Electronic Journal of Statistics
Vol. 10 (2016) 1596–1607
ISSN: 1935-7524
DOI: 10.1214/16-EJS1150

Linear scoring rules for probabilistic binary classification

Matthew Parry

Dept of Mathematics & Statistics
University of Otago
P.O. Box 56
Dunedin 9054
New Zealand
e-mail: mparry@maths.otago.ac.nz

Abstract: Probabilistic binary classification typically calls for a vector of marginal probabilities where each element gives the probability of assigning the corresponding case to class 1. Scoring rules are principled ways to assess probabilistic forecasts about any outcome that is subsequently observed. We develop a class of proper scoring rules called linear scoring rules that are specifically adapted to probabilistic binary classification. When applied in competition situations, we show that all linear scoring rules essentially balance the needs of organizers and competitors. Linear scoring rules can also be used to train classifiers. Finally, since scoring rules have a statistical decision theoretic foundation, a linear scoring rule can be constructed for any user-defined misclassification loss function.

MSC 2010 subject classifications: Primary 62C99.
Keywords and phrases: Scoring rules, binary classification, probabilistic forecast.

Received November 2015.

Contents

1. Introduction

Classification challenges have become an exciting and useful feature of the statistical and machine learning community. Given a labelled training dataset, contestants are invited to submit their classifications for a test dataset. In order to make the challenge more interesting, challenge organizers typically publish a ranked list of the leading submissions and, ultimately, announce the winner of the challenge. However, in order for such a competition to be considered worth entering, the challenge organizers must be seen to evaluate the submissions in a fair and open manner.

Perhaps the most common classification challenge involves probabilistic binary classification. Suppose there are n test cases and that $y = (y_i \in \{0, 1\} \,|\, i = 1, \ldots, n)$ is the vector of labels known only to the challenge organizers. Contestants are asked to submit a vector of probabilities $\omega = (\omega_i)$ with the interpretation that $\omega_i = \mathbb{P}(Y_i = 1)$. Note that the contestant is being asked for marginal probabilities only. How can such probabilistic classifications be evaluated?

Scoring rules were devised precisely to answer this kind of question. Scoring rules are a principled way to assess probabilistic forecasts about any outcome that is subsequently observed. Crucially, *proper* scoring rules elicit honest statements of belief about the outcome. In the context of the probabilistic classification challenge, if the challenge organizers use a proper scoring rule to evaluate submissions, a competitor's expected score under their true belief about the class labels will be minimized[1] by actually quoting that belief to the organizers. A proper scoring rule therefore rules out any possibility of a competitor gaming the challenge.

Scoring rules have long been applied to forecasts of binary outcomes. Indeed, in one of the first papers on the subject, Brier (Brier, 1950) explicitly considered the case of a sequence of weather forecasts for rain or no rain. Almost all discussion, however, has centered on sequential or online evaluation of forecasters. Here our focus is on batch evaluation. Some of our results are anticipated in the technical report by Buja *et al.* (Buja, Stuetzle and Shen, 2005), though they implicitly assume additive scoring rules – see section 2.3. Banerjee *et al.* (Banerjee, Guo and Wang, 2005) considered loss functions that are minimized in expectation by quoting the expected value of the outcome in question. When restricted to binary outcomes, their loss functions can be recast as scoring rules that are essentially given by eq. 3. The excellent review article by Gneiting and Raftery (Gneiting and Raftery, 2007) includes discussion of scoring rules for categorical outcomes; these are superficially similar to the scoring rules introduced here but the quoted probabilities are constrained to lie on the simplex. Recently, Byrne (Byrne, 2016) has written about area-under-the-curve (AUC) measures for probabilistic forecasting. In his elegant formulation of the problem, when only marginal probabilities are quoted, the concept of a scoring function is invoked, as opposed to a scoring rule. Finally, Frongillo and Kash (Frongillo and Kash, 2015) have recently considered the general problem of devising proper

[1] Scoring rules are typically taken to be negatively oriented.

scoring rules to elicit vector-valued properties of a distribution. In their terminology, a property is linear if it is a linear function on the space of distributions. The linear scoring rules in this paper can be understood as eliciting the vector-valued linear property that is the marginal probabilities.

In section 2, we introduce the class of linear scoring rules and contrast them with more general but more complicated scoring rules. We find three useful sub-classes of linear scoring rules: additive, homogeneous and rank-based. We also make a connection between a particular rank-based linear scoring rule and the AUC measure. Section 3 is an aside on using linear scoring rules to train probabilistic classifiers. In section 4, we show how linear scoring rules fit within statistical decision theory. We are able to show that there is a linear scoring rule which accounts correctly for any user-defined misclassification loss function. Finally, in section 5, we show that all linear scoring rules essentially achieve the same balance between the organizers' need for discriminative power and the competitors' wish not to be penalized unduly by outliers.

2. The class of linear scoring rules

To fix notation, let $y \in \mathcal{Y} := \{0,1\}^n$ be an observed outcome of class labels and let $\omega \in \mathcal{P} := [0,1]^n$ be a vector of probabilities. We will refer to these probabilities as *marginal* probabilities to emphasize the fact that ω is not a joint probability from $\mathcal{P}_\mathcal{Y}$, the class of all distributions on $\mathcal{Y}$. Note that the restriction to $\mathcal{P}$ is not done for convenience but rather to fit in with the framework of the classification challenge: competitors are asked to quote marginal probabilities, not a joint distribution. We will be interested in scoring rules $S : \mathcal{Y} \times \mathcal{P} \to \mathbb{R} \cup \{\infty\}$ and will say $S(y,\omega)$ is the *score* for quoting ω and observing y.

The fact that $\mathcal{P}$ is convex is crucial to what follows. For $p \in \mathcal{P}_\mathcal{Y}$, let $\mathcal{M}p$ denote the product of its marginal probabilities. More precisely, $(\mathcal{M}p)(y) = \prod_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$, where $\pi = \mathbb{E}_{Y\sim p}\, Y$. Then, as is well known, $\mathcal{M}\mathcal{P}_\mathcal{Y}$ is not convex: for $p, q \in \mathcal{P}_\mathcal{Y}$ and $\lambda \in (0,1)$, typically there is no $p(\lambda) \in \mathcal{P}_\mathcal{Y}$ such that $\mathcal{M}p(\lambda) = (1-\lambda)\mathcal{M}p + \lambda\mathcal{M}q$. For this reason, we do not look to define scoring rules on $\mathcal{Y} \times \mathcal{M}\mathcal{P}_\mathcal{Y}$. However, this does mean a competitor's quote need not be derived from a joint distribution.

We overload the notation for scoring rules by defining the *expected score* for each $\pi \in \mathcal{P}$,

$$S(\pi,\omega) = \mathbb{E}_{Y\sim\pi}\, S(Y,\omega), \tag{1}$$

where $Y \sim \pi$ is shorthand for $Y_i \overset{\text{ind}}{\sim} \text{Bern}(\pi_i)$, $i = 1,\ldots,n$. So defined, $S(\pi,\omega)$ is *affine* in its first argument. A scoring rule is said to be *proper* in $\mathcal{P}$, if $S(\pi,\omega) \geq S(\pi,\pi)$, for all $\pi,\omega \in \mathcal{P}$. A scoring rule is said to be *strictly* proper if equality holds for $\omega = \pi$ only. Note that the scoring rules we discuss in this paper remain proper in $\mathcal{P}_\mathcal{Y}$, though not strictly proper.

As indicated previously, a proper scoring rule will elicit an honest statement of a competitor's belief. To see this, suppose π represents the competitor's actual belief about the class labels. Then $S(\pi,\omega)$ will be their expected score

under their actual belief if they quote ω. But, if the scoring rule is proper, their expected score cannot be less than $S(\pi, \pi)$, hence they should quote π.

It is convenient to define $d(\pi, \omega) := S(\pi, \omega) - S(\pi, \pi)$, a quantity commonly called the *divergence*. The divergence cannot be negative for a proper scoring rule. For a strictly proper scoring rule, $d(\pi, \omega) = 0$ only if $\omega = \pi$.

2.1. Linear scoring rules

A great deal is known about how to generate proper scoring rules (McCarthy, 1956; Hendrickson and Buehler, 1971; Gneiting and Raftery, 2007). For our situation, Theorem 1 of Gneiting and Raftery (2007) ensures that under mild regularity conditions and for convex $\mathcal{P}$, $S(\cdot, \omega)$ will be a (strictly) proper scoring rule iff there exists a (strictly) concave function $H(\omega)$ such that

$$S(y, \omega) = H(\omega) + (y - \omega) \cdot \partial H(\omega), \tag{2}$$

where $\partial H(\omega)$ is a supergradient to $H(\cdot)$ at ω. The converse result is that $H(\omega) = S(\omega, \omega)$. To reflect its importance, $H(\omega)$ is called the *entropy*.

When $H(\omega)$ is differentiable, the expression for the scoring rule simplifies to

$$S(y, \omega) = H(\omega) + (y - \omega) \cdot \nabla H(\omega), \tag{3}$$

where $\nabla H(\omega) := (\partial H / \partial \omega_i)$.

We call a scoring rule that is derived from eq. (2) a *linear* scoring rule. This is motivated by the fact that such a scoring rule is a linear function of the class labels y. For convenience, we will always assume $H(\omega)$ is concave so that a linear [2] scoring rule is also necessarily proper. Eq. (3) is also anticipated in Banerjee, Guo and Wang (2005). Banerjee *et al.* (Banerjee, Guo and Wang, 2005) found the necessary form of loss functions $L(y, \omega)$ that are minimized in expectation by predicting $\omega = \mathbb{E}\, Y$ for the outcome y. When y is restricted to binary outcomes, their loss functions are our linear scoring rules, since $\pi = \mathbb{E}_{Y \sim p}\, Y$.

A useful consequence of linearity is that even though in truth $Y \sim p \in \mathcal{P}_{\mathcal{Y}}$, still $S(p, \omega) = S(\pi, \omega)$, where π is the resulting vector of marginal probabilities for the class labels.

2.2. Connection to other scoring rules

It is important to realize that linear scoring rules do not exhaust the forms of scoring rules that can be applied to probabilistic binary classification. Indeed, the obvious approach is the indirect one: take existing proper scoring rules $S(y, q)$ on $\mathcal{Y} \times \mathcal{P}_{\mathcal{Y}}$, and then restrict consideration to probability distributions

[2] We use linear in the sense of a linear function rather than a linear map. As a mapping, the scoring rule is an affine transformation of the class labels y. However, we don't want to confuse this use of affine with the common use of affine in connection with scoring rules to refer to the fact that $S(\pi, \omega)$ is affine in its first argument.

of the form $q(y) = \prod_{i=1}^{n} \omega_i^{y_i}(1-\omega_i)^{1-y_i}$. However, apart from the logarithmic scoring rule, $S(y,q) = -\log q(y)$, the resulting scoring rules are rather unwieldy.

Consider for example, the Brier scoring rule, which for $q \in \mathcal{P}_{\mathcal{Y}}$ takes the form $S(y,q) = -q(y) + \frac{1}{2}\sum_{z\in\mathcal{Y}} q(z)^2$. When $q(y) = \prod_{i=1}^{n} \omega_i^{y_i}(1-\omega_i)^{1-y_i}$, this becomes

$$S(y,\omega) = -\prod_{i=1}^{n} \omega_i^{y_i}(1-\omega_i)^{1-y_i} + \tfrac{1}{2}\prod_{i=1}^{n}\left\{\omega_i^2 + (1-\omega_i)^2\right\}. \tag{4}$$

Thus linear scoring rules have the appeal of simplicity and tractability.

Having said that, linear scoring rules have a slightly reduced flexibility under certain additive transformations. Typically, if $S(y,q)$ is a scoring rule then so is $S(y,q) + k(y)$. For linear scoring rules, however, $k(y)$ must take the form $k(y) = k \cdot y + c$.

2.3. Additive sub-class

We call a scoring rule *additive* if it is generated by an entropy function of the form $H(\omega) = \sum_i h_i(\omega_i)$, where each $h_i(\cdot)$ is a concave function of its argument, so that

$$S(y,\omega) = \sum_i S_i(y_i,\omega_i). \tag{5}$$

Note that Frongillo and Kash (Frongillo and Kash, 2015) refer to additivity as separability.

In most applications, we expect that $h_i(s) = h(s)$, for each i. Common examples include $h(s) = -s\log s - (1-s)\log(1-s)$, which leads to the *logarithmic scoring rule*,

$$S(y,\omega) = \sum_i \left\{-y_i \log\omega_i - (1-y_i)\log(1-\omega_i)\right\}, \tag{6}$$

and $h(s) = \frac{1}{2}s(1-s)$, which leads to the linear class version of the *Brier scoring rule*,

$$S(y,\omega) = \tfrac{1}{2}\|y-\omega\|^2. \tag{7}$$

Additive scoring rules also have a "local" property: the score for test case i depends on ω_i but not on the quoted probability for any other case. Note that this is a different type of locality to that of local scoring rules (Parry, Dawid and Lauritzen, 2012); there locality refers to the (relative) lack of dependence on the quoted probability for *unrealized* outcomes.

An interesting twist on the usual additive scores comes from considering $h_i(s) = w_i h(s)$, where the w_i are weights satisfying $w_i > 0$ and $\sum_i w_i = 1$. The ensuing scoring rule then weights the test cases differently. While this is a proper scoring rule, if competitors are to make use of the weighting scheme, they should also be given the weighting scheme for the training data.

2.4. Homogeneous sub-class

Recall that a function $f(s)$ is said to be homogeneous of order k or k-homogeneous, if $f(\lambda s) = \lambda^k f(s)$, for $\lambda > 0$. We call a scoring rule *homogeneous* if it is generated by an entropy function that is 1-homogeneous (up to an irrelevant additive constant). When $H(\omega)$ is also differentiable, 1-homogeneity implies $H(\omega) = \omega \cdot \nabla H(\omega)$, and the associated scoring rule takes the very simple form,

$$S(y, \omega) = y \cdot \nabla H(\omega), \tag{8}$$

and is 0-homogeneous.

Pseudospherical scoring rules are examples of homogeneous scoring rules. They arise from the fact that the L^α-norm, $\|\omega\|_\alpha = (\sum_i |\omega_i|^\alpha)^{1/\alpha} = (\sum_i \omega_i^\alpha)^{1/\alpha}$, is convex for $\alpha \geq 1$. Letting $H(\omega) = -\|\omega\|_\alpha$, we have

$$S(y, \omega) = -\frac{\sum_{i=1}^n y_i \omega_i^{\alpha-1}}{\|\omega\|_\alpha^{\alpha-1}}. \tag{9}$$

In the limit $\alpha \to \infty$, i.e. the L^∞-norm, we have *the zero-one scoring rule*

$$S(y, \omega) = -\frac{1}{\#M(\omega)} \sum_{i \in M(\omega)} y_i, \tag{10}$$

where $M(\omega) = \{j \,|\, \omega_j = \max\{\omega\}\}$ and $\#A$ denotes the cardinality of A. In slightly different contexts, this is sometimes referred to as the misclassification loss.

The only scoring rule that is both additive and homogeneous is the trivial scoring rule, $S(y, \omega) = k \cdot y$.

2.5. Rank-based sub-class

A scoring rule is said to be *rank-based* if it depends only on the ranks of the quoted probabilities ω. As a consequence, a rank-based scoring rule cannot be strictly proper. A rank-based scoring rule is also a homogeneous scoring rule.

Here we give only an important example of a rank-based scoring rule. Let

$$\psi_i(\omega) = \#\{j \,|\, \omega_j < \omega_i\} - \#\{j \,|\, \omega_j > \omega_i\}, \tag{11}$$

so that $\psi_i(\omega)$ is the net number of elements of ω that are exceeded by ω_i. Then $H(\omega) = -\omega \cdot \psi(\omega)$ is both 1-homogeneous and concave, where $\psi(\omega) := (\psi_i(\omega))$. One-homogeneity is immediate since $\psi(\omega)$ 0-homogeneous. To show concavity, first note that because $H(\omega)$ is a collection of planar surfaces essentially indexed by the rank sets of ω, it suffices to consider what happens on either side of $\omega_i = \omega_k$, for an arbitrary pair (i, k). Without loss of generality, fix k and let $M_k(\omega) = \{j \,|\, \omega_j = \omega_k\}$. Letting superscript 0 indicate the value of a quantity when $\omega_i = \omega_k$, and superscript $\pm$ indicate the value of a quantity when $\omega_i = \omega_k^\pm$, we have $\psi_i^\pm = \psi_i^0 \pm \#M_k$, $\psi_j^\pm = \psi_j^0 \mp 1$, for $j \in M_k$, and $\psi_j^\pm = \psi_j^0$ otherwise.

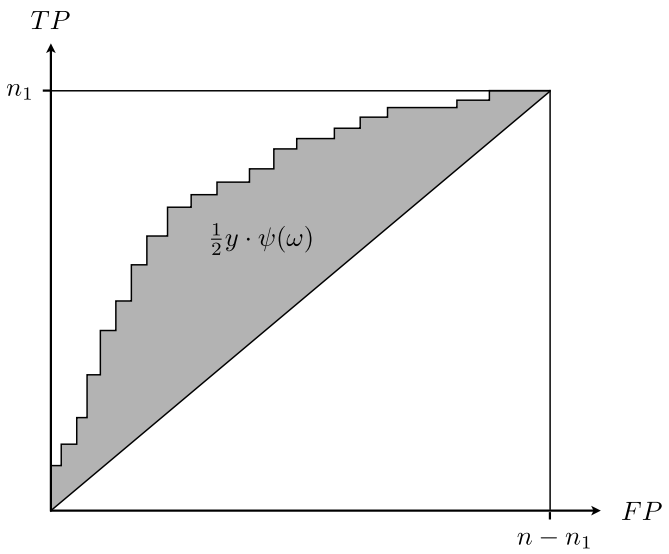

FIG 1. *Geometric interpretation of the ranked-based scoring rule given in eq. (12). In a sense, the scoring rule averages the performance of the thresholded probabilistic classifier over all thresholds.*

Continuity of $H(\omega)$ follows since $H^{\pm} = H^0 \mp \omega_k^{\pm} \cdot \#M_k \pm \sum_{j \in M_k} \omega_j = H^0 \mp \omega_k \cdot \#M_k \pm \omega_k \sum_{j \in M_k} 1 = H^0$. Finally, $H(\omega)$ is concave because $\partial H / \partial \omega_i |_{-}^{+} = -2 \#M_k < 0$. The resulting scoring rule is

$$S(y, \omega) = -y \cdot \psi(\omega). \tag{12}$$

Byrne (Byrne, 2016) has shown that this is related to the Wilcoxon-Mann-Whitney U-statistic and to the area-under-the-curve (AUC) measure, which is very commonly used in classification challenges. Specifically, if we define $n_1 = \sum_i y_i$ to be the number of test cases of class 1, then

$$\mathrm{AUC}(y, \omega) = \begin{cases} \frac{1}{2}\left[1 - \frac{1}{n_1(n-n_1)} \cdot (-y \cdot \psi(\omega))\right] & n_1 \neq 0, n, \\ \frac{1}{2} & \text{otherwise,} \end{cases} \tag{13}$$

where we follow Byrne (Byrne, 2016) and define the AUC to be $\frac{1}{2}$, when $n_1 = 0, n$. Although the AUC appears to be a scaled, positively oriented scoring rule, in general it is not. However, if n_1 is known beforehand – sometimes this information is provided to challenge contestants – then the AUC is in the class of linear scoring rules. Byrne (Byrne, 2016) also shows that the AUC is a proper scoring rule in the (unrealistic) case that $Y \sim \omega \in \mathcal{P}$.

The obvious connection between eq. (12) and eq. (13), however, enables us to give a simple geometric picture of the rank-based scoring rule introduced here. Figure 1 is a plot of false positive counts (FP) vs. true positive counts (TP) for all thresholds between 0 and 1. Then $S(y, \omega) = -2 \times$ (area above the diagonal).

3. Training with linear scoring rules

Given a set of features or predictors $x \in \mathcal{X} \subseteq \mathbb{R}^p$, a rather general approach to probabilistic classification is to let

$$\omega(x) = F(x \cdot \theta), \tag{14}$$

where $F : (-\infty, \infty) \to [0, 1]$ is a cumulative distribution function and $\theta \in \mathbb{R}^p$ is a parameter vector to be estimated from the training data. This framework includes logistic and probit regression as special cases. We now show that there is a natural additive scoring rule associated with each continuous cdf $F(\cdot)$.

Lemma 1. *If $Q(\cdot)$ is the quantile function associated with the cdf $F(\cdot)$ then $h(\omega) := -\int^{\omega} \mathrm{d}z\, Q(z)$ is concave in ω for $\omega \in [0, 1]$.*

Proof. Since $Q(\cdot)$ is a non-negative, non-decreasing function, $h(\pi) - h(\omega) = -\int_{\omega}^{\pi} \mathrm{d}z\, Q(z) \leq -(\pi - \omega)Q(\omega) = (\pi - \omega)h'(\omega)$. □

Consequently, $H(\omega) = \sum_{i=1}^{n} h(\omega_i)$ is concave and generates an additive scoring rule. Using this in eq. (3) and after integrating by parts and a change of variables, we obtain the scoring rule

$$S(y, \omega) = \sum_{i=1}^{n} \left\{ -y_i\, x_i \cdot \theta + \int^{x_i \cdot \theta} F(\zeta)\, \mathrm{d}\zeta \right\}. \tag{15}$$

Note that in the case of logistic regression, this scoring rule is exactly the log score.

The *perceptron scoring rule* is an interesting example connected to the perceptron neural network that arises as a limiting case of eq. (15). Letting $F(\zeta) = \mathbb{1}\{\zeta > 0\}$, then

$$S(y, \omega) = \sum_{i=1}^{n} \{-y_i + F(x_i \cdot \theta)\}\, x_i \cdot \theta. \tag{16}$$

3.1. Estimating equations

The system of (unbiased) estimating equations (Dawid and Lauritzen, 2005) for θ is

$$\frac{\partial S}{\partial \theta_\alpha} = \sum_{i=1}^{n} \{-y_i + F(x_i \cdot \theta)\}\, x_{i\alpha} = 0, \tag{17}$$

where $x_{i\alpha}$ denotes the α-component of feature vector x_i and $\alpha = 1, \ldots, p$. The simple form of these equations has useful consequences for back propagation in neural net-type applications.

4. Deterministic classification and connection to decision theory

In some challenges, the organizers require definite class predictions and will rank the competitors in terms of a loss function $L(y, y^\star)$, where $y^\star = (y_i^\star)$ and $y_i^\star$ is

the predicted class for test case i. The question is then how to turn a probabilistic classification into a deterministic classification. The obvious approach is by thresholding:

$$y_i^\star = \mathbb{1}\{\omega_i > s\}, \tag{18}$$

for some threshold $s \in [0,1]$. Another approach is to suppose the probabilities ω are the basis for the *randomized* classification

$$Y^\star \sim \omega. \tag{19}$$

(Recall this is shorthand for $Y_i^\star \overset{\text{ind}}{\sim} \text{Bern}(\omega_i), i = 1, \ldots, n$.) We now show that neither approach corresponds to a proper scoring rule but that there is, nevertheless, a linear scoring rule naturally associated with the loss function $L(y, y^\star)$.

Following Hand (Hand, 2009), let $c_\ell \in [0, \infty]$ denote the cost of misclassifying an object that is in class ℓ. If we assume that there is no cost in correctly assigning an object to its class and that the loss is additive in the cases, then

$$L(y, y^\star) = c_0\, y^\star \cdot (\mathbf{1} - y) + c_1(\mathbf{1} - y^\star) \cdot y, \tag{20}$$

where $\mathbf{1} = (1, \ldots, 1)$.

4.1. Thresholding

Under thresholding, the implied scoring rule is $S(y, \omega) = c_0 \mathbb{1}\{\omega > s\} \cdot (\mathbf{1} - y) + c_1(\mathbf{1} - \mathbb{1}\{\omega > s\}) \cdot y$. The associated entropy is therefore

$$H(\omega) = c_0 \mathbb{1}\{\omega > s\} \cdot (\mathbf{1} - \omega) + c_1(\mathbf{1} - \mathbb{1}\{\omega > s\}) \cdot \omega. \tag{21}$$

We now show that for $s \neq c_0/(c_0 + c_1)$, the entropy is not a continuous function of ω, and hence cannot be a generator of a proper scoring rule. Choose i and compare the left and right limits as $\omega_i \to s$, with ω otherwise fixed. Then $H|_-^+ = c_0(1 - s) - c_1 s \neq 0$. The case $s = c_0/(c_0 + c_1)$ is a special case that we will return to shortly.

4.2. Random classification

Randomized classification implies the scoring rule $S(y, \omega) = \mathbb{E}_{Y^\star \sim \omega} L(y, Y^\star) = c_0\, \omega \cdot (\mathbf{1} - y) + c_1(\mathbf{1} - \omega) \cdot y$. We can see that this is not a proper scoring rule in two different ways. The more direct way is via the implied entropy:

$$H(\omega) = (c_0 + c_1)\, \omega \cdot (\mathbf{1} - \omega) \tag{22}$$

and this actually generates the Brier scoring rule and not $S(y, \omega)$ above. The more explicit way comes from noting that the divergence

$$d(\pi, \omega) = S(\pi, \omega) - S(\pi, \pi) = (\omega - \pi) \cdot (c_0(\mathbf{1} - \pi) - c_1 \pi) \tag{23}$$

can be negative. For if we have $\pi_i \neq 0, 1$ for some i, then there exists $\epsilon = (\epsilon_i)$ such that $\pi \pm \epsilon$ are interior points of $[0,1]^n$, and it follows that $d(\pi, \pi \pm \epsilon)$ will be of opposite sign.

4.3. Proper scoring rule

Grünwald and Dawid (Grünwald and Dawid, 2004) give a decision theoretic approach for turning any loss function into a proper scoring rule. The key is to consider optimal acts in light of the expected loss, where the expectation is over possible outcomes y. In this formulation, $y^\star$ is the act of classification. Again overloading the notation, the expected loss is $L(\pi, y^\star) = c_0\, y^\star \cdot (\mathbf{1} - \pi) + c_1(\mathbf{1} - y^\star) \cdot \pi$. Then the *Bayes act* a^π against π is the choice

$$a_i^\pi = \mathbb{1}\left\{\pi_i > \frac{c_0}{c_0 + c_1}\right\}, \tag{24}$$

which ensures $L(\pi, y^\star) \geq L(\pi, a^\pi)$. Following Grünwald and Dawid, we have that

$$S(y, \omega) := L(y, a^\omega) = c_0\, a^\omega \cdot (\mathbf{1} - y) + c_1(\mathbf{1} - a^\omega) \cdot y \tag{25}$$

is a proper scoring rule. Given the previous discussion, we immediately see that this corresponds to converting the probabilistic classifier into a deterministic classifier by choosing the particular threshold $s = c_0/(c_0 + c_1)$. This is the only threshold that is appropriate.

The entropy associated with the scoring rule is

$$H(\omega) = c_0\, a^\omega \cdot (\mathbf{1} - \omega) + c_1(\mathbf{1} - a^\omega) \cdot \omega, \tag{26}$$

which is continuous and concave since $H(\omega) = \sum_i h(\omega_i)$, where

$$h(\omega_i) = \begin{cases} c_0(1 - \omega_i), & \omega_i > \frac{c_0}{c_0 + c_1} \\ c_1\omega_i, & \text{otherwise} \end{cases}. \tag{27}$$

5. Discriminative ability and robustness: choosing the best scoring rule

Given the large number of linear scoring rules that could be used, it is natural to wonder whether there is an optimal scoring rule or a set of criteria for selecting an appropriate proper scoring rule. We argue in this section that all linear scoring rules are essentially on a par when it comes to balancing the requirements of the organizers and the competitors.

Let ω be a competitor's probabilistic classification and $\pi = \mathbb{E}_{Y\sim p}\, Y$, the true marginal distribution resulting from $p \in \mathcal{P}_\mathcal{Y}$. Organizers will value discriminative power in the scoring rule, i.e. the ability to discriminate between classifications that are "close to" π. This will be achieved if $(\mathbb{E}_{Y\sim p}[S(Y, \omega) - S(Y, \pi)])^2$ is large. On the other hand, contestants will not want their score to be sensitive to outliers or unusual cases, i.e. the scoring rule should have a degree of robustness. This will be achieved if $\mathrm{var}_{Y\sim p}[S(Y, \omega) - S(Y, \pi)]$ is small. These two desiderata can be combined by seeking to maximize

$$\frac{(\mathbb{E}_{Y\sim p}[S(Y, \omega) - S(Y, \pi)])^2}{\mathrm{var}_{Y\sim p}[S(Y, \omega) - S(Y, \pi)]}. \tag{28}$$

Importantly, this combination is invariant under a multiplicative rescaling of the scoring rule.

To make more concrete statements about the objective function in eq. (28), we will assume that the generating entropy $H(\omega)$ for the scoring rule is strictly concave and sufficiently differentiable. Consequently, $-\nabla^2 H(\omega) := -(\partial^2 H/\partial\omega_i\,\partial\omega_j)$ exists and is symmetric positive definite. Now suppose ω is close to π, i.e. $\omega = \pi + \epsilon\,\eta$, where ϵ is a small real number and $\eta^\top\eta = 1$. Expanding the competitor's score around π, we have

$$S(y,\omega) = S(y,\pi) + \epsilon\,S_1(y,\pi;\eta) + \tfrac{1}{2}\,\epsilon^2 S_2(y,\pi;\eta) + \mathcal{O}(\epsilon^3), \tag{29}$$

where $S_k(y,\pi;\eta) := (\mathrm{d}/\mathrm{d}\epsilon)^k S(y,\pi+\epsilon\eta)|_{\epsilon=0}$.

We now have the following lemma:

Lemma 2. *For any $p \in \mathcal{P}_\mathcal{Y}$, if $H(\omega)$ is strictly concave and sufficiently differentiable, then $\mathbb{E}_{Y\sim p}\,S_1(Y,\pi;\eta) = 0$, where $\pi = \mathbb{E}_{Y\sim p}\,Y$.*

Proof. $S_1(y,\pi;\eta) = (y-\pi)^\top\nabla^2 H(\pi)\,\eta$, and the result follows. □

As a consequence of lemma 2,

$$\mathrm{var}_{Y\sim p}\,S_1(Y,\pi;\eta) = \mathbb{E}_{Y\sim p}\,S_1(Y,\pi;\eta)^2 = \eta^\top\nabla^2 H(\pi)\,\Sigma\,\nabla^2 H(\pi)\,\eta, \tag{30}$$

where Σ is the unknown covariance matrix $\mathrm{cov}_{Y\sim p}(Y)$. Furthermore, $S_2(y,\pi;\eta) = -\eta^\top\nabla^2 H(\pi)\,\eta$, so that (trivially)

$$\mathbb{E}_{Y\sim p}\,S_2(Y,\pi;\eta) = -\eta^\top\nabla^2 H(\pi)\,\eta. \tag{31}$$

To lowest order in ϵ, the objective function becomes

$$\tfrac{1}{4}\epsilon^2\frac{\left(-\eta^\top\nabla^2 H(\pi)\,\eta\right)^2}{\eta^\top\nabla^2 H(\pi)\,\Sigma\,\nabla^2 H(\pi)\,\eta}. \tag{32}$$

We now argue that all linear scoring rules have similar discriminatory and robustness properties, at least for predictions close to the truth. The usual method of Lagrange multipliers shows that the objective function achieves its worst case when η is an eigenvector of $\Sigma\nabla^2 H(\pi)$, where it evaluates to $\frac{1}{4}\epsilon^2\eta^\top\Sigma^{-1}\eta$. Thus the best worst-case scenario is controlled by the data-generating distribution alone – specifically the smallest eigenvalue of Σ^{-1} – and cannot be targeted by any linear scoring rule.

6. Discussion and future work

We have introduced linear scoring rules that can be used in binary classification challenges that call for a vector of class probabilities. We have illustrated important sub-classes of these scoring rules and have shown that they balance the needs of the organizers and the contestants. We have also shown how linear scoring rules can be used to train a classifier. An important question for future work is, given the scoring rule that will be used on the test cases, what is the optimal way to train the classifier?

Acknowledgments

The author gratefully acknowledges the constructive feedback from referees, many useful discussions with Simon Byrne and the hospitality of the Statistical Laboratory at the University of Cambridge.

References

Banerjee, A., Guo, X. and Wang, H. (2005). On the Optimality of Conditional Expectation as a Bregman Predictor. *IEEE Transactions on Information Theory* **51** 2664–2669. MR2246384

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

Buja, A., Stuetzle, W. and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: structure and applications. Technical report available at http://www-stat.wharton.upenn.edu/~buja/.

Byrne, S. (2016). A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electron. J. Statist.* **10** 380–393. MR3466187

Dawid, A. P. and Lauritzen, S. L. (2005). The Geometry of Decision Theory. In *Proceedings of the Second International Symposium on Information Geometry and its Applications* 22–28. University of Tokyo.

Frongillo, R. and Kash, I. (2015). Vector-Valued Property Elicitation. *JMLR: Workshop and Conference Proceedings* **40** 1–18.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** 359–378. MR2345548

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics* **32** 1367–1433. MR2089128

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning* **77** 103–123.

Hendrickson, A. D. and Buehler, R. J. (1971). Proper scores for probability forecasters. *Ann. Math. Statist.* **42** 1916–1921. MR0314430

McCarthy, J. (1956). Measures of the value of information. *Proc. Nat. Acad. Sci.* **42** 654–655.

Parry, M., Dawid, A. P. and Lauritzen, S. (2012). Proper local scoring rules. *Annals of Statistics* **40** 561–592. MR3014317