A Comment on the ROC Curve and the Area Under it as Performance Measures

Caren Marzban[*]

The Applied Physics Laboratory, and the Department of Statistics,
University of Washington, Seattle, WA 98195

and

Center for Analysis and Prediction of Storms
University of Oklahoma, Norman, OK 73019

June 28, 2004

[*] http://www.nhn.ou.edu/~marzban

Abstract

The Receiver Operating Characteristic (ROC) curve is a two dimensional measure of classification performance. The area under the ROC curve (AUC) is a scalar measure gauging one facet of performance. In this note, five idealized models are utilized to relate the shape of the ROC curve, and the area under it, to features of the underlying distribution of forecasts. This allows for an interpretation of the former in terms of the latter. The analysis is pedagogical in that many of the findings are already known in more general (and more realistic) settings; however, the simplicity of the models considered here allows for a clear exposition of the relation. For example, although in general there are many reasons for an asymmetric ROC curve, the models considered here clearly illustrate that an asymmetry in the ROC curve can be attributed to unequal widths of the distributions. Furthermore, it is shown that AUC discriminates well between "good" and "bad" models, but not between "good" models.

1 Introduction

Consider the problem of assessing the quality of forecasts produced for binary observations (here labeled 0 and 1). The forecast quantity may be a continuous quantity ranging from $-\infty$ to $+\infty$, or it may be a probability, ranging from 0 to 1. It was shown by Murphy and Winkler (1987, 1992) that this problem is best cast into a framework based on the joint probability distribution of the forecasts and observations. Figure 1 depicts the general situation, where L_0 and L_1 are the likelihoods for the two classes. In other words, $L_i(x)$ is the probability of the forecast x, given that the observation is from the ith class.[1] This figure illustrates an example of what Murphy and Winkler call a discrimination diagram. There, it was shown that the quality of forecasts can be assessed with complete generality in terms of several such diagrams; other diagrams gauge different facets of that quality, e.g. refinement, resolution, reliability, etc.

Meteorologists ((Harvey et al. 1992; Mason 1982, Mason and Graham 1999; Stephenson 2000, Wilks 2001) have also become interested in a procedure heavily utilized in medical circles (Dorfman et al. 1969, 1997; Metz, Herman, and Shen 1998; Shapiro 1999; Zhou, McClish, and Obuchowski 2002; Zou 2001). The procedure is based on the Receiver Operating Characteristic (ROC) curve. In its simplest form it is a parametric plot of the hit rate (or probability of detection) vs. the false alarm rate, as a decision threshold is varied across the full range of a continuous forecast quantity. The diagonal line corresponds to random forecasts, and the amount of concavity is taken to be a measure of performance. The area under the ROC curve (AUC) is often taken as a scalar measure (Hanley and McNeil, 1982). An AUC of 0.5 reflects random forecasts, while AUC=1 implies perfect forecasts. It has also been shown by Mylne (1999) and Richardson (2000, 2001) that AUC is closely related to the economic value of a forecast system.

The hit rate and the false alarm rate can be computed from the likelihoods:

$$H = \int_t^\infty \; L_1(x) \; dx \quad , \quad F = \int_t^\infty \; L_0(x) \; dx \; , \tag{1}$$

where t is the decision threshold. The upper limit of the integral corresponds to the maximum allowed value of x. For probabilistic forecasts, that limit is 1.

[1] For a given data set, a normalized histogram of x is the best of way of visualizing the likelihood.

The ROC framework is somewhat different from the Murphy-Winkler framework. For example, for probabilistic forecasts the Murphy/Winkler framework does not require, and indeed discourages, the reduction of the forecasts into categorical. The ROC analysis, by contrast, is based on the contingency table, and therefore, requires the introduction of a decision threshold for the purpose of reducing the continuous forecasts into binary forecasts. Of course, the introduction of a threshold does not imply that ROC analysis is in any way inferior to the Murphy/Winkler framework; it is simply another method of assessing performance, with an emphasis on different facets of performance. The Murphy-Winkler framework is more suitable for comparing different sets of forecasts (e.g., from two forecasters), while the explicit presence of a decision threshold in ROC analysis lends itself to the situation where a decision must be made, or action must be taken, in response to forecasts.

In this note, a number of questions are addressed regarding the shape of ROC curves. A few examples are provided to motivate the questions, and five toy models are utilized to answer the questions. The toy models, although somewhat unrealistic, are designed to be progressively better approximations to the general problem depicted in Figure 1. The primary aim of this study is to introduce an awareness of the connections between the Murphy-Winkler framework and ROC analysis. As such, the results reported here are specific to the toy models considered, and are unlikely to be generally true. Although one model - based on gaussians - is likely to be generally valid, all the considered examples are sufficiently flexible to allow for a number of ROC behaviors observed in realistic situations. The simplicity of the models offers a transparent environment wherein observed ROC behaviors can be explained in terms of more basic quantities, namely the parameters of the class-conditional distribution of forecasts (i.e., the likelihoods).

Fig. 2a displays 16 ROC curves representing different levels of performance. These curves gauge the performance of a markov chain model for forecasting tornadic activity in 4 different regions of the U.S. during 4 seasons (Drton, et al., 2003). The behavior of these curves is canonical in that they do what they are expected to. They all begin from the point (0,0) and end at (1,1). But note the high degree of symmetry about the diagonal(s). Fig. 2b, displays another set of 16 ROC curves, this time from a statistical model for predicting hail size (Marzban and Witt 2001). Although, these curves are not pathological in any sense, they do display a few features that are common to many ROC curves. The lowest performing models have symmetric ROC curves, but the mid-range models begin to loose that symmetry. A natural question is if this asymmetry can be explained in terms of the underlying distributions?

Another feature that often emerges is the extensive overlap of the ROC curve with one (or two) of the axes of the diagram. In Fig. 2b, this can be seen in the most concave curves (i.e., corresponding to the best-performing models). These yield ROC curves that overlap the top axis for all false alarm rates higher than 0.4. What is the explanation for this type of overlap? And what about an overlap with the y-axis?

Another type of asymmetry (not shown here) arises when the ROC curve cross the diagonal at some (usually one) point. What causes this type of cross-over?

Many users of ROC curves observe that in dealing with a wide range of forecasts in different situations, most forecasts appear to lead to highly concave ROC curves, or equivalently high AUC values. AUC values of, say 0.9995, are not uncommon. Fig. 2c displays 8 sets of ROC curves with extreme concavity. These are related to a neural network developed for the prediction of ceiling and visibility (Marzban, Leyton, Colman, 2003). The forecasts underlying the curves have different forecast characteristics (in terms of the various attributes of probabilistic forecasts computed within the Murphy-Winkler framework), yet they all lead to very concave ROC curves. The AUC values for these curves vary from 0.990 to 0.996. Why are these AUC values exceedingly near 1? Is it because the forecasts are of extraordinary quality? Or is it an artifact of AUC itself? If the former is true, then a histogram of all AUC values would be right-peaked (or show a heavy tail to the left). This is difficult to test, because the necessary data would be difficult to compile. On the other hand, if the culprit is the measure itself, then testing that hypothesis would be unnecessary, for an explanation would then be at hand. And what sort of artifact would lead to near-one AUC values?

As mentioned above, although the two approaches have different emphasis, they are related. After all, the quantities from which an ROC curve is derived - hit rate and false alarm rate - are areas under the conditional distributions, above some decision threshold. Moreover, although the computation of ROC curves does not require knowledge of these distributions, an assessment of the statistical significance of ROC curves does (Dorfman et al, 1969, Hanley and McNiel, 1982; Stephenson 200). For example, in order to compute standard errors for ROC or AUC (in a parametric approach) one makes some assumptions regarding these underlying distributions. It is natural, then, to utilize the connection between the ROC curve and the underlying distributions to answer the above questions. The answers, then, offer a means of interpreting ROC curves at a more fundamental level.

In summary, here, several toy models are utilized to relate some characteristic features of ROC curves with features of the underlying distributions. As such, the shape of the ROC curve can be interpreted or "explained." Knowledge of the underlying distributions can guide the development of better forecasts. AUC is also examined within the toy models. It is important to emphasize that the distributions examined here are toy-models and mostly of pedagogical value. The five distributions considered are shown in Figures 3a-7a. They are referred to as 1) Uniform, 2) Triangular with unconstrained support, 3) Gaussian, 4) Triangular with constrained support, and 5) Beta distributions. The first three are appropriate for cases where the forecast quantity varies over the real line from $-\infty$ to $+\infty$, while the last two apply to probabilistic forecasts.

2 Uniform Distribution

A generic situation involving forecasts with uniform distributions is shown in Fig. 3a. There are four parameters involved - two means, c_0 and c_1, and two half-widths, w_0 and w_1.[2] Without loss of generality, it is assumed that $c_1 \geq c_0$. It is then straightforward to show that the false alarm rate and the hit rate are given by

$$F = \frac{c_0 + w_0 - t}{2w_0}, \quad H = \frac{c_1 + w_1 - t}{2w_1}, \tag{2}$$

[2] Throughout this paper, the symbols c and w refer to measures of central tendency and half-width, respectively, of the respective distribution. For the case of the gaussian, they coincide with the mean and the standard deviation of the distribution.

where t is the threshold above (below) which a case is classified into class 1 (0).[3]

The equation for the ROC curve follows immediately from (2):

$$H = \frac{w_0}{w_1} \; F \; + \; \frac{\delta c + \delta w}{2w_1} \quad , \tag{3}$$

where $\delta c = c_1 - c_0$ and $\delta w = w_1 - w_0$. Fig. 3b displays the situation. It can be seen that the ROC curve consists of three line segments, with the equation for the middle segment given by (3).

Several observations can be made. First, (3) implies that two models with different means and widths can yield the same ROC curve if they have the same slope and intercept (see Fig. 3b). As such, the ROC curve does not uniquely specify the underlying parameters. In other words, there is a family of underlying distributions that give rise to the same ROC curve. This is a known fact even for more general distributions (Zhou, McClish, Obuchowski 2002).

Second, the length of the vertical segment overlapping the y-axis is determined by two quantities, δc and w_0/w_1. This is sensible since the goodness of the underlying model is determined by both quantities. By contrast, the slope of the middle segment depends only on the ratio of the half-widths (and not δc). As such, the inequality of w_0 and w_1 reflects itself as an asymmetric ROC curve.

Given the analytic expression for the ROC curve (3), it is then possible to compute the area under the curve:[4]

$$AUC = 1 - \frac{1}{8} (\frac{\Delta}{\sqrt{w_0 w_1}})^2 \tag{4}$$

where

$$\Delta = \delta c - (w_0 + w_1). \tag{5}$$

Since $\delta c \leq w_0 + w_1$ for the arrangement displayed in Fig. 3a, it can be seen that increasing δc leads to better performance. Furthermore, decreasing w_0 or w_1 can also yield better performance. In short, model selection based on AUC selects for

[3] The expressions in (2) are specific to Fig. 3a; changing the relative position of c_0 and c_1, or the magnitudes of the widths, yields different expressions.

[4] Again, this equation is specific to the arrangement considered in Fig. 3a.

sharp (i.e., narrow width) and well separated class-conditional distributions. Note that in terms of the underlying distributions, each of the quantities $\delta c, w_0, w_1$ can be interpreted as a performance measure.

As a function of the measure δc, AUC is a parabola. Fig. 3c shows an instance for $w_0 = w_1 = 0.4$ and $w_0 = 0.4,\ w_1 = 0.6$. The AUC curve rises rapidly and then flattens off. It is this nonlinear behavior that explains the appearance of near-one AUC values in practice. For example, in Fig. 3c, as a model improves in terms of δc, its AUC value increases quickly to 0.99 at around $\delta c \sim 0.8$. And the infinity of better models with $\delta c \geq 0.8$ will result in only comparable AUC values, still around 0.99. In other words, the frequent appearance of high AUC values in practice suggests that the corresponding models are all in the "good" range of the AUC curve. One can say that AUC discriminates well between "good" and "bad" models, but not between "good" models, where those adjectives are gauged in terms of the underlying distributions.[5] Similar arguments apply to the performance measures w_0 and w_1; AUC flattens-off for sharper distributions.

3 Triangular Distribution with Unconstrained Support

A better, but still crude, approximation is shown in Fig. 4a. For this case one has

$$F = \frac{1}{2}(\frac{c_0 + w_0 - t}{w_0})^2, \quad H = 1 - \frac{1}{2}(\frac{t - c_1 + w_1}{w_1})^2 \ . \tag{6}$$

The ROC curve is given by

$$H = 1 - \frac{1}{2}(\frac{\Delta - w_0\sqrt{2F}}{w_1})^2 \ , \tag{7}$$

and is shown in Fig. 4b. Evidently, this ROC curve is more realistic than that of the previous section. A common feature, however, is the overlap with the axes.

From the endpoints of the middle segment (Fig. 4b), it follows that the ROC curve is asymmetric if and only if $w_0 \neq w_1$. Specifically, if the concavity is mostly

[5] This is not a problem in model selection, because the standard error of the AUC converges to zero, as AUC approaches 1 (Hanley and McNeil 1983).

to the left, then $w_0 < w_1$. Bowing to the right suggests $w_0 > w_1$. Note that the asymmetry is independent of c_i.

Also, the two extremes of the curves - $F = 0$ and $H = 1$ - convey some useful information as well. Note that if $\delta w = \delta c$, then the right extreme of the curve meets the (1,1) point without overlapping the $H = 1$ line. Similarly, $\delta w = -\delta c$ implies that the left extreme of the curve meets the (0,0) point without overlapping the $F = 0$ axis. Therefore, the amount of overlap of the curve and the two axes is a measure of the distance between the two means *relative to* the difference between the half-widths.

AUC can be computed to be

$$AUC = 1 - \frac{1}{8}(\frac{\Delta}{\sqrt{w_0 w_1}})^4 \tag{8}$$

Like the expression for AUC in the previous case (eq. 4) this expression also displays an affinity for the quantity Δ. Furthermore, noting the quartic power of Δ, in comparison with the quadratic power in (4), it is clear that this AUC is more nonlinear in that it rises faster and has a broader plateau. Fig. 4c displays this quartic dependence. This further flattening of the AUC curve exacerbates AUC's inability to discriminate between good models.

4 Gaussian Distribution

Among the three distributions dealing with unbounded forecast quantities, the Gaussian offers the most realistic approximation. However, the expression for ROC and AUC are not as transparent because of the appearance of certain integrals. The likelihood for the forecasts in the i^{th} class is written as (see Fig 5a)

$$L_i(x) = \frac{1}{\sqrt{2\pi w_i^2}} exp^{-\frac{1}{2}(\frac{x-c_i}{w_i})^2} . \tag{9}$$

Then, (1) implies

$$F = \Phi(\frac{c_0 - t}{w_0}) \ , \ \ H = \Phi(\frac{c_1 - t}{w_1}), \tag{10}$$

where $\Phi(x)$ is the standard normal cumulative distribution,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp^{-\frac{1}{2}z^2} \, dz \ . \tag{11}$$

Eliminating the threshold t from these equations leads to a formal expression for the ROC curve:

$$H = \Phi[\frac{\delta c}{w_1} - \frac{w_0}{w_1}\Phi^{-1}(F)], \tag{12}$$

where Φ^{-1} is defined by $\Phi^{-1}\Phi = 1$. This expression is not too illuminating, but it does allow one to compute some useful quantities. For example, it implies that if plotted on a double-probability paper, the ROC curve will be a straight line with slope w_0/w_1 and intercept $\delta c/w_1$. Note the similarity to eqn. (3) for the case of uniform distributions. It also allows one to compute the slope of the ROC curve to be $L_1(t)/L_0(t)$.[6] Substituting (9) into this expression yields a formula (not shown) that implies that the slope of the ROC curve at its ends is either 0 or ∞. In other words, the ROC curve is always tangent to the axes.

A common error is to assume that a theoretical ROC curve based on gaussian distributions is constrained to obey the canonical ROC behavior, i.e., concave either above or below the diagonal. Although this is true for the symmetric case where $w_0 = w_1$, in general the ROC curve is not strictly concave. It is easy to show that if $w_0 \neq w_1$ then the ROC curve crosses the diagonal at precisely one point (other than the end-points). Proof: The ROC curve will cross the diagonal where $\Phi(\frac{c_0-t}{w_0}) = \Phi(\frac{c_1-t}{w_1})$, i.e., when $\frac{c_1}{w_1} - \frac{c_0}{w_0} = (\frac{1}{w_1} - \frac{1}{w_0})\ t$. This equation has only one nontrivial solution when $w_0 \neq w_1$. The value of F at this crossing point is given by $\Phi(\frac{\delta c}{\delta w})$. Fig. 5b illustrates this cross-over.

This result must be interpreted cautiously. Specifically, it does not imply that an apparently concave empirical ROC curve suggests $w_0 = w_1$. Even if $w_0 \neq w_1$, the ROC curve can still *appear* to be mostly concave (i.e., without a cross over). This is because $\Phi(x)$ is a rapidly increasing function of x. In fact, it is nearly 0 or 1, when x is nearly +2 or -2, respectively. Therefore, a concave empirical ROC curve suggests one of two possibilities: Either $w_0 = w_1$, or $w_0 \neq w_1$ but with $|\frac{\delta c}{\delta w}| \geq\sim 2$

[6] In decision theoretic applications where one seeks an "optimal" decision threshold, this expression is often given to argue for the threshold at which slope=1. However, that choice assumes that the two classes have equal prior probabilities, p_i. p_1 and p_0 are sometimes referred to as the base rate and its complement. The optimal threshold should be the one corresponding to slope = p_0/p_1.

AUC can be computed to be

$$AUC = \Phi(\frac{\delta c}{\sqrt{w_0^2 + w_1^2}}) \ . \tag{13}$$

Again the AUC is a nonlinear function of all the underlying parameters that assess performance - $\delta c, w_0$, and w_1. The functional dependence on the former is shown in Fig 5c. Clearly, the nonlinearity of the curve is present even in this realistic example. Again, two "good" models, with one distinctly superior to the other (e.g., with different values of δc) can have comparable and high AUC values. Eq. (13) also explains why empirical AUC values in practice are often in the 0.9 or higher range. The reason can be traced again to the behavior of $\Phi(x)$. As mentioned previously, modestly large values of x, e.g., 2, correspond to near-one values for Φ.

5 Triangular Distribution with Constrained Support

In some situations the forecast quantity is a probability, calling for distributions that are restricted to that range. The first of the two such distributions considered here is shown in Fig. 6a. This model does assume that the forecasts do span the full range of possibilities (i.e., 0 to 1). In the language of Murphy and Winkler (1987, 1992), the forecasts are assumed to be well-refined. Also note that in this approximation, the only parameters are the two modes, c_0 and c_1.[7]

Three different regions must be considered: $t \leq c_0, c_0 \leq t \leq c_1$, and $t \geq c_1$. Unlike the previous examples, here there exists no region that overlaps with the axes; this is a consequence of the aforementioned assumption about the refinement of the forecasts. The respective ROC curves are

$$H = 1 - \frac{c_0}{c_1}(1-F) \ \ , \ \ H = 1 - \frac{1}{c_1}[1 - \sqrt{(1-c_0) \ F} \ ]^2 \ \ , \ \ H = (\frac{1-c_0}{1-c_1}) \ F \ \ . \tag{14}$$

Note that the ROC curve for the first and third regions are linear, while that of the middle section is not. Fig. 6b displays the ROC curve.

[7] First, note that in this section, c stands for the mode (not mean) of the distribution. Also, the widths of the distributions are not independent quantities. The mean is given as $(1+c)/3$, and the variance as $(1-c+c^2)/18$.

From an expression of the slope it follows that a symmetric ROC curve implies $c_0 + c_1 = 1$. Any other combination of c_0 and c_1 will result in an asymmetric curve. It is also easy to show that there does not exist a cross-over; a nontrivial curve is either always above or always below the diagonal. It also follows that the ROC curve will bow to the left if $c_0 \sim 0.5$, and to the right if $c_1 \sim 0.5$.

Finally, the AUC can be computed to be

$$\mathrm{AUC} = \frac{1}{2} + \frac{1}{2}(c_1 - c_0) - \frac{(c_1 - c_0)^3}{6c_1(1 - c_0)}. \tag{15}$$

First, note that AUC depends on two independent quantities - $(c_1 - c_0)$ and $c_1(1 - c_0)$. For small values of the former, i.e., low performance, the first two terms in (10) dominate the expression, leading to a linear dependence on $c_1 - c_0$. However, for higher performance values, the last term begins to penalize (because of the negative sign) AUC in a nonlinear fashion. This nonlinear penalty again leads to a flattening of the AUC curve for better models. Fig. 6c displays AUC as a function of the measure $c_1 - c_0$. The reason the flattening is not evident in this figure is that the simplicity of the model does not allow high values of AUC. In fact, according to (15) the highest allowed value of AUC is only 5/6 or 0.83.

6 Beta Distribution

A more realistic likelihood for probabilistic forecasts is the beta distribution

$$L_i(x) = \frac{1}{B(a_i, b_i)} x^{a_i - 1} \, (1 - x)^{b_i - 1} \ , \tag{16}$$

where $B(a_i, b_i) = \int_0^1 \ x^{a_i - 1} \ (1 - x)^{b_i - 1}$. An instance is shown in Fig. 7a. Note that in this example, the distributions themselves are possibly asymmetric (or skewed). If a_i, b_i are integers, then one can write $B(a_i, b_i) = \frac{(a_i - 1)!(b_i - 1)!}{(a_i + b_i - 1)!}$. The mean, mode, and variance can be computed by

$$c_i = \frac{a_i}{a_i + b_i} \ , \ m_i = \frac{a_i - 1}{a_i + b_i - 2} \ , \ w_i^2 = \frac{a_i \ b_i}{(a_i + b_i)^2 \ (a_i + b_i + 1)}. \tag{17}$$

In this case, given that the likelihoods are written in terms of a_i and b_i, it is natural to ask what combination of these quantities constitutes a measure of performance.

From a decision theoretic point of view, the natural quantity is $L_1(x)/L_0(x)$, and this ratio is a function of $(a_1 - a_0)$ and $(b_1 - b_0)$.[8] Therefore, these two differences are natural measures of performance. Note that each of these measures depends on both c and w. For example,

$$a_1 - a_0 = (c_1 - c_0) + [\frac{c_1^2(1-c_1)}{w_1^2} - \frac{c_0^2(1-c_0)}{w_0^2}] \ . \tag{18}$$

The corresponding ROC curve is shown in Fig. 7b. The analytic expressions for F and H are not illuminating, but the slope of the ROC curve is

$$\text{slope}(t) = \frac{B(a_0, b_0)}{B(a_1, b_1)} \ t^{a_1 - a_0} \ (1-t)^{b_1 - b_0}. \tag{19}$$

A symmetric ROC curve requires the product of the slopes at the end points of the curve to be inversely proportional. And for that to occur one must have $(a_1 + b_1) = (a_0 + b_0)$. It follows that the ROC curve is symmetric if $(a_0 + b_0) = (a_1 + b_1)$, which in terms of the means and variances translates to

$$\frac{c_1(1-c_1)}{w_1^2} = \frac{c_0(1-c_0)}{w_0^2}. \tag{20}$$

An apparent asymmetry in an empirical ROC curve, then, implies that this equation is violated. Note that in the symmetric ROC case, the two performance measures $a_1 - a_0$ and $b_1 - b_0$ differ only in sign.

It also follows that a cross-over occurs when $a_1 > a_0$ and $b_1 > b_0$, because the slopes at the two extremes are then both less than 1. These two inequalities together imply

$$\frac{c_1(1-c_1)}{w_1^2} > \frac{c_0(1-c_0)}{w_0^2}. \tag{21}$$

Compare with (20) which is the condition for a symmetric ROC curve. The quantity $c(1-c)/w^2$ determines both the symmetry and the cross-over of the ROC curve. The cross-over is displayed in Fig. 7b.

The expression for AUC is somewhat tedious to derive, but for the case of integer parameters can be computed as

$$AUC = \frac{1}{(a_1 + b_1)} \frac{1}{B(a_0, b_0)} \sum_{k=1}^{a_1} \frac{B(a_0 + a_1 - k, b_0 + b_1 + k - 1)}{B(a_1 - k + 1, b_1 + k)} \ . \tag{22}$$

[8] Technically, this expression should be multiplied by the ratio of the respective prior probabilities as well. They are neglected here because they are not functions of x.

Fig. 7c displays a plot of AUC as a function of the measure $(a_1 - a_0)$ for a few different values of the parameters. The nonlinearity is now evident when AUC reaches near-1 values.

7 Summary and Conclusion

Several models are examined for the purpose of explicitly illustrating some features of ROC curves and the area under the curve (AUC). The findings aid in interpreting the shape of the ROC curve in terms of the parameters defining the class-conditional distributions of the forecast quantity. In addition to providing a pedagogical exposition of ROC analysis, the work also offers some guidance for interpreting ROC curves and AUC. The guidance is based on only the models examined here. As such, the generality of the results is not assured by any means. Nevertheless, all of the examples shown in Fig. 2 are found to be completely consistent with the findings here. The following statements should be interpreted only as qualitative guidance. More quantitative statements are found in the text.

For unbounded forecasts, an asymmetric ROC curve suggests unequal widths for the underlying distributions. If the class with the larger mean is labeled as 1, then a concavity to the top suggests $w_0 > w_1$, and concavity to the bottom suggests $w_0 < w_1$. In other words, in attempting to explain any asymmetry in an empirical ROC curve, it is advisable to examine the widths of the underlying distributions. The amount of overlap with the axes is also a measure of the difference in the widths. The crossing of the diagonal by an ROC curve suggests that the quantity $|\frac{\delta c}{\delta w}|$ is smaller than some critical value. For example, if the distributions are gaussian, then that critical value is approximately 2.

For bounded forecasts, the distributions examined here do not generate an overlap with the axes. The existence of a significant overlap in an empirical ROC plot suggests that the underlying distributions are different from the ones examined here in some significant way. The symmetry and cross-over of ROC are determined by a combination of means and variances, e.g. eqn. (20).

For both bounded and unbounded forecasts, AUC increases nonlinearly with respect to natural measures of forecast quality derived from parameters of the underlying distributions. Moreover, in the examples considered here, the more realistic models display more of this nonlinearity. The nonlinearity is such as to reduce the effectiveness of AUC in assessing performance, as performance increases. As such, the frequent occurrence of near-one AUC values observed empirically is an indication that many forecasts are of "reasonable" quality.

Acknowledgement

The author is grateful to Rich Caruana for invaluable discussions and a reading of an early version of this article.

Figure Caption

Figure 1. A generic situation involving a forecast of two classes.

Figure 2. Examples of ROC curves representing different levels of performance quality. The diagonal line corresponds to random forecasts (i.e., poor performance), while the curves away from the diagonal represent higher levels of performance. The following features are noted: Symmetric ROC curves (a), symmetric and asymmetric curves, also overlapping one axis (b), and extremely concave curves (c).

Figure 3. The schematics of uniform class-conditional distributions (top), the corresponding ROC curve (middle), and the AUC curve as a function of $\delta c = c_1 - c_0$.

Figure 4. Same as Fig. 3, but for triangular distributions over unbounded forecasts.

Figure 5. Same as Fig. 3, but for Gaussian Distributions.

Figure 6. Same as Fig. 3, but for bounded (e.g., probabilistic) forecasts.

Figure 7. Same as Fig. 3, but for beta distributions. The corresponding parameters are $b_0 = 2,\ b_1 = 3,\ a_0 = 2$, with a_1 taking values 2, 3, 4, 5 (from top to bottom).

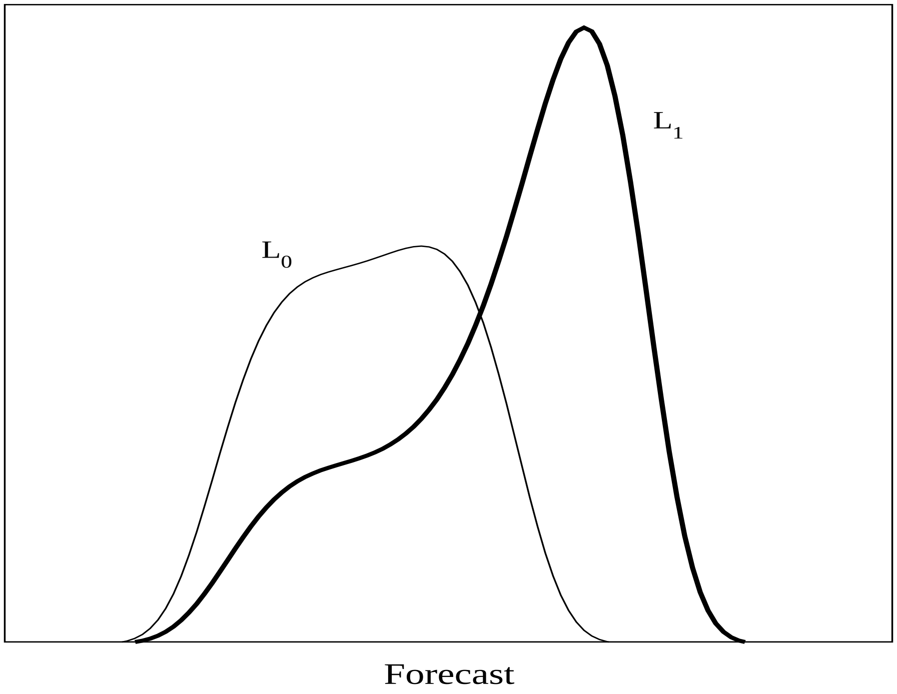


Figure 1.

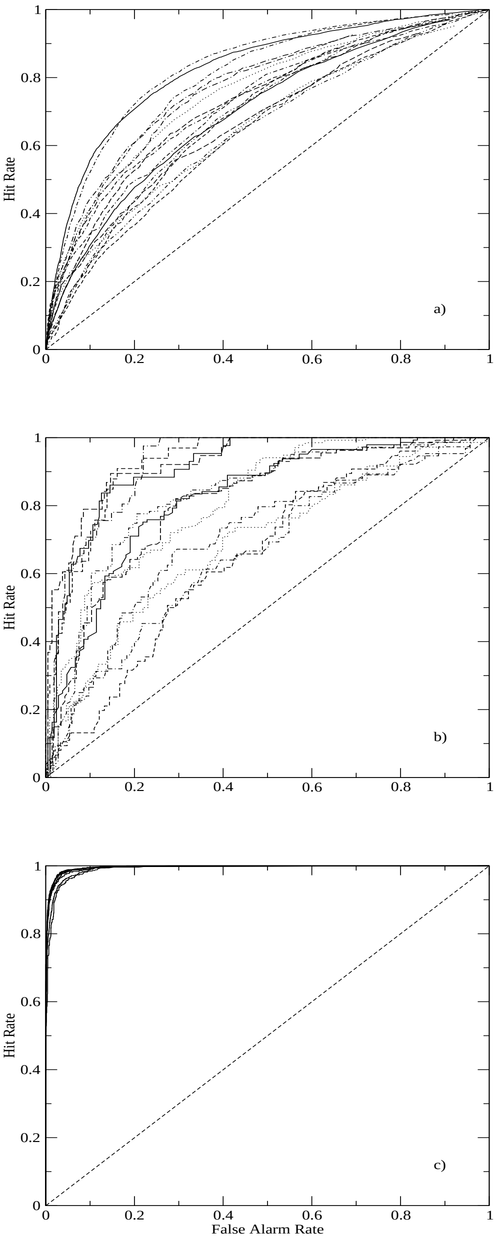


Figure 2.

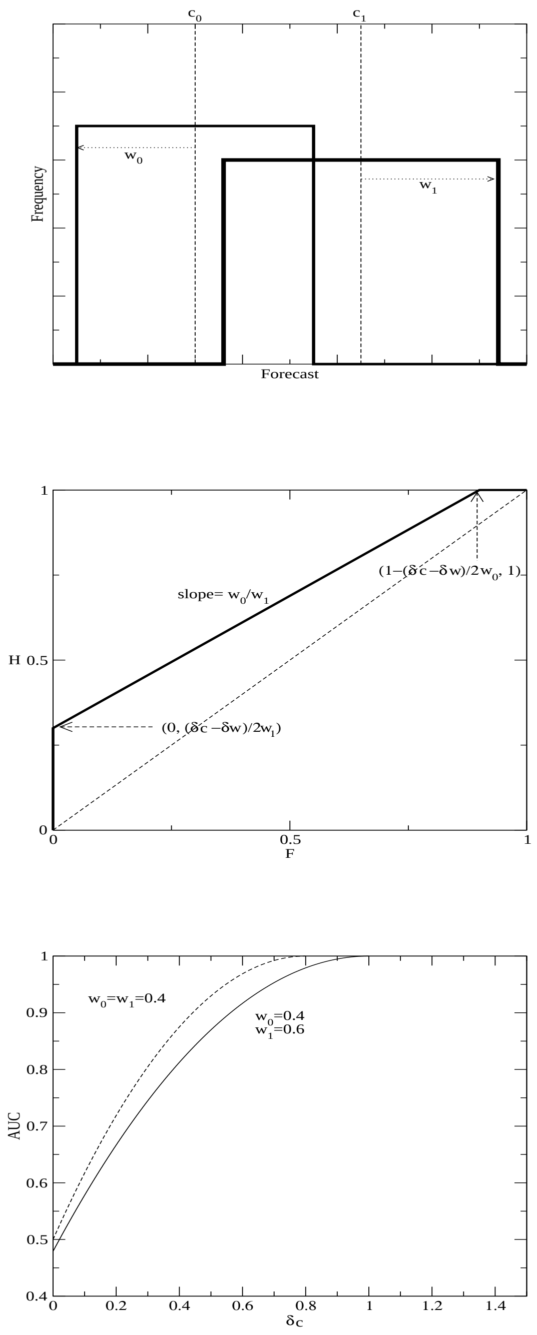

c_0
c_1
w_0
w_1
Frequency
Forecast
1
H 0.5
0
0
0.5
1
F
slope= w_0/w_1
$(1-(\delta c-\delta w)/2w_0, 1)$
$(0, (\delta c-\delta w)/2w_1)$
1
0.9
0.8
0.7
0.6
0.5
0.4
AUC
0
0.2
0.4
0.6
0.8
1
1.2
1.4
δ_c
w_0=w_1=0.4
w_0=0.4
w_1=0.6

Figure 3.

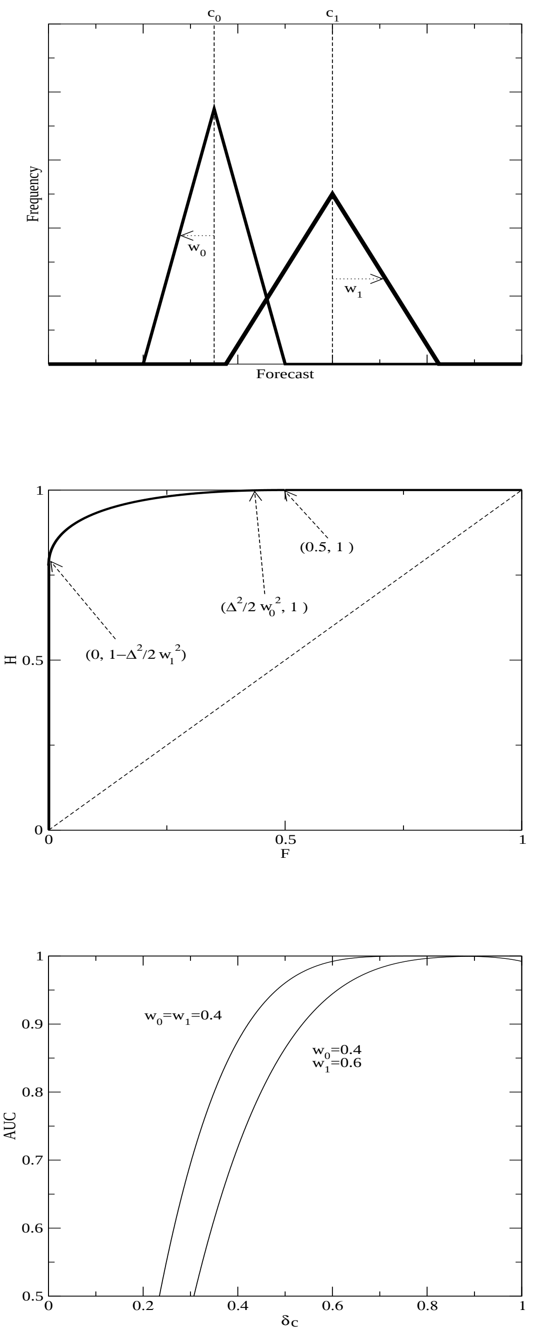

c_0
c_1
Frequency
w_0
w_1
Forecast
1
0.5
0
H
0
0.5
1
F
(0.5, 1)
$(\Delta^2/2\,w_0^2,\ 1\ )$
$(0,\ 1-\Delta^2/2\,w_1^2)$
1
0.9
0.8
0.7
0.6
0.5
AUC
0
0.2
0.4
0.6
0.8
1
δ_c
w_0=w_1=0.4
w_0=0.4
w_1=0.6

Figure 4.

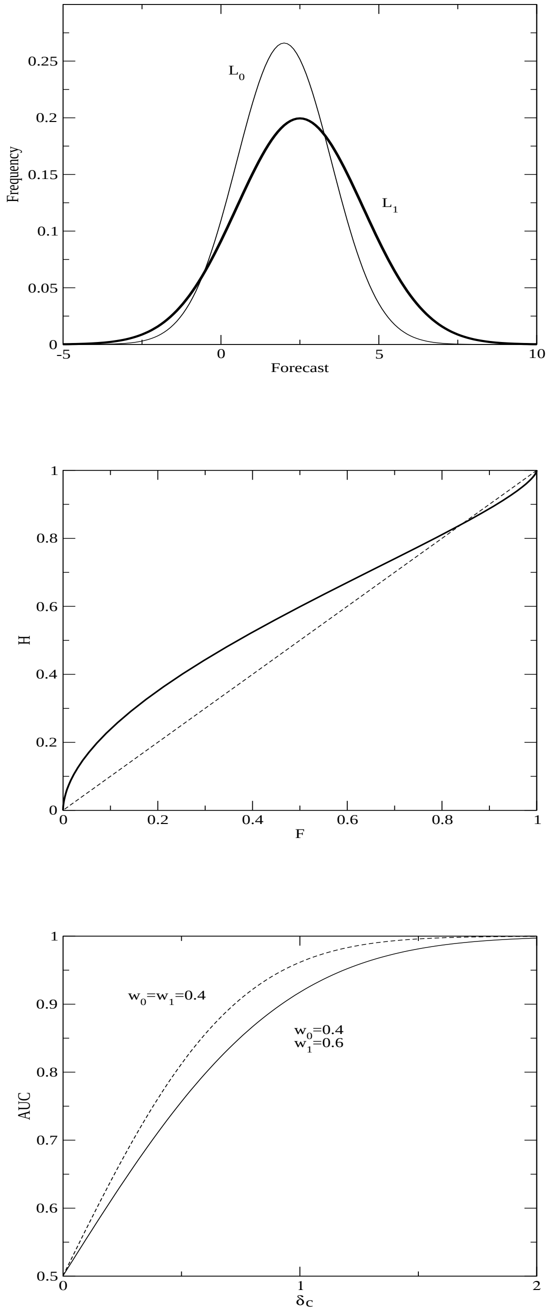
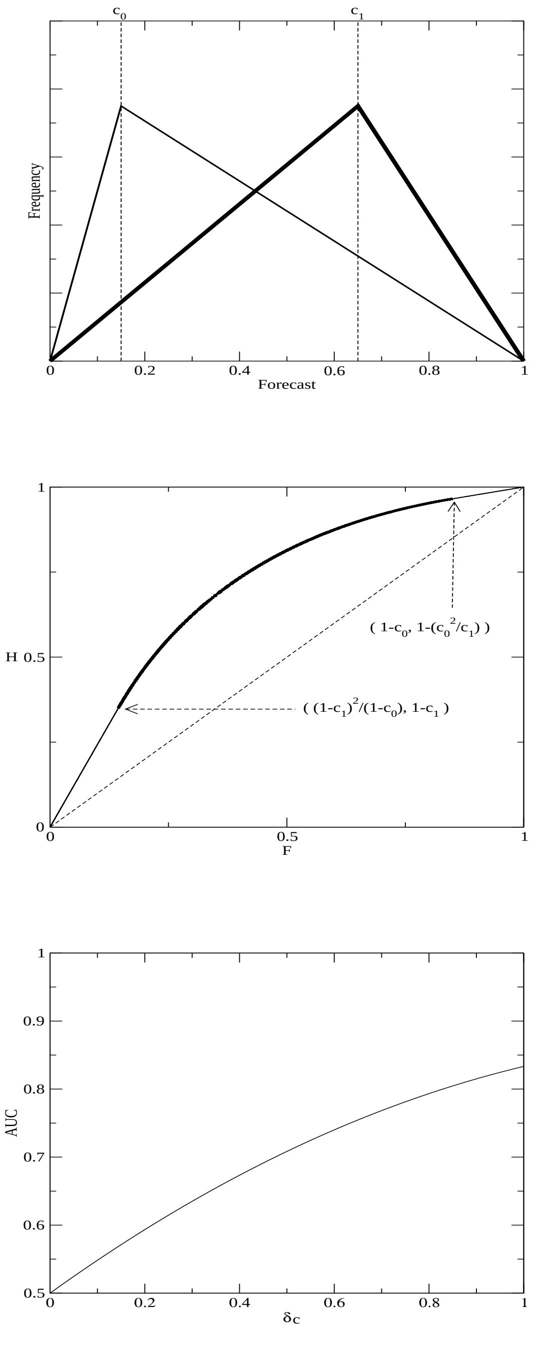

0.25
0.2
0.15
0.1
0.05
0
Frequency
L_0
L_1
-5
0
5
10
Forecast
1
0.8
0.6
0.4
0.2
0
H
0
0.2
0.4
0.6
0.8
1
F
1
0.9
0.8
0.7
0.6
0.5
AUC
w_0=w_1=0.4
w_0=0.4
w_1=0.6
0
1
2
δ_c

Figure 5.



c_0
c_1
Frequency
0
0.2
0.4
0.6
0.8
1
Forecast
1
H 0.5
0
0
0.5
1
F
$(1\text{-}c_0, 1\text{-}(c_0^2/c_1))$
$((1\text{-}c_1)^2/(1\text{-}c_0), 1\text{-}c_1)$
1
0.9
0.8
0.7
0.6
0.5
AUC
0
0.2
0.4
0.6
0.8
1
δ_c

Figure 6.

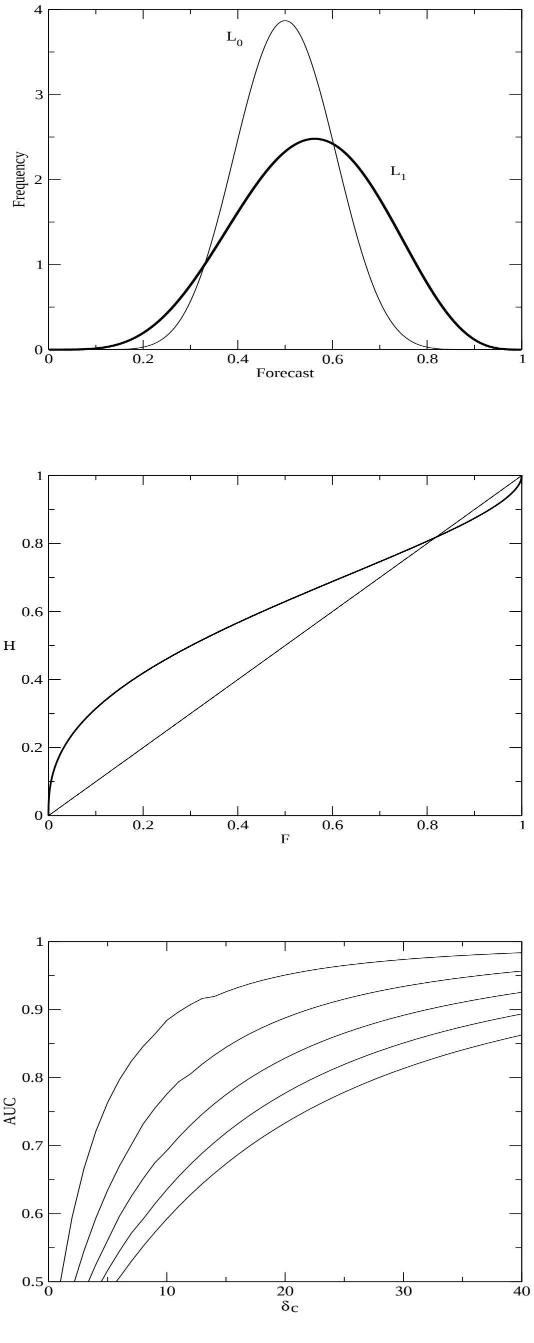

L_0
L_1
Frequency
Forecast
0
1
2
3
4
0
0.2
0.4
0.6
0.8
1
H
F
0
0.2
0.4
0.6
0.8
1
AUC
δ_c
0.5
0.6
0.7
0.8
0.9
1
0
10
20
30
40

Figure 7.

References

Coffin M., S. Sukhatme, 1997: Receiver operating characteristic studies and measurement errors. *Biometrics* **53**, 82337.

Dorfman, D.D., K. S. Berbaum, C. E. Metz, R. V., Lenth, J. A. Hanley, and H. Abu Dagga, 1997: Proper receiver operating characteristic analysis: The bigamma model. *Academic Radiology*, **4**, 138-149.

Dorfman, D.D. and Alf, E. Jr., 1969: Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals". *J. Math. Psychol.* **6**, 487-496.

Drton, M., C. Marzban, P. Guttorp, and J. T. Schaefer, 2003: A Markov Chain Model of Tornadic Activity. *Mon. Wea. Rev.*, **131**, 2941-2953.

Hanley J. A., and B. J. McNeil, 1982: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 2936.

Hanley J. A. and B. J. McNeil, 1983: A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology*, **148**, 839-43.

Harvey, L. O. Jr., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992. Application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863-883.

Marzban, C., S. Leyton, B. Colman, 2003: Nonlinear Post-processing of Model Output: Ceiling and Visibility. NWS/COMET report.
http://www.nhn.ou.edu/˜marzban/comet_report1.pdf
http://www.nhn.ou.edu/˜marzban/comet_talk.pdf

Marzban, C., and A. Witt, 2001: A Bayesian Neural Network for Hail Size Prediction. *Wea. Forecasting*, **16**, 600-610.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291303.

Mason, S. J., and N. E. Graham, 1999: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels. *Wea. Forecasting*, **14**, 713-725.

Metz, C. E., B. A. Herman, J. H., Shen, 1998: Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, **17**, 1033-1053.

Miettinen, K., 1999: Nonlinear Multiobjective Optimization, Kluwer Academic, Boston, 300 pp.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435-455.

———, and ———-, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Mylne, K. R., 1999: The use of forecast value calculations for optimal decision making using probability forecasts. Preprints, 17th Conf. on Weather Analysis and Forecasting, Denver, CO, Amer. Meteor. Soc., 235239.

Richardson, 2000: Applications of costloss models. Proc. Seventh ECMWF Workshop on Meteorological Operational Systems, Reading, United Kingdom, ECMWF, 209213.

———, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 24732489.

Shapiro, D. E., 1999: The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, **8**, 113-134.

Stephenson, D. B., 2000: Use of the “Odds Ratio” for Diagnosing Forecast Skill. *Wea. Forecasting*, **15**, 221-232.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, NY. 467 pp.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteorol. Appl.*, **8**, 209-219.

Zhou, X-H., D. K. McClish, and N. A. Obuchowski, 2002: *Statistical Methods in Diagnostic Medicine.* John Wiley & Sons, pp. 464.

Zou, K. H., 2001: http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html