Cost-Sensitive Classification Meets Proper Scoring Rules

Peter Flach

Intelligent Systems Laboratory, University of Bristol, UK

Abstract. In cost-sensitive classification we penalise false negatives and false positives with distinct cost parameters, jointly referred to as the *cost context*. In this paper I argue that the standard analysis carries an implicit assumption that costs are expressed on an additive scale, summing up to a fixed budget. It is then natural to investigate what happens when we assume alternate scales. The main technical result of the paper is that, for cost contexts expressed on a harmonic scale, expected loss of a probabilistic cost-sensitive classifier is equal to the model's Log-Loss (as opposed to additive cost contexts where expected loss is equal to the model's Brier score for additive cost contexts, as proved by Hernández-Orallo et al. (2012)). Both Brier score and Log-Loss are proper scoring rules used to evaluate probability estimators. I argue that the cost-based perspective allows to enumerate a family of candidate proper scoring rules, and give a preliminary analysis of some of these using cost curves.

Keywords: ROC analysis, cost curve, operating context, classification performance metrics, Brier score, Log-Loss.

1 Introduction and motivation

In cost-sensitive classification we penalise false positives with a cost c_0 and false negatives with a cost c_1, jointly referred to as the *cost context*. It is usually assumed that only the *cost proportion* $c = c_0/(c_0+c_1)$ matters, so that $c_0 = 1, c_1 = 3$; $c_0 = 2/3, c_1 = 2$; and $c_0 = 1/2, c_1 = 3/2$ are all equivalent. Furthermore, the latter cost context has the advantage of leading to cost-sensitive loss being expressed on a scale commensurate with error rate, which has cost context $c_0 = 1, c_1 = 1$.

In this paper I argue that this carries an implicit assumption that costs are expressed on an additive scale, summing up to a fixed budget. It is then natural to investigate what happens when we assume different scales. For example, the second cost context above is commensurate with error rate if we measure costs on a harmonic scale, since $1/c_0 + 1/c_1 = 2$. The main technical result of the paper is that, for harmonic cost contexts, expected loss of a probabilistic classifier which sets its decision threshold equal to c, averaged over uniform c, is equal to the model's Log-Loss (while it is equal to the model's Brier score for additive cost contexts (Hernández-Orallo et al., 2012)). Both Brier score and Log-Loss are so-called *proper scoring rules* used to evaluate probability estimators (Dawid and Musio, 2014). I argue that the cost-based perspective allows to enumerate a family of candidate proper scoring rules, and give a preliminary analysis of some of these using cost curves (Drummond and Holte, 2006).

The outline of the paper is as follows. In Section 2 I give general definitions and notation pertaining to cost-sensitive classification. Section 3 recalls some of the main results of Hernández-Orallo et al. (2012), and Section 4 extends these results to Log-Loss. In Section 5 I discuss other possible cost contexts, and Section 6 concludes with a short discussion.

2 Cost-sensitive classification

This section mostly follows Hernández-Orallo et al. (2012).

A (single-label) *classifier* is a function that maps instances x from an instance space X to classes y from an output space Y. For this paper I will assume binary classifiers, i.e., $Y = \{0,1\}$. A *model* is a function $m : X \to \mathbb{R}$ that maps examples to real numbers (scores) on an unspecified scale. I use the convention that higher scores express a stronger belief that the instance is of class 1. A *probabilistic model* is a function $m : X \to [0,1]$ that maps examples to estimates $\hat{p}(1|x)$ of the probability of example x to be of class 1. In order to make predictions in the Y domain, a model can be converted to a classifier by fixing a decision threshold t on the scores. Given a predicted score $s = m(x)$, the instance x is classified in class 1 if $s > t$, and in class 0 otherwise. Given a dataset $D \subset \langle X,Y \rangle$ of size $n = |D|$, I denote by D_k the subset of examples in class $k \in \{0,1\}$, and set $n_k = |D_k|$. The (positive) *class proportion* or *class context* is then $\pi = n_0/n$.

For a given, unspecified model and population from which data are drawn, I denote the score density for class k by f_k and the cumulative distribution function by F_k. Thus, $F_0(t) = \int_{-\infty}^{t} f_0(s)ds = P(s \leq t|0)$ is the proportion of class 0 points correctly classified if the decision threshold is t, which is the sensitivity or true positive rate at t. Similarly, $F_1(t) = \int_{-\infty}^{t} f_1(s)ds = P(s \leq t|1)$ is the proportion of class 1 points incorrectly classified as 0 or the false positive rate at threshold t. The average score of actual class k is $\bar{s}_k = \int_0^1 s f_k(s)ds$.[1]

In classification a typical loss function is the error rate, which can be defined in terms of class distribution and true and false positive rates as follows:

$$Q(t;\pi) \triangleq \pi(1 - F_0(t)) + (1-\pi)F_1(t) \tag{1}$$

In cost-based classification we can denote the cost of misclassifying a positive as c_0 and of misclassifying a negative as c_1, leading to the following cost-sensitive loss:

$$Q(t;\pi,c_0,c_1) \triangleq c_0\pi(1 - F_0(t)) + c_1(1-\pi)F_1(t) \tag{2}$$

3 Brier score and Brier curve

Setting $b \triangleq c_0 + c_1$ for the cost associated with misclassifying one positive and one negative, and $c \triangleq c_0/b$ for the relative cost of misclassifying a positive, we obtain the

[1] Note that I use 0 for the positive class and 1 for the negative class, but scores increase with $\hat{p}(1|x)$. That is, $F_0(t)$ and $F_1(t)$ are monotonically non-decreasing with t. This has some notational advantages and is the same convention as used by, e.g., Hand (2009).

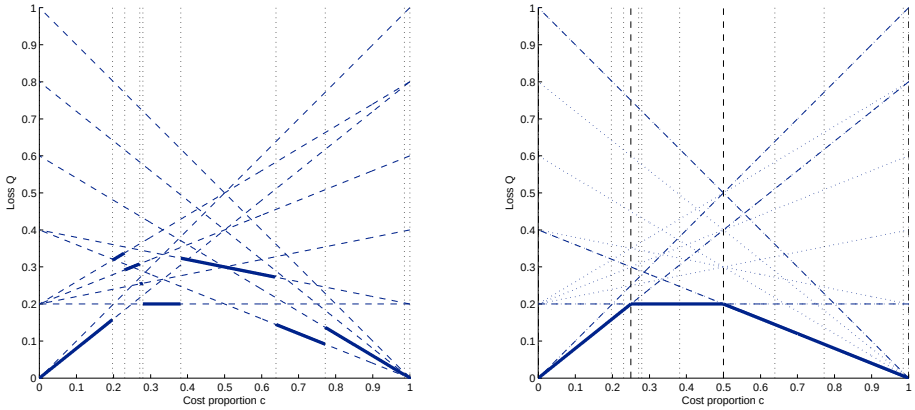

Fig. 1. (left) The Brier curve is a piecewise linear cost curve (solid lines) jumping between different cost lines (dashed lines). The vertical dotted lines denote actual scores assigned by the model, and hence determine the values of c where the operating point changes. Score-driven thresholds are sub-optimal whenever the Brier curve departs from the lower envelope. The area under the Brier curve is the Brier score. **(right)** Optimal Brier curve resulting from perfectly calibrated scores (dashed verticals).

following alternative parametrisation:

$$Q(t;\pi,b,c) \triangleq b\{c\pi(1-F_0(t))+(1-c)(1-\pi)F_1(t)\} \tag{3}$$

Hernández-Orallo et al. (2012) argued that it makes sense in many situations to assume b and c independent (I will revisit this later). If we also assume π fixed this leads to the following expected loss:

$$L = \mathbb{E}\{b\}\int_0^1 \{c\pi(1-F_0(t))+(1-c)(1-\pi)F_1(t)\}\,w(c)dc \tag{4}$$

This means that for the expected loss the variability in b is only captured through its expected value, which can be rescaled to $\mathbb{E}\{b\}=2$ to ensure commensurability with error rate.

Hernández-Orallo et al. (2011) introduced the *score-driven* threshold choice method which sets the threshold of a probabilistic classifier equal to c in order to account for unequal misclassification costs. They obtained the following result: assuming probabilistic scores and a decision threshold equal to the cost context c, expected loss under a uniform distribution of cost contexts is equal to the model's Brier score

$$BS = \pi\int_0^1 s^2 f_0(s)ds + (1-\pi)\int_0^1 (1-s)^2 f_1(s)ds \tag{5}$$

Using integration by parts we can rewrite the per-class components of the Brier score in terms of true and false positive rate:

$$BS_0 = \int_0^1 s^2 f_0(s)ds = \int_0^1 2s(1-F_0(s))ds \tag{6}$$

$$BS_1 = \int_0^1 (1-s)^2 f_1(s)ds = \int_0^1 2(1-s)F_1(s)ds \tag{7}$$

The right-most expressions under the integral sign are linear in s, and so the Brier score can be constructed as the area under a piecewise linear cost curve known as the *Brier curve* (Hernández-Orallo et al., 2011). Example Brier curves are given in Figure 1.

4 Harmonic cost contexts and Log-Loss

In the previous section we had $c = c_0/(c_0+c_1)$ and b equal to the sum of c_0 and c_1, i.e., twice their arithmetic mean. Alternatively, we can take d to be half the harmonic mean of c_0 and c_1: i.e., $1/d \triangleq 1/c_0 + 1/c_1$.[2] Keeping the definition of c the same, it now follows that $c_1 = d/c$ and $c_0 = d/(1-c)$. Instead of assuming b and c independent we may wonder what happens if we assume c and d independent – we will call these *harmonic* cost contexts (contrasting with the additive cost contexts in the previous section).

The expression for loss is now

$$Q(t;\pi,d,c) \triangleq d\left\{\frac{1}{1-c}\pi(1-F_0(t)) + \frac{1}{c}(1-\pi)F_1(t)\right\} \tag{8}$$

and hence expected loss becomes

$$L = \mathbb{E}\{d\}\int_0^1 \left\{\frac{1}{1-c}\pi(1-F_0(t)) + \frac{1}{c}(1-\pi)F_1(t)\right\} w(c)dc \tag{9}$$

To keep commensurability with error rate I will assume $\mathbb{E}\{d\} = 1/2$.

We can then obtain the following novel result.

Theorem 1. *Assuming probabilistic scores, harmonic cost contexts and a decision threshold equal to the cost parameter c, expected loss under a uniform distribution of c is equal to* Log-Loss*, defined as*

$$LL = \pi/2\int \ln\frac{1}{1-s}f_0(s)ds + (1-\pi)/2\int \ln\frac{1}{s}f_1(s)ds = \pi LL_0/2 + (1-\pi)LL_1/2 \tag{10}$$

Proof. Use integration by parts to obtain

$$\int \frac{1}{1-s}(1-F_0(s))ds = \int \ln\frac{1}{1-s}f_0(s)ds = LL_0 \tag{11}$$

$$\int_0^1 \frac{1}{s}F_1(s)ds = \int \ln\frac{1}{s}f_1(s)ds = LL_1 \tag{12}$$

and use this to rewrite Equation 9, noting that $\mathbb{E}\{d\} = 1/2$, $w(c) = 1$ and $t = c = s$.

[2] This is in fact the choice made by (Zhao et al., 2013, p.1048) – they write '[...] we find it most convenient to normalize the two costs c_0 and c_1 (by multiplying them by a common factor) such that $\frac{c_0 c_1}{c_0+c_1} = 0.5$, that is, $c_0^{-1} + c_1^{-1} = 2$.'

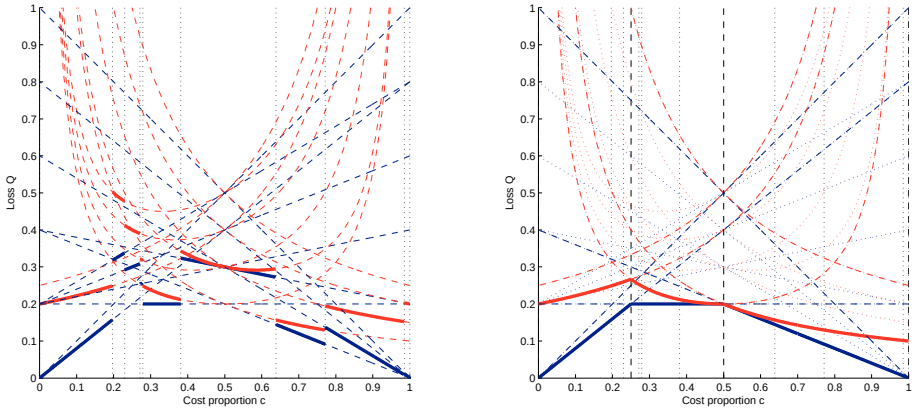

Fig. 2. (left) Comparison between Brier curve (in blue) and Log-Loss curve (in red). Comparatively, the Log-Loss curve is more affected by the model's behaviour at extreme values of c. **(right)** Optimal Brier and Log-Loss curves.

We can thus draw *Log-Loss curves* in a similar way to Brier curves, such that the area under the Log-Loss curve is equal to Log-Loss. Figure 2 shows this for the running example. Note that cost lines for harmonic cost contexts are not linear but hyperbolic. This difference arises since, for example, in additive cost contexts $c = 0$ means $c_0 = 0$ and $c_1 = b$; whereas in harmonic cost contexts it means $c_1 = \infty$ and $c_0 = d$. Consequently, the always-positive classifier has zero loss in an additive cost context with $c = 0$, whereas it has non-zero loss $d\pi$ in a harmonic cost context with the same cost proportion. Conversely, if $c = 1/2$ the two losses are equal. In conclusion, Log-Loss emphasises the extreme values of c when evaluating a model.

5 A family of cost contexts

In this section I will generalise the idea of a cost context with associated scale further. We have seen two examples where the cost parameters are governed by a fixed budget: in the additive scenario we have $c_0 + c_1 = b$, whereas in the harmonic case we have $1/c_0 + 1/c_1 = 1/d$. In both cases this gives a functional relationship between c_0 and c_1 which is symmetric in c_0 and c_1 and monotonically decreasing. I will now consider a few other choices which are visualised in Figure 3.

5.1 Geometric cost contexts

Define $e^2 \triangleq c_0 c_1$ to be the square of the geometric mean of c_0 and c_1, then

$$c_0 = e\sqrt{\frac{c}{1-c}} \tag{13}$$

$$c_1 = e\sqrt{\frac{1-c}{c}} \tag{14}$$

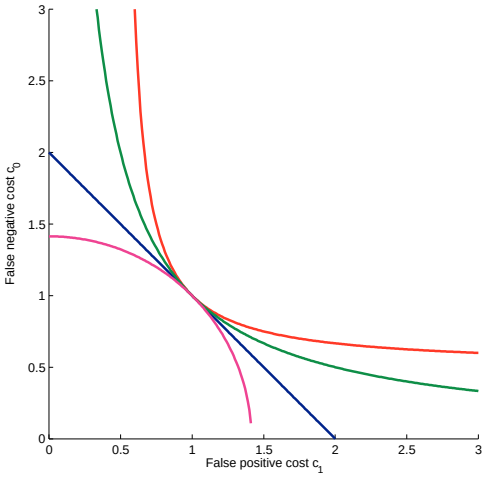

Fig. 3. Different cost contexts: additive (in blue), harmonic (in red), geometric (in green) and Euclidean (in violet). All of these are scaled to be commensurate to error rate, which means that they all go through $(c_0 = 1, c_1 = 1)$.

The expression for loss becomes

$$Q(t;\pi,e,c) \triangleq e\left\{\sqrt{\frac{c}{1-c}}\pi(1-F_0(t)) + \sqrt{\frac{1-c}{c}}(1-\pi)F_1(t)\right\} \tag{15}$$

and hence expected loss is now

$$L = \mathbb{E}\{e\}\int_0^1 \left\{\sqrt{\frac{c}{1-c}}\pi(1-F_0(t)) + \sqrt{\frac{1-c}{c}}(1-\pi)F_1(t)\right\} w(c)dc \tag{16}$$

To keep commensurability with error rate I will assume $\mathbb{E}\{e\} = 1$.

Equation 16 is sufficient to draw a cost curve for geometric cost contexts (Figure 4). I leave the derivation of a corresponding expected loss metric, which requires rewriting the integral in terms of score densities, for future work.

5.2 A general case

I will now derive a general formulation covering all cost contexts considered in this paper except the geometric one. Put $f^k \triangleq c_0^k + c_1^k$ for arbitrary real $k \neq 0$, then

$$c_0 = f\left(\frac{c^k}{(1-c)^k + c^k}\right)^{1/k} \tag{17}$$

$$c_1 = f\left(\frac{(1-c)^k}{(1-c)^k + c^k}\right)^{1/k} \tag{18}$$

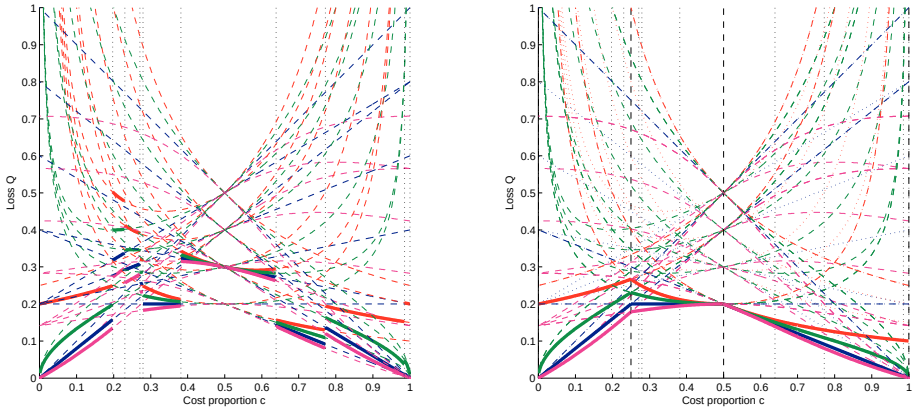

Fig. 4. (left) In addition to the Brier curve for additive cost contexts (in blue) and the Log-Loss curve for harmonic contexts (in red), this plot shows cost curves for geometric cost contexts (in green) and Euclidean cost contexts (in violet). **(right)** Corresponding optimal cost curves.

The expression for loss becomes

$$Q(t;\pi,e,c) \triangleq f\left\{\left(\frac{c^k}{(1-c)^k+c^k}\right)^{1/k}\pi(1-F_0(t))+\left(\frac{(1-c)^k}{(1-c)^k+c^k}\right)^{1/k}(1-\pi)F_1(t)\right\} \tag{19}$$

and hence expected loss becomes

$$L=\mathbb{E}\{f\}\int_0^1\left(\frac{c^k}{(1-c)^k+c^k}\right)^{1/k}\pi(1-F_0(t))+\left(\frac{(1-c)^k}{(1-c)^k+c^k}\right)^{1/k}(1-\pi)F_1(t)\Bigg\}w(c)dc \tag{20}$$

Commensurability with error rate requires $\mathbb{E}\{f\}=2^{1/k}$. This general expression covers additive cost contexts ($k=1$) and harmonic cost contexts ($k=-1$), as well as, e.g., Euclidean cost contexts ($c=2$), all three of which are plotted in Figure 4 together with geometric cost contexts.

6 Discussion

A scoring rule is designed to evaluate probability estimates. A wide variety of such scoring rules exists (Dawid and Musio, 2014), yet the concept occurs only sporadically in the machine learning literature. Exceptions include Hernández-Orallo et al. (2012), which connects cost-sensitive classification to the Brier score, and Zhao et al. (2013), which makes an alternative connection to Log-Loss and information entropy. This paper attempts to lift the discussion to a more general level by distinguishing different scales on which cost contexts can be expressed. Besides the additive and harmonic scales

on which the previous two results are based, and which can be seen as instantiations of a more general case, I have considered the geometric case. Deriving the geometric expected loss is an interesting open problem as the form of the loss suggests a possible connection with boosting.

Putting forward these alternatives raises the question whether some may be better suited than others. This question can be addressed in different ways. One possible answer would be that it is ultimately the application context which dictates the scale on which costs are expressed – in the absence of such information, additive costs seem most intuitive. Another way to address this question is to look at the shape of the cost curves. I have already remarked that Log-Loss puts more emphasis than the Brier score on extreme values of c, as can be clearly seen in Figure 2. One can argue, with Hand (2009), that such extreme values are much less likely and should rather be de-emphasised. This would then be an argument against the use of Log-Loss (another argument against Log-Loss is that in a multi-class setting it penalises uncertainty regarding the true class only). Conversely, it can be seen in Figure 4 that Euclidean cost contexts de-emphasise extreme c values a bit more than the additive context, which suggests that this may be another interesting candidate to consider.

Acknowledgments

This work was partly supported by the REFRAME project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences and Technologies ERA-Net (CHISTERA), and funded by the Engineering and Physical Sciences Research Council in the UK under grant EP/K018728/1. Discussions with Meelis Kull, José Hernández-Orallo and other members of the Intelligent Systems Lab and the REFRAME project have helped to shape the ideas in this paper. Constructive comments by the anonymous reviewers are gratefully acknowledged.

Bibliography

Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *METRON*, 72(2):169–183, 2014.

Chris Drummond and Robert C. Holte. Cost curves: an improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006.

David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77(1):103–123, 2009.

José Hernández-Orallo, Peter A. Flach, and Cèsar Ferri. Brier curves: a new cost-based visualisation of classifier performance. In *28th International Conference on Machine Learning*, pages 585–592, 2011.

José Hernández-Orallo, Peter A. Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, 2012.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14(1):1033–1090, 2013.