Mach Learn (2009) 77: 103–123
DOI 10.1007/s10994-009-5119-5

Measuring classifier performance: a coherent alternative to the area under the ROC curve

David J. Hand

Received: 21 August 2008 / Revised: 24 March 2009 / Accepted: 4 May 2009 /
Published online: 16 June 2009

Abstract The area under the ROC curve (*AUC*) is a very widely used measure of performance for classification and diagnostic rules. It has the appealing property of being objective, requiring no subjective input from the user. On the other hand, the *AUC* has disadvantages, some of which are well known. For example, the *AUC* can give potentially misleading results if ROC curves cross. However, the *AUC* also has a much more serious deficiency, and one which appears not to have been previously recognised. This is that it is fundamentally incoherent in terms of misclassification costs: the *AUC* uses different misclassification cost distributions for different classifiers. This means that using the *AUC* is equivalent to using different metrics to evaluate different classification rules. It is equivalent to saying that, using one classifier, misclassifying a class 1 point is p times as serious as misclassifying a class 0 point, but, using another classifier, misclassifying a class 1 point is P times as serious, where $p \neq P$. This is nonsensical because the relative severities of different kinds of misclassifications of individual points is a property of the problem, not the classifiers which happen to have been chosen. This property is explored in detail, and a simple valid alternative to the *AUC* is proposed.

Keywords ROC curves · Classification · *AUC* · Specificity · Sensitivity · Misclassification rate · Cost · Loss · Error rate

1 Introduction

A large number of problems fall into the framework of supervised classification. In such problems the aim is to construct a decision rule which will allow one to assign new objects

Editor: Johannes Fürnkranz.

D.J. Hand (✉)
Department of Mathematics, Imperial College London, London, UK
e-mail: d.j.hand@imperial.ac.uk

D.J. Hand
Institute for Mathematical Sciences, Imperial College London, London, UK

to one of a prespecified set of classes, using descriptive information about those objects. The rule is constructed from a 'training set' of data which consists of descriptive information for a sample of objects for which one also knows the true class labels.

Many approaches to constructing such classification rules have been explored, including tree classifiers, random forests, neural networks, support vector machines, nearest neighbour methods, naive Bayes methods, linear and quadratic discriminant analysis, and many others. Reviews are given in Hand (1997), Hastie et al. (2001), and Webb (2002). Since there are so many methods from which to choose, the question naturally arises as to which method is appropriate or 'best' for any particular application. This is a question which defies a simple answer, because what is best will depend on many factors. Such issues have been discussed in Hand (2006), which argues that comparative studies have often overlooked important aspects of real problems, so casting doubt on their conclusions, and Jamain and Hand (2008), which draws attention to failure to address the diversity of issues when comparing classification methods.

One of the important issues in performance evaluation is that of which criterion to choose to measure classifier performance. Once again, many such criteria are used. They include misclassification (or error) rate, the Kolmogorov-Smirnov (*KS*) statistic, likelihood ratios, the area under the ROC curve (or, equivalently, the Gini coefficient, defined below), pairs of measures such as specificity and sensitivity or precision and recall, measures of accuracy of probability estimates such as Brier or log score, and many others (see, for example, Flach 2003; Hand 1997; Pepe 2003).

Ideally, of course, one would choose a measure which properly reflected one's aims. Indeed, if the aims have been precisely specified, choosing a measure which does not reflect them could lead to incorrect conclusions, as the different measures need not lead to the same rank-order of performance of classifiers. Often, however, it is difficult to choose a measure, perhaps because the aims are not precisely specified (e.g. perhaps the future circumstances under which the classifier will be used cannot be known precisely) or because it is impossible to give precise values to parameters (e.g. the costs of misclassifications). In such cases, either arbitrary choices are made (e.g. the assumption of equal misclassification costs implicit in the definition of error rate) or aggregate measures are used, which combine measures of performance under different circumstances (e.g. the log-likelihood, which can be viewed as a mean of the log-likelihoods of each data point in the training set).

The aim of this paper is to look at one particular, very popular, such aggregate measure, to demonstrate in detail that it is based on a choice which is not merely arbitrary, but which is typically inappropriate, and to suggest a superior alternative. This measure is the *area under the ROC curve* (*AUC*) and its equivalent, the *Gini coefficient*. The widespread use of the *AUC* in assessing performance of classification rules, especially in medicine, radiology, psychology, credit scoring, and bioinformatics, as well as more generally in statistics, machine learning, data mining, and other areas, indicates that the serious shortcoming of the *AUC* described here deserves to be better known. Discussions of ROC curves and the *AUC* are given in many places, including Bradley (1997), Fawcett (2006), Hanley and McNeil (1982), Hanley (1989), and Krzanowski and Hand (2009).

The *AUC* has many merits. It is a single number derived from a classification rule, so that comparisons of classification rules can be made in a straightforward way. It is objective, requiring no choices of parameter values to be made by the user, so that different researchers would obtain the same results from the same data. And it has a number of attractive intuitive interpretations, some of which are described below. However, it also has some well-known weaknesses. For example, if ROC curves cross then it is possible that one curve has a larger

AUC (and so is apparently better) even though the alternative may show superior performance over almost the entire range of values of the classification threshold (defined below) for which the curve will be used.

In many practical applications, it is likely that the ROC curves being compared will cross. One reason for this is that comparisons are likely to be between classifiers with similar performance. In many situations, an empirical process of classifier improvement is undertaken, adjusting the classifier a small step at a time so as to gradually improve the *KS*, *AUC*, or whatever performance measure is being used. The result is a series of comparisons between similar classifiers, which are therefore likely to have similar ROC curves. When curves are similar, it is unlikely that one will dominate another—unlikely that one will have a superior sensitivity for all choices of specificity. Indeed, there is empirical evidence supporting this supposition: Provost et al. (1998) compared a variety of classifiers on ten datasets from the UCI repository and found that 'for *only one* (Vehicle) of these ten domains was there an absolute dominator' (their italics).

The risks of comparing classifiers on the basis of simple summary measures which fail to take account of the potential for ROC curves to cross are well-known. However, underlying this is a much more fundamental weakness of the *AUC* which appears not to have been previously recognised. This is that, as we show below, the *AUC* is equivalent to measuring the performance of classification rules using metrics which depend on the rules being measured. In particular, instead of regarding the relative severities of different kinds of misclassifications (i.e., misclassifying a class 0 object as class 1, and a class 1 as class 0) as the same for different classifiers, the *AUC* takes these relative severities themselves to depend on the classifier being measured. This is, of course, nonsensical, since the relative severities, even if one does not know them precisely, are independent of the classifiers themselves and must be determined by factors describing the problem external to the classifiers. It is as if one chose to compare the heights of two people using rulers in which the basic units of measurement themselves depended on the heights.

Having noted this weakness of the *AUC* as a measure of classifier performance, the paper then goes on to examine its source, and then to present an alternative measure which does not suffer from it.

The next section sets the context and defines the *AUC*. Section 3 presents a non-technical overview of the fundamental incoherency of the *AUC*, before going into the mathematical detail in Sect. 4. A key part of this incoherency lies in the relationship between misclassification costs and optimal choice of classification threshold, and this is explored in Sect. 5. Section 6 then goes on to describe the H measure, an alternative measure of performance which overcomes the intrinsic incoherence of the *AUC*. Estimating the H measure raises various peripheral issues, not central to its definition and the incoherency problem, so estimation of the H measure is discussed separately, in Sect. 7. Section 8 gives three examples, two artificial and one real, showing that the *AUC* and H measure are not monotonically related, so that a classifier which appears superior under one measure may appear inferior under the other. Finally, Sect. 9 draws some conclusions.

2 Background

This paper assumes that we have only two classes, labelled 0 and 1. A classification rule might produce an estimate $\hat{p}(1|x)$ of the probability that a point with the vector x of descriptive values belongs to class 1, or, more generally, it might simply produce a score $s = s(x)$, an unspecified monotonic increasing transformation of an estimate $\hat{p}(1|x)$. Let the probability density function of the scores $s = s(x)$ for class k points be $f_k(s)$, $k = 0, 1$, with

corresponding cumulative distribution functions (CDFs) $F_k(s)$. For purposes of exposition, we will suppose that class 0 points tend to have smaller scores than class 1 points. This has no material effect on the argument, and if it is not true for any particular problem it can be made so by interchanging the class labels. We say a little more about this below. We will take π_k to be the prior probability of class k (the 'size' of class k)—that is, the probability that a randomly drawn object, about which no further information is available, will belong to class k. It follows that $\pi_0 + \pi_1 = 1$.

A classification of a new object is obtained by comparing the score, s, of the object with a 'classification threshold' t. If $s > t$ the object is classified as coming from class 1, and if $s \leq t$ as coming from class 0.

The sensitivity of a classifier is the proportion of 'cases' (which we take to be class 0) which are correctly classified as cases. The sensitivity at a classification threshold t is thus $F_0(t)$. Similarly, the specificity of a classifier is the proportion of 'non-cases' (class 1) which are correctly classified as non-cases: $1 - F_1(t)$. As the classification threshold t changes, so different values of sensitivity and specificity are produced (in general, varying inversely with each other). The ROC curve is then a plot of $F_0(t)$ on the vertical axis against $F_1(t)$ on the horizontal axis. A classifier which perfectly separates the two classes would produce a curve which began at the lower left, with $F_0(t) = F_1(t) = 0$, and consisted of a vertical line to $(F_1(t) = 0, F_0(t) = 1)$ followed by a horizontal line to $(F_1(t) = 1, F_0(t) = 1)$. A classifier which completely failed to separate the classes would produce a curve lying on the diagonal line from $(F_1(t) = 0, F_0(t) = 0)$ to $(F_1(t) = 1, F_0(t) = 1)$. Given these properties of the ROC curve, a natural measure of the performance of the classifier producing the curve is the area under the curve—the *AUC*. This will range from 0.5 for a perfectly random classifier to 1.0 for a perfect classifier. Areas less than 0.5 are possible, but in that case areas greater than 0.5 can be obtained by changing the predictions—predicting 0 instead of 1 and vice versa.

Letting $v = F_1(s)$, from the definition of the ROC curve, we see that the area beneath it is

$$\int_0^1 F_0(F_1^{-1}(v))dv$$

and by a simple change of variable, in terms of the distribution of scores, the area under the ROC curve is

$$AUC = \int_{-\infty}^{\infty} F_0(s) f_1(s) ds. \tag{1}$$

As noted above, and as can be seen from (1), the *AUC* has the particular attraction that it does not require the user to specify any value of t. It is also clear that, given the same distributions f_0 and f_1, all researchers will obtain the same *AUC*. Furthermore, (1) can also be seen to be the probability that a randomly drawn member of class 0 will produce a score lower than the score of a randomly drawn member of class 1. This is the Mann-Whitney-Wilcoxon U statistic, and it provides a natural intuitive interpretation of the *AUC*. A variant of this interpretation is as follows: suppose we randomly choose a value from the mixture distribution of the two scores to be the classification threshold, t, and then randomly choose two scores, one from each class, following the class score distributions $f_k(s)$, with the restriction that the mean of the two scores equals t. Then the probability of correctly classifying both scores (i.e. that the class 0 score is less than t and the class 1 score is greater than t) is given by the *AUC*.

Yet another natural interpretation is that the *AUC* is the average sensitivity, regarding all values of the specificity as equally likely: $AUC = \int F_0(s) dF_1(s)$. This sort of interpretation

has also led to modified versions of the *AUC* which recognise that perhaps not *all* values of specificity (or sensitivity) will be regarded as of equal interest or relevance, and so restrict the range of the integration (e.g. McClish 1989; Dodd and Pepe 2003). Another variant arises in screening applications, where one might choose to accept for further investigation a specified proportion p of the overall population. If one regards, *a priori*, each proportion of the population as equally likely to be chosen as the proportion to be accepted, then it follows that the average sensitivity is $\pi_0/2 + \pi_1 AUC$.

Other interpretations which might be useful in any particular practical application are also possible.

A chance-standardised variant of the *AUC* is also in widespread use, taking values between 0 (no difference between the score distributions of the two classes) and 1 (complete separation between the two distributions). This is the *Gini coefficient*, G, defined as twice the area between the ROC curve and the chance diagonal: $G = 2AUC - 1$.

3 Incoherency of the *AUC*: an outline

The aim of this section is to outline the cause of the incoherency of the *AUC*, before exploring it rigorously in the two following sections.

In this paper, we suppose that correct classifications incur no cost, and that the two different kinds of misclassifications (misclassifying a class 0 point as class 1, and misclassifying a class 1 point as class 0) incur potentially different costs. Choosing a classification threshold t typically results in some objects from each class being misclassified, so that an overall loss is made. For a given pair of misclassification costs, one can choose the threshold $t = T$ to yield minimum overall loss.

This is fine if one knows what the two misclassification costs will be when the classifier is actually used, or, at least, their ratio, but in most problems one does not know these. On the other hand, one often has some idea about the *likely values* of the ratio of the misclassification costs: one might, for example, believe that misclassifying a fraudulent credit card transaction as legitimate will be regarded as more serious than the reverse. Generalising this, one can try to construct a distribution showing how likely one thinks are the different values of the misclassification cost ratio. Since each value of the cost ratio corresponds to a value of the optimal classification threshold T (i.e. that threshold which minimises the overall loss), this leads to a corresponding distribution of the classification threshold. Now, since each choice of T leads to an overall minimum misclassification loss for those costs, we can integrate the loss, weighted by the chosen distribution of T, to give an overall measure of classification performance. In fact, it turns out that, for a particular choice of distribution for the classification threshold, this gives the *AUC*. In particular, we obtain the *AUC* when the chosen weighting distribution of T is the mixture distribution of the scores from the two classes.

Since the distribution over T corresponds to a distribution over the cost ratios, one implication of this is that *the AUC is equivalent to averaging the misclassification loss over a cost ratio distribution which depends on the score distributions*. Since the score distributions depend on the classifier, this means that, when evaluating classifier performance, the *AUC* evaluates a classifier using a metric which depends on the classifier itself. That is, *the AUC evaluates different classifiers using different metrics*. It is in that sense that the *AUC* is an incoherent measure of classifier performance.

This problem is a deep one. It does not hinge on concavity or convexity of the ROC curve. It is clearly an important one, since it says that the order of merit of classifiers produced by the *AUC* in any comparative study is based on measurement procedures which are

different for each classifier being compared. This contravenes the fundamental principle of comparison, namely that when things are compared, one should compare them using the same measure (we do not compare the height of one person with the weight of another, and on obtaining a larger number for the first then say that therefore the first person is 'larger').

In an ideal world, for any particular problem, researchers would specify the weighting distribution for the costs used in the integration. In practice, however, such distributions can seldom be specified. In view of this, it seems sensible to propose a standard which can be used as the default. Such a proposal is made in Sect. 6, leading to a new measure, the H measure. The practical implementation of the H measure requires some discussion of issues unrelated to the core concept—such as how to handle non-convex curves, so estimation is described separately, in Sect. 7. Of course, if a researcher does have opinions about the appropriate weight distribution for the costs, then this should be used—although we recommend that the standard H measure is also reported so that other researchers can make comparisons.

Section 8 presents some empirical results, showing that the *AUC* and H measure can rank classifiers differently. When they do produce different rankings, it would be foolish to adopt the *AUC* ordering, because of the different metrics implicit in this measure.

R software to calculate the H measure is available from the author's personal website (see Sect. 9).

4 The *AUC* as an averaged minimum loss measure

As noted above, we suppose that correct classifications incur no cost, and that the cost of misclassifying a class k point is $c_k \in [0, \infty]$, $k = 0, 1$. Note, in particular, that this means that the cost arising from misclassifying an object does not depend on how far its score s is from the threshold t, but only on whether it is greater than or less than t. This is not an unreasonable assumption: very often the reason for making the classification is in order to take some action, and the choice of action will simply depend on whether the score is above or below t. Hand (2005) discusses this point and its implications in more detail.

As we show below, choosing the score threshold value t is equivalent to specifying what one believes are the relative costs of misclassifying a class 0 object as class 1 compared with misclassifying a class 1 object as class 0. Now, such assessment of relative costs has to come from the context of the problem, and is typically extremely difficult to determine. This is true even of areas where one might have expected it to be straightforward, such as financial services, as well as areas where it clearly might be expected to be difficult, such as medicine.

Because of this difficulty, two choices are especially popular: (i) Taking the relative costs to be equal: this choice leads to the classifier's misclassification rate (or error rate) being the performance metric; (ii) Taking the cost of misclassifying a class k point to be inversely proportional to π_k: this choice leads to the Kolmogorov-Smirnov statistic being the performance metric. I have argued elsewhere (e.g. Hand 2005, 2006) that these choices are almost certainly inappropriate, precisely because they are made not on the basis of consideration of the relative severity of misclassifications in the presenting problem, but simply on grounds of convenience.

With classification threshold t, and prior class probabilities as above, the overall misclassification loss is

$$c_0\pi_0(1 - F_0(t)) + c_1\pi_1 F_1(t).$$

The value of t which minimises this is $T(c_0, c_1)$ given by

$$T(c_0, c_1) \triangleq \arg\min_t \{c_0\pi_0(1 - F_0(t)) + c_1\pi_1 F_1(t)\}.$$

It is clear that this minimising value of t will be the same for cost pairs (c_0, c_1) and $(C_0, C_1) = (Kc_0, Kc_1)$, where K is an arbitrary positive constant: that is, the optimal threshold depends only on the ratio of the costs, and not on their absolute value. For this reason it is convenient to transform the pair (c_0, c_1) to the pair (b, c), defined by $b = (c_0 + c_1)$ and $c = c_0/(c_0 + c_1)$, so that only c depends on the ratio of the costs: $c = (1 + c_1/c_0)^{-1}$. We can then simplify the argument of T to write

$$T(c) = \arg\min_t \{c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)\}, \tag{2}$$

and we can write the loss for arbitrary choice of t as

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

If the score distributions are differentiable, we can find T by differentiating (3), leading to the minimising T satisfying

$$c\pi_0 f_0(T) = (1 - c)\pi_1 f_1(T) \tag{4}$$

and $d^2Q/dt^2 > 0$. There may, of course, be more than one value of t satisfying these conditions. Such multiple values of t arise if the ROC curve has concave regions. If the ROC curve is everywhere convex, then the minimising t is unique. [At this point, it is useful to make a brief parenthetical comment on the words 'convex' and 'concave', since different intellectual communities use these terms in different ways. In particular, mathematicians define a function g to be *convex* if $g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y)$ for $0 < \lambda < 1$ (see, for example, Rudin 1964, p. 88), and *concave* if $g(\lambda x + (1 - \lambda)y) \geq \lambda g(x) + (1 - \lambda)g(y)$. In contrast, the machine learning community typically defines these terms the other way round, at least as far as ROC curve analysis goes. The genesis of this usage in the machine learning community is not simply contrariness, but derives from the notion of the convex hull *enclosing* the points beneath the ROC curve. Since this paper is targeted at the machine learning community, it will adopt the machine learning usage.]

This convexity condition is equivalent to requiring the gradient of the ROC curve to be monotonically decreasing. On the other hand, if the ROC curve has discontinuities in the (monotonically decreasing) gradient, the threshold value at such a discontinuity will be associated with a range of values of c. These issues are explored in detail below.

To gain insight into the shortcomings of the *AUC*, it is convenient first to examine the simple case in which the relationship (2) between T and c is one-to-one, so we make this assumption in this section, relaxing it in the next. Under the assumption of a one-to-one relationship, (4) leads to

$$c = P(1|T) = \pi_1 f_1(T)/\{\pi_0 f_0(T) + \pi_1 f_1(T)\} \tag{5}$$

relating a given cost ratio c and the optimising classification threshold T, where $P(1|T)$ is the conditional probability of belonging to class 1, given the score T. It is convenient to write $P_1(T) = P(1|T)$, and then the one-to-one assumption means that $c = P_1(T)$ is invertible. Given the relative misclassification costs in terms of c, we can use $T = P_1^{-1}(c)$ to give the appropriate classification threshold.

The prior probabilities π_0 and $\pi_1 = 1 - \pi_0$ can be estimated from the training data or some other source (perhaps the training data have been deliberately undersampled because of unbalanced class sizes). In any case, these parameters are properties of the distributions defining the problem, and are subject to empirical investigation. The values of the misclassification costs c_0 and c_1, and hence of b and c, are, however, another matter. Since they represent the severities of the different kinds of misclassification they are not things which can be discerned by looking at the score distributions. Rather, as we saw above, their values must come from outside the mathematics—from the context of the problem—and they are typically very difficult to choose. Indeed, in many applications they are likely to vary from time to time (e.g. in making bank loan decisions, where the relative costs, and hence the classification threshold, is likely to depend on future economic conditions). However, even if one is unable to specify the pair (c_0, c_1) precisely, one may be able to say something about their likely values. For example, one might feel that misdiagnosing, as healthy, someone suffering from a potentially fatal disease which can be easily treated by a harmless medicine, is (or will be regarded as) more serious than the reverse, so that $c > 1/2$ (taking class 0 as the diseased class) and possibly $c \gg 1/2$. Or that misclassifying a fraudulent bank transaction (class 0) as legitimate is more serious than the reverse, so that again $c > 1/2$. In terms of b and c, we denote a subjective distribution of likely values of the unknown pair (b, c) by $v(b, c)$. In the rare cases when one can give precise misclassification costs, v will be a delta function.

The overall expected minimum loss is then

$$\begin{aligned} L &= \int_0^1 \int_0^\infty Q(T(c); b, c) v(b, c) db dc \\ &= \int_0^1 \int_0^\infty \{c\pi_0(1 - F_0(T(c))) + (1 - c)\pi_1 F_1(T(c))\} b v(b, c) db dc \\ &= \int_0^1 \{c\pi_0(1 - F_0(T(c))) + (1 - c)\pi_1 F_1(T(c))\} w(c) dc, \end{aligned} \tag{6}$$

where $w(c) = \int b v(b, c) db$ and we have used (3). $w(c)$ serves as a weight function over the losses associated with different values of c (equivalently, over different cost ratios) when calculating the overall expected minimum loss.

Still under the one-to-one assumption, we can change the variable of the integral in (6) from c to T. Thus

$$L = \int_{-\infty}^{\infty} \{c(T)\pi_0(1 - F_0(T)) + (1 - c(T))\pi_1 F_1(T)\} W(T) dT, \tag{7}$$

where the function W includes the Jacobian of the transformation.

The classification threshold T is the threshold which minimises the loss for a particular value of misclassification cost c, and the expression in parentheses in (7) corresponds to the loss when this value of T, or equivalently, this value of c, is used. In (7), then, this loss is thus weighted by $W(T)$ and integrated over the range of T. $W(T)$ can thus also be regarded as reflecting the user's beliefs about the likely values and importance of c, and by the transformation $T = P_1^{-1}(c)$, the likely values and importance of T. It follows that, as noted above, $W(T)$ must be based on the extra-mathematical context of the problem.

Now let us consider the particular choice

$$W(T) = W_G(T) \triangleq \pi_0 f_0(T) + \pi_1 f_1(T). \tag{8}$$

Plugging $W_G(T)$ into L in (7), and using $c(T) = \pi_1 f_1(T)/\{\pi_0 f_0(T) + \pi_1 f_1(T)\}$ from (5), we obtain

$$L_G = \int_{-\infty}^{\infty} \{\pi_0\pi_1\{f_1(T)(1 - F_0(T)) + f_0(T)F_1(T)\}\}dT,$$

which gives

$$\begin{aligned}
L_G &= \pi_0\pi_1\left\{\int_{-\infty}^{\infty}\int_{T}^{\infty} f_1(T)f_0(s)dsdT + \int_{-\infty}^{\infty}\int_{-\infty}^{s} f_1(T)f_0(s)dTds\right\} \\
&= \pi_0\pi_1\left\{\int_{-\infty}^{\infty}\int_{T}^{\infty} f_1(T)f_0(s)dsdT + \int_{-\infty}^{\infty}\int_{T}^{\infty} f_1(T)f_0(s)dsdT\right\} \\
&= 2\pi_0\pi_1\int_{-\infty}^{\infty}\int_{T}^{\infty} f_1(T)f_0(s)dsdT \\
&= 2\pi_0\pi_1\left\{1 - \int_{-\infty}^{\infty}\int_{-\infty}^{T} f_0(s)dsf_1(T)dT\right\} \\
&= 2\pi_0\pi_1\{1 - AUC\}.
\end{aligned}$$

This is just a linear transformation of the *AUC* (and hence also of the Gini coefficient). What this means is that using the *AUC* or Gini coefficient to compare classifiers is equivalent to taking an average of the losses at different classification thresholds, using the distribution $W_G(T)$ as a weighting function. It follows that, in terms of c, the *AUC* is equivalent to taking an average of the losses corresponding to different cost ratios c, where the average is calculated according to the distribution

$$w(c) = w_G(c) \triangleq \pi_0 f_0(P_1^{-1}(c))\left|\frac{dP_1^{-1}(c)}{dc}\right| + \pi_1 f_1(P_1^{-1}(c))\left|\frac{dP_1^{-1}(c)}{dc}\right|. \tag{9}$$

The implication of this is that the weight distribution over cost ratios c, implicitly used in calculating the *AUC*, depends on the empirical score distributions f_k. That is, *the weight distribution used to combine different cost ratios c, will vary from classifier to classifier*. But this is absurd. The beliefs about likely values of c must be obtained from considerations separate from the data: they are part of the problem definition. One cannot change one's mind about how important one regards a misclassification according to which tool one uses to make that classification. Nevertheless, this is effectively what the *AUC* does—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, and decided who was the taller by merely comparing the numbers, ignoring the fact that different units of measurement had been used (see Hand 2004, for further discussion of such measurement scale issues).

5 Transformation between cost ratio and threshold

So as to gain intuitive insight, the exposition has so far been under the assumption that the relationship between the cost ratio given by c and the threshold T which minimised the overall loss when c was used, was one-to-one. This assumption can break down in two ways.

The first kind of breakdown arises when the ROC curve has a discontinuity in its first derivative, since then the threshold value T at the discontinuity will correspond to a range of

c values. This will turn out to be important when we come to map the scores to costs based on an empirical ROC curve, since (unsmoothed) empirical ROC curves typically have many discontinuities in their first derivatives. This is because such curves are estimated directly from the empirical CDFs, and so consist of a series of vertical and horizontal line segments (and diagonals where there are ties, as explained below). In place of unique values of the cost c associated with each value of T, whenever there is a discontinuity of this kind we have a range of values of c which has to be integrated over.

The second breakdown occurs when the ROC curve is not everywhere convex. In this case there will be (concave) regions of the curve in which the values of the classification threshold t do not minimise the loss for any choice of c. That is, such choices of t lead to larger losses than some other choice for *all* possible values of c. We can overcome this by using (2) to define an upper convex hull for the ROC curve (defined as the convex function of minimum sensitivity which bounds the ROC curve from above), and then using this hull in place of the ROC curve, as described below. This replaces the suboptimal choice of t by an optimal choice. This convex hull is defined as follows.

If a particular value of t on the ROC curve does not minimise $Q(t;c) = c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)$ for any value of c, then this t will lie in an interval of values, only the end points of which minimise Q for some value of c. Let T_U and T_L represent the upper and lower end points of such an interval on the ROC curve. These points both lie on the ROC curve and define a straight line segment of the upper convex hull which lies above the ROC curve. Clearly the value of c for which T_U is the minimising threshold is the same as the value of c for which T_L is the minimising threshold. Thus, for any $T \in [T_L, T_U]$, $c(T)$ is given by solving $Q(T_U;c) = Q(T_L;c)$.

This gives an upper convex hull defined by

(i) $\{F_1(t), F_0(t)\}$ for all t values satisfying

$$T = \arg\min_T \{c\pi_0(1-F_0(T)) + (1-c)\pi_1 F_1(T)\},$$

for some c as c ranges over the interval $[0,1]$.

(ii) The set of $\{F_1(t), F_0(t)\}$ points on the line intervals connecting $\{F_1(T_U), F_0(T_U)\}$ to $\{F_1(T_L), F_0(T_L)\}$, for those values of c such that

$$c = \frac{\pi_1(F_1(T_L) - F_1(T_U))}{\pi_0(F_0(T_L) - F_0(T_U)) + \pi_1(F_1(T_L) - F_1(T_U))}. \tag{10}$$

The piecewise linear nature of the convex hull means that, in practice, instead of the smooth transformation given by (5) for the idealised case in which $c = P(1|T)$ is invertible, ranges of values of t map to discrete values of c. For example, all those values of t lying in an interval $\{F_1(T_U), F_0(T_U)\}$ to $\{F_1(T_L), F_0(T_L)\}$ will map to c given in (10). Thus, in place of a continuous weighting distribution $w(c)$, we have a discrete distribution—and one which depends on the score distributions and the resulting convex hull. The discrete nature of this distribution is another example of the inappropriateness of the standard *AUC* approach, since it will be a very rare problem for which one's subjective beliefs about the likely values of c form a discrete distribution.

So far in this section the discussion has focused on how real ROC curves will depart from the idealised ROC curve which is everywhere differentiable with a negative second derivative. However, the notion of the convex hull also has deeper implications.

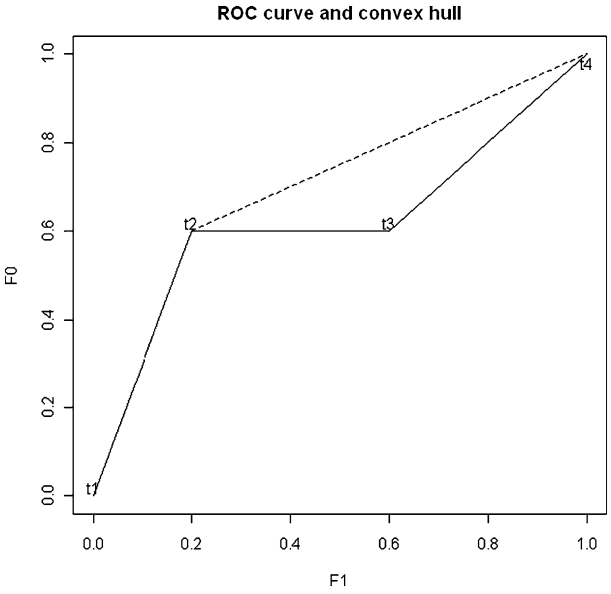


Fig. 1 The ROC curve and the convex hull

The fact that choices of the classification threshold t in concave intervals of the ROC curve do not minimise the loss for any choice of c means that one can produce a classifier superior to that summarised by such a ROC curve (Provost and Fawcett 1997; Scott et al. 1998; Fawcett 2004). For example, in Fig. 1, the continuous line shows a ROC curve with a concave region. A point on the curve in this region, such as that indicated by the classification threshold t_3, will yield an overall loss of

$$c\pi_0(1 - F_0(t_3)) + (1 - c)\pi_1 F_1(t_3). \tag{11}$$

However, for all values of c, at least one of the thresholds t_1, t_2, or t_4 will produce a loss smaller than this. In fact, when c is such that

$$c\pi_0(1 - F_0(t_1)) + (1 - c)\pi_1 F_1(t_1) = c\pi_0(1 - F_0(t_2)) + (1 - c)\pi_1 F_1(t_2)$$

any choice of threshold on the ROC curve line segment (t_1, t_2) will yield the same loss, smaller than (11), so one has a choice of (sensitivity, specificity) pairs, all giving the same overall loss. A similar point applies to the interval (t_2, t_4), where all (specificity, sensitivity) pairs on the broken line segment (t_2, t_4) will yield the same loss. Since this line segment does not lie on the ROC curve, to obtain a (specificity, sensitivity) pair on it, we cannot simply choose a particular threshold value. Instead it is necessary to randomly choose thresholds t_2 or t_4 in respective proportions p and $1 - p$. When $p = 1$ we are at the (specificity, sensitivity) pair corresponding to threshold t_2 and when $p = 0$ we are at the (specificity, sensitivity) pair corresponding to threshold t_4. When $p = 1/2$ we are at the (specificity, sensitivity) pair corresponding to the midpoint of the line segment (t_2, t_4). By this means a classifier superior to the original one is produced. As well as having a smaller loss for all values of c, its superiority is reflected by it having a larger *AUC*—the area under the convex hull—whereas the original classifier has an *AUC* given by the area under the ROC curve itself. We denote the area under the convex hull of a ROC curve by *AUCH*. Of course, $AUCH \geq AUC$.

We can also imagine the convex hull of a ROC curve being mapped back to two score distributions. This will be important later on. The mapping is not unique, since ROC curves

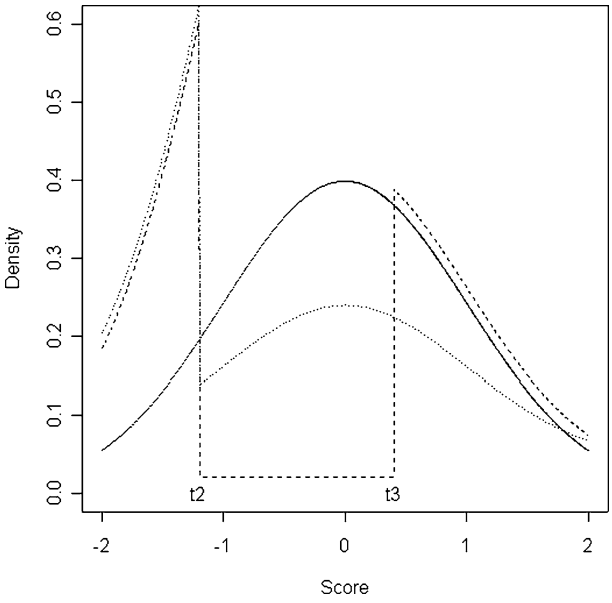


Fig. 2 Distributions of scores yielding the ROC curve and convex hull of Fig. 1

preserve only ordinal relationships. Such a mapping, corresponding to the ROC curve in Fig. 1, is illustrated in Fig. 2, where the plots of the distributions have been rescaled so that their shapes can be conveniently compared. The continuous line in Fig. 2 shows the distribution for class 1, which we have taken to be a standard normal distribution. The broken line, with a mode at t_2 and having zero value between t_2 and t_3 is the distribution for class 0. These two distributions produce the ROC curve shown in Fig. 1. The dotted line in Fig. 2 shows an alternative distribution for class 0 which, with the same class 1 distribution, yields the convex hull ROC curve in Fig. 1.

In summary, if the ROC curve of a classifier has any concave regions, then superior classifiers can easily be produced.

6 An alternative measure of performance

We saw, in Sect. 4, that the *AUC* implicitly uses a weight function $W_G(T)$, corresponding to a $w_G(c)$ function which varies from classifier to classifier. That is, we saw that using the *AUC* is equivalent to measuring classifier performance using an instrument which varies from classifier to classifier. Hand and Till (2001) have previously pointed out that the weight function implicitly used in calculating the *AUC* in terms of the classification threshold is the mixture distribution of the scores for the two classes, as defined in (8), and Hand (2005) explored the consequences of the fact that this meant that the *AUC* depended on the empirical data. He suggested replacing the mixture by an objective measure $W(s)$, independent of the f_k. However, even if the same weight function is adopted for the *scores* from different classifiers, then, because the scores are related to the costs via the empirical distributions, this does not completely solve the problem of using different distributions over the costs. That is, Hand (2005) tackles the arbitrariness of using the mixture distribution, but does not go far enough in mapping things back to the misclassification costs. At the other extreme, Adams and Hand (1999), in considering how to compare classifiers, focused attention directly on the costs, ignoring the relationship between costs and scores.

So that classifiers are compared using the same metric, we must choose a function $w(c)$ which does not depend on the score distributions. Ideally, since $w(c) = \int bv(b, c)db$, it will be chosen using expert knowledge about likely values of b and c (or, equivalently, of c_0 and c_1). However, that leaves open the possibility that different users would choose different $w(c)$ functions. This is equivalent to the choice of different priors in Bayesian analysis, and may not be straightforward. Recall also that one of the attractions of the *AUC* is that it is objective, in the sense that different researchers would obtain the same result on the same data sets. With this in mind, on the grounds of simplicity and objectivity it is therefore desirable to define some function $w(c)$ which can be used as a universal standard.

If b and c are independent, $w(c)$ simplifies to $w(c) = u(c)E(b)$, a function proportional to the marginal distribution of c. Without loss of generality (merely by changing the units in which b is measured), we can set $w(c) = u(c)$.

A simple and obvious choice would be to let $u(c)$ be a uniform distribution, so that $u(c) = 1$ for $c \in [0, 1]$ and 0 otherwise. This satisfies both desiderata of objectivity (everyone would obtain the same results from the same data) and of being a metric independent of the empirical score distributions (so that the same measurement scale is being used regardless of the score distributions). However, although the uniform choice satisfies these criteria, it might be regarded as unappealing on other grounds. In particular, it weights the very extreme values of c and the moderate values equally. For example, it treats a c value of $1/2$ as if it were as likely as c values of 10^9 and 10^{-9}. It seems unlikely that such a range of equally probable values would be contemplated in many real problems; a c distribution which decays towards the more extreme values might be regarded as more satisfying.

A simple weight function satisfying this is the Beta distribution, with form

$$u(c) = u_{\alpha,\beta}(c) \triangleq beta(c; \alpha, \beta) = c^{\alpha-1}(1-c)^{\beta-1}/B(1; \alpha, \beta), \tag{12}$$

with α, β restricted to be greater than 1, and where $B(x; \alpha, \beta) = \int_0^x c^{\alpha-1}(1-c)^{\beta-1}dc$ is an incomplete beta function normalising constant. This leads to the general loss

$$L_{\alpha,\beta} = \int Q(T(c); b, c)u_{\alpha,\beta}(c)dc.$$

If one believes that misclassifying class 1 points is likely to be more serious than misclassifying class 0 points, so that $c_1 > c_0$, so that c is likely to be less than $1/2$, then α and β should be chosen so that $u(c)$ is larger for values of c less than $1/2$. Suitable parameter values are $\alpha = 2$ and $\beta = 4$, which yields a unimodal distribution with mode at $(\alpha - 1)/(\alpha + \beta - 2) = 1/4$, and which places most probability between 0 and 0.5. In contrast, if one believes that misclassifying class 0 points is more serious, suitable parameters are $\alpha = 4$ and $\beta = 2$. Such asymmetric distributions hinge on the researcher deciding which type of misclassification is the more serious, and it is possible (though, I think, unlikely for most problems) that different researchers might have different opinions on this. With this in mind, and so that we have a criterion which requires no decisions from the user, we propose as the basic standard default a symmetric beta distribution with $\alpha = \beta$. Without additional knowledge of the likely values of c, there seems no way to choose between alternative such symmetric distributions. We therefore arbitrarily propose the use of $\alpha = \beta = 2$ as the default values, yielding $beta(x; 2, 2)$ as the default weight distribution. While this does have an arbitrary aspect, its general use will mean that different researchers will be using the same measure, so that they can legitimately compare classifiers—unlike the *AUC*. Once again we note that, if a researcher does have particular knowledge about the shape of the $u(c)$ distribution, then this should be used—but, even so, we recommend also reporting the H measure based on $beta(x; 2, 2)$ so that other researchers can make comparable statements.

The maximum values that the *AUC* and Gini coefficient can take are both 1, corresponding to score distributions for which there exists a threshold which yields perfect separation between the sets of scores for the class 0 and class 1 training data. The minimum of the Gini coefficient of 0 corresponds to identical score distributions. On grounds of consistency, it would be satisfying for our index also to take larger values for superior performance, and to range between 0 for identical score distributions (worst case) and 1 for perfect classification (best case).

For general u, in the worst case (when the class score distributions are identical, so that the ROC curve is diagonal), this leads to the maximum loss, L_{Max}, of

$$L_{Max} = \pi_0 \int_0^{\pi_1} cu(c)dc + \pi_1 \int_{\pi_1}^1 (1-c)u(c)dc.$$

At the other extreme, if it is possible to choose a threshold yielding perfect classification, we obtain a minimum loss of 0.

Standarising for the maximum, and subtracting from 1 so that large values correspond to good performance, we obtain the general measure

$$H = 1 - \frac{L}{L_{Max}} = 1 - \frac{\int Q(T(c); b, c)u(c)dc}{\pi_0 \int_0^{\pi_1} cu(c)dc + \pi_1 \int_{\pi_1}^1 (1-c)u(c)dc}$$

and, for the particular case of $u(c) = u_{\alpha,\beta}(c)$

$$H = 1 - \frac{L_{\alpha,\beta}}{L_{Max}} = 1 - \frac{\int Q(T(c); b, c)u_{\alpha,\beta}(c)dc}{\pi_0 \int_0^{\pi_1} cu_{\alpha,\beta}(c)dc + \pi_1 \int_{\pi_1}^1 (1-c)u_{\alpha,\beta}(c)dc}. \tag{13}$$

It 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.

7 Estimating the H measure

The empirical ROC curve is often defined by plotting the empirical CDF of class 0 on the vertical axis against the empirical CDF of class 1 on the horizontal axis. The result is a sequence of line segments, vertical ones of length $1/n_0$ and horizontal ones of length $1/n_1$. If tied scores belong to more than one class, then there is ambiguity about the empirical ROC curve, because there is no natural ordering to the tied scores. This ambiguity reflects itself in uncertainty about the values of both the *AUC* and L. To overcome it, the multiple segments in the ROC curve corresponding to the separate tied entities should be transformed into a single diagonal segment. For example, if one class 1 object ties in score with three class 0 objects, the empirical ROC curve should show a single diagonal segment corresponding to a vertical step of $3/n_0$ and a horizontal step of $1/n_1$.

A simple search strategy can be used to construct the upper convex hull. This strategy can be accelerated, but this will probably be unnecessary for most practical problems. The simple strategy is as follows.

First, construct the ROC curve. To do this, rank the scores of training points from both classes combined, and order the class labels in the same way. Let i index the different score

values, so $i = 1, \ldots, S$, where S is the number of unique score values. (If there are no ties, then $S = n_0 + n_1$.) Let σ_{0i} be the number of class 0 points with the ith score value. This will be 0 if the ith score value is taken only by a class 1 point or points. Let σ_{1i} be the number of class 1 points with the ith score value. Let $(r_{10}, r_{00}) = (0, 0)$ be the starting coordinates of the ROC curve and define

$$(r_{1i}, r_{0i}) = (r_{1(i-1)}, r_{0(i-1)}) + (\sigma_{1i}/n_1, \sigma_{0i}/n_0), \quad i = 1, \ldots, S.$$

The empirical ROC curve is then given by the sequence of straight line segments connecting (r_{10}, r_{00}) to (r_{1S}, r_{0S}). Note that $(r_{1S}, r_{0S}) = (1, 1)$.

To construct the convex hull, begin with point (r_{10}, r_{00}), and consider all points (r_{1i}, r_{0i}), $i = 1, \ldots, S$. The segment of the upper convex hull going through (r_{10}, r_{00}) is the straight line interval which passes through (r_{10}, r_{00}) and (r_{1j}, r_{0j}) where j is the value of $i \in \{1, \ldots, S\}$ which leads to the minimum value of

$$c = \frac{\pi_1(r_{1i} - r_{10})}{\pi_0(r_{0i} - r_{00}) + \pi_1(r_{1i} - r_{10})}.$$

The next segment of the convex hull is then the straight interval which passes through (r_{1j}, r_{0j}) and (r_{1k}, r_{0k}), where k is the value of $i \in \{(j+1), \ldots, S\}$ which minimises

$$c = \frac{\pi_1(r_{1i} - r_{1j})}{\pi_0(r_{0i} - r_{0j}) + \pi_1(r_{1i} - r_{1j})}.$$

This is repeated: having identified some point (r_{1j}, r_{0j}) as defining the end of a segment, the next segment begins at (r_{1j}, r_{0j}) and ends at (r_{1k}, r_{0k}), where k is the value of $i \in \{(j+1), \ldots, S\}$ which minimises

$$c = \frac{\pi_1(r_{1i} - r_{1j})}{\pi_0(r_{0i} - r_{0j}) + \pi_1(r_{1i} - r_{1j})}.$$

Let m be the number of such segments in this upper convex hull. (In fact, all of these calculations can be simplified somewhat because

$$\frac{\pi_1 X}{\pi_0 Y + \pi_1 X} > \frac{\pi_1 U}{\pi_0 V + \pi_1 U} \quad \Longleftrightarrow \quad \frac{X}{Y + X} > \frac{U}{V + U}$$

so that the π_i can be dropped.)

To evaluate $\hat{L}$ from this upper convex hull, consider the line segment of the hull corresponding to the end points (r_{1j}, r_{0j}) and $(r_{1(j+1)}, r_{0(j+1)})$. Denote the score corresponding to point (r_{1j}, r_{0j}) by s_j and that corresponding to $(r_{1(j+1)}, r_{0(j+1)})$ by s_{j+1}. Then if any score $s \in [s_j, s_{j+1}]$ is selected as a threshold, it will minimise the loss for c satisfying

$$\{c\pi_0(1 - r_{0j}) + (1 - c)\pi_1 r_{1j}\} = \{c\pi_0(1 - r_{0(j+1)}) + (1 - c)\pi_1 r_{1(j+1)}\}.$$

It is convenient if we index this c using the index corresponding to the upper end of the segment, so that

$$c_{(j+1)} = \frac{\pi_1(r_{1(j+1)} - r_{1j})}{\pi_0(r_{0(j+1)} - r_{0j}) + \pi_1(r_{1(j+1)} - r_{1j})}. \tag{14}$$

Also define $c_{(0)} = 0$ and $c_{(m+1)} = 1$.

Then

$$\hat{L} = \sum_{i=0}^{m} \int_{c_{(i)}}^{c_{(i+1)}} \{c\pi_0(1-r_{0i}) + (1-c)\pi_1 r_{1i}\} u(c)dc$$

$$= \sum_{i=0}^{m} \left\{ \pi_0(1-r_{0i}) \int_{c_{(i)}}^{c_{(i+1)}} cu(c)dc + \pi_1 r_{1i} \int_{c_{(i)}}^{c_{(i+1)}} (1-c)u(c)dc \right\}. \quad (15)$$

Using the form in (12), we obtain

$$\hat{L}_\beta = \sum_{i=0}^{m} \{\pi_0(1-r_{0i})\{B(c_{(i+1)}; 1+\alpha, \beta) - B(c_{(i)}; 1+\alpha, \beta)\}/B(1; \alpha, \beta)$$

$$+ \pi_1 r_{1i}\{B(c_{(i+1)}; \alpha, 1+\beta) - B(c_{(i)}; \alpha, 1+\beta)\}/B(1; \alpha, \beta)\} \quad (16)$$

so that

$$\hat{H} = 1 - \frac{\hat{L}_\beta B(1; \alpha, \beta)}{\{\pi_0 B(\pi_1; 1+\alpha, \beta) + \pi_1 B(1; \alpha, 1+\beta) - \pi_1 B(\pi_1; \alpha, 1+\beta)\}}. \quad (17)$$

8 Examples

Example 1 Figure 3 shows three ROC curves (and the diagonal in each case). That in (a) has a concave region, and the convex hull is indicated by the broken line. The *AUC*, Gini index G, *AUCH*, and H values are shown in Table 1. The first striking thing is that the *AUC* and H are not simply monotonically related; whereas the *AUC* is (substantially) smaller in (a) than in (b) and (c), H is (again, substantially) larger. The *AUCH* row enables us to see that not all of this difference is attributable to the concave region of the ROC curve in (a). In particular, although the *AUCH* values are the same for all three curves, the H values are all different.

Figure 4 presents the weight functions $w(c)$ implicitly used by the AUC for the two ROC curves shown in Figs. 3(b) and (c). These functions are discrete, which is unlikely to be appropriate for most real situations. Furthermore, they are strikingly different. For example, Fig. 4(a) puts a weight of about 0.7 on $c = 0.5$ while Fig. 4(b) puts a weight of 0 on this value. That means that if the classifier producing the ROC curve in Fig. 3(b) is used then

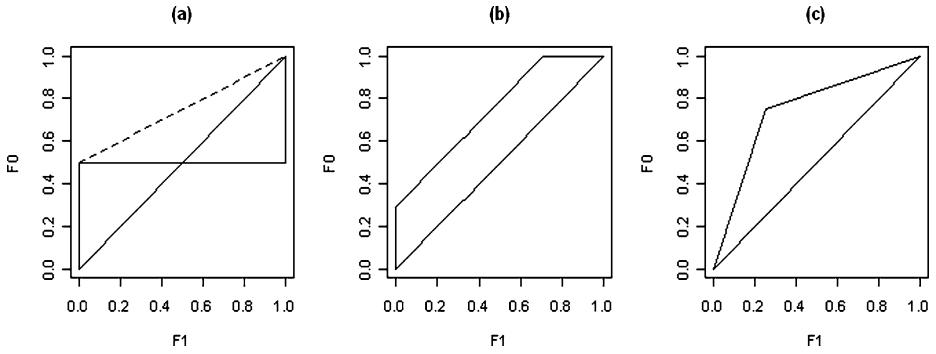


Fig. 3 Examples of three ROC curves

Table 1 Performance measures for the ROC curves in Fig. 3

	(a)	(b)	(c)
AUC	0.5	0.75	0.75
G	0	0.5	0.5
AUCH	0.75	0.75	0.75
H	0.348	0.293	0.288

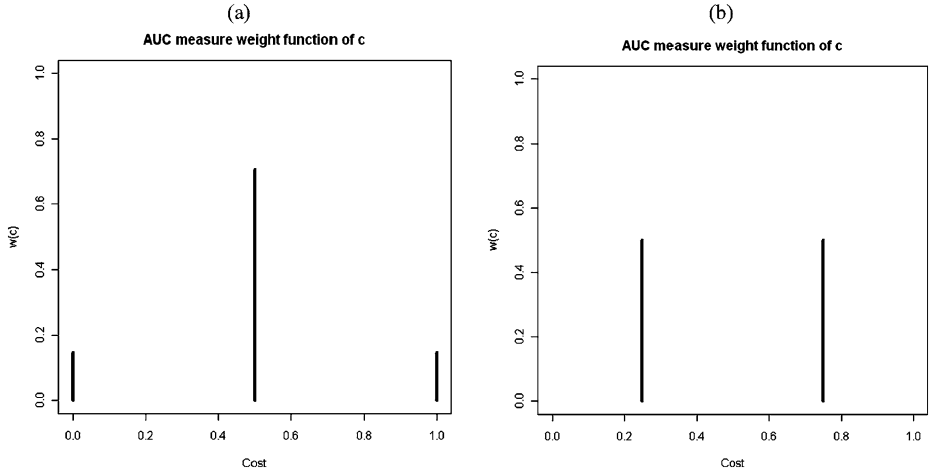


Fig. 4 The $w(c)$ functions corresponding to the ROC curves in Figs. 3(b) and (c)

one believes that there is a probability of about 0.7 that the two types of misclassification will be regarded as equally serious ($c = 0.5$), but if the classifier producing the ROC curve in Fig. 3(c) is used then one believes that there is no probability at all that the two types of misclassification will be regarded as equally serious. This is absurd: the relative probability that the two types of misclassification will be regarded as equally serious cannot depend on the choice of classifier!

Example 2 Consider two classifiers, in the first of which (logistic regression, say) using the *AUC* implies one is using a $w(c)$ function which puts a probability of 0.8 at $c = 0.1$, and in the second of which (a tree classifier, say) using the *AUC* implies one is using a $w(c)$ function which puts a probability of 0.2 at $c = 0.1$. This would mean that if one used logistic regression one would believe that there was a probability of 0.8 that misclassifying a class 1 point was 9 times as serious as the reverse, but that if one instead used a tree classifier one would believe that there was a probability of 0.2 that misclassifying a class 1 point was 9 times as serious as the reverse. One's beliefs about the distribution of probabilities over the possible values of the relative severities of the two kinds of misclassification cannot depend on which classifier one happens to use. But this is exactly what happens if one uses the *AUC*.

Example 3 Thomas et al. (2002) provide a small data set describing a number of bank customers, along with a good/bad outcome indicator. A simple logistic classifier predicting the outcome variable from age, number of children, number of dependents other than children, and ownership of a phone yields $\hat{H} = 0.038$ and $AUC = 0.567$. Figure 5 illus-

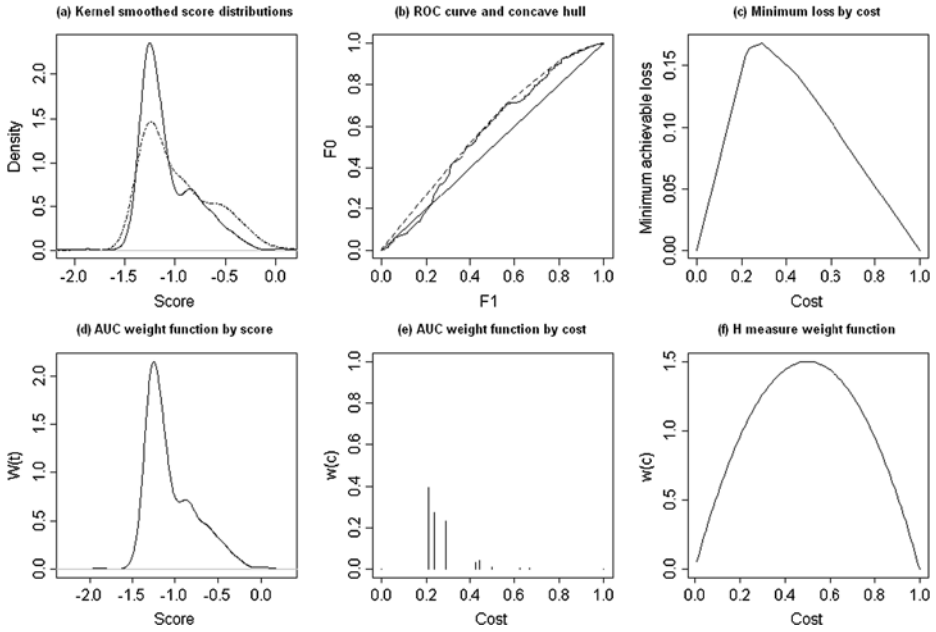


Fig. 5 Plots for first classifier on data from Thomas et al. (2002)

trates the results of this analysis. Working from left to right across the top row, the figures show (a) the kernel smoothed score distributions for the two classes; (b) the ROC curve and convex hull; and (c) the minimum loss achieved by choosing the appropriate threshold for each c. The second row of figures shows (d) the score weight function $W_G(T)$ implicitly used by the *AUC* (which is the mixture distribution of the two scores in the plot above); (e) the cost weight function $w_G(c)$ implicitly used by the *AUC*; and (f) the cost weight function $beta(x; 2, 2)$ used by the H measure. Note the discrete nature of the *AUC* cost weight function.

Figure 6 shows the corresponding plots for a logistic classifier using value of home, mortgage balance outstanding, outgoings on mortgage or rent, outgoings on loan, outgoings on hire purchase, and outgoings on credit cards as predictor variables. We can see that both the score and cost weight functions used in the *AUC* calculations differ between the two classifiers, but that the H cost weight function (lower right) is, of course, the same. This second classifier gives $\hat{H} = 0.035$ and $AUC = 0.591$. Thus the *AUC* of the second classifier is higher than that of the first, suggesting superior performance for the second classifier. However, the H value is lower, suggesting superior performance for the first classifier. That is, the *AUC* and H values lead to different performance orderings of the two classifiers.

For completeness, the two ROC curves are shown superimposed in Fig. 7.

9 Conclusion

The *AUC* is an attractive measure for comparing classification rules and diagnostic instruments. It is objective, so that, given the same set of scores, two researchers will obtain the same *AUC*. It also has various natural intuitive interpretations, one of which is that it is the average sensitivity of a classifier under the assumption that one is equally likely to choose

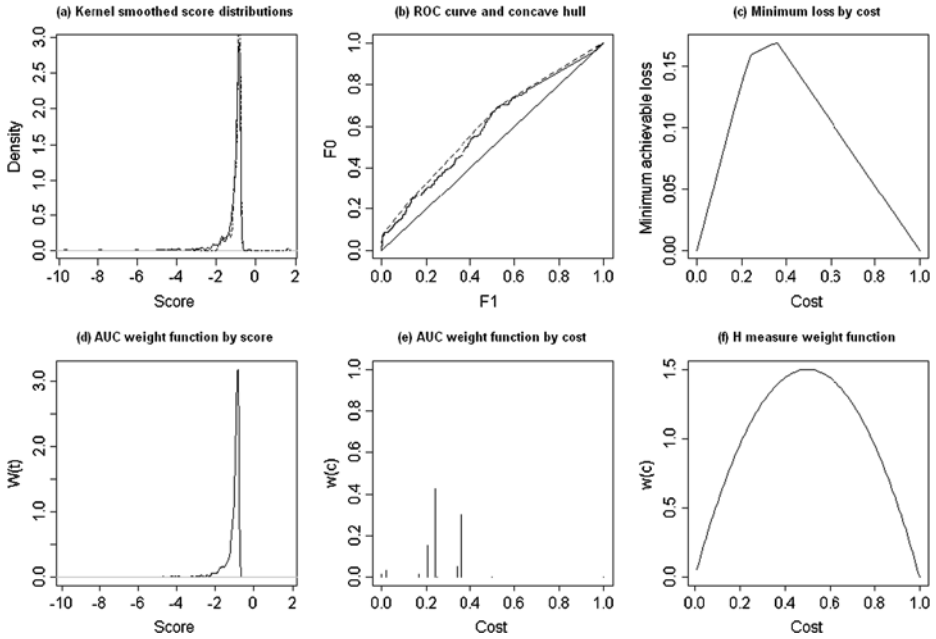
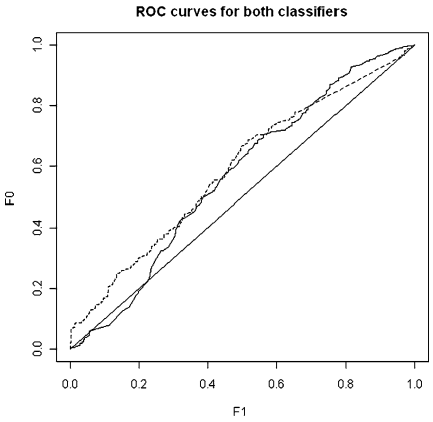


Fig. 6 Plots for second classifier on data from Thomas et al. (2002)

Fig. 7 ROC curves for the two classifiers using the data from Thomas et al. (2002)



any value of the specificity—under the assumption of a uniform distribution over specificity. (As well as, of course, the symmetric interpretation of being the average specificity under a uniform choice of sensitivity.)

Leaving aside the improbability of such a uniform distribution being appropriate for any real problem, the fact is that fixing the distribution of specificity over which the averaging takes place translates into averaging the minimum misclassification loss of the classifier over a distribution of the relative misclassification costs which differs from classifier to classifier. This is a consequence of the relationship between cost and minimum loss: this relationship depends on the empirical score distributions. Conversely, of course, fixing the choice of the distribution of relative misclassification costs (for example, to a beta distribution, as in this

paper), translates into calculating the mean sensitivity over distributions of the specificity which vary between classifiers. It is not possible (except in some special artificial cases) to have both the same cost distribution and the same specificity distribution.

This means that one must choose between these two approaches: either one must choose a specificity distribution over which to average the sensitivity (or a sensitivity distribution over which to average the specificity) or one must choose a relative misclassification cost distribution. This paper takes the view that the specificity is a matter of choice, not a fundamental feature of the problem, but that the relative cost is a fundamental aspect of the problem, and not subject to the whim of the researcher. Of course, one might not know the relative misclassification cost, and hence be forced to adopt some distribution, and these distributions may differ between researchers, but that is a different matter. In particular, the fact that the cost ratio is a property of the problem means that it would be incoherent to choose different cost ratio distributions for different classifiers. Cost should dominate specificity in the choice of measure.

The *AUC* avoids the choice of particular values for the relative cost—a choice which is implicit in misclassification rate and the Kolmogorov-Smirnov statistic—by averaging over all possible values for the relative cost. But in calculating this average it uses a distribution which depends on the classifier being evaluated: the classifier being evaluated determines the choice of measurement scale. This is incoherent: to make valid comparisons, the same 'ruler' must be used on each object being compared.

Unless a cost ratio distribution is specified by the researcher, there is inevitably an arbitrary aspect involved in any measure which integrates over such a distribution. For the H measure this lies in the choice of a beta distribution and equal parameter values. This arbitrariness is, however, far less worrying than the intrinsic incoherence implicit in the *AUC* measure, which uses different distributions to evaluate different classifiers. It means, at least, that the H measure is making fair comparisons.

An R program for calculating the H measure is available on http://stats.ma.ic.ac.uk/d/djhand/public_html/ or directly from the author. This program also gives the *AUC*, the Gini coefficient, and the Kolmogorov-Smirnov statistic, and also the *AUCH*, for comparative purposes. To aid exploration and diagnosis it also produces an array of six plots, as in Figs. 5 and 6 above: (1) the kernel smoothed score distributions for the two classes; (2) the ROC curve and convex hull; (3) the minimum loss function by cost; (4) the score weight distribution the *AUC* implicitly uses; (5) the cost weight function the *AUC* implicitly uses; and (6) the weight function used in the H measure.

Acknowledgements This work was supported by a Wolfson Research Merit Award from the Royal Society. I am indebted to the four anonymous reviewers who clearly spent considerable time and effort on the paper, and whose comments led to material improvements.

References

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

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Dodd, L. E., & Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics*, *59*, 614–623.

Fawcett, T. (2004). *ROC graphs: notes and practical considerations for researchers*. Palo Alto: HP Laboratories.

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

Flach, P. A. (2003). The geometry of ROC space: understanding machine learning metrics through isometrics. In *Proc. 20th international conference on machine learning (ICML'03)* (pp. 194–201).

Hand, D. J. (1997). *Construction and assessment of classification rules*. New York: Wiley.
Hand, D. J. (2004). *Measurement theory and practice: the world through quantification*. London: Arnold.
Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, *56*, 1109–1117.
Hand, D. J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science*, *21*, 1–34.
Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, *45*, 171–186.
Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging*, *29*, 307–335.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under an ROC curve. *Radiology*, *143*, 29–36.
Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.
Jamain, A., & Hand, D. J. (2008). Mining supervised classification performance studies: a meta-analytic investigation. *Journal of Classification*, *25*, 87–112.
Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. London: Chapman and Hall.
McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*, 190–195.
Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.
Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *KDD-97—third international conference on knowledge discovery and data mining*.
Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th international conference on machine learning, ICML-98*.
Rudin, W. (1964). *Principles of mathematical analysis* (2nd edn.). New York: McGraw-Hill.
Scott, M. J. J., Niranjan, M., & Prager, R. W. (1998). *Parcel: feature subset selection in variable cost domains* (Technical Report CUED/F-INFENG/TR. 323). Cambridge University Engineering Department, UK.
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Philadelphia: Society for Industrial and Applied Mathematics.
Webb, A. (2002). *Statistical pattern recognition* (2nd edn.). New York: Wiley.