On the Uniqueness of Gandin and Murphy's Equitable Performance Measures

Caren Marzban[1,2,3] & V. Lakshmanan[1,2]

[1] National Severe Storms Laboratory, Norman, OK 73069
[2] Cooperative Institute for Mesoscale and Meteorological Studies, and
[3] Department of Physics, University of Oklahoma, Norman, OK 73019

Abstract

Gandin and Murphy have shown that if a skill score is linear in the scoring matrix, and if the scoring matrix is symmetric, then in the 2-event case there exists a unique, "equitable" skill score, namely the True Skill Score (or Kuipers' performance index). As such, this measure is treated as preferable to other measures because of its equitability. However, in most practical situations the scoring matrix is not symmetric due to the unequal costs associated with false alarms and misses. As a result, GM's considerations must be re-examined without the assumption of a symmetric scoring matrix. In this note, it will be proven that if the scoring matrix is nonsymmetric, then there does not exist a unique performance measure, linear in the scoring matrix, that would satisfy *any* constraints of equitability. In short, there does not exist a unique, equitable skill score for 2-category events that have unequal costs associated with a miss and a false alarm.

1 Introduction

Performance measures or skill scores are often required to be "equitable" in that their use must not induce the forecasters to make forecasts that differ from their best judgments. Gandin and Murphy [3], hereafter referred to as GM, considered measures that are linear in the scoring matrix, and derived the constraints that must be placed on the scoring matrix in order to assure the equitability of the measure. In the 2-event case, if in addition to the derived constraints the scoring matrix is assumed symmetric, then the number of constraints is equal to the number of elements of the scoring matrix. This, in turn, allows for the determination of a unique scoring matrix and, consequently, a unique, equitable measure - the True Skill Score (or

Kuipers' performance index, among other names). Gerrity [4] expands on the multi-category considerations of GM and finds a closed formula for a symmetric, equitable scoring matrix in terms of the marginal probabilities of the various categories.

However, in most practical situations the scoring matrix is not symmetric. This occurs not only when the two events have different a priori (climatological) probabilities, but also when the cost (or loss) associated with a false alarm is different from that of a miss. Therefore, to check for the existence of a unique, equitable measure GM's analysis must be re-examined without assuming that the scoring matrix is symmetric. It is, in fact, possible to generalize further: GM also assume that constant forecasts of two events must be assigned equal scores (e.g. zero). However, this assumption is too restrictive and so it, too, can be relaxed (see the Discussion section).

Specifically, GM asked the following question: what scoring matrix will yield a performance measure, S, satisfying the following three constraints:

$$S(\text{constant forecasts}) = S(\text{random forecasts}) = \alpha,$$

and

$$S(\text{perfect forecasts}) = \beta,$$

where α and β are constants defining a scale for the score? They defined such a measure as "equitable" and showed that in the 2-event case if the scoring matrix is further assumed to be symmetric, then there is a unique measure that satisfies these constraints. The uniqueness of the solution may be anticipated based on the number of constraints (3) and unknowns (3), but it is not automatically implied. Unless these constraints yield an independent set of simultaneous linear equations, an unique solution will not exist.

Relaxing both the assumption of a symmetric scoring matrix and the equality of constant forecasts, the equitability constraints translate to 4 equations for 4 unknowns. As such, one might anticipate a unique solution. However, in this note it will be argued that without the assumption of a symmetric scoring matrix (or any other constraint on the scoring matrix) there does not exist a unique measure satisfying these or any set of (four) constraints placed on S. In other words, in most practical situations there is no unique, equitable measure. Or said differently, there exists no definition of equitability that would yield a unique measure of performance linear in the scoring matrix.

2 Preliminaries

Many measures of performance are defined in terms of the elements of the contingency table. For dichotomous forecasts of two events, labeled as 0 and 1, the contingency table, C, is

$$\begin{aligned} \mathrm{C} = \begin{pmatrix} N_{00} & N_{01} \\ N_{10} & N_{11} \end{pmatrix} &= \begin{pmatrix} \text{\# of 0's predicted as 0} & \text{\# of 0's predicted as 1} \\ \text{\# of 1's predicted as 0} & \text{\# of 1's predicted as 1} \end{pmatrix}, \\ &= \begin{pmatrix} . & \text{false alarms} \\ \text{misses} & \text{hits} \end{pmatrix}. \end{aligned}$$

Note that $N_{i0} + N_{i1}$, represented by $N_{i.}$, is simply the sample size of the i^{th} *observation*, and $N_{0i} + N_{1i} \equiv N_{.i}$ is the number of *forecasts* of the i^{th} type $(i = 0, 1)$.

The joint probability of forecasts, f, and observations, o, is [6]

$$P_{ij} = N_{ij}/N_{..} \quad ,$$

where $N_{..}$, is the total sample size. The two relevant conditional probabilities are the probability of assigning (forecasting) an observed event from the i^{th} class $(i = 0, 1)$ into the j^{th} class $(j = 0, 1)$

$$p(f = j|o = i) = N_{ij}/N_{i.} \equiv Q_{ij} \quad ,$$

and the belief that a class-i event was assigned to class j,

$$p(o = i|f = j) = N_{ij}/N_{.j} \equiv B_{ij} \quad .$$

The matrices P, Q, and B are sometimes called the performance matrix, the percent confusion matrix, and the belief matrix, respectively.

The class-conditional risks, $R_i(C)$, are given by

$$R_i(C) = \sum_j L_{ij} Q_{ij} \quad ,$$

where L_{ij} are the elements of the loss matrix, and Bayes risk [1, 7] is then defined as

$$R(C) = \sum_i R_i(C) P_i = \sum_{i,j} L_{ij} Q_{ij} P_i \quad , \tag{1}$$

where $P_i = N_{i.}/N_{..}$ are the a priori (i.e. climatological) probabilities.

Bayes risk and the loss function are quantities central to the analysis of performance, and it is evident that the quantity that GM call the expected score is in fact equal to Bayes risk, and what they call the scoring matrix is equal to (the transpose) of the loss matrix. In other words, $R = S$, if the loss matrix L is identified with (the transpose of) the scoring matrix. [1]

3 Theorem

The question asked by GM can be asked at a more general level: what loss matrix yields a measure satisfying constraints of the form

$$R(C^{(k)}) = \alpha^{(k)}, \quad k = 0, ..., n-1 \quad , \tag{2}$$

where n is the number of constraints, and $C^{(k)}$ and $\alpha^{(k)}$ are the contingency table and the value of the measure associated with the k^{th} constraint?

[1] The contingency table in this article is the transpose of that of GM.

Since the (2×2) loss matrix has 4 degrees of freedom one would require 4 equations in order to uniquely solve for L_{ij}. For example,

$$R(f = \text{constant``0''}) = \alpha^{(0)} \quad , \quad R(f = \text{constant``1''}) = \alpha^{(1)},$$
$$R(f = \text{random}) = \alpha^{(2)} \quad , \quad R(f = \text{perfect}) = \alpha^{(3)} \quad ,$$

constitute four such constraints. The special case $\alpha^{(0)} = \alpha^{(1)} = \alpha^{(2)}$ is the one considered by GM.

More generally, however, equation (1) and (2) imply

$$\sum_{i,j} L_{ij} Q^{(k)}_{ij} P_i = \alpha^{(k)}, \quad k = 0, ..., n \ ,$$

where $Q^{(k)}_{ij}$ is the percent confusion matrix for the k^{th} constraint, and n is the number of constraints. For $n = 4$, this yields

$$\left| \begin{array}{cccc} Q^{(0)}_{00} P_0 & Q^{(0)}_{01} P_0 & Q^{(0)}_{10} P_1 & Q^{(0)}_{11} P_1 \\ Q^{(1)}_{00} P_0 & Q^{(1)}_{01} P_0 & Q^{(1)}_{10} P_1 & Q^{(1)}_{11} P_1 \\ Q^{(2)}_{00} P_0 & Q^{(2)}_{01} P_0 & Q^{(2)}_{10} P_1 & Q^{(2)}_{11} P_1 \\ Q^{(3)}_{00} P_0 & Q^{(3)}_{01} P_0 & Q^{(3)}_{10} P_1 & Q^{(3)}_{11} P_1 \end{array} \right| \left| \begin{array}{c} L_{00} \\ L_{01} \\ L_{10} \\ L_{11} \end{array} \right| = \left| \begin{array}{c} \alpha^{(0)} \\ \alpha^{(1)} \\ \alpha^{(2)} \\ \alpha^{(3)} \end{array} \right| .$$

Noting the identity $Q^{(k)}_{i0} + Q^{(k)}_{i1} = 1, \forall k$, it is then straightforward to show that the determinant of the 4×4 matrix is zero.

Consequently, the system of 4 equations and 4 unknowns is under-determined. As such, for a general scoring matrix the True Skill Score is no longer uniquely equitable. Therefore, in practical cases where the scoring matrix has no particular symmetry, there exists no unique, equitable score. Note that this is true for any definition of equitability based on the four constraints of the aforementioned type (equation 2).

Another family of risk functions, or performance measures, can be defined in terms of the belief matrix, B (instead of percent confusion matrix, Q) [2]:

$$R' = \sum_{i,j} L_{ij} B_{ij} P_i \quad ,$$

but it can be shown that the above theorem applies to R' as well, because $B_{0i} + B_{1i} = 1$.

4 Discussion

It is important to point out that the above result is not contained in GM's findings. The question asked in this article does partially reduce to that of GM in the limit $\alpha^{(0)} = \alpha^{(1)} = \alpha^{(2)}$. However, the number of equations and unknowns in this article (i.e. 4) is different from that of GM (i.e. 3), and so there is no smooth limit in which the two questions are related.

The findings herein generalize GM's results in that several assumptions made by GM are no longer invoked. Specifically, two independent assumptions are made by GM: 1) $\alpha^{(0)} = \alpha^{(1)}$, and 2) $L_{ij} = L_{ji}$ (i.e. that the loss matrix is symmetric). [2]

GM motivate the first assumption by arguing that it precludes the forecaster from over- or under-forecasting all observations as one event (i.e. $f = 0$) or the other event (i.e. $f = 1$). It is true that such an assumption would be necessary if the performance measure, R, behaved like the dotted curve in Figure 1, wherein the behavior of a measure is plotted against a quantity, Q, that at its extremes coincides with $R(f = 0)$ and $R(f = 1)$. Two examples of Q are 1) the percentage of class 1 forecasts that the forecaster has issued, and 2) the decision threshold that a forecaster must place on a probabilistic forecast in order to dichotomize the forecasts. More generally, however, the plot of a measure would have a shape similar to the dashed line in Figure 1, wherein there is a local maximum marking the "optimal performance." Indeed, in the case of the second Q example it has been shown

[2] As pointed out by GM, $\alpha^{(0)} = \alpha^{(1)}$ implies $\alpha^{(0)} = \alpha^{(1)} = \alpha^{(2)}$.

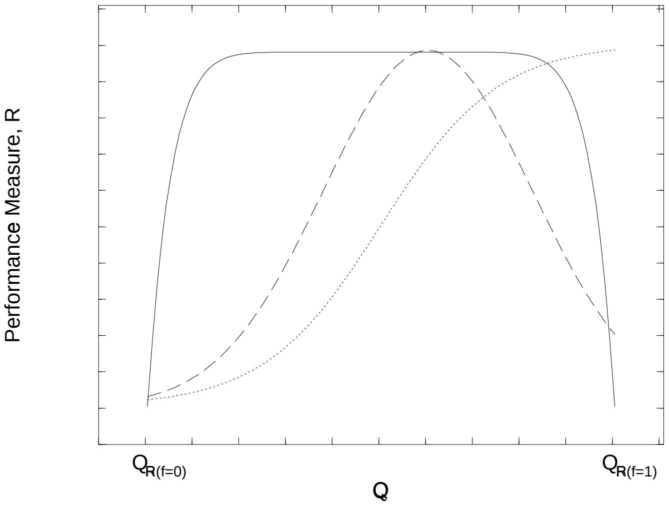

Figure 1: The generic behavior of R as a function of Q.

that most measures behave as such [5]. Of course, the use of such a measure will still influence the forecaster's judgment via his attempts to reach the optimum critical point; however, the asymptotic limits of the measure are no longer of any concern. Hence, it is not necessary for the constant-forecast scores to be equal. It is important to emphasize that a truly "equitable" measure would behave like the solid curve in Figure 1 in that its use would not induce the forecaster to significantly affect her judgement.

GM motivate the second assumption by emphasising the issue of accuracy, and requiring the generality of the ultimate results. However, in practice it is much more likely that the misclassification costs are unequal for the two classes. For example, the cost associated with missing a tornado is rarely the same as that associated with making a false tornado forecast. Therefore, the second assumption is far too stringent to be of any utility in most practical situations. For these reasons, neither assumption was invoked here.

5 Acknowledgements

C.M. is grateful to Gregory Heath of M.I.T. Lincoln Laboratory for an invaluable discussion on Bayes risk. Partial support was provided by the FAA and the NWS/OSF.

6 References

1. Bishop, C. M., 1996: *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, pp. 482.

2. Duda, R. O., and P. E. Hart, (1973): *Pattern Classification and Scene Analysis.* Wiley, New York, pp. 482.

3. Gandin, L. S., and A. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.

4. Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709-2712.

5. Marzban, C., 1998: Scalar measures of performance in rare-event situations. To appear in the September issue of *Wea. Forecasting.*

6. Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

7. Ripley, B. D., 1996: *Pattern Recognition and Neural Networks.* Cambridge: University Press.