HOW GOOD WERE THOSE PROBABILITY PREDICTIONS? THE EXPECTED RECOMMENDATION LOSS (ERL) SCORING RULE

David B. Rosen
Center for Biomedical Modeling Research
University of Nevada, Reno
Present address:
Department of Medicine
New York Medical College
Valhalla, NY 10595 USA
Internet: `d.rosen@ieee.org` (or `rosen@unr.edu`)

To appear in *Maximum Entropy and Bayesian Methods. (Proceedings of the Thirteenth International Workshop, August 1993.)* G. Heidbreder, ed. Kluwer, Dordrecht, The Netherlands, circa 1995.

First version submitted November 1993; preprint placed in Neuroprose archive as rosen.scoring.ps.Z , with title "Scoring the Forecaster. . . ".

Superseded by this version, October 1995; available via
http://www.scs.unr.edu/~cbmr/people/rosen/erl/ or as
ftp://ftp.scs.unr.edu/pub/rosen/exp-rec-loss.ps.gz or
ftp://archive.cis.ohio-state.edu/pub/neuroprose/rosen.exp-rec-loss.ps.Z .

Contents

HOW GOOD WERE THOSE PROBABILITY PREDICTIONS?
THE EXPECTED RECOMMENDATION LOSS (ERL) SCORING RULE

David B. Rosen
Center for Biomedical Modeling Research
University of Nevada, Reno
Present address:
Department of Medicine
New York Medical College
Valhalla, NY 10595 USA
Internet: `d.rosen@ieee.org` (or `rosen@unr.edu`)

ABSTRACT. We present a new way to understand and characterize the choice of scoring rule (probability loss function) for evaluating the performance of a supplier of probabilistic predictions after the outcomes (true classes) are known. The ultimate value of a prediction (estimate) lies in the actual utility (loss reduction) accruing to one who *uses* this information to make some decision(s). Often we cannot specify with certainty that the prediction will be used in a particular decision problem, characterized by a particular loss matrix (indexed by outcome and decision), and thus having a particular decision threshold. Instead, we consider the more general case of a distribution over such matrices. The proposed scoring rule is the *expectation*, with respect to this distribution, of the loss that is actually incurred when following the decision *recommendation*, the latter being the decision that would be considered optimal *if* we were to *assume* the predicted probabilities. Logarithmic and quadratic scoring rules arise from specific examples of these distributions, and even common single-threshold measures such as the ordinary misclassification score obtain from degenerate special cases.

1. Introduction

1.1. Purpose

One of two outcomes (events or classes) will occur in an observation or experiment. We consider a forecaster providing an assessment, i.e. estimate, opinion, or *prediction,* of the probability that one of them (say outcome 1) will occur. We use the term *forecaster* broadly: this could be a human expert, maximum-likelihood fit of some parametric model, classifier / learning machine, or Bayesian inference procedure given a particular prior, to name some possibilities. We are not concerned here with how this prediction was or should have been generated from some set of available information (such as training sample data and prior knowledge), but rather with the question of what figure of merit, i.e. scoring rule or *probability loss function*, we should assign to the (probabilistic) prediction in hindsight once the true outcome is known.

1.2. Probability Loss Function

As an example, consider a weather forecaster who states that "the probability of *rain* today is $\hat{p}$", and of course, perhaps implicitly, "the probability of *no rain* today is $(1 - \hat{p})$".

We wish to choose a function L in order to assign a score to today's prediction by the forecaster as $L(i, \hat{p})$, where i is the actual outcome: 0 for no rain or 1 for rain. Examples of such a function include the logarithmic loss $L(i, \hat{p}) = -i \log(\hat{p}) - [1 - i] \log(1 - \hat{p})$ [6], the quadratic loss (squared error) $L(i, \hat{p}) = [i - \hat{p}]^2$ [3, 15, 5], and of course the binary misclassification loss, which is zero or one depending merely on whether $\hat{p}$ is on the appropriate side of $\frac{1}{2}$.

1.3. Importance of the Probability Loss Function

To place this problem into a particular practical setting for concreteness, suppose that some *user* of weather forecasts wishes to hire one forecasting consultant (or purchase one computer-based weather forecasting system) from among several available. The user has access to the forecasters' respective predictions and the true weather every day for the past year, and wishes to decide which one to hire (or buy) based on performance on this set of test data[1]. The relevant measure of performance is the expected benefit (i.e. utility) that *would have* accrued to the user during the test period if he had relied entirely on a given forecaster's predictions.

Then the user might hire the forecaster providing the best performance (after subtracting off the consulting fee each charges, if these differ).

We consider the predictions of a single forecaster. Since (expected) utilities are additive, it suffices to consider each test point (this forecaster's prediction and the actual outcome, for a single day) separately, and then sum these later. Thus, we are back to the problem of choosing a loss function for a single probability prediction, but now with the idea that it should perhaps represent the actual loss to the user of a prediction $\hat{p}$ when the ith outcome occurs.

1.4. Overview of Paper

Section 2. explains how a probability prediction can be viewed as a mapping from possible decision problems the user may face, each characterized by a decision loss (utility or regret) matrix, to corresponding *decision recommendations*. For any one such decision problem, the loss actually incurred after making the recommended decision defines the quality of the probability prediction. Then in Section 3. we consider the case in which we might use the probability prediction in any of a continuum of decision problems, described by a *distribution* of decision loss matrices. We show that the recommendation loss approach to scoring or loss functions can be generalized by using as our scoring function the *expected* recommendation loss (ERL), where the expectation is over (only) the decision loss matrix distribution. We explain how the commonly used scoring functions mentioned above arise in this recommendation loss approach. Table 1 summarizes the quantities and notation used in this paper.

Section 4. briefly discusses some of the literature on probability loss functions based on truth- or honesty-rewarding properties, and relates this to the ERL results of the present paper.

2. Decision, Prediction, and Recommendation

A decision problem concerns the choice of a course of action having real-world consequences (costs) that depend on both the action (decision) and an outcome event (true class). In the case of weather, examples would include deciding whether to water the lawn, bring an umbrella, or cancel some outdoor social event. The "cost" of cancelling a social event unnecessarily (vs. having the event rained on) may be quite different from the "cost" of the lawn being neither watered nor rained on (vs. being both watered and rained on), so you might make a different decision for each based on the same probability prediction.

[1] This is an oversimplification since the user should be interested in expected future (generalization) performance rather than empirical past (test set) performance. But we certainly cannot determine the expectation (over outcomes) of performance if we cannot determine performance even when the true outcome is known.

Table 1: Quantities and functions appearing in this paper.

i	=	observed outcome event (true class)
$\hat{p}$	=	prediction (judgment or estimate) of probability of outcome $i = 1$
$L(i, \hat{p})$	=	probability loss function (prediction scoring rule)
j	=	decision index (in some decision problem where i is relevant)
$C = \{c_{ij}\}$	=	decision loss (regret or cost matrix) characterizing a decision problem
t	=	decision threshold $= \frac{c_{01}}{c_{01}+c_{10}}$
s	=	stakes $= c_{01} + c_{10}$
$\hat{j}$	=	decision recommendation $= I(\hat{p} > t)$
$L_C(i, \hat{p})$	=	recommendation loss $= c_{i\hat{j}} = c_{i,I(\hat{p}>t)}$
$L_{\mathsf{ERL}}(i, \hat{p})$	=	expected recommendation loss $= E^C L_C(i, \hat{p})$
$h(t)$	=	threshold importance density $=$ pr$(t)E^s\{s\|t\}$
p	=	"true" outcome probability or that believed by expert (Sec. 4.)

Table 1a: Notation.

I(inequal.)	=	1 if inequality is true; 0 otherwise	
$E^z\{g(z)\|A\}$	=	expectation over z of $g(z)$ given A	$= \int_{\mathcal{Z}} dz$ pr$(z\|A)g(z)$

Probability predictions per se have no real-world consequences—until used to make decisions. Thus the true measure of a system's performance is of course the actual (or expected) gains or losses to those who use its predictions to make one or more decisions.

2.1. Decision Loss (Cost Matrix)

A decision problem is characterized by the *decision loss* c_{ij} for each observed outcome i and decision j; these c_{ij} can be said to form the elements of a cost matrix C. [2] In general, the number of decision alternatives need not be equal to the number of outcomes (classes), but we assume the two-outcome (i.e. two-class) two-alternative case for simplicity, giving

	$j = 0$	$j = 1$
$i = 0$	0	$c_{01} > 0$
$i = 1$	$c_{10} > 0$	0

,

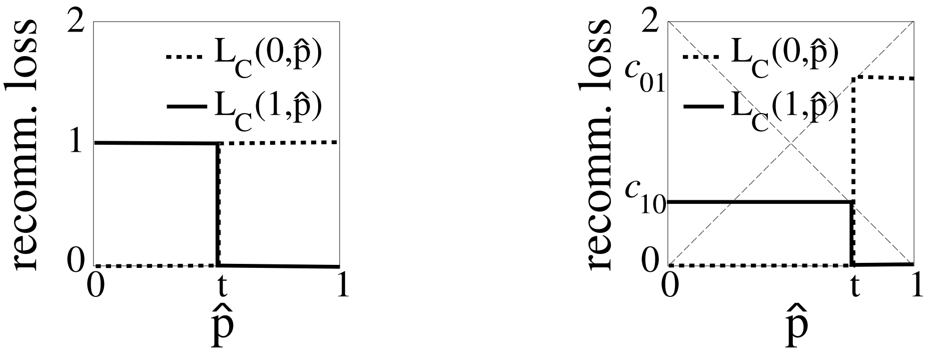
where we have defined decision $j = 0$ as that most favorable when outcome $i = 0$, and ignored nonzero c_{00} or c_{11} (diagonal elements) as merely leading to overall offsets (Section 3.3.).

2.2. Decision Recommendation Implicit in Prediction

Presumably, if our system were designed for (or a human expert were apprised of) a particular known decision loss matrix C, it could simply plug this and its prediction $\hat{p}$ into elementary decision theory, and thus recommend the course of action $j = \hat{j}$ that minimizes the expected decision loss

$$E^i\{c_{ij}|\hat{p}\} \quad = \quad \hat{p}c_{1j} + (1-\hat{p})c_{0j} \quad = \quad \begin{cases} \hat{p}c_{10} & \text{if } j = 0 \\ (1-\hat{p})c_{01} & \text{if } j = 1 \end{cases} .$$

[2] Note that the decision loss is a function of (i.e. is indexed by) outcome and decision made, while a probability loss is a function of outcome and probability prediction.



(a): $c_{10} = c_{01} = 1 \;\Leftrightarrow\; t = .5, s = 2$ (b): $c_{10} = .5, c_{01} = 1.5 \;\Leftrightarrow\; t = .75, s = 2$

Figure 1: Recommendation loss $L_C(i, \hat{p}) = c_{i,I(\hat{p}>t)}$ vs. prediction $\hat{p}$ for two fixed decision problems indicated. In (a) this gives the ordinary "0–1" misclassification loss.

The solution is given by

$$\begin{aligned} \hat{\jmath} &= \begin{cases} 1 & \text{if } (1-\hat{p})c_{01} < \hat{p}c_{10} \\ 0 & \text{otherwise} \end{cases} \\ &\equiv I((1-\hat{p})c_{01} < \hat{p}c_{10}) \\ &= I(\hat{p} > t), \end{aligned}$$

where *decision threshold* t is defined as $\frac{c_{01}}{c_{01}+c_{10}}$ and lies between 0 and 1 (inclusively). Of course this recommendation may be poor if the prediction is poor.

In the work of Thomas Bayes, one can interpret a personal probability as merely a convenient summary of one's decision rule, which is a function mapping the cost matrix to a decision preference. Similarly, we consider a probability *prediction* $\hat{p}$ to be merely a convenient summary of the function mapping cost matrix C to decision *recommendation* $\hat{\jmath}$.

2.3. Decision Recommendation Loss

We rewrite the cost matrix in terms of threshold t (Section 2.2.) and overall *stakes* $s = c_{01} + c_{10}$ as

$$c_{01} = ts, \qquad c_{10} = (1-t)s.$$

Since we consider a probability prediction to represent an implicit decision recommendation to the user, this user's question "how much would the prediction $\hat{p}$ (by itself) be worth to me?" naturally becomes equivalent to "what would be my losses if I followed the corresponding recommendation $\hat{\jmath} = I(\hat{p} > t)$". The user need not make this recommended decision, but then the user's decision loss would not be a measure of the value of the prediction $\hat{p}$ by itself in decision problem C, since we would then presume that the decision was based (at least in part) on different believed/assumed outcome probabilities or other information.

The actual decision loss in a given decision problem, when following the recommendation implicit in $\hat{p}$ and the actual outcome is i, is given by the *recommendation loss*

$$\begin{aligned} L_C(i, \hat{p}) = c_{i\hat{\jmath}} = c_{i,I(\hat{p}>t)} &= \begin{cases} tsI(\hat{p} > t) & \text{if } i = 0 \\ (1-t)sI(\hat{p} < t) & \text{if } i = 1 \end{cases} \\ &= [1-i]tsI(\hat{p} > t) + i[1-t]sI(\hat{p} < t), \end{aligned} \tag{1}$$

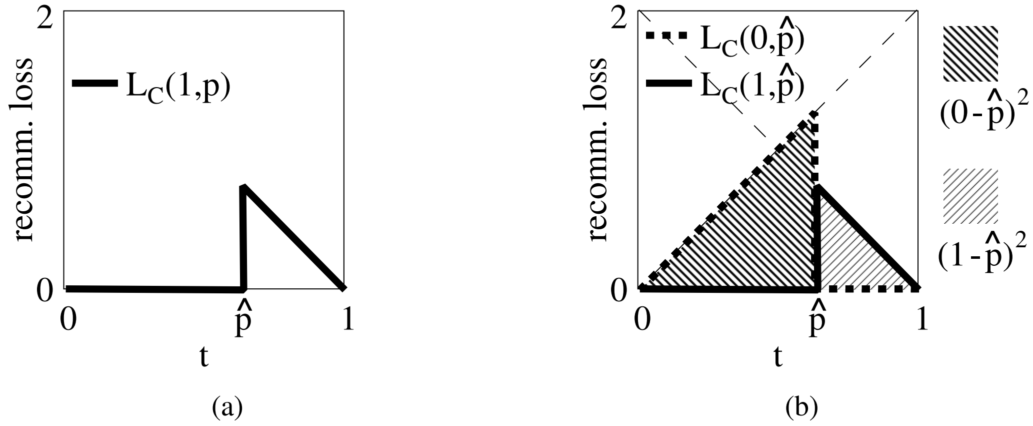


Figure 2: Recommendation loss (with constant stakes $s = 2$) vs. decision threshold t for fixed prediction $\hat{p}$ and (a) $i = 1$; (b) both values of i. In (b), the shaded areas represent $L(i, \hat{p})$ = quadratic loss (squared error), which is the expected recommendation loss (ERL) when importance $h(t) = 2$ (Section 3.1.).

which is plotted as a function of the prediction in figure 1, for two particular decision problems. *If* we are given a single decision problem of interest (with a specific cost matrix C determining a specific threshold), then our final performance measure (probability loss) should simply be given by this recommendation loss. We call this type of probability loss function *single-decision* or *single-threshold*, since it depends only on whether $\hat{p}$ is above or below a particular t.

2.4. Cost Matrix Unknown

If we don't know what the cost matrix, and thus the decision threshold, will be, we can plot the recommendation loss *as a function of this unknown threshold* for a given prediction and outcome, as in figure 2. By summing or averaging such curves over a data set consisting of many prediction-outcome pairs, we can completely characterize the performance on this data. We call this a(n empirical) *recommendation loss characteristic* (RLC)[12] curve, as it is an alternative to the widely-used *receiver operating characteristic* (ROC)[8] curve.

To measure and compare the overall performance of forecasters, we need in general a probability loss function to assign a single numeric score to each. A single-threshold loss is crude and not appropriate unless we are certain that the predictions are never to be used in any other decision problem. Other probability loss functions have been proposed and used historically, but which should we use, and what relationship (if any) will it have to the recommendation loss in decision problems?

3. Expected Recommendation Loss (ERL)

When it is uncertain in which decision problem a prediction will be used, we can describe the situation by a probability distribution over possible decision problems. Suppose pr(C) gives the probability density that our actual decision problem will be described by cost matrix C [3]. The

[3] C will be known exactly by the decision-maker; we simply do not know it (and thus the decision threshold) now. In contrast, if pr(C) were to remain when the decision itself were made, the decision problem would simply be characterized

natural probability loss function is then the *expectation over cost matrices* of the recommendation loss, i.e. $E^C L_C(i,\hat{p}) = \int dC\, \mathrm{pr}(C) L_C(i,\hat{p})$. We call this choice of probability loss function the *expected recommendation loss* $L_{\mathsf{ERL}}(i,\hat{p})$.

It is convenient to express the distribution of cost matrices as a joint probability density $\mathrm{pr}(t,s)$ over threshold and stakes, instead of over c_{01} and c_{10}. From (1), we have

$$\begin{aligned} L_{\mathsf{ERL}}(i,\hat{p}) &= \int_0^1 dt \int_0^\infty ds\, \mathrm{pr}(t,s) \left\{[1-i]tsI(\hat{p}>t) + i[1-t]sI(\hat{p}<t)\right\} \\ &= [1-i]\int_0^{\hat{p}} dt\, t h(t) + i\int_{\hat{p}}^1 dt\, [1-t]h(t), \end{aligned} \tag{2}$$

where the *threshold importance density* is

$$h(t) = \mathrm{pr}(t) E^s\{s|t\} \quad \geq 0,$$

i.e. the probability that a decision problem will use this threshold, times the expected stakes of decision problems having such a threshold.

The ERL probability loss function (2) can be described as an average of the decision recommendation loss per stakes (or recommendation loss for unit stakes), *weighted* by the importance $h(t)$ of each decision threshold. It reduces to a single-threshold loss (for example the misclassification loss) when $h(t)$ is concentrated at a single t (for example $\frac{1}{2}$).

3.1. Quadratic Loss

For uniform threshold importance density $h(t) = 2$ (for $t \in [0,1]$), the ERL is given by the area under the recommendation loss curve, which is the quadratic loss $L_{\mathsf{ERL}}(i,\hat{p}) = [i-\hat{p}]^2$, as indicated by the shaded areas in figure 2b.

3.2. Logarithmic Loss

Let $h(t) = \{t[1-t]\}^{-1}$, sometimes called the Haldane density. From eqn. (2), $L_{\mathsf{ERL}}(i,\hat{p}) = -i\log(\hat{p}) - [1-i]\log(1-\hat{p})$, which is simply the logarithmic loss mentioned earlier. This assigns an unbounded penalty when the prediction $\hat{p}$ is near 0 (1) when the true outcome $i = 1$ ($i = 0$). This is due to the unbounded importance given to thresholds near zero or one by this improper (non-normalizable) density. It can be argued[12] that the Haldane density may be an appropriate noninformative prior when all that is known about the decision problem is that it is nontrivial, thus leading to the choice of the logarithmic loss as a performance measure in such a situation.

The logarithmic and quadratic probability loss functions are members of a particular one-parameter family[7] of ERL loss functions. Even the misclassification loss (recommendation loss when $c_{01} = c_{10} = 1$) is obtained [11] in a limiting case of this parameter.

3.3. Arbitrary offsets in loss functions

If we add to the costs in a decision problem an amount depending on the outcome i but not on the decision j, i.e. $ai + b[1-i]$ with arbitrary real a and b, we then have nonzero diagonal elements c_{00} and c_{11}, but the decision analysis does not change. Similarly, if we add such arbitrary offsets to a probability loss function, the comparison of two forecasters is never affected, even after summing over a data set or taking expectations[13]. If we had considered a distribution over a full cost matrix without assuming diagonal elements of zero, the effect in Section 3. would have been to add to the

by the expected cost matrix $E\{C\}$, with a *single* resulting decision threshold.

resulting probability loss function (2) terms that could simply be incorporated into its own arbitrary offsets. In addition, of course, multiplying all loss functions by a constant has no substantive effect.

The logarithmic loss, also called empirical cross-entropy, is related to Kullback-Liebler distance and mutual information by such arbitrary offsets and overall scale[5].

4. Truth-Rewarding Loss Functions

A *strictly proper* probability loss function $L(i, \hat{p})$ can be defined as one that is *truth-rewarding*, i.e. its expectation (over i) is minimized when and only when $\hat{p}$ is equal to the true probability p (input-conditional class probability[4] $\mathrm{pr}(i = 1|x)$ for classification from input features/predictors).

Equivalently, a strictly proper loss function is *honesty-rewarding*: if we dock an expert's pay by $L(i, \hat{p})$ for prediction $\hat{p}$ and outcome i, then her expected pay (conditional on her *belief*) will be maximized if and only if she gives us $\hat{p}$ equal to the probability p implied by her belief[13].

The form (2) of our ERL loss functions (or variants of it, or generalizations to continuous outcomes [2] or their expectation [13, 9], or to more than two discrete outcomes[5]) has previously been proposed as either an objective function to be optimized in parameter estimation [7, 9] or as a device to elicit honest predictions from an expert [1, 13, 10, 4], in contrast to our interpretation as the expected recommendation loss to the user. Also in those treatments, $h(t)$ is an arbitrary non-negative function, in contrast to our interpretation in terms of a probability distribution over cost matrices. For greater than two outcomes (classes), the logarithmic loss is often advocated based on its *locality*, i.e. that it depends only on the predicted probability of the outcome that did in fact occur, rather than on the entire predicted probability distribution. The characterization of this assumption as a kind of "likelihood principle" for probability loss functions is attributed by Bernardo [2] to one of his manuscript's referees. Locality would not necessarily seem appropriate in the ERL context, since the decisions one makes, and thus the value of predictions, would *in general* depend on the entire predicted distribution. Locality alone also leaves the choice of $h()$ completely arbitrary in the two-discrete-outcome case, which is precisely the case considered in the present paper.

Single-threshold loss functions are not strictly proper, since they have the same expected loss for any $\hat{p}$ on the same side of the threshold, not just the true p. They are however *loosely proper*, meaning that the true p does indeed minimize the expected loss, even if other values do as well. A loosely proper loss function never rewards an expert for lying, but it may not always *penalize* the expert for doing so.

It follows from the work cited above that all ERL loss functions are (at least loosely) proper. An ERL with $h(t) > 0$ almost everywhere (on $[0, 1]$) is strictly proper. In addition, at least one of those authors [13] showed that any absolutely continuous strictly proper loss function can be written in the form (2), or to restate this in our present framework, there exists some importance $h()$ that generates such a loss function as an ERL.

5. Conclusion

It has been said [14][6] that

[4] Often called a posterior probability in the pattern recognition literature, where typically Bayes' Theorem is used to calculate it from a class prior and the input features' class-conditional probabilities.

[5] In some cases the results are given in differential form instead of integral form.

[6] Parentheses added and clauses rearranged; "[strictly]" added; references can be found in the cited paper.

> ...in most situations, rewarding the assessor according to the value of his forecasts with respect to some decision-making problem (if such a value can be determined) (Murphy, 1968) would conflict with the desire to use a [strictly] proper scoring rule (Roberts, 1968).

The present paper can be viewed as a way around this conflict, where we replace "some decision-making problem" by "some *generalized* decision-making problem", the latter meaning the situation described by a distribution over ordinary decision problems.

Acknowledgments

Thanks to David Wolpert for several useful discussions and criticism, and to José Bernardo, Chris Fuchs, John Miller, Padhraic Smyth, and Michael Stutzer for useful discussions and references. Supported in part by the (U.S.) Agency for Health Care Policy and Research (HS06830), and by the American College of Surgeons, AJCC prognostic systems project.

References

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

[2] José M. Bernardo. Expected information as expected utility. *Ann. Stat.*, 7(3):686–691, 1979.

[3] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[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 Int'l Conf. on Acoustics, Speech and Signal Processing*, pages 1361–1364, April 1990.

[6] I. J. Good. Rational decisions. *J. of the Royal Stat. Soc. 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: Proc. of the 1990 Summer School*, pages 159–172. San Mateo, CA: Morgan Kaufmann Publishers, 1991.

[8] Charles E. Metz. Basic principles of ROC analysis. *Seminars Nuclear Med.*, 8(4):283–298, 1978.

[9] John W. Miller, Rod Goodman, and Padhraic Smyth. On loss functions which minimize to conditional expected values and posterior probabilities. *IEEE Tr. Information Theory*, 39(4):1404–1408, 1993.

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

[11] David B. Rosen. Cross-entropy vs. squared error vs. misclassification: On the relationship among loss functions. Submitted. For preprint info., e-mail d.rosen@ieee.org with Subject: QUERY PAPER CSM .

[12] David B. Rosen. Issues in selecting empirical performance measures for probabilistic classifiers. In Kenneth Hanson and Richard Silver, editors, *Maximum Entropy and Bayesian Methods (Proceedings of the Fifteenth International Workshop, July 1995)*. Kluwer, Dordrecht, The Netherlands, 1996. Paper title subject to revision. To appear. For preprint info., e-mail d.rosen@ieee.org with Subject: QUERY PAPER ISEPM .

[13] Leonard J. Savage. Elicitation of personal probabilities and expectations. *J. of the American Stat. Assoc.*, 66(336):783–801, 1971.

[14] Robert L. Winkler. Scoring rules and the evaluation of probability assessors. *J. of the American Stat. Assoc.*, 64:1073–1078, 1969.

[15] J. Frank Yates. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30:132–156, 1982.