A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance

Peter Flach PETER.FLACH@BRISTOL.AC.UK
Intelligent Systems Laboratory, University of Bristol, UK

José Hernández-Orallo JORALLO@DSIC.UPV.ES
Cèsar Ferri CFERRI@DSIC.UPV.ES
Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain

Abstract

The area under the ROC curve (*AUC*), a well-known measure of ranking performance, is also often used as a measure of classification performance, aggregating over decision thresholds as well as class and cost skews. However, David Hand has recently argued that *AUC* is fundamentally incoherent as a measure of aggregated classifier performance and proposed an alternative measure (Hand, 2009). Specifically, Hand derives a linear relationship between *AUC* and expected *minimum* loss, where the expectation is taken over a distribution of the misclassification cost parameter that depends on the model under consideration. Replacing this distribution with a Beta$(2,2)$ distribution, Hand derives his alternative measure H. In this paper we offer an alternative, coherent interpretation of *AUC* as linearly related to expected loss. We use a distribution over cost parameter and a distribution over data points, both uniform and hence model-independent. Should one wish to consider only optimal thresholds, we demonstrate that a simple and more intuitive alternative to Hand's H measure is already available in the form of the area under the cost curve.

1. Introduction and Motivation

The area under the ROC curve (*AUC*) is a well-known measure of ranking performance, estimating the probability that a random positive is ranked before a random negative, without committing to a particular decision threshold. It is also often used as a measure of aggregated classification performance, on the grounds that *AUC* in some sense averages over all possible decision thresholds. David Hand criticises this practice in a recent paper (Hand, 2009). One version of his argument runs as follows (the details will be given in Section 2.3). *AUC* can be interpreted as the expected true positive rate, averaged over all false positive rates. For any given classifier we don't have direct access to the false positive rate, and so we average over possible decision thresholds. While there is a relationship between decision thresholds and cost parameters under which this threshold is optimal, this relationship is model-specific, and so the way *AUC* aggregates performance over possible cost parameters is model-specific. Expectations over the cost parameter should be task-specific and not dependent on the model, and so *AUC* may make a model's classification performance look better or worse than it actually is.

The conclusions Hand draws from his analysis appear devastating for *AUC*: it is "fundamentally incoherent in terms of misclassification costs", he writes, as it "evaluates different classifiers using different metrics. It is as if one measured person A's height using a ruler calibrated in inches and person B's using one calibrated in centimetres". It is worth emphasising that Hand's criticism is specifically directed towards interpreting *AUC* as a measure of aggregated *classification* performance. The standard interpretation of *AUC* as an estimator of the probability that a random positive is ranked before a random negative is undoubtedly model-independent and therefore coherent. It may appear, then, that the perceived incoherence of *AUC* as aggregated classification performance results from the misguided interpretation of a ranking performance measure as a classification performance measure.

However, we demonstrate in this paper that Hand's model-dependent interpretation of *AUC* arises from restricting attention to thresholds that are *optimal* under given cost parameters. We argue that this is not a natural interpretation

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

of *AUC*, and offer an alternative interpretation of *AUC* as expected classification performance averaged over cost parameters as well as data points. This interpretation is coherent because both the distribution over cost parameter and over examples are uniform and hence model-independent. Should one wish to consider only optimal thresholds, we demonstrate that a simple and more intuitive alternative to Hand's H measure is already available in the form of the area under the cost curve. The main contributions of the paper can thus be summarised as follows. First, we offer a novel, model-independent interpretation of *AUC* as an aggregation of macro-accuracy over all possible decision thresholds and cost parameters. Second, in doing so we provide a unifying framework for classifier performance evaluation. Thirdly, we offer a natural interpretation of the H measure as the area under the cost curve.

The outline of the paper is as follows. The following section introduces some notation, and gives a detailed summary of Hand's argument. Section 3 presents our new and classifier-independent interpretation of *AUC*. In Section 4 we show that Hand's alternative H measure is a variation of the area under the cost curve (Drummond & Holte, 2006). Finally, Section 5 concludes the paper.

2. Preliminaries

We follow notation from (Hand, 2009) to a large extent. (Much of this section is shared with a related paper proposing a new visualisation of classifier performance in cost space (Hernández-Orallo et al., 2011).)

The instance space is denoted X and the output space Y. Elements in X and Y will be referred to as x and y respectively. For this paper we will assume binary classifiers, i.e., $Y = \{0,1\}$. A crisp or categorical classifier is a function that maps examples to classes. A soft or scoring classifier is a function $m : X \to \Re$ that maps examples to real numbers, where the outputs can be interpreted as estimates $\hat{p}(1|x)$ of the probability of example x to be of class 1 or, more generally, as scores that are monotonically related to $\hat{p}(1|x)$. (Hand's notation is slightly non-standard in that he uses 0 for the positive class and 1 for the negative class, but scores increase with $\hat{p}(1|x)$. That is, a ranking on decreasing score proceeds from strongest negative prediction to strongest positive prediction.) In order to make predictions in the Y domain, a soft classifier can be converted to a crisp 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 classifier 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 positive rate or sensitivity. We then have that the proportion of correctly classified examples is (micro-average) accuracy $Acc(t) = \pi_0 F_0(t) + \pi_1(1 - F_1(t))$; macro-average accuracy is defined as the unweighted mean of true positive rate and true negative rate, $MAcc(t) = \frac{1}{2}(F_0(t) + 1 - F_1(t))$.

Given a dataset $D \subset \langle X, Y \rangle$ 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 $\pi_k = n_k/n$. 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 any given strict order for a dataset 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. Given a dataset and a classifier, we can define empirical score distributions for which we will use the same symbols as the population functions. We then have $f_k(s) = \frac{1}{n_k}|\{\langle x,y \rangle \in D_k | m(x) = s\}|$, which is non-zero only in n'_k points, where $n'_k \leq n_k$ is the number of unique scores assigned to instances in D_k (when there are no ties, we have $n'_k = n_k$). Furthermore, the cumulative distribution functions $F_k(t) = \sum_{s \leq t} f_k(s)$ are piecewise constant with $n'_k + 1$ segments.

2.1. Expected Loss

An operating condition or deployment context is usually defined by a class distribution and a way to aggregate misclassification cost over examples. 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 classifiers we can describe the cost matrix by two values $c_k \geq 0$, representing the misclassification cost of an example of class k. Additionally, we can normalise the costs by setting $b = c_0 + c_1$ and $c = c_0/b$; we will refer to c as the *cost proportion*.

The loss which is produced at a decision threshold t and a cost proportion c is then given by the formula:

$$\begin{aligned} Q_c(t;c) &\triangleq c_0\pi_0(1 - F_0(t)) + c_1\pi_1 F_1(t) \\ &= b\{c\pi_0(1 - F_0(t)) + (1-c)\pi_1 F_1(t)\} \end{aligned} \quad (1)$$

We are often interested in analysing the influence of class proportion and cost proportion at the same time. Since the relevance of c_0 increases with π_0, an appropriate way to consider both at the same time is by the definition of *skew*,

which is a normalisation of their product:

$$z \triangleq \frac{c_0\pi_0}{c_0\pi_0 + c_1\pi_1} = \frac{c\pi_0}{c\pi_0 + (1-c)(1-\pi_0)} \quad (2)$$

From Eq. (1) we obtain

$$\frac{Q_c(t;c)}{c_0\pi_0 + c_1\pi_1} = z(1-F_0(t)) + (1-z)F_1(t) \triangleq Q_z(t;z) \quad (3)$$

This gives an expression for loss at a threshold t and a skew z. We then have the following simple but useful result.

Lemma 1. *If $\pi_0 = \pi_1$ then $z = c$ and $Q_z(t;z) = \frac{2}{b}Q_c(t;c)$.*

Proof. If classes are balanced we have $c_0\pi_0 + c_1\pi_1 = b/2$, and the result follows from Eq. (2) and Eq. (3). □

This justifies taking $b = 2$, 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 $c_0 = c_1 = 1$. The upshot of Lemma 1 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 $z = c$.

In many real problems, when we evaluate or compare classifiers, we do not know the cost proportion or skew that will apply at application time. One general approach is to evaluate the classifier on a range of possible operating conditions and report expected loss over that range. In order to do this, we have to set a weight or distribution on cost proportions or skews. Following (Adams & Hand, 1999) we can define the expected loss over a range of situations as follows:

$$L_c \triangleq \int_0^1 Q_c(T_c(c);c)w_c(c)dc \quad (4)$$

where T_c is a threshold choice method which maps cost proportions to decision thresholds, and $w_c(c)$ is a distribution for cost proportions over $[0,1]$. By using Q_z instead of Q_c we can define expected loss over a range of skews:

$$L_z \triangleq \int_0^1 Q_z(T_z(z);z)w_z(z)dz \quad (5)$$

Here we use T_z as a function which converts skews into decision thresholds and w_z as a distribution over skews.

2.2. ROC Curves and Cost Curves

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. We then have that the Area Under the ROC curve (*AUC*) can be defined as

$$AUC = \int_0^1 F_0(s)dF_1(s) = \int_{-\infty}^{+\infty} F_0(s)f_1(s)ds \quad (6)$$

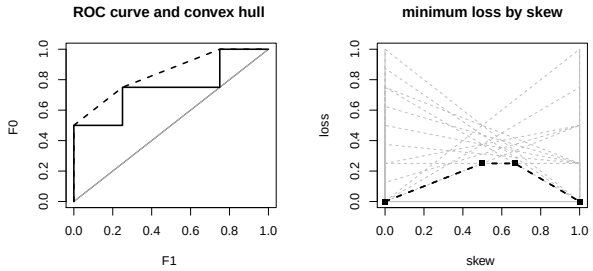

Figure 1. ROC curve and convex hull (left), and cost lines and cost curve (right) for a classifier with scores (0.95,0.9,0.8,0.7,0.65,0.6,0.5,0.4,0.3,0.2,0.1,0.05) and corresponding true classes (1,1,0,1,1,1,1,0,1,1,0,0) where 1 stands for the negative class. (*AUC*: 0.7500, *AUCH*: 0.8438).

When dealing with empirical distributions the integral is replaced by a sum.

The convex hull of a ROC curve (ROCCH) is a construction over the ROC curve in such a way that all the points on the ROCCH have minimum loss for some choice of c or z. It is defined in terms of the *optimal* threshold choice method for a given cost proportion c:

$$\begin{aligned} T_c^o(c) &\triangleq \arg\min_t \{Q_c(t;c)\} \\ &= \arg\min_t \; 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\} \end{aligned} \quad (7)$$

which matches the optimal threshold choice method for a given skew z:

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

The convex hull is obtained by linear interpolation between the points $\{F_1(t), F_0(t)\}$ where $t = T_c^o(c)$ for some c. The Area Under the ROCCH (denoted by *AUCH*) can be computed in a similar way as the *AUC* with modified versions of f_k and F_k. Obviously, $AUCH \geq AUC$, with equality implying the ROC curve is convex. Figure 1 (left) shows an example ROC curve and its convex hull.

Cost curves (Drummond & Holte, 2006) are a graphical technique to analyse the behaviour and performance of classifiers. A cost plot has $Q_z(t;z)$ on the y-axis against skew z on the x-axis (Drummond and Holte use the term 'probability cost' rather than skew). Since $Q_z(t;z) = z(1-F_0(t)) + (1-z)F_1(t)$, cost lines for a given decision threshold t are straight lines $Q_z = a + bz$ with intercept $a = F_1(t)$ and slope $b = 1 - F_0(t) - F_1(t)$. A cost line, denoted by CL_t, visualises how cost at that threshold changes between $F_1(t)$ for $z = 0$ and $1 - F_0(t)$ for $z = 1$. The cost curve is then the lower envelope of all the cost lines, obtained by only considering the optimal threshold for each skew; an explicit definition of the cost curve as a function of z in our notation is

$$CC(z) \triangleq Q_z(T_z^o(z);z) \quad (8)$$

The cost lines and cost curve of the classifier whose ROC curve is depicted in Figure 1 (left) is shown on the right-hand side of the same figure.

There is a clear duality between ROC plots and cost plots. Points in a ROC plot are obtained by fixing the decision threshold, and hence correspond to cost lines. Conversely, points in a cost plot arise from intersecting two cost lines, and hence correspond to a line segment in ROC space connecting two points. Furthermore, the convex hull of a classifier in ROC space corresponds to the lower envelope of the cost lines in the cost curve space. The cost curve has as many points as there are segments in the ROCCH, and as many segments as there are non-trivial points in the ROCCH. The main difference with ROC curves is that cost curves focus on classification performance rather than ranking performance. Furthermore, cost curves portray an optimistic view of the expected loss of a classifier, since they only consider optimal thresholds. For a new type of curve in cost space corresponding to a non-convex ROC curve the reader is referred to (Hernández-Orallo et al., 2011).

2.3. A Summary of Hand's Argument

In this section we summarise the main steps in Hand's argument that *AUC* is incoherent as a measure of aggregated classification performance. Combining L_c as defined in Eq. (4) and Q_c as defined in Eq. (1) we obtain an expression for expected loss:

$$L_c = \int_0^1 b\{c\pi_0(1-F_0(T_c(c)))+(1-c)\pi_1 F_1(T_c(c))\}w_c(c)dc \tag{9}$$

As in Section 2.1 we choose $b=2$ and multiply Hand's formulae for expected loss with a factor 2 below.

Hand now uses a specific choice for $T_c(c)$, namely the optimal one as defined in Eq. (7); this means that we switch to expected *minimum* loss, which we indicate as L_c^o. Under the assumption of a convex and continuously differentiable ROC curve, the mapping from c to $T_c^o(c)$ is one-to-one and invertible. (Later on in his paper Hand shows how to relax this assumption, but it is sufficient to consider this special case for the moment.) By differentiating $Q_c(t;c)$ we find that the minimising T for cost proportion c satisfies $c\pi_0 f_0(T) = (1-c)\pi_1 f_1(T)$, and hence the inverse of $T_c^o(c)$ is

$$c(T) = \pi_1 f_1(T)/\{\pi_0 f_0(T)+\pi_1 f_1(T)\} \tag{10}$$

We now change the variable of integration in Eq. (9) from c to T, which gives (c goes to $c(T)$, dc goes to $c'(T)dT$, and $T_c^o(c)$ goes to T)

$$L_c^o = \int_{-\infty}^{\infty} 2\{c(T)\pi_0(1-F_0(T))+(1-c(T))\pi_1 F_1(T)\}W(T)dT \tag{11}$$

Here, $W(T) = w_c(c(T))c'(T)$ is a distribution over optimal thresholds, translating prior beliefs over c expressed by $w_c(c)$ into prior beliefs over T expressed by $W(T)$.

Hand then considers a particular choice for $W(T)$ which is the mixture distribution of the two score densities defining the model, and hence model-dependent:

$$W_G(T) \triangleq \pi_0 f_0(T)+\pi_1 f_1(T) \tag{12}$$

We can recover the individual score densities as $W_G(T)c(T) = \pi_1 f_1(T)$ and $W_G(T)(1-c(T)) = \pi_0 f_0(T)$; this simplifies the expression for expected loss to

$$L_{c,G}^o = \int_{-\infty}^{\infty} 2\{\pi_0\pi_1\{f_1(T)(1-F_0(T))+f_0(T)F_1(T)\}\}dT$$

Since $AUC = \int_{-\infty}^{+\infty} F_0(s)f_1(s)ds$ and $1-AUC = \int_{-\infty}^{+\infty} F_1(s)f_0(s)ds$, we finally arrive at

$$L_{c,G}^o = 4\pi_0\pi_1(1-AUC) \tag{13}$$

In other words, optimising *AUC* means minimising expected minimum loss under threshold distribution W_G. As there is a one-to-one mapping from optimal thresholds to scores, this can be traced back to a score distribution

$$w_G(c) = \{\pi_0 f_0(T_c^o(c))+\pi_1 f_1(T_c^o(c))\}\left|\frac{dT_c^o(c)}{dc}\right| \tag{14}$$

which depends on the score densities and hence on the classifier. (Here, Hand writes $P_1^{-1}(c)$ instead of $T_c^o(c)$.)

To summarise, Hand derives a linear relationship between expected minimum loss and *AUC under a classifier-dependent distribution over cost proportions* ($w_G(c)$ in Eq. (14)).[1] So, two classifiers may have the same *AUC*, but that doesn't imply that they have equal expected minimum loss if a different distribution over cost proportions was used that was the same for both classifiers. This appears to be a fatal blow to the credibility of *AUC* as a measure of aggregated classification performance. However, in the next section we derive an alternative interpretation of *AUC* as linearly related to expected loss (rather than expected minimum loss), where the expectation is taken uniformly over all cost proportions. The key to our analysis is that we do not assume that thresholds are chosen optimally – a very strong assumption, as we will argue.

[1] To deal with non-continuously differentiable ROC curves, such as the empirical ROC curves obtained from test samples, Hand shows in his paper how to deal with a many-to-one relationship between cost proportions and optimal thresholds. To deal with non-convex ROC curves, Hand proposes to use *AUCH* instead of *AUC*. So, strictly speaking, what Hand has derived is a linear relationship between *AUCH* and expected minimum loss under a classifier-dependent distribution over cost proportions.

3. AUC is Coherent When Including Non-Optimal Thresholds

So far, we have followed Hand, and Drummond and Holte, in restricting attention to optimal thresholds T_c^o as defined by Eq. (7). But one cannot always expect the user of a classifier to be aware of ROC curves and/or convex hulls. Even if s/he knows, it is possible that the optimal threshold for the validation dataset is sub-optimal for the test dataset. Drummond and Holte (Drummond & Holte, 2006) are conscious of this problem and are willing to 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".

In our view, basing performance metrics on optimal thresholds is overly optimistic. Other threshold choice methods are possible. One is to choose the threshold such that $\hat{p}(1|x) = z$ where z is the operating condition. If scores are calibrated estimates of the class posterior, this means setting $T(z) = z$, i.e., making the score threshold equal to the skew. Drummond and Holte criticise this approach because many classifiers are not well-calibrated. In what follows we take a third approach by considering *as many thresholds as there are examples*. This leads to an alternative, and coherent, interpretation of *AUC* as a measure of aggregate classification performance.

In order to derive our alternative interpretation of *AUC*, we return to Eq. (1) which gives an expression for the loss produced at threshold t and cost proportion c:

$$Q_c(t;c) = b\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}$$

We again assume $b = 2$ to ensure that loss is commensurate with error rate. We want to obtain an expression for expected loss *without assuming a one-to-one mapping between cost proportion and threshold*, which means that we need to integrate over these separately:

$$L_c^t \triangleq \int_0^1 \int_{-\infty}^{\infty} Q_c(t;c)W(t)dt\ w_c(c)dc \qquad (15)$$

If we assume a uniform distribution over cost proportions (i.e., $w_c(c) = 1$) this reduces to

$$\begin{aligned} L_{U(c)}^t &\triangleq \int_0^1 \int_{-\infty}^{\infty} 2\{c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\}W(t)dt\ dc \\ &= 2\left[\int_{-\infty}^{\infty} \{\frac{c^2}{2}\pi_0(1-F_0(t)) + (c-\frac{c^2}{2})\pi_1 F_1(t)\}W(t)dt\right]_0^1 \\ &= \int_{-\infty}^{\infty} \{\pi_0(1-F_0(t)) + \pi_1 F_1(t)\}W(t)dt \qquad (16) \end{aligned}$$

Now, it may seem natural to choose a uniform distribution also for $W(t)$. It can be shown that this gives us an expression for expected loss in terms of the per-class mean scores. However, this does not give us a connection between expected loss and *AUC*. In order to obtain such a link, we need to choose thresholds indirectly: we uniformly select an instance x, and set the threshold to the score of that instance: $t = m(x)$. If x is a positive example then the resulting threshold distribution is $f_0(t)$, and if it's a negative it is $f_1(t)$; since positives are chosen with probability π_0 and negatives with probability π_1 the overall distribution over thresholds is exactly Hand's mixture distribution $W_G(t) = \pi_0 f_0(t) + \pi_1 f_1(t)$. We denote the resulting loss function as $L_{U(c)}^{U(i)}$ for uniform cost proportion and uniform instance selection. We then have the following result.

Theorem 2. *Expected loss for uniform cost proportion and uniform instance selection decreases linearly with AUC as follows:*

$$L_{U(c)}^{U(i)} = \frac{L_{c,G}^o}{2} + \frac{\pi_0^2+\pi_1^2}{2} = 2\pi_0\pi_1(1-AUC) + \frac{\pi_0^2+\pi_1^2}{2}$$

Proof.

$$\begin{aligned} L_{U(c)}^{U(i)} &\triangleq \int_{-\infty}^{\infty} \{\pi_0(1-F_0(t)) + \pi_1 F_1(t)\}\{\pi_0 f_0(t) + \pi_1 f_1(t)\}dt \\ &= \pi_0\pi_1 \int_{-\infty}^{\infty} \{(1-F_0(t))f_1(t) + F_1(t)f_0(t)\}dt \\ &+ \pi_0^2 \int_{-\infty}^{\infty} (1-F_0(t))f_0(t)\ dt + \pi_1^2 \int_{-\infty}^{\infty} F_1(t)f_1(t)\ dt \end{aligned}$$

The first term is equal to half the expected minimum loss obtained by Hand selecting only optimal thresholds (Eq. (13)). For the other two terms, $\int_{-\infty}^{\infty} F_1(t)f_1(t)\ dt = \int_{-\infty}^{\infty} F_1(t)\ dF_1(t) = 1/2$ and the result follows. □

Corollary 3. *Expected loss for uniform skew and uniform instance selection decreases linearly with AUC as follows:*

$$L_{U(z)}^{U(i)} = \frac{L_{z,G}^o}{2} + \frac{1}{4} = \frac{1-AUC}{2} + \frac{1}{4}$$

These results demonstrate the effect of including non-optimal thresholds very clearly: it weakens the influence of *AUC* by a factor 2, and introduces a constant term depending only on the class distribution. As a result, 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.

3.1. The Case of Empirical ROC Curves

So far we have concentrated on the case where we have access to population densities $f_k(s)$ and distribution functions $F_k(t)$. In practice this is unrealistic, and we have to work with empirical estimates (in fact, the example in Figure 1

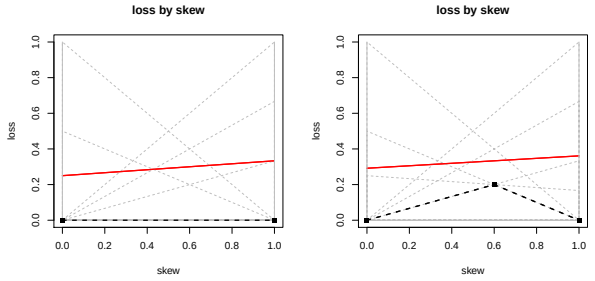

Figure 2. The cost line plots for the classifier in Example 1 (Left) and in Example 2 (Right). The red line shows the average of all the cost lines, the area under which is $L^{\mathscr{U}(i)}_{U(z)}$.

used empirical estimates). In this section we provide an alternative formulation of our main result, relating empirical loss to the *AUC* of the empirical ROC curve. We introduce the main ideas by means of examples.

Example 1. *Consider a dataset with 4 examples with classes* $(1,1,0,0,0)$ *and scores* $(0.9,0.8,0.7,0.2,0.1)$*. Using the example scores as thresholds* $t_1,\ldots,t_5$*, we can calculate the loss for each threshold as follows:*

	t_1	t_2	t_3	t_4	t_5	*Avg*
$F_0(t_i)$	1	1	1	$2/3$	$1/3$	$4/5$
$F_1(t_i)$	1	$1/2$	0	0	0	$3/10$
$Q_z(t_i;z)$	$1-z$	$\frac{1-z}{2}$	0	$z/3$	$2z/3$	$\frac{3-z}{10}$
$\int_0^1 Q_z(t_i;z)dz$	$1/2$	$1/4$	0	$1/6$	$1/3$	$\mathbf{1/4}$

Notice that each of the $Q_z(t_i;z)$ *represents one of the cost lines associated with the classifier. Averaging the* Q_z *terms gives* $\frac{3-z}{10}$*; averaging over all* z *(i.e., taking* $z=1/2$*) gives an expected loss of* $1/4$*.*

However, the problem with this loss calculation is that it does not take account of one of the default classifiers, namely the one with $F_0(t_6)=F_1(t_6)=0$ and $Q_z(t_6;z)=z$. Including this one in the average gives $\frac{3+z}{12}$ and a slightly higher expected loss of $7/24$ (denoted by $L^{\mathscr{U}(i)}_{U(z)}$ since now we use a discrete uniform distribution).

The second issue we have to deal with concerns tied scores.

Example 2. *We now change the scores in the previous example to* $(0.9,0.7,0.7,0.2,0.1)$*; i.e., the lowest scoring negative and the highest scoring positive get the same score. This reduces the AUC to* $11/12$*. We obtain the following losses for each threshold:*

	t_1	t_2	t_3	t_4	t_5	t_6	*Avg*
F_0	1	1	1	$2/3$	$1/3$	0	$2/3$
F_1	1	$1/2$	$1/2$	0	0	0	$1/3$
Q_z	$1-z$	$\frac{1-z}{2}$	$\frac{1-z}{2}$	$z/3$	$2z/3$	z	$1/3$
$\int_0^1 Q_z$	$1/2$	$1/4$	$1/4$	$1/6$	$1/3$	$1/2$	$\mathbf{1/3}$

Averaging the Q_z *terms gives* $1/3$*, i.e., a horizontal cost line, which is therefore also the expected loss.*

However, this turns out to be a biased estimate, as we see when we swap the classes and take the complement of the scores. We now have a ranking $(1,1,1,0,0)$ *and scores* $(0.9,0.8,0.3,0.3,0.1)$*, which produces:*

	t'_1	t'_2	t'_3	t'_4	t'_5	t'_6	*Avg*
F_0	1	1	1	1	$1/2$	0	$3/4$
F_1	1	$2/3$	$1/3$	$1/3$	0	0	$7/18$
Q_z	$1-z$	$\frac{2(1-z)}{3}$	$\frac{1-z}{3}$	$\frac{1-z}{3}$	$z/2$	z	$\frac{14-5z}{36}$
$\int_0^1 Q_z$	$1/2$	$1/3$	$1/6$	$1/6$	$1/4$	$1/2$	$\mathbf{23/72}$

The Q_z *terms average to* $\frac{14-5z}{36}$ *with an expected loss of* $\frac{23}{72}$*.*

The reason for this discrepancy – which only manifests itself when we have tied scores between positives and negatives – is that setting thresholds on examples, combined with a decision rule that classifies examples as class 1 if they exceed the threshold, biases tied examples in favour of class 0. It follows that swapping the classes exhibits a bias in favour of the original class 1, and the right thing to do is to average these two cases. That is, the expected loss in Example 2 is actually $47/144$. The previous calculations can be generalised according to the following result (proofs of this theorem and the next can be found in (Flach et al., 2011)):

Theorem 4. *Let AUC be an empirical estimate obtained from a dataset with n examples, then the expected loss for uniform skew and (discrete) uniform instance selection is*

$$L^{\mathscr{U}(i)}_{U(z)} = \left(\frac{n}{n+1}\right)\frac{1-AUC}{2} + \left(\frac{n+2}{n+1}\right)\frac{1}{4}$$

Clearly, $L^{\mathscr{U}(i)}_{U(z)}$ converges to $L^{U(i)}_{U(z)}$ when $n\to\infty$.

Interestingly, considering the n examples plus an extra default classifier in both directions is equivalent to considering the threshold placed between examples. And each of these $n+1$ cases corresponds to one cost line CL_{t_i}, as argued before. Since each cost line CL_{t_i} goes from $F_1(t_i)$ to $1-F_0(t_i)$, its area is $1/2(F_1(t_i)+1-F_0(t_i))$. But this is exactly $1-MAcc(t_i)$. This leads to the following result:

Theorem 5.

$$\begin{aligned} L^{\mathscr{U}(i)}_{U(z)} &= \frac{1}{(n+1)}\sum_{i=1}^{n+1}\int_0^1 CL_{t_i}(s)ds \\ &= \frac{1}{(n+1)}\sum_{i=1}^{n+1}(1-MAcc(t_i)) \end{aligned}$$

The average of all the cost lines can itself be drawn as a cost line, as shown in Figure 2 for Examples 1 and 2. For Example 1 the red line stretches from 1/4 to 1/3 and the area beneath it is $7/24$. This line, which we call *loss line* and denote by *LL*, is defined as follows

$$LL(s) \triangleq \frac{1}{(n+1)}\sum_{i=1}^{n+1} CL_{t_i}(s)$$

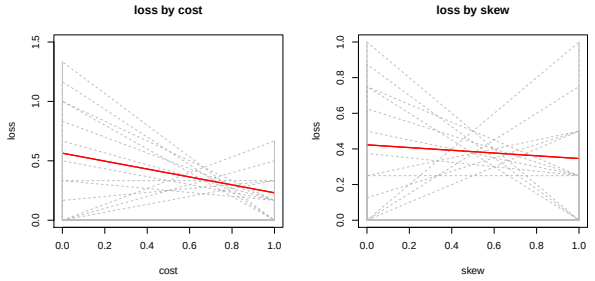

Figure 3. Cost lines for the classifier in Figures 1 and 4 for cost proportions (left) and for skews (right). The red line shows the average of all the lines, the area under which is the expected loss.

We then have that

$$L_{U(z)}^{\mathscr{U}(i)} = \int_0^1 LL(s)ds$$

Since $L_{U(z)}^{\mathscr{U}(i)}$ is linearly related to *AUC*, this allows us to say that *the loss line is a graphical representation of AUC in cost space*.

We thus reach a natural and straightforward interpretation of *AUC* for classification. Classification performance for a single crisp classifier is measured by macro-accuracy (when working with skews). By placing the thresholds between the examples, we obtain an instance-based average of the macro-accuracies. As our results show, and as demonstrated in Figure 2 and Figure 3 (right), this is exactly what $L_{U(z)}^{\mathscr{U}(i)}$ is. And *AUC* is just a linear scaling of it. When working with cost proportions, we obtain a similar interpretation with accuracy instead of macro-accuracy, as shown in Figure 3 (left). The red line in this case represents a variant of *AUC* which is sensitive to the class proportion.

4. Expected Minimum Loss is Measured by the Area under the Cost Curve

In this section we show that Hand's new measure H – which is to use expected minimum loss under a Beta distribution for $w(c)$ – is a variation of a well-known measure, namely the area under cost curve as introduced by (Drummond & Holte, 2006).

Hand's alternative to *AUC* is an explicit expected minimum loss measure with the cost distribution $w_c(c)$ equal to the beta distribution $B(c,\alpha,\beta)$:

$$L_{\alpha,\beta} \triangleq \int_0^1 Q_c(T_c^o(c);c)B(c,\alpha,\beta)dc \qquad (17)$$

Hand defines his new measure H as a linear transformation of $L_{\alpha,\beta}$:

$$H \triangleq 1 - \frac{L_{\alpha,\beta}}{L_{Max}} \qquad (18)$$

where L_{Max} is the expected loss of the worst case (a classifier whose ROCCH is diagonal). This makes H comparable to *AUC*: both are expressed on a 0-1 scale, with higher numbers meaning better performance. Hand suggests to use $\alpha = \beta = 2$.

The argument that Hand's proposal is closely related to the area under the cost curve is straightforward. From Eq. (5) we obtain the following definition of expected minimum loss in terms of skews:

$$L_z^o \triangleq \int_0^1 Q_z(T_z^o(z);z)w_z(z)dz = \int_0^1 CC(z)w_z(z)dz \qquad (19)$$

The equation on the right uses the functional form of the cost curve as defined in Eq. (8). Clearly, this calculates the area under the cost curve if $w_z(z)$ is the uniform distribution. Drummond and Holte give essentially the same formula on p.106, writing x for z and $prob(x)$ for $w_z(z)$ and calling the resulting loss 'total expected cost' or TEC. They note that "[t]he area under a cost curve is the expected cost of the classifier assuming all possible probability-cost values [i.e., skews] are equally likely, i.e. that $prob(x)$ is the uniform distribution."

It follows that *the evaluation measure H as introduced by Hand is a linear transformation of the area under the cost curve*, with two small variations: using cost proportions instead of skews (so being sensitive to class priors); and replacing the uniform distribution with a Beta$(2,2)$ distribution. Figure 4 visualises these variations for the classifier in Figure 1. Each plot has loss ($Q_c(T_c^o(c);c)$ or $Q_z(T_z^o(z);z)$) on the y-axis and weighted cost proportion or skew ($w_c(c)$ or $w_z(z)$) on the x-axis. The bottom left plot shows the cost curve as introduced by Drummond and Holte; this is the same curve as in Figure 1 (right). The other curves show the effect of using the cost proportion instead of the skew (top row) and using the Beta$(2,2)$ distribution instead of the uniform distribution (right column). Hand's proposal $L_{2,2}$ is at the top right. We denote these four loss measures $L_{U(c)}^o$, $L_{U(z)}^o$, $L_{B_{2,2}(c)}^o$ and $L_{B_{2,2}(z)}^o$. From Lemma 1 it follows that $L_{U(c)}^o = L_{U(z)}^o$ and $L_{B_{2,2}(c)}^o = L_{B_{2,2}(z)}^o$ when the classes are balanced. For the uniform distribution cost lines are straight, while for Beta$(2,2)$ they are curved. It is worth mentioning that using *AUC*, *AUCH*, $L_{U(z)}^o$, $L_{B_{2,2}(z)}^o$, $L_{U(c)}^o$ and $L_{B_{2,2}(c)}^o$ for model selection may produce different choices, since none of these six measures is a monotonic function of another.

The fact that Hand's new measure is dependent on the class priors is openly recognised by him when he writes: "[i]t is worth noting that, whereas the *AUC*, the Gini coefficient, and the *AUCH* measure are independent of the class priors, π_0 and π_1, the H measure depends on the priors. This is clearly necessary since H is a measure of the (complement of) misclassification loss, and this depends on the relative proportion of objects belonging to each class" ((Hand, 2009), page 116). We disagree with Hand as this being

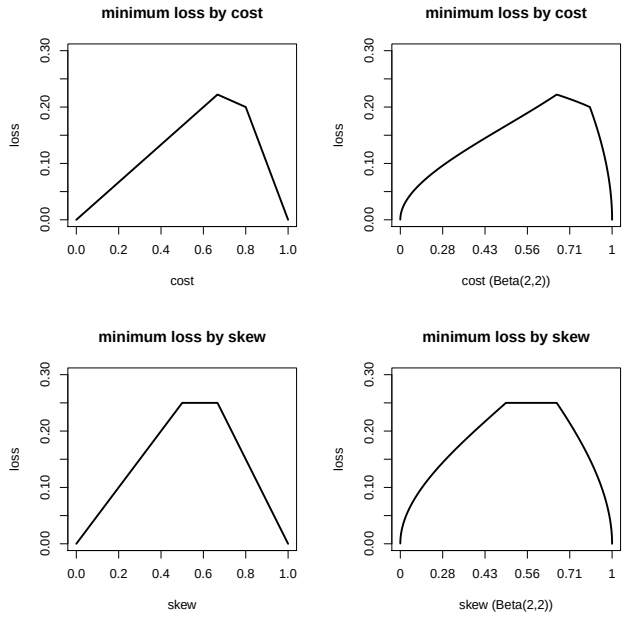

Figure 4. A visualisation of the classifier in Figure 1 using several plots which change the distribution on cost proportions or skews. Top Left: Q_c using a uniform distribution ($L^o_{U(c)} = 0.0611$). Top right: Q_c using a Beta$(2,2)$ distribution ($L^o_{B_{2,2}(c)} = 0.0727$, corresponding to $H = 0.4653$). Bottom left: Q_z using a uniform distribution (Area under the cost curve, $L^o_{U(z)} = 0.0708$). Bottom right: Q_z using a Beta$(2,2)$ distribution, $L^o_{B_{2,2}(z)} = 0.0900$.

"necessary", since we have seen that a similar derivation can be obtained with skews. In our opinion, the notion of skew is a more general and more useful formalisation of operating condition, as has been vindicated in ROC analysis and also in cost curves. On the other hand, the use of a Beta$(2,2)$ distribution instead of a uniform distribution is an "arbitrary" choice ((Hand, 2009), page 115). In fact, this choice makes cost plots more difficult to draw and interpret since cost lines become curved.

5. Concluding Remarks

In this paper we have shown that *AUC* can be a coherent measure of aggregated classification performance when, in lieu of Hand's interpretation which restricts attention to optimal thresholds, we consider all scores that have been assigned to data points as thresholds. In this way, we have been able to derive an expression of *AUC* which integrates loss over a skew distribution that is independent of the classifier. This makes Hand's objections regarding the incoherence of *AUC* vanish.

We have also derived a visualisation of (a linear transformation of) *AUC* in cost space by means of averaging cost lines, which we call loss line. This loss line leads to a view of *AUC* in terms of the average of the macro-accuracies between examples. As a result, the interpretation of *AUC* as an aggregated classification performance measure becomes crystal clear.

Regarding the H measure, we show that it is a linear transformation of the area under the cost curve, with two small variations: using cost proportions instead of skews and replacing the uniform distribution with a Beta$(2,2)$ distribution. Neither of these variations appears more strongly justified than the area under the cost curve): the first makes H sensitive to class priors, and the second is, ultimately, not less arbitrary than a uniform distribution which has the advantage that cost lines are straight.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work has been partially supported by the EU (FEDER) and the Spanish MICINN, under grant TIN2010-21062-C02-02, the Spanish project 'Agreement Technologies' (Consolider Ingenio CSD2007-00022) and the GVA project PROMETEO/2008/051.

References

Adams, N.M. and Hand, D.J. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999. ISSN 0031-3203.

Drummond, C. and Holte, R.C. Cost Curves: An Improved Method for Visualizing Classifier Performance. *Machine Learning*, 65(1):95–130, 2006.

Elkan, C. The foundations of Cost-Sensitive learning. In Nebel, Bernhard (ed.), *17th Intl. Conference on Artificial Intelligence (IJCAI-01)*, pp. 973–978, 2001.

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

Flach, P. A., Hernández-Orallo, J., and Ferri, C. On the coherence of AUC. Technical report, 2011. URL http://users.dsic.upv.es/~flip/papers/TRCoherentAUC.pdf.

Hand, D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123, 2009. ISSN 0885-6125.

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

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