Electronic Journal of Statistics
Vol. 10 (2016) 380–393
ISSN: 1935-7524
DOI: 10.1214/16-EJS1109

A note on the use of empirical AUC for evaluating probabilistic forecasts

Simon Byrne[*]

Cambridge, United Kingdom
e-mail: simonbyrne@gmail.com

Abstract: Scoring functions are used to evaluate and compare partially probabilistic forecasts. We investigate the use of rank-sum functions such as empirical Area Under the Curve (AUC), a widely used measure of classification performance, as a scoring function for the prediction of probabilities of a set of binary outcomes. It is shown that the AUC is not generally a proper scoring function, that is, under certain circumstances it is possible to improve on the expected AUC by modifying the quoted probabilities from their true values. However with some restrictions, or with certain modifications, it can be made proper.

MSC 2010 subject classifications: Primary 62C99.
Keywords and phrases: Rank-sum, area under the curve, probabilistic prediction, scoring rule, scoring function.

Received August 2015.

Contents

[*] This work was completed while the author was a Research Fellow at the Department of Statistical Science, University College London, supported by EPSRC fellowship EP/K005723/1

1. Introduction

Predicting the outcome of a vector of binary variates is a common problem across a variety of application domains, such as fraud detection, credit risk evaluation, medical diagnostics and weather forecasting. Such forecasts typically carry some information describing the uncertainty of the forecaster, such as assigning explicit probabilities or some other numerical value to each variable that allows the variables to be ranked in order of relative probability of occurrence.

This note investigates numerical measures for evaluating and comparing the accuracy of such forecasts. Although such measures have always been important for comparing algorithms, their role has become increasingly important with the popularity of prediction competitions, where it is necessary to precisely quantify the accuracy of various predictions. In particular, we extend the framework of *scoring functions*, developed by Gneiting [6], which maps the prediction and subsequent observation to a single real number, the *score*, representing the reward to the forecaster. The aim of the forecaster is then to maximise this reward.

Scoring functions can be viewed as extensions of *scoring rules* (section 2.1), which require that the forecast be fully probabilistic, providing a full joint probability distribution over the set of all possible outcomes, which can be infeasible and unnecessary in many situations. Scoring functions (section 2.2) on the other hand can make use of partial probabilistic information such as marginal distributions, or rankings of expected values. One desirable feature of both scoring rules and scoring functions is that they be *proper*: that the forecaster always has the incentive to be honest, in that the forecast which maximises their expected score matches their true belief.

The central contribution of this note is on a class of scoring functions termed *rank-sum functions* (section 3), the most well-known of which is the *area under the curve* (AUC), the curve in question being the receiver operating characteristic (ROC). The ROC and AUC describe the usefulness of the forecast in terms of its ability to discriminate between positive and negative outcomes. We note that we specifically interested in the empirical AUC, and not the theoretical quantity that is perhaps more often studied: this distinction is explained in detail in section 3.1.

The main results (section 3.2) identify sufficient conditions for rank-sum scoring functions to be proper for evaluating the accuracy of forecasts of the marginal probabilities of a sequence of binary forecasts. In general, the AUC is not of this class, and two counterexamples are provided of cases in which the AUC is *not* a proper scoring function, in that there exist distributions under which the forecaster might improve their expected score by quoting probabilities different than their true belief.

We consider various ways in which the framework can be extended (section 4), such as a sequential setting (section 4.1), and the case where the forecaster is required to provide a mapping that indirectly makes predictions from an as-yet unobserved covariate (section 4.2). Finally, we discuss some open questions and possible future research directions (section 5).

2. Scoring of forecasts for binary outcomes

2.1. Scoring rules

Consider the setting where one is eliciting forecasts about some future outcome Y that takes values in an *outcome space* $\mathcal{Y}$. A *probabilistic forecast* is a distribution Q for Y that describes the forecasters uncertainty of Y. We define $\mathcal{F}$ to be a family of distributions over $\mathcal{Y}$ that are under consideration.

After the actual outcome $Y = y$ is observed, the reward to the forecaster is determined by a *scoring rule* [7], a function $S : \mathcal{Y} \times \mathcal{F} \to \mathbb{R}$, that maps the quoted Q and observed outcome y to a real number $S(y, Q)$ termed the *score*. We take scoring rules to be *positively oriented*, that is the score represents the reward to the forecaster, who therefore aims to maximise this quantity. In a decision theoretic context, the negation of the score can be considered a *loss function*. Mathematically, the problem can be precisely phrased in the form of a game between a Forecaster and Nature [4].

For any $P \in \mathcal{F}$, the *expected score* is $\mathbb{E}_P[S(Y, Q)]$, where Y is generated from P. A scoring rule S is defined to be *proper* if an optimal strategy for the forecaster is to quote a distribution that matches their actual uncertainty, that is, if for all $Q, P \in \mathcal{F}$,

$$\mathbb{E}_P[S(Y, Q)] \leq \mathbb{E}_P[S(Y, P)]. \tag{2.1}$$

Additionally, S is termed *strictly proper* if this is the only optimal strategy, i.e. (2.1) is an equality only if $Q = P$. Proper scoring rules for discrete variables have been extensively studied [e.g. 4]; common examples include the Brier, spherical and the log scores.

In this note, we will consider the outcome space to be a vector of binary variables,

$$Y = (Y_1, \ldots, Y_n) \in \mathcal{Y} = \{0, 1\}^n.$$

In this case, the distribution Q takes values on Δ_{2^n-1}, the $(2^n - 1)$-dimensional unit simplex. If the family $\mathcal{F}$ is the set of all such distributions, then for large values of n this can place a large burden in terms of time and resources in constructing, communicating and evaluating the score of the forecast. This motivates a more flexible framework.

2.2. Scoring functions

Gneiting [6] introduced the concept of a *scoring function* for evaluating point forecasts. We now show how this idea can be extended to a more general setting, of what we term *partially probabilistic forecasts*, which aims to capture some of the uncertainty.

Suppose that instead of supplying a full probability distribution Q from a family $\mathcal{F}$, we require forecaster to quote an element from an arbitrary set $\mathcal{Z}$, which we will term the *prediction space*. Then a *scoring function* is a mapping of the form $s : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ (Gneiting [6] considered only the case where $\mathcal{Z} = \mathcal{Y}$, though much of that work extends directly to the more general setting).

The price of this generality is that we now need an additional object to specify the aspects of the forecasters uncertainty that we want to capture. This can be described by a *(statistical) functional*, a possibly set-valued function, $T : \mathcal{F} \to \mathcal{Z}$ or $T : \mathcal{F} \to \wp\mathcal{Z}$, where $\wp\mathcal{Z}$ denotes the power set of $\mathcal{Z}$.

A scoring function s is then said to be *T-proper* (Gneiting [6] uses the term *consistent*) if for all $P \in \mathcal{F}$, and all $u \in \mathcal{Z}$,

$$\mathbb{E}_P[s(Y, u)] \leq \mathbb{E}_P[s(Y, T(P))] \tag{2.2}$$

for $\mathcal{Z}$-valued functional T, or for a set-valued functional T,

$$\mathbb{E}_P[s(Y, u)] \leq \mathbb{E}_P[s(Y, t)] \quad \text{for all } t \in T(P). \tag{2.3}$$

Furthermore, we can define s to be *strictly T-proper* if equality holds only if $u = T(P)$ or $u \in T(P)$, respectively. Note that the condition in (2.3) implies that for any T-proper scoring function s of a set-valued functional, the expected score $\mathbb{E}_P[s(Y, t)]$ must be constant for all $t \in T(P)$.

The functional can be interpreted as a summary of a full probabilistic forecast. Indeed, there is a strong link between scoring functions and scoring rules, in that a functional and (strictly) proper scoring function defines a (strictly) proper scoring rule [6, Theorem 3].

Our central focus is on two specific classes of functionals for distributions on $\mathcal{Y} = \{0, 1\}^n$.

2.2.1. Marginal scoring

Definition 1. The *marginal functional M* maps a joint distribution to the marginal probabilities of each element of Y,

$$M(P) = \mathbb{E}_P[Y] = \big(P[Y_1 = 1], \ldots, P[Y_n = 1]\big).$$

This functional reduces the $(2^n - 1)$-dimensional distribution space to the n-dimensional prediction space $\mathcal{Z} = [0, 1]^n$.

We can easily construct scoring functions for the marginal functional as functions of scoring rules for the individual elements of Y.

Theorem 1. *Let $S_i : \{0, 1\} \times [0, 1] \to \mathbb{R}$ be a scoring rule for a single binary outcome, such as the logarithmic, quadratic or Brier score. Then the scoring function*

$$s(y, m) = \sum_{i=1}^{n} S_i(y_i, m_i)$$

is (strictly) M-proper if each of the S_i are (strictly) proper.

Proof. Each S_i can be maximised independently by choosing $m_i = \mathbb{E}[Y_i]$. □

2.2.2. Rank scoring

Recall that a *total preorder* is a transitive and reflexive relation $\precsim$ such that for any pair i, j, at least one of $i \precsim j$ or $j \precsim i$. Given such a $\precsim$, we can define $i \sim j$ as the symmetric relation $i \precsim j$ and $i \succsim j$ and $i \prec j$ as the asymmetric relation $i \not\succsim j$ (which due to totality, implies $i \precsim j$). Note $\precsim$ also implies a total ordering of the equivalence classes under $\sim$.

Define Ξ_n to be the set of total preorders on the set of indices $I = \{1, \dots, n\}$, then any vector $v \in \mathbb{R}^n$ *induces* an element of $\precsim_v \in \Xi_n$ by

$$i \precsim j \quad \Leftrightarrow \quad v_i \leq v_j.$$

Definition 2. The *exact rank functional* $R : \mathcal{F} \to \Xi_n$ maps a joint distribution to the total preorder induced by the marginal functional M.

The exact rank functional can also be characterised in terms of pairwise comparisons.

Proposition 2. *Let* $\precsim = R(P)$ *for some distribution* P *on* $\mathcal{Y}$. *Then*

$$i \precsim j \quad \Leftrightarrow \quad P[Y_i > Y_j] \leq P[Y_i < Y_j].$$

Proof. By adding $P[Y_i = 1, Y_j = 1]$ to both sides, we have that

$$P[Y_i = 1, Y_j = 0] \leq P[Y_i = 0, Y_j = 1] \quad \Leftrightarrow \quad P[Y_i = 1] \leq P[Y_j = 1] \qquad \square$$

In the case where all the elements of $M(P)$ are unique, $R(P)$ is a total order. We define $\Omega_n \subseteq \Xi_n$ to be the set of all total orders on I.

Note that the exact rank functional requires that ties ($\mathbb{E}[Y_i] = \mathbb{E}[Y_j]$) be identified exactly. We define a weaker notion under which the ties can be ignored. A relation $\precsim'$ is *contained* in a relation $\precsim$ if $\precsim' \subseteq \precsim$, that is, if $i \precsim' j$ implies that $i \precsim j$.

Definition 3. The *weak rank functional* $R^* : \mathcal{F} \to \wp\Xi_n$ is the set-valued functional that maps a probability distribution to the set of total preorders contained in the exact rank functional:

$$R^*(P) = \{\precsim \in \Xi_n : \precsim \subseteq R(P)\}.$$

As a result, if all elements of $M(P)$ are unique, then $R^*(P) = \{R(P)\}$, and conversely if all the elements of $M(P)$ are equal, then $R^*(P) = \Xi_n$.

Given an R^*-proper scoring function s, we can construct a M-proper scoring function s', via $s'(y, m) = s(y, \precsim_m)$. Of course, such a scoring function can never be strictly M-proper, as $\precsim_m$ is preserved under any monotonic increasing transformation.

An advantage of rank-based scoring functions is that they allow the use of more abstract measures of propensity other than probability, and make it possible to compare forecasts generated by a wide variety of algorithms, whose outputs need not necessarily have a direct probabilistic interpretation. The downside is that we lose the ability to say anything about the *calibration* of the forecaster.

3. Rank-sum scoring functions

We now consider a particular class of rank-based scoring functions. For any total preorder $\precsim$, we define its *rank vector* $\rho : \Xi_n \to \mathbb{R}^n$ to be the net number of elements that precede each element,

$$\rho_i(\precsim) = \sum_{j=1}^{n} \mathbb{1}_{j \precsim i} - \mathbb{1}_{j \succsim i}$$

We will consider the class *rank-sum* scoring functions, of the form

$$s(y, \precsim) = g(y) + \sum_{i=1}^{n} \sigma_i(y)\rho_i(\precsim). \tag{3.1}$$

for some functions g and $\sigma = (\sigma_i)_{i=1,\dots,n}$

Example 1 (Wilcoxon–Mann–Whitney u)**.** The most well-known example of such a function is the *Wilcoxon–Mann–Whitney* u, commonly used as a non-parametric test statistic for comparing magnitude of two random variables. It is defined as the number of times observations where $y_i = 0$ precede observations where $y_i = 1$, with ties counting as half

$$u(y, \precsim) = \sum_{i:y_i=0} \sum_{j:y_j=1} \mathbb{1}_{i \prec j} + \tfrac{1}{2}\mathbb{1}_{i \sim j}. \tag{3.2}$$

The term inside the summation is equal to $\frac{1}{2}[1 + \mathbb{1}_{i \precsim j} - \mathbb{1}_{i \succsim j}]$, and so

$$u(y, \precsim) = \tfrac{1}{2} n_0(y) n_1(y) + \tfrac{1}{2} \sum_{i,j=1}^{n} y_i(1 - y_j)(\mathbb{1}_{i \precsim j} - \mathbb{1}_{i \succsim j}).$$

where $n_1(y) = \sum_{i=1}^{n} y_i$, and $n_0(y) = n - n_1(y)$. By symmetry, we have that $\sum_{i,j}(\mathbb{1}_{i \precsim j} - \mathbb{1}_{i \succsim j}) = 0$, and hence,

$$u(y, \precsim) = \tfrac{1}{2} n_0(y) n_1(y) + \tfrac{1}{2} \sum_{i=1}^{n} y_i \rho_i(\precsim).$$

For a fixed y, u will take values on the half-integers $0, \frac{1}{2}, 1, \dots, n_0(y) n_1(y)$.

Example 2 (Area under the curve)**.** The *receiver operating characteristic* (ROC) describes the trade-off of sensitivity and specificity (or type I and type II error) of a preorder, and is calculated by plotting the true positive rate against the false positive rate that would be obtained by taking different elements of the preorder as the cutoff.

It can be described as the parametric curve on $[0,1] \times [0,1]$, starting at $(1,1)$, then linearly connecting the points

$$\left(\sum_{j:y_j=0} \frac{\mathbb{1}_{j \succ i}}{n_0(y)}, \sum_{j:y_j=1} \frac{\mathbb{1}_{j \succ i}}{n_1(y)} \right), \tag{3.3}$$

for each equivalence class i under $\sim$, in the order of $\prec$.

The *area under the curve* (AUC) is then the total area under this curve, which will take values on $[0,1]$. It is well-established [e.g. 9] that this is in fact equal to the Wilcoxon–Mann–Whitney u, standardised by dividing by $n_0(y)n_1(y)$.

Note that if the outcomes are identical (i.e. $y = \mathbf{0}$ or $\mathbf{1}$), then the ROC and AUC are not properly defined. For convenience, we can define the AUC to be $1/2$ in both these cases, however the choice of this constant does not affect any of the results other than Theorem 3.

As a result, we can write

$$\mathrm{AUC}(y, \precsim) = \tfrac{1}{2} + \tfrac{1}{2}\sum_{i=1}^{n} \alpha_i(y)\rho_i(\precsim) \quad \text{where } \alpha_i(y) = \begin{cases} \dfrac{y_i}{n_0(y)n_1(y)} & n_1(y) \neq 0, n, \\ 0 & \text{otherwise.} \end{cases}$$

Also related is the *Gini coefficient*, $g(y, \precsim) = 2\,\mathrm{AUC}(y, \precsim) - 1$, which is twice the net area of the ROC above the diagonal, and takes values on $[-1, 1]$.

3.1. Relation to theoretical AUC

Although the AUC has been widely explored in the literature, much of this work [e.g. 1, 3, 8, 5] focuses on a related but distinct quantity, which we will term the theoretical AUC.

Let θ be a joint distribution for a random pair (X_i, Y_i), where X_i, taking values in some set $\mathcal{X}_{\bullet}$, is termed the *covariate* or *feature*, and Y_i is a single binary response. For some mapping $f : \mathcal{X}_{\bullet} \to \mathbb{R}$, we define the conditional CDFs $F_y(z) = \theta[f(X_i) < z \mid Y_i = y]$. Then the *theoretical ROC* replaces the empirical quantities of (3.3) with their theoretical equivalents,

$$\big(1 - F_0(z), 1 - F_1(z)\big), \quad z \in \mathbb{R}$$

which again, describes a curve over $[0,1] \times [0,1]$. Similarly, the *theoretical AUC*, denoted $\mathrm{tAUC}(\theta, f)$, is the area under this curve.

The theoretical AUC can be rewritten as the conditional expectation [e.g. 3, Proposition B.2],

$$\mathrm{tAUC}(\theta, f) = \mathbb{E}\left[\mathbb{1}_{f(X_1)>f(X_2)} + \tfrac{1}{2}\mathbb{1}_{f(X_1)=f(X_2)} \mid Y_1 = 1, Y_2 = 0\right], \tag{3.4}$$

where the expectation is with respect to the product measure of $\theta \times \theta$ for $[(X_1, Y_1), (X_2, Y_2)]$.

The relationship between the empirical and theoretical AUCs is well-established, though for completeness we clarify the usual presentation [e.g. 1, Lemma 2].

Theorem 3. *Let the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed as θ, then the expected empirical AUC,*

$$\mathbb{E}[\mathrm{AUC}(Y, \precsim_{f(X)})] = (1 - \pi_0^n - \pi_1^n)\,\mathrm{tAUC}(\theta, f) + \tfrac{1}{2}(\pi_0^n + \pi_1^n)$$

where $\pi_c = \theta(Y_i = c)$.

Proof. For any vector $y \neq \mathbf{0}, \mathbf{1}$, the expectation of (3.2) conditional on $Y = y$ gives an expression of the form of (3.4), and hence $\mathbb{E}[\mathrm{AUC}(Y, \precsim_{f(X)}) \mid Y = y] = \mathrm{tAUC}(\theta, f)$. □

We emphasise several key differences between the empirical and theoretical AUC. Firstly, the theoretical AUC is a function of the mapping f from X_i that is used to induce a ranking on Y_i (confusingly, this is itself referred to as a "scoring function" in the literature).

Another distinction is that the distribution θ is now a hypothetical sampling model for a single pair (X_i, Y_i), whereas the previous distribution P describes the forecasters uncertainty for a set $(Y_1, \dots, Y_n)$. We emphasise that these are distinct concepts: whereas the i.i.d. assumption is typically reasonable in a sampling context, it is extremely unrealistic for describing uncertainty, in that it would imply that there is absolutely no information to be gained about Y_n from the other $Y_1, \dots, Y_{n-1}$.

Additionally, although the negation of $\mathrm{tAUC}(\theta, f)$ can still be interpreted as a loss function in the standard decision-theoretic sense (e.g. for deriving minimax procedures), $\mathrm{tAUC}(\theta, f)$ cannot be used as a scoring function as θ is typically never observed directly.

3.2. Proper rank-sum scoring functions

To determine the propriety of such scoring functions, we utilise the following key lemma.

Lemma 4. *For any fixed vector $v \in \mathbb{R}^n$, the quantity*

$$\sum_{i=1}^{n} v_i \rho_i(\precsim) \tag{3.5}$$

is maximised over $\precsim \in \Xi_n$ if and only if $\precsim$ is contained in $\precsim^{(v)}$, the preorder induced by v.

Proof. Firstly, note that if we were to consider only total orders $\precsim \in \Omega_n$, then the statement is a direct result of the rearrangement inequality. For any total preorder $\precsim \in \Xi_n$, define $A(\precsim)$ to be the set of total orders contained in $\precsim$, that is $A(\precsim) = R^*(\precsim) \cap \Omega_n$. Then for any i, j, by symmetry we have that

$$\mathbb{1}_{i \precsim j} = \frac{1}{|A(\precsim)|} \sum_{\precsim' \in A(\precsim)} \mathbb{1}_{i \precsim' j}.$$

Therefore $\rho(\precsim)$ is the average of all $\rho(\precsim')$ for $\precsim' \in A(\precsim)$. It follows then that (3.5) is is maximised if and only if all such $\precsim'$ are themselves contained $\precsim^{(v)}$, which in turn implies that $\precsim$ itself is contained in $\precsim^{(v)}$. □

This then leads to our main result.

Theorem 5. *A rank-sum scoring function s of the form in (3.1) is strictly R^*-proper if and only if $\precsim_{P\sigma}$, the preorder induced by $\mathbb{E}_P[\sigma_i(Y)]$, is an element of $R^*(P)$ for all $P \in \mathcal{F}$.*

Proof. By the linearity of expectation, we have that

$$\mathbb{E}_P[s(Y,\precsim)] = \mathbb{E}_P[g(Y)] + \sum_{i=1}^{n} \mathbb{E}_P[\sigma_i(Y)]\rho_i(\precsim).$$

By Lemma 4, this can be maximised by any $\precsim$ contained in $\precsim_{P\sigma}$. These are all elements of $R^*(P)$ if and only if $\precsim_{P\sigma}$ itself is in $R^*(P)$. □

Consequently, the Wilcoxon–Mann–Whitney u function is a strictly R^*-proper scoring function, however the same cannot be said of the AUC.

Example 3. Define the distribution P on (Y_1, Y_2, Y_3, Y_4) with the following non-zero probabilities:

$$P(1,1,0,0) = \tfrac{1}{2}, \quad P(0,0,1,0) = \tfrac{7}{16}, \quad P(0,0,0,1) = \tfrac{1}{16}.$$

Then defining α as in Example 2, we have that

$$\mathbb{E}[Y] = \left(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{7}{16}, \tfrac{1}{16}\right) \quad \text{and} \quad \mathbb{E}[\alpha(Y)] = \left(\tfrac{1}{8}, \tfrac{1}{8}, \tfrac{7}{48}, \tfrac{1}{48}\right).$$

Define $\precsim_P$ and $\precsim_\alpha$ as the preorders induced by $\mathbb{E}[Y]$ and $\mathbb{E}[\alpha(Y)]$, respectively. Then $\rho(\precsim_P) = (2, 2, -1, -3)$ and $\rho(\precsim_\alpha) = (0, 0, 3, -3)$, with expected AUCs

$$\mathbb{E}[\mathrm{AUC}(Y, \precsim_P)] = \tfrac{31}{48} < \mathbb{E}[\mathrm{AUC}(Y, \precsim_\alpha)] = \tfrac{33}{48}.$$

This rather contrived example is illustrative of how the problem arises, namely the denominator of α can alter the relative importance of certain outcomes. Nevertheless, there exist certain families $\mathcal{F}$ under which AUC is indeed proper.

Theorem 6. *If the number of positive outcomes $n_1(Y)$ is almost surely constant for all $P \in \mathcal{F}$, then AUC is a strictly R^*-proper scoring function.*

Proof. If $n_1(Y) = r$ almost surely, then $\mathbb{E}_P[\alpha_i(Y)] = \mathbb{E}_P[Y_i]/\big((n-r)r\big)$. □

This justifies the use of AUC as a scoring function in cases where the forecaster is informed of the number of positive outcomes beforehand. This means that the forecaster is able to use this information to rule out extreme tail events that might otherwise have provided a windfall score. For example, in the IJCNN Social Network Challenge by Kaggle (https://www.kaggle.com/c/socialNetwork) competitors were required to estimate 8960 binary outcomes (corresponding to presence/absence of an edge), of which they were informed that exactly half were positive.

Theorem 7. *If the Y_i's are mutually independent under all $P \in \mathcal{F}$, then AUC is a strictly R^*-proper scoring function.*

Proof. Note that if $y_i \neq y_j$, then $n_1(y) = 1 + n_1^{\neg(i,j)}(y)$, where $n_1^{\neg(i,j)}(y) = \sum_{k \neq i,j} y_k$, and similarly for n_0. Then

$$\alpha_i(y) - \alpha_j(y) = \frac{y_i - y_j}{n_0(y) n_1(y)} = \frac{y_i - y_j}{[1 + n_0^{\neg(i,j)}(y)][1 + n_1^{\neg(i,j)}(y)]},$$

since if $y_i = y_j$, the numerator is zero. Then by mutual independence,

$$\mathbb{E}[\alpha_i(Y)] - \mathbb{E}[\alpha_j(Y)] = (\mathbb{E}[Y_i] - \mathbb{E}[Y_j])\, \mathbb{E}\left[\frac{1}{[1 + n_0^{\neg(i,j)}(Y)][1 + n_1^{\neg(i,j)}(Y)]}\right].$$

As the latter expectation is strictly positive, it follows that $\mathbb{E}[\alpha_i(Y)] \leq \mathbb{E}[\alpha_j(Y)]$ if and only if $\mathbb{E}[Y_i] \leq \mathbb{E}[Y_j]$. □

As noted in section 3.1, mutual independence is a somewhat unrealistic condition for scoring functions. Nevertheless, it can be useful when combined with the following result.

Theorem 8. *Let $\mathcal{F}$ consist of distributions P such that there is a latent variable Z whereby*

(i) for almost all Z, $\mathbb{E}_P[Y \mid Z]$ induces the same preordering as $\mathbb{E}_P[\alpha(Y) \mid Z]$, and

(ii) this preordering is the same for almost all Z,

then AUC is a strictly proper scoring function for R^*.

Proof. Condition (i) implies that

$$\mathbb{E}_P[Y_i - Y_j \mid Z] \geq 0 \quad \Leftrightarrow \quad \mathbb{E}_P[\alpha_i(Y) - \alpha_j(Y) \mid Z] \geq 0,$$

and by condition (ii) then,

$$\mathbb{E}_P[\mathbb{E}[Y_i - Y_j \mid Z]] = \mathbb{E}_P[Y_i - Y_j] \geq 0 \quad \Leftrightarrow \quad \mathbb{E}_P[\alpha_i(Y) - \alpha_j(Y)] \geq 0. \qquad \square$$

This provides a means for showing AUC is proper in more general contexts, by combining it with one of the previous two theorems to satisfy condition (i). For example, if θ is a parameter in a Bayesian model, conditional on which the outcomes are independent (e.g. a logistic regression model), then AUC is proper for the predictive distributions if (ii) holds.

However these conditions can fail if there is significant uncertainty in the ordering of the outcomes, which may arise in problems such as out-of-sample prediction.

Example 4. Suppose that there are two candidate models, A and B, each weighted with probability 1/2, and the forecaster is to rank 100 outcomes, of which 10 have a particular feature U present. Suppose that the forecast probabilities are

$$\mathbb{E}[Y_i \mid U_i, A] = 0.4 \qquad \mathbb{E}[Y_i \mid \neg U_i, A] = 0.5$$
$$\mathbb{E}[Y_i \mid U_i, B] = 0.95 \qquad \mathbb{E}[Y_i \mid \neg U_i, B] = 0.9,$$

and that outcomes are independent within each model. Then the resulting marginal probabilities are

$$\mathbb{E}[Y_i \mid U_i] = 0.675 \qquad\qquad \mathbb{E}[Y_i \mid \neg U_i] = 0.7$$

However using the induced ranking will result in an expected AUC of 0.496, whereas the opposite ranking will give an expected AUC of 0.504 (see supplementary material [2]).

This example is admittedly rather extreme in the difference between the candidate models. The fact that such lengths were required suggests that there may exist some additional weak yet realistic condition under which AUC is indeed proper.

4. Extensions

4.1. Sequential scoring

We have also only considered the batch prediction setting where the forecaster is required to provide the preordering for all Y before any outcomes have been observed. One alternative is a sequential framework, where at each point in time the forecaster is required to provide a forecast for Y_{t+1}, having already observed $Y_1, \dots, Y_t$. In the ranking case, this requires the forecaster to provide a total preorder $\precsim_{t+1}$ on I_{t+1} that is compatible with the one $\precsim_t$ provided on I_t. Unfortunately, rank-sum scoring functions are essentially useless in this setting.

Example 5. Let s be any rank-sum scoring rule of the form in (3.1), where $\sigma_i(y) = \sigma_j(y)$ if $y_i = y_j$, and $\sigma_i(y) \geq \sigma_j(y)$ if $y_i > y_j$ (both u and the AUC satisfy this property). Then in the sequential setting, it is possible to maintain an optimal score by choosing $\precsim_{t+1}$ such that

$$i \prec_{t+1} t+1 \prec_{t+1} j \quad \text{for all } i,j \leq t\text{: } Y_i = 0 \text{ and } Y_j = 1.$$

By a straightforward application of induction, it is easy to see that such a sequence exists, and that it will maintain this “perfect separation”, in that all i where $Y_i = 1$ will always be ranked above all j where $Y_j = 0$. Therefore, by Lemma 4, this will result in the largest possible score (i.e. an AUC of 1): note that unlike the previous sections, we refer to *actual* score, not just the expected score.

In other words, it is possible to construct an optimal procedure with absolutely no information whatsoever about the process of Y_t, and consequently the the scoring function cannot be proper for any reasonable functional. This problem will persist in the analogous mapping problem, where the forecaster is free to choose the mapping $f_t : \mathcal{X}_\bullet \to \mathbb{R}$ at each iteration.

4.2. Scoring functions for mappings

In many forecasting settings, each variable Y_i has a corresponding *covariate* or *feature* X_i taking values in some measurable space $\mathcal{X}_\bullet$, which can be used to inform the prediction. In the case where the forecaster is able to observe the covariates directly, we can assume any relevant information is taken into account, and thus no additional consideration is required.

However we can also consider the setting in which the forecaster does not observe the covariates, but is instead required to provide some sort of mapping from the covariate space $\mathcal{X} = (\mathcal{X}_\bullet)^n$ to the original prediction space $\mathcal{Z}$ for Y (we use the term *mapping* so as to distinguish from scoring functions). In other words, the forecaster is required to make a prediction in the *mapping prediction space*

$$\vec{\mathcal{Z}} = \{f : \mathcal{X} \to \mathcal{Z}\}.$$

Furthermore, any scoring function $s : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ has a corresponding *mapping form* $\vec{s} : (\mathcal{X} \times \mathcal{Y}) \times \vec{\mathcal{Z}} \to \mathbb{R}$ which is simply s evaluated using the mapping applied to the observed covariates,

$$\vec{s}\big((x, y), f\big) = s\big(Y, f(X)\big).$$

Similarly, given any statistical functional $T : \mathcal{F} \to \mathcal{Z}$, we can define the corresponding *mapping functional* $\vec{T} : \mathcal{F}_{XY} \to \vec{\mathcal{Z}}$ as the mapping of the conditional expectation

$$\vec{T}(P_{XY})(x) = T(P_{Y|X=x}),$$

where $P_{Y|X=x}$ denotes the conditional distribution of Y given $X = x$ under P. That is, the optimal mapping should map each $x \in \mathcal{X}$ to the optimal prediction under the conditional distribution $P_{Y|X=x}$.

Theorem 9. *Let s be a T-proper scoring function for a family $\mathcal{F}$, then $\vec{s}$ is a $\vec{T}$-proper scoring function for $\mathcal{F}_{XY}$ if for each $P_{XY} \in \mathcal{F}_{XY}$, there exists a family of conditional distributions $\{P_{Y|X=x}\}_x$ which is a subset of $\mathcal{F}$.*

Proof. The expected mapping score is

$$\mathbb{E}\big[\vec{s}\big((x, y), f\big)\big] = \mathbb{E}\big[\mathbb{E}\big[s\big(Y, f(X)\big) \mid X\big]\big].$$

The inner expectation can be maximised for each value of $X \in \mathcal{X}$ by choosing $f(x) = \arg\max_z \mathbb{E}[s(Y, z) \mid X]$, which, as s is T-proper, will be (an element of) $T(P_{Y|X=x})$. □

However we typically don't want to consider all possible mappings $f : \mathcal{X} \to \mathcal{Z}$. Instead, we typically are only interested in mappings that can be applied coordinate-wise,

$$f(x) = \big(f_\bullet(x_1), \ldots, f_\bullet(x_n)\big), \quad \text{where } f_\bullet : \mathcal{X}_\bullet \to \mathbb{R}.$$

In other words, we constrain the mapping such that the forecast for each Y_i depends only on its corresponding covariate X_i, and require that this mapping be the same for all i. Of course, we also need to constrain the family of distributions to ensure that the marginal mapping is coordinate-wise.

Theorem 10. *Let $\vec{\mathcal{F}}$ be the set of distributions for (X, Y) such that*

(i) Y_i are conditionally independent of X given X_i, and
(ii) the distribution of $Y_i \mid X_i$ is the same for all i.

Then for any M-proper scoring function s for a family $\mathcal{F}$, $\vec{s}$ is a $\vec{M}$-proper scoring function for the set of coordinate-wise mappings if the conditional distributions $P_{Y|X=x}$ are in $\mathcal{F}$.

Proof. By (i) we have that $\mathbb{E}[Y_i \mid X = x] = \mathbb{E}[Y_i \mid X_i = x_i]$, and by (ii) it follows that this quantity is the same for all i. Therefore the mapping $f(x) = \vec{M}(P_{Y|X=x})$ is coordinate-wise, which by Theorem 9, implies that $\vec{s}$ is $\vec{M}$-proper. □

Consequently $\vec{u}$, the mapping form of u is $\vec{M}$-proper for any $\vec{\mathcal{F}}$ satisfying (i) and (ii). For AUC to be $\vec{M}$-proper, additional conditions are required, such as mutual independence of elements of Y conditional on X.

5. Discussion

We have shown that AUC is not generally a proper scoring function. However our counterexamples, Examples 3 and 4, both exhibit quite extreme dependence between outcomes. Therefore, we conjecture that it might be possible to establish a more relaxed criteria for establishing propriety of AUC, for example, bounds on correlation or other measures of dependence.

Acknowledgements

The author is grateful to Matthew Parry for comments and feedback.

Supplementary Material

Supplement to "A note on the use of empirical AUC for evaluating probabilistic forecasts"
(doi: 10.1214/16-EJS1109SUPP; .zip). Supplementary material contains the calculations for Example 4.

References

[1] Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S. and Roth, D. (2005). Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research* **6** 393–425. MR2249826

[2] Byrne, S. (2016). Supplement to "A note on the use of empirical AUC for evaluating probabilistic forecasts". doi:10.1214/16-EJS1109SUPP

[3] Clémençon, S., Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *Annals of Statistics* **36** 844–874. doi:10.1214/009052607000000910. MR2396817 (2009d:68069)

[4] Dawid, A. P., Lauritzen, S. and Parry, M. (2012). Proper local scoring rules on discrete sample spaces. *Annals of Statistics* **40** 593–608. doi:10.1214/12-AOS972. MR3014318

[5] Flach, P., Hernandez-Orallo, J. and Ferri, C. (2011). A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance. In *Proceedings of the 28th International Conference on Machine Learning* (L. Getoor and T. Scheffer, eds.) 657–664. ACM, New York, NY, USA.

[6] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106** 746–762. doi:10.1198/jasa.2011.r10138. MR2847988 (2012j:62355)

[7] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. doi:10.1198/016214506000001437. MR2345548

[8] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* **77** 103–123. doi:10.1007/s10994-009-5119-5

[9] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36. doi:10.1148/radiology.143.1.7063747