SCORING THE FORECASTER BY MEAN RESULTING PAYOFF OF A DISTRIBUTION OF DECISION PROBLEMS

David B. Rosen
Center for Biomedical Modeling Research
University of Nevada School of Medicine
77 Pringle Way, H1-166 WMC
Reno, Nevada 89520 USA
Internet: `rosen@unr.edu` (or `d.rosen@ieee.org`)

Presented at the Thirteenth International Workshop on Maximum Entropy and Bayesian Methods, Santa Barbara, California, 2–5 August 1993.

Revised 15 November 1993.

Submitted for inclusion in *Maximum Entropy and Bayesian Methods. Proceedings of the Thirteenth International Workshop.* G. Heidbreder, ed. Kluwer, Dordrecht, The Netherlands, 1994.

SCORING THE FORECASTER BY MEAN RESULTING PAYOFF OF A DISTRIBUTION OF DECISION PROBLEMS

David B. Rosen
Center for Biomedical Modeling Research
University of Nevada School of Medicine
77 Pringle Way, H1-166 WMC
Reno, Nevada 89520 USA
Internet: `rosen@unr.edu` (or `d.rosen@ieee.org`)

ABSTRACT. We present a new way to understand and characterize the choice of scoring, objective, or loss function for a model, machine, or expert estimating (i.e. predicting) the probabilities of events or outcomes (such as class membership). The ultimate value of a probability estimate lies in the actual payoff (utility) accruing to those who *use* this information to make a decision. We allow that we often cannot specify with certainty that the estimate will be used in a particular decision problem, characterized by a particular decision-outcome payoff matrix (cost schedule), and thus by a particular decision threshold. Instead, we consider the more general case of a distribution over such matrices. The proposed scoring function is the *expectation*, with respect to this distribution, of the payoff that will actually be received. Square-error loss (or Brier score) and log-likelihood (or various entropic measures) arise from specific examples of such distributions, and even common single-threshold measures such as the misclassification rate obtain from degenerate special cases. Our scoring functions are always "honesty-rewarding," or maximized in expectation when the estimate equals the true outcome probability. There is still no single "universal" choice of scoring function appropriate in all cases, but specific choices correspond to specific assumptions about the probability that the predictions will be thresholded in various ways when used to make a decision.

1. Introduction

One of a set of events

$$\{y_i\}$$

will occur in an experiment or observation. We consider a forecaster F who gives us his prediction of the experiment's outcome as corresponding probability estimates

$$\{q_i\} \qquad \left(q_i \geq 0, \quad \sum_i q_i = 1\right)$$

i.e.

$$\Pr(y_i|F) = q_i.$$

We make no assumptions about how the forecaster arrived at such predictions.

If event y_i occurs, we wish to assign a figure of merit, i.e. a score or *value*

$$v_i(q_i)$$

to the forecaster's prediction. (v_i can be considered the negative of a corresponding loss function.) What form of value functions v_i should we use?

A classic example of this problem is the weather forecaster who states that "the probability of rain today is q_r". The two possible events corresponding to rain and shine are

$$y_r, y_s,$$

and the forecaster's corresponding predictions are

$$q_r, \qquad q_s = 1 - q_r.$$

We seek functions v_r and v_s in order to assign the value of today's prediction by the weatherman as

$$v_r(q_r) \text{ if it rains,} \qquad v_s(q_s) \text{ if it shines.}$$

Then the total value of this forecaster's predictions over many days is the sum of such daily scores. This sum might be compared to that of another forecaster.

We consider the primary purpose of a forecaster's prediction to be to help us make better decisions. In the case of weather, examples would include deciding whether to water the lawn, bring an umbrella, or cancel an outdoor social event. The ultimate *value* of prediction $\{q_i\}$ is the actual utility (gain or loss) accruing to those who use it to make such decisions. Ideally, we would survey the *consumers* of the forecast, to find their actual losses or gains using forecaster's predictions.

In general, we would use whatever information we have about the probability that the prediction will be used for various decision problems, each such decision problem being characterized by a decision payoff matrix. Given such information, we can calculate the value functions v_i as the mean payoff over these decision problems.

In the next section, we consider the single decision problem as a preliminary. Then in further sections we show how to generalize this to the mean payoff over a continuum of decision problems, obtaining our main result. We relate these scoring functions obtained by considering the utility of predictions to their user, to those obtained by considering the utility of payment to an expert supplying the predictions, and show that commonly used scoring functions arise as special cases. Throughout this paper, we consider dichotomous (two-class) outcomes and dichotomous decisions.

2. Single Decision Based on $\{q_i\}$

Suppose we make a decision d_j ($j = 0$ or 1) and then the experiment is performed and event (outcome or class) y_i ($i = 0$ or 1) is observed. Our net gain in such a decision-observation trial is given by $\{b_{ij}\}$, defining a payoff matrix B as follows:

	d_0	d_1
y_0	b_{00}	b_{01}
y_1	b_{10}	b_{11}

We assume the forecaster's predictions can be taken literally as probability estimates, and thus be plugged into decision theory to make our decision. [1]

[1] To relax this restriction, the "forecaster" could include both a generator of *uncalibrated* predictions, along with an external calibration of these predictions mapping them to a probability scale. This calibration would typically have been estimated from a large sample of predictions and corresponding true outcomes.

We make decision d_{j^*}, where

$$j^* \equiv D(q_1, B) = \mathrm{argmax}_j \sum_i q_i b_{ij}$$

Our payoff in this decision is

$$\beta_i(q_1, B) = b_{i,D(q_1,B)}, \tag{1}$$

which is our actual net gain when we take F at his word in making this decision.

In order to help evaluate D, define

$$a_0 = b_{00} - b_{01}, \qquad a_1 = b_{11} - b_{10}.$$

Then

$$D(q_1, B) = \Theta(q_1 a_1 - q_0 a_0),$$

where $\Theta(x)$ is the unit step function (0 if $x < 0$, 1 otherwise).

For present purposes, we can ignore trivial cases where a_0, a_1 have opposite sign (decision constant, independent of $\{q_i\}$) or both are zero (payoff independent of decision). Because the decision index j is an arbitrary label, we assume (WOLOG) that it has been chosen so that $a_i > 0$. We have

$$D(q_1, B) = \Theta(q_1 - \frac{a_0}{a_0 + a_1}).$$

Defining

$$r_1(B) = \frac{a_0}{a_0 + a_1},$$

our payoff (1) in this decision becomes

$$\beta_i(q_1, B) = b_{i,\Theta(q_1 - r_1(B))}. \tag{2}$$

3. Proposed Value Function

Suppose we have some information I about the probability that each possible payoff matrix B will be used in a decision, represented by

$$\Pr(B|I).$$

The proposed v_i is just the expectation over B of our single-decision payoff β_i:

$$\begin{aligned} v_i^{(I)}(q_i) &= E\{\beta_i(q_1, B)|I\} \\ &= \int_{\Re^4} \mathrm{d}B \Pr(B|I)\beta_i(q_1, B) \end{aligned} \tag{3}$$

A degenerate special case corresponding to a *single* decision is

$$\Pr(B|I) = \delta(B - B^{(1)}),$$

from which we get the single-decision payoff

$$v_i^{(I)}(q_i) = \beta_i(q_1, B^{(1)}) \tag{4}$$

as our value function.

4. Evaluating the Proposed Value Function

For observation y_i, our proposed value function (3) is

$$v_i(q_i) = \int_{\Re^4} \mathrm{d}B \, \Pr(B) b_{i,\Theta(q_1 - r_1(B))}, \tag{5}$$

where we have used (2) and made the conditioning on I implicit for notational brevity.

We change variables from b_{ij} to r_1 and three others as follows.

$$\begin{array}{rcll} s &=& a_0 + a_1 = b_{00} - b_{01} + b_{11} - b_{10} & \in (0, \infty) \\ r_1 &=& a_0/s = (b_{00} - b_{01})/s & \in [0, 1] \\ t_0 &=& b_{10}/s & \in \Re \\ t_1 &=& b_{01}/s & \in \Re \end{array}$$

Defining $r_0 \equiv 1 - r_1$ for convenience,

$$\begin{array}{ll} b_{00} = (r_1 + t_0)s & b_{01} = t_0 s \\ b_{10} = t_1 s & b_{11} = (r_0 + t_1)s. \end{array}$$

Thus, r_1 (or r_0) determines the decision made, s is an overall scale factor of the payoff matrix, and t_i is an offset in the payoff when y_i is observed.

The value function (5) for $i = 1$ becomes

$$\begin{aligned} v_1(q_1) &= \int\int\int\int \mathrm{d}r_1 \mathrm{d}s \mathrm{d}t_0 \mathrm{d}t_1 \, \Pr(r_1, s, t_0, t_1) b_{1,\Theta(q_1 - r_1)} \\ &\quad + \int \mathrm{d}s \int \mathrm{d}t_1 \, \Pr(s, t_1) t_1 s \\ &= \int_0^{q_1} \mathrm{d}r_1 h_1(r_1)(1 - r_1) + k_1, \end{aligned}$$

where we define the *weighted threshold density* by

$$\begin{aligned} h_1(r_1) &= \int \mathrm{d}s \, \Pr(r_1, s) s \\ &= \Pr(r_1) E\{s|r_1\} \quad \geq 0, \end{aligned} \tag{6}$$

and the *mean offsets* by

$$k_i = E\{t_i s\}. \tag{7}$$

Doing likewise for $i = 0$ and dropping the subscript in dummy variable r_i, we have

$$v_i(q_i) = \int_0^{q_i} \mathrm{d}r h_i(r)(1 - r) + k_i, \tag{8}$$

where we have defined $h_0(\cdot) \equiv h_1(1 - \cdot)$ for notational convenience and symmetry (as with r_i), so that

$$h_i(r_i) = \Pr(r_i) E\{s|r_i\} \quad \geq 0. \tag{9}$$

Thus, the proposed value functions (8) are simple integrals involving the (weighted) probability density of decision thresholds.

5. Transformations of Value Functions

Before proceeding, we consider the question of what transformations of a value function $v_i(q_i)$ may be considered arbitrary and be discounted.

We wish to be able to sum value functions over multiple trials or events, in order to obtain the total value of a set of predictions. We may consider these *sums* to be ordinal scores, that is, arbitrary to within an order-preserving transformation. Thus the value function $v_i(q_i)$ for individual event predictions can be considered arbitrary to within any transformation preserving the ordering of arbitrary *sums* of of such value functions.

In general, we may have a different value function $v_i^{(m)}(q_i)$ for each observation (indexed by m). V is total value summed over m. Order among V for different forecasters is preserved if we transform individual value functions by a common scale factor, and each event i by a separate offset, as

$$v_i'^{(m)}(q_i) = \alpha v_i^{(m)}(q_i) + \gamma_i.$$

Allowing these transformations, the proposed value functions (8) would become

$$v_i^{(m)}(q_i^{(m)}) \quad = \quad \alpha \int_0^{q_i^{(m)}} \mathrm{d}r h_i^{(I_m)}(r)(1-r) + k_i^{(I_m)} + \gamma_i,$$

where I_m may be different for different observations m.

However, throughout this paper, we consider for simplicity only a single observation, or assume all I_m are the same, and so can combine $k_i^{(I)}$ with the γ_i term. When we ignored trivial payoff matrices B, i.e. those independent of q_i, the integrals over these were implicitly absorbed into γ_i.

If also s is independent of r_i, then

$$h_i(r_i) = \overline{s} Pr(r_i),$$

and $\overline{s} \equiv E\{s\}$ can be absorbed into α, leaving

$$h_i(r_i) = \Pr(r_i)$$

In this case, h_i is precisely the probability density of the decision threshold, without any "weighting".

6. Expectation Maximization and Keeping the Expert Honest

If an expert agrees to be *paid* an amount[2] $f_i(q_i)$ if event y_i is observed after he predicts it with probability q_i, and the expert believes (for whatever reason) that the probability of y_i is p_i, then he expects to receive

$$\sum_i p_i f_i(q_i). \tag{10}$$

If this expression attains its maximum value at $q_i = p_i$, then the set of functions f_i is said to keep the expert honest, or be *honesty-rewarding* (HR) (or *proper*), since then the expert

[2]Such payment should be measured by the *utility* it provides to the expert, which is defined as that quantity whose expectation he seeks to maximize.

never has an incentive to lie by providing $q_i \neq p_i$. Since this concept was first introduced [6], the class of pay scale functions having this property has been studied extensively [8, 1, 11, 10, 4, 2, 3].

If expression (10) attains its maximum value at and *only* at $q_i = p_i$, for all $p_i \in [0, 1]$, we call f_i *strongly honesty-rewarding* (SHR)[3]. If a set of pay-scale functions is HR but not SHR, we call it weakly honesty-rewarding (WHR), which means that for at least some true belief p_i, there is at least one other value of q_i (other than p_i) that the expert might give as his prediction and still have the same expected pay.

Estimation of a single set of parameters of sufficiently powerful statistical models by maximizing SHR functions (e.g. maximum-likelihood or least-squares estimation) is known to guarantee convergence (in the limit of infinite available data) of the probability estimates obtained to the true probability [4] [7, 9].

Incentive considerations in the expert's pay scale form a distinct problem from that of the *value* of predictions to us as users of such predictions[5]. However, if we do pay an expert according to the value (to us) of the predictions he supplies, i.e. choose $f_i(q_i) = v_i(q_i)$, then we have coupled his (expected) payoff to ours, and we would therefore suspect that it would never be in his interest to lie to us. Indeed, not only are our proposed value functions (8) always HR as pay scale functions, but they have previously been derived as constituting the set of absolutely continuous HR functions [11]. Thus for every such pay scale function there exists a corresponding density of payoff matrices giving this as our value function, and the choice of scoring functions within this class is seen not to be arbitrary.

If $h_i(r) > 0$ for (almost) all $r \in (0, 1)$, then our scoring functions are SHR. Thus, if all decision thresholds have nonzero probability density, there are no additional values q_i (other than p_i), that would give equal expected payment to the expert. On the other hand, an example of an h_i leading to a WHR value function is the delta function. This leads to equation (4), which is neutral to changes of q_i that do not cross the decision threshold r_i. If $r_i = .5$, this is the often-quoted misclassification rate. Such a scoring function is only appropriate when one is certain that this single decision threshold is the only one to be used, in which case we didn't really need the full probability estimate $\{q_i\}$, and could have merely asked the forecaster for a 1-bit decision (in our two-outcome case).

7. Deriving h_i From a Proposed Scoring Function

Suppose someone proposes scoring function f_i, differentiable on $(0, 1)$ and HR (honesty-rewarding, i.e. (10) is minimized at $q_i = p_i$). It follows that

$$\frac{f_1'(q_1)}{1 - q_1} = \frac{f_0'(q_0)}{1 - q_0}$$

[3]We use the HR/SHR terminology here because there seems to be some confusion in the literature as to which of these two are meant by terms such as *proper*.

[4]Occasionally one will see in the literature the use of a scoring function that does not even meet this minimal requirement, such as the summed absolute (L_1) error, whose expectation is optimized if the forecaster thresholds all his predictions at 1/2, thus lying by always saying the probability is 0 or 1.

[5]See, for example, ref. [8], which pointed this out in the context of considering the value of deterministic predictions, especially under information loss due to quantization.

[1, 9], for $q_i \in (0,1)$. Then functions h_i defined by

$$h_i(q_i) \equiv \frac{f_i'(q_i)}{1-q_i} \tag{11}$$

may be plugged into (8), giving $v_i(q_i) = f_i(q_i)$. One density of payoff matrices corresponding to a given h_i is that where s is independent of r_i, in which case we can take

$$\Pr(r_i) = h_i(q_i) = \frac{v_i'(q_i)}{1-q_i},$$

and $\Pr(s, t_0, t_1)$ just determines an overall scale and offsets.

8. Square Error

Consider a *uniform* weighted threshold density

$$h_i(r) = \text{constant}, \qquad r \in [0,1]$$

From (8), we have

$$v_i(q_i) \propto -(1-q_i)^2.$$

Thus, when we have, for example, a uniform distribution $\Pr(r_1)$ of decision thresholds, and independent overall scale distribution $\Pr(s)$, we obtain the square error (or Brier score) loss function.

9. Log Likelihood

Let

$$v_i(q_i) = \log(q_i).$$

This is a log likelihood because when we think of the forecaster as a statistical model F_θ having parameters θ, q_i is the likelihood $\Pr(y_i|\theta)$, i.e. the probability of the observed event given a hypothesis. The expectation of this expression gives cross-entropy, mutual information, etc. [5], which differ only by the affine transformations of section 5.

Using (11), we have

$$h_i(r) \propto \frac{1}{r(1-r)}.$$

Thus, the log-likelihood scoring function arises from this non-normalizable threshold distribution, whose divergence at 0 and 1 corresponds to the unbounded penalty assigned to a near-zero prediction of the event that actually occurs.

10. Conclusion

By considering the utility of probability estimates in their actual uses to be the ultimate measure of their value, we have seen that scoring and loss functions can be directly related to the distribution of such uses (each use defining a decision threshold). Different existing scoring functions simply correspond to different such distributions.

We have considered the *value* of predictions (and related this to incentives for experts), rather than the problem of probabilistic inference (how best to estimate the outcome probability) itself. If we have two models or experts, we should consider combining their estimates (using Bayesian marginalization or perhaps some other form of weighted average) rather than using only the one having the greatest value by itself. Loss or scoring functions are perhaps best used in situations where we are forced to choose just one expert, machine, or parameter vector among those evaluated, perhaps because one is all we can afford.

We view the distribution of decision matrices, rather than the scoring function to which it leads, as being primitive. Choosing such a distribution is an inference problem on a different event space — that of uses to which the predictions will be put. Within a Bayesian approach to *this* inference problem may certainly arise familiar issues such as the use of appropriate reference or ignorance priors.

Acknowledgements

Thanks to José Bernardo, Chris Fuchs, Phil Goodman, John Miller, Carlos Rodriguez, Michael Stutzer, and David Wolpert for useful discussions and references.

References

[1] J. Aczél. Remarks on the measurement of subjective probability and information. *Metrika*, 11(2):91–105, 1966.

[2] J. Aczél and Z. Daróczy. *On Measures of Information and Their Characterization*, volume 115 of *Mathematics in Science and Engineering*. Academic Press, New York, 1975.

[3] José M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(3):686–691, 1979.

[4] P. Fischer. On the inequality $\sum p_i f(p_i) \geq \sum p_i f(q_i)$. *Metrika*, 18:199–208, 1972.

[5] H. Gish. A probabilistic approach to the understanding and training of neural network classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1361–1364, April 1990.

[6] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society B*, 14:107–114, 1952.

[7] John B. Hampshire II and Barak Pearlmutter. Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In *Connectionist Models: Proceedings of the 1990 Summer School*, pages 159–172. San Mateo, CA: Morgan Kaufmann Publishers, 1991.

[8] Jacob Marschak. Remarks on the economics of information. In *Contributions to Scientific Research in Management*, pages 79–98. University of California Press, 1960.

[9] John W. Miller, Rod Goodman, and Padhraic Smyth. On loss functions which minimize to conditional expected values and posterior probabilities. To appear in *IEEE Transactions on Information Theory*, 1993.

[10] Gy. Muszély. On continuous solutions of a functional inequality. *Metrika*, 19:65–69, 1973.

[11] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.