_1(T)} = c(T) = s_i.$$

■

B.3 Main Result

Finally, we are ready to prove the theorem treating the three kinds of intervals.

Theorem 48 *(Theorem 29 in the paper) For every convex model m, we have that:*

$$L^o_{U(c)}(m) = RL(m).$$

Proof

Let us start from Lemma 46:

$$L^o_{U(c)}(m) = \acute{\Lambda}(m) + \dot{\Lambda}(m).$$

working with Equation (22) first for the bijective intervals:

$$\acute{\Lambda}(m) = \sum_{]\tau_i,\tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} 2c(T)\pi_0(1-F_0(T)) + 2(1-c(T))\pi_1 F_1(T)\}c'(T)dT.$$

Since this only includes the bijective intervals, we can use the correspondence between the H and the F, and making the change $s = c(T)$.

$$\begin{aligned}
\acute{\Lambda}(m) &= \sum_{]\tau_i,\tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} 2c(T)\pi_0(1-H_0(c(T))) + 2(1-c(T))\pi_1 H_1(c(T))\}c'(T)dT \\
&= \sum_{]c(\tau_i),c(\tau_{i+1})[\in \acute{I}_\sigma} \int_{c(\tau_i)}^{c(\tau_{i+1})} 2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds \\
&= \sum_{]\sigma_i,\sigma_{i+1}[\in \acute{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds.
\end{aligned}$$

and now working with Equation (23) for the singular intervals and also using the correspondence between the H and the F:

$$\begin{aligned}
\dot{\Lambda}(m) &= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2c\pi_0(1-F_0(\tau_i)) + 2(1-c)\pi_1 F_1(\tau_i)\}dc \\
&= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2c\pi_0(1-H_0(c(\tau_i))) + 2(1-c)\pi_1 H_1(c(\tau_i))\}dc \\
&= \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2s\pi_0(1-H_0(\sigma_i)) + 2(1-s)\pi_1 H_1(\sigma_i)\}ds.
\end{aligned}$$

The last step also uses the renaming of the variable. But since $h_0(s) = h_1(s) = 0$ for the singular intervals, we have that $H_0(s)$ and $H_1(s)$ are constant in these intervals, so this can be rewritten as:

$$\dot{\Lambda}(m) = \sum_{]\sigma_i,\sigma_{i+1}[\in \dot{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds.$$

Putting $\acute{\Lambda}(m)$ and $\dot{\Lambda}(m)$ together, because the constant intervals ($\bar{I_\sigma}$) have length 0 (and loss 0), we have:

$$L^o_{U(c)}(m) = \sum_{]\sigma_i,\sigma_{i+1}[\in I_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} 2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds.$$

We can join the integrals into a single one, even though the whole integral has to be calculated by intervals if it is discontinuous:

$$\begin{aligned}
L^o_{U(c)}(m) &= \int_{\sigma_0}^{\sigma_n} \{2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds \\
&= \int_0^1 \{2s\pi_0(1-H_0(s)) + 2(1-s)\pi_1 H_1(s)\}ds.
\end{aligned}$$

By Theorem 14 (Equation 10) in the paper (and also because this theorem holds pointwise) we have that the last expression equals the Brier score, so this leads to:

$$L^o_{U(c)}(m) = BS(m^{(c)}).$$

And now we have that using Definition 8 for the *BS*:

$$BS(m^{(c)}) = \int_0^1 \{\pi_0 s^2 h_0(s) + \pi_1 (1-s)^2 h_1(s)\} ds.$$

This is 0 when $h_0(s) = h_1(s) = 0$, so we can ignore the singular intervals for the rest of the proof. The calibration loss for model $m^{(c)}$ can be expanded as follows, and using Lemma 47 (which is applicable except for non-singular intervals) we have:

$$\begin{aligned} CL(m^{(c)}) &= \int_0^1 \left(s - \frac{\pi_1 h_1(s)}{\pi_0 h_0(s) + \pi_1 h_1(s)} \right)^2 (\pi_0 h_0(s) + \pi_1 h_1(s))\, ds \\ &= \int_0^1 (s-s)^2 (\pi_0 h_0(s) + \pi_1 h_1(s))\, ds = 0. \end{aligned}$$

So, we have that:

$$L^o_{U(c)}(m) = RL(m^{(c)}). \tag{28}$$

And now we need to work with *RL*:

$$\begin{aligned} RL(m^{(c)}) &= \int_0^1 \frac{\pi_1 h_1(s) \pi_0 h_0(s)}{\pi_0 h_0(s) + \pi_1 h_1(s)} ds = \int_0^1 \pi_0 h_0(s) \frac{\pi_1 h_1(s)}{\pi_0 h_0(s) + \pi_1 h_1(s)} ds \\ &= \int_0^1 \pi_0 h_0(s) c_{m^{(c)}}(s) ds = \int_0^1 \pi_0 h_0(s) s ds. \end{aligned}$$

The last step applies Lemma 47 again.

We now need to treat the bijective and the constant intervals separately, otherwise the integral cannot be calculated when h_0 and h_1 are discontinuous.

$$RL(m^{(c)}) = \sum_{]\sigma_i, \sigma_{i+1}[\in \acute{I}_\sigma} \int_{\sigma_i}^{\sigma_{i+1}} \pi_0 h_0(s) s ds + \sum_{]\sigma_i, \sigma_{i+1}[\in \bar{I}_\sigma} \pi_0 h_0(\sigma_i) \sigma_i.$$

We apply the variable change $s = c(T)$ for the expression on the left:

$$\begin{aligned} \sum_{]c(\tau_i), c(\tau_{i+1})[\in \acute{I}_\sigma} \int_{c(\tau_i)}^{c(\tau_{i+1})} \pi_0 h_0(s) s ds &= \sum_{]\tau_i, \tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} \pi_0 h_0(c(T)) c(T) \frac{dc(T)}{dT} dT \\ &= \sum_{]\tau_i, \tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} \pi_0 h_0(c(T)) \frac{\pi_1 f_1(T)}{\pi_1 f_1(T) + \pi_0 f_0(T)} \frac{dc(T)}{dT} dT \\ &= \sum_{]\tau_i, \tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} \pi_0 h_0(c(T)) \frac{dc(T)}{dT} \frac{\pi_1 f_1(T)}{\pi_1 f_1(T) + \pi_0 f_0(T)} dT \\ &= \sum_{]\tau_i, \tau_{i+1}[\in \acute{I}_\tau} \int_{\tau_i}^{\tau_{i+1}} \pi_0 f_0(T) \frac{\pi_1 f_1(T)}{\pi_1 f_1(T) + \pi_0 f_0(T)} dT. \end{aligned}$$

We now work with the expression on the right using Equation (26):

$$\begin{aligned}\sum_{]c(\tau_i),c(\tau_{i+1})[\in\bar{I}_\sigma}\pi_0 h_0(c(\tau_i))c(\tau_i) &= \sum_{]\tau_i,\tau_{i+1}[\in\bar{I}_\tau}\pi_0[F_0(T)]_{\tau_i}^{\tau_{i+1}}c(\tau_i) = \sum_{]\tau_i,\tau_{i+1}[\in\bar{I}_\tau}\pi_0\int_{\tau_i}^{\tau_{i+1}} f_0(T)dT\,c(\tau_i)\\ &= \sum_{]\tau_i,\tau_{i+1}[\in\bar{I}_\tau}\int_{\tau_i}^{\tau_{i+1}}\pi_0 f_0(T)c(T)dT\\ &= \sum_{]\tau_i,\tau_{i+1}[\in\bar{I}_\tau}\int_{\tau_i}^{\tau_{i+1}}\pi_0 f_0(T)\frac{\pi_1 f_1(T)}{\pi_1 f_1(T)+\pi_0 f_0(T)}dT.\end{aligned}$$

The change from $c(\tau_i)$ to $c(T)$ inside the integral can be performed since $c(T)$ is constant, because here we are working with the constant intervals.

Putting everything together again:

$$\begin{aligned}RL(m^{(c)}) &= \sum_{]\tau_i,\tau_{i+1}[\in\acute{I}_\tau}\int_{\tau_i}^{\tau_{i+1}}\pi_0 f_0(T)\frac{\pi_1 f_1(T)}{\pi_1 f_1(T)+\pi_0 f_0(T)}dT + \sum_{]\tau_i,\tau_{i+1}[\in\bar{I}_\tau}\int_{\tau_i}^{\tau_{i+1}}\pi_0 f_0(T)\frac{\pi_1 f_1(T)}{\pi_1 f_1(T)+\pi_0 f_0(T)}dT\\ &= \int_{\tau_0}^{\tau_n}\frac{\pi_0 f_0(T)\pi_1 f_1(T)}{\pi_1 f_1(T)+\pi_0 f_0(T)}dT = \int_{-\infty}^{\infty}\frac{\pi_0 f_0(T)\pi_1 f_1(T)}{\pi_1 f_1(T)+\pi_0 f_0(T)}dT = RL(m).\end{aligned}$$

This and Equation (28) complete the proof. ■

References

N. M. Adams and D. J. Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999.

T. A. Alonzo, M. S. Pepe, and T. Lumley. Estimating disease prevalence in two-phase studies. *Biostatistics*, 4(2):313–326, 2003.

M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 5:641–647, 1955.

A. Bella, C. Ferri, J. Hernandez-Orallo, and M.J. Ramirez-Quintana. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications*, pages 128–146. IGI Global, 2009.

A. Bella, C. Ferri, J. Hernández-Orallo, and M.J. Ramírez-Quintana. Quantification via probability estimators. In *2010 IEEE International Conference on Data Mining*, pages 737–742. IEEE, 2010.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

N. Brümmer. Measuring, refining and calibrating speaker and language information extracted from speech. *Ph.D. Dissertation, Department of Electrical and Electronic Engineering, University of Stellenbosch*, 2010.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation: structure and applications. http://www-stat.wharton.upenn.edu/∼buja/PAPERS/paper-proper-scoring.pdf, 2005.

I. Cohen and M. Goldszmidt. Properties and benefits of calibrated classifiers. *Knowledge Discovery in Databases: PKDD 2004*, pages 125–136, 2004.

C. Drummond and R. C. Holte. Explicitly representing expected cost: an alternative to ROC representation. In *Knowledge Discovery and Data Mining*, pages 198–207, 2000.

C. Drummond and R. C. Holte. Cost curves: an improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

C. Elkan. The foundations of cost-sensitive searning. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCAI-01)*, pages 973–978, San Francisco, CA, 2001.

T. Fawcett. Using rule sets to maximize ROC performance. In *2001 IEEE International Conference on Data Mining (ICDM-01)*, pages 131–138, 2001.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

T. Fawcett and A. Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1): 97–106, July 2007.

T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3): 291–316, 1997.

C. Ferri, P.A. Flach, J. Hernández-Orallo, and A. Senad. Modifying ROC curves to incorporate predicted probabilities. In *Second Workshop on ROC Analysis in ML, ROCML*, pages 33–40, 2005.

C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009. ISSN 0167-8655.

P. A. Flach. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pages 194–201, 2003.

P. A. Flach. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, to appear, 2012.

P. A. Flach and E. T. Matsubara. A simple lexicographic ranker and probability estimator. In *18th European Conference on Machine Learning*, pages 575–582. Springer, 2007.

P. A. Flach, J. Hernández-Orallo, and C. Ferri. A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning, ICML2011*, 2011.

G. Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

J. H. Friedman. On bias, variance, 0/1loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.

M. Gebel. *Multivariate Calibration of Classifier Scores into the Probability Space*. PhD thesis, University of Dortmund, 2009.

T. Gneiting and A.E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. ISSN 0162-1459.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114, 1952.

D. J. Hand. *Construction and Assessment of Classification Rules*. John Wiley & Sons Inc, 1997.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

D. J. Hand. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in Medicine*, 29(14):1502–1510, 2010.

D. J. Hand and C. Anagnostopoulos. When is the area under the ROC curve an appropriate measure of classifier performance? *Technical report, Department of Mathematics, Imperial College, London*, 2011.

D. J. Hand and C. Anagnostopoulos. A better Beta for the H measure of classification performance. *arXiv:1202.2564v1 [stat] 12 Feb 2012*, page 9, 2012.

J. Hernández-Orallo, P. A. Flach, and C. Ferri. Brier curves: a new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on Machine Learning, ICML2011*, 2011.

J. Hernández-Orallo, P. Flach, and C. Ferri. ROC curves in cost space. In *Submitted*, 2012.

N. Lachiche and P. Flach. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using roc curves. In *International Conference on Machine Learning*, pages 416–423, 2003.

G. Lebanon and J. D. Lafferty. Cranking: combining rankings using conditional probability models on permutations. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)*, pages 363–370, 2002.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Fifth European Conference on Speech Communication and Technology*, 1997.

A. H. Murphy. A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology*, 5:534–536, 1966. ISSN 0894-8763.

A. H. Murphy. Measures of the utility of probabilistic predictions in cost-loss ratio decision situations in which knowledge of the cost-loss ratios is incomplete. *Journal of Applied Meteorology*, 8:863–873, 1969. ISSN 0894-8763.

A. H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12: 595–600, 1973.

A. H. Murphy and R. L. Winkler. Scoring rules in probability assessment and evaluation. *Acta Psychologica*, 34:273–286, 1970.

J. M. Murphy, D. M. Berwick, M. C. Weinstein, J. F. Borus, S. H. Budman, and G. L. Klerman. Performance of screening and diagnostic tests: application of receiver operating characteristic analysis. *Archives of General Psychiatry*, 44(6):550, 1987.

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938.

A. Niculescu-Mizil and R. Caruana. Obtaining calibrated probabilities from boosting. In *The 21st Conference on Uncertainty in Artificial Intelligence (UAI 05), AUAI Press*, pages 413–420. AUAI Press, 2005.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 625–632. ACM, 2005.

G. Piatetsky-Shapiro and B. Masand. Estimating campaign benefits and modeling lift. In *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 193. ACM, 1999.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Boston, 1999.

R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard. A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 23:1601–1618, 2011.

F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42 (3):203–231, 2001.

M. D. Reid and R. C. Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *The Journal of Machine Learning Research*, 12:731–817, 2011.

T. Robertson, F. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. John Wiley and Sons, New York, 1988.

S. Rüping. Robust probabilistic calibration. *Machine Learning: ECML 2006*, pages 743–750, 2006.

J. A. Swets, R. M. Dawes, and J. Monahan. Better decisions through science. *Scientific American*, 283(4):82–87, October 2000.

A. Tenenbein. A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, pages 1350–1361, 1970.

P. Turney. Types of cost in inductive concept learning. *Canada National Research Council Publications Archive*, 2000.

S. Wieand, M.H. Gail, B.R. James, and K.L. James. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3):585–592, 1989.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 609–616, 2001.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.