UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Diana Grygorian

Classifier Evaluation With Proper Scoring Rules

Master's Thesis (30 ECTS)

Supervisor: Meelis Kull, PhD

Tartu 2019

Classifier Evaluation With Proper Scoring Rules

Abstract: Classification is a fundamental task in machine learning, which involves predicting the class of a data instance based on a set of features. Performance of a classifier can be measured using a loss function, which assigns a loss value for each classification error.

Classification error happens when the predicted and the actual class differ. In the simplest case, all combinations resulting in a classification error are considered equal in terms of cost. However, some problems demand different types of misclassification to be of different importance, which forms a cost context.

Depending on the properties of the cost contexts, different loss functions can be applied. For example, if the arithmetic mean of costs for one false positive and one false negative is fixed and these costs are uniformly distributed, then Brier score is the suitable loss function. If their harmonic mean is fixed, then log loss should be used instead. These two functions belong to a larger family of loss functions known as proper scoring rules. Scoring rules are loss functions which deal specifically with probabilistic classification, where the classifier is required to predict probability for each class, indicating prediction confidence.

In this thesis, a new cost context for binary classification is presented, where both costs have their own uniform distributions. A corresponding new loss function for this cost context is proposed, named Inverse Score, and is subsequently proven to be a proper scoring rule.

The experiments confirm that the total cost when using said cost context and expected loss when using the new loss function are the same.

Keywords: machine learning, classifier evaluation, probabilistic classification, proper scoring rules, cost-sensitive learning

CERCS: P176 – Artificial intelligence

Klassifikaatorite hindamine kohaste skoorimisreeglitega

Lühikokkuvõte: Üks põhilisi ülesandeid masinõppes on klassifitseerimine, mis seisneb andmepunktile kategoorse väärtuse ennustamises teatud tunnuste alusel. Klassifitseerija sooritusvõimet saab mõõta kaofunktsiooni abil, mis omistab igale klassitsifeerimisel tehtud veale mingi väärtuse.

Klassifitseerimisveaks nimetatakse olukorda, kus ennustatud kategoorne väärtus on erinev sellest, mis peaks olema tegelik väärtus. Kõige lihtsam on käsitleda kõikvõimalikke klassifitseerimisvigu võrdse kuluga. Siiski, mõndade probleemide lahendamine nõuab erinevat tüüpi klassifitseerimisvigadele erineva kaalu omistamist, ning see moodustab kaokonteksti.

Olenevalt kaokontekstist on võimalik rakendada erinevaid kaofunktsioone. Näiteks, kui ühe valepositiivse ja ühe valenegatiivse hindade aritmeetiline keskmine on fikseeritud ning mõlemad on ühtlaselt jaotunud, sobib kaofunktsiooniks Brier'i skoor. Kui nende harmooniline keskmine on fikseeritud, sobib selle asemel kasutada logaritmilist kaofunktsiooni. Need kaks funktsiooni kuuluvad suuremasse kaofunktsioonide perekonda, mida tuntakse kohaste skoorimisreeglite nime all. Skoorimisreeglid on kaofunktsioonid mis tegelevad spetsiifiliselt tõenäosusliku klassifitseerimisega, kus klassifitseerijalt on oodatud iga kategooria tõenäosuse ennustamist, kus tõenäosus omakorda näitab kindlust ennustatud kategoorias.

Antud magistritöös esitletakse uut kaokonteksti binaarsele klassifitseerimisele, kus kummalgi klassil on sõltumatult ühtlane jaotus. Nimetatud kaokontekstile pakutakse välja uus kaofunktsioon nimega Pöördskoor ning selle puhul tõestatakse, et see on kohane skoorimisreegel.

Eksperimendid kinnitavad, et kogukulu vastavas kaokontekstis ning oodatud kadu kasutades uut kaofunktsiooni on samad.

Võtmesõnad: masinõpe, klassifikaatorite hindamine, tõenäosuslik klassifitseerimine, kohased skoorimisreeglid, hinnatundlik õpe

CERCS: P176 – Tehisintellekt

Contents

1 Introduction

Classification is the task of given an instance of data predicting its class from a fixed set of options. An instance is defined through its *features*, denoted by x, an its corresponding *class*, denoted by y. For example, one can consider a classification task of predicting whether a given image is of a chihuahua or of a blueberry muffin. The set of features are all the pixels in the image. Feature space, denoted by X, is composed by all the possible images of given image size. Label space, denoted by Y, consists of two labels: 'chihuahua' and 'muffin' [Yao, 2017].

Classifier evaluation is a task of measuring how good classifier is at labelling data. According to the 'no free lunch' (NFL) theorem [Wolpert, 1996], there is no universal learning algorithm that would work best for every problem. Therefore, for different datasets different algorithms (classifiers) perform with variable success. This means that for each task and dataset classifiers need to be reviewed and evaluated separately. That is why classifier evaluation is a crucial task in machine learning [Tharwat, 2018].

Loss function is a type of classification evaluation measurement. It is a non-negative function that maps a pair of predicted and actual class to a real number, where higher values correspond to worse predictions while 0 indicates perfect classification.

Binary classification is a particular type of classification problems where there are only 2 classes, which are usually referred to as *positive* and *negative*. This allows for 2 cases of misclassification: *false negative* (predicting negative when the actual class is positive) and *false positive* (predicting positive when the actual class is negative). Most of the time, these errors are of different importance, and to highlight these differences parameters called costs are used. For binary classification, there are only 2 costs: cost of false negative, denoted by c_0, and cost of false positive, denoted by c_1 [Elkan, 2001].

Evaluation is complicated by the fact that models learn and work in different contexts. A context is defined by a set of parameters, such as cost magnitude, proportion of different costs, proportion of instances of different classes. Such contexts are called *operating condition*. Evaluating model in a concrete operating condition is a trivial task. In case information about operating condition is unavailable, a model needs to be evaluated over a range of operating conditions [Hernández-Orallo et al., 2012]. Loss functions applicable to such problems are called *proper scoring rules*. The most commonly used examples of proper scoring rules are Brier score and log-loss [Brier, 1950] [Good, 1952].

An operating condition, where the sum of costs is fixed, one cost, c_0, has uniform distribution and the other one, c_1, can be calculated from the sum and c_0, is called *additive*. It was shown that Brier score produces correct expected loss when used with additive context [Hernández-Orallo et al., 2011]. A similar operating

condition where instead the harmonic mean of costs is fixed, one cost, c_0, has uniform distribution and another one, c_1, can be calculated from the harmonic mean and c_0, is called *harmonic*, and it was proven that log-loss is equal to expected loss under harmonic context [Flach, 2015].

In this thesis, a new context is proposed, where both costs c_0 and c_1 have uniform distribution. Since in this case there is no correlation between c_0 and c_1, this context will be called *independent uniform*. A score was found that is equal to total cost under *independent uniform* cost context and its correctness and properness is proven. This score will be called *Inverse Score*.

In Section 2, types of classification and their evaluation will be introduced, and cost-sensitive learning will be discussed.

In Section 3, proper scoring rules will be reviewed, some examples of them are shown and their properness is proven, and Bregman divergences and their connection to proper scoring rules are introduced with some examples.

In Section 4, independent cost context and its appropriate proper scoring rule Inverse Score are defined and the theorems about correctness and properness are proven.

In Section 5, the results of experiments are displayed and analyzed.

In Section 6, the thesis is concluded.

2 Classifier Evaluation

A *classifier*, or *classification model*, is a function that implements classification, so its input is a list of instance's features, and its output is predicted class [Hernández-Orallo et al., 2012].

In this section, different types of classification models will be discussed and their assessment methods will be introduced.

2.1 Types of Classifiers

Labelling Classifier *Labelling classifier* model is a function that produces a predicted class, given an instance with set of values called features. For a concrete instance, its features are denoted by x and its label as y. Then, feature space is set of possible features, denoted by $\mathscr{X}$, and label space is set of all possible labels to chose from, denoted by $\mathscr{Y}$. In this thesis, classification with 2 labels will be discussed unless stated otherwise. Such classification is called *binary*. For binary classification, class 0 will represent positive class and class 1 will represent negative one. Following [Hernández-Orallo et al., 2012] and [Hand, 2009], it has notational advantages. Then $\mathscr{Y} = \{0, 1\}$.

Probabilistic Classifier A *probabilistic classifier* is a mapping $\hat{\mathbf{p}} : X \to [0, 1]$, where $\hat{\mathbf{p}}$ is a vector of probabilities of getting each class, $\hat{\mathbf{p}} = (\hat{p}_1, ..., \hat{p}_k)$. Probabilities of all classed add up to one, $\sum_{i=1}^{k} \hat{p}_i = 1$. In binary classification, a *score* $\hat{p}$ specifies how likely instance x is of class 1, with higher value of $\hat{p}$ meaning more likely, $\hat{p} \approx P(Y = 1|X = x)$. And $1 - \hat{p}$ specifies how likely instance x is of class 0, $1 - \hat{p} \approx P(Y = 0|X = x)$.

Probabilistic classifier gains labels from scores using a *decision rule*. The most common decision rule uses a *threshold* t, so instances with scores $\hat{p} > t$ would be predicted as of class 1, and as of class 0 otherwise [Hernández-Orallo et al., 2011]. To discuss what value threshold should have, we need to review calibration first.

For example, for a fixed instance model produces probability $\hat{p} = 0.9$ of being of negative class. A model is called well-calibrated, if for all such instances with probability $\hat{p} = 0.9 = 90\%$ approximately 90% of them are actually of negative class. It is very unlikely for the proportion of negative instances to be exactly 0.9, but the closer proportion of negative instances is to probability, the better calibrated it is.

So, probability estimator is calibrated if among all instances where the model predicts $(\hat{p}_1, ..., \hat{p}_k)$ the actual class distribution is also $(\hat{p}_1, ..., \hat{p}_k)$.

If model is calibrated, then a threshold should be such value that means that class that has a bigger probability should be predicted. To predict negative class that has probability $\hat{p}$, the following inequality needs to be satisfied:

$$\hat{p} > 1 - \hat{p}$$
$$2\hat{p} > 1$$
$$\hat{p} > 0.5$$

Thus, model will predict class 1 if probability $\hat{p}$ is more than 0.5, which means that decision threshold t is $t = 0.5$. So, model will predict class 1 if $\hat{p} > t$ and class 0 otherwise.

We will call probabilistic classifier *calibrated* because it produces a value that is always in range $[0;1]$. A score s does not have such restrictions; it can have any range, where higher value will mean more likely of negative class 1 and lower value will mean more likely of positive class 0. Such score will be termed *uncalibrated*. We will discuss uncalibrated scores, their calibration to probabilistic scores and decision rules in Section 2.2.

2.2 Basic Evaluation Measurements

Accuracy and Error Rate *Accuracy* is primary tool used for classification evaluation and is equal to proportion of correctly predicted instances. It is obvious that accuracy has values in range $[0;1]$, and higher accuracy (closer to 1) means good classification and lower accuracy (closer 0) means worse.

Error rate is another measurement and it is equal to proportion of incorrectly predicted instances. It is a complement of accuracy, $err = 1 - acc$, and lower error rate means better than higher error rate.

Accuracy and error rate are not the best tool, because they do not evaluate classes separately. In real world, classes are sometimes highly imbalanced. It is common that negative class represents ordinary instances, and positive class means in some way 'outstanding' and therefore interesting instance. In such cases negative instances outweigh positive instances, and developer is interested in predicting positive instances correctly.

Consider online advertisement as an example, where a developer needs to predict which users will click on the ad. Then, negative instance means that user did not click on ad and positive click means user did. Most people do not click on the ads, so a good click-through rate (proportion of positive class) is only 2%, which means that classes are highly imbalanced [Volovich, 2019]. It is obvious that a developer is more interested to find users who would click rather than not, so positive class is more important and prediction for positive class needs to be more precise.

And accuracy fails this task: if classes are imbalanced enough, positive prediction can be neglected altogether without losing high accuracy. Similarly, error rate can ignore a rare class as well.

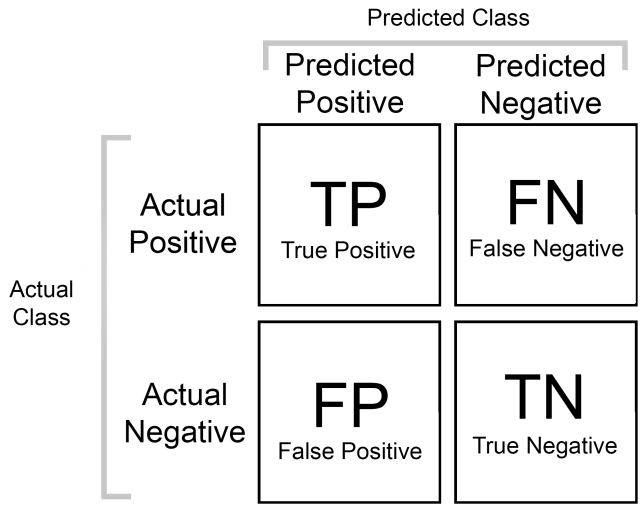

Figure 1. Confusion matrix for binary classification. Each entry in a cell is a count of instances with such prediction.

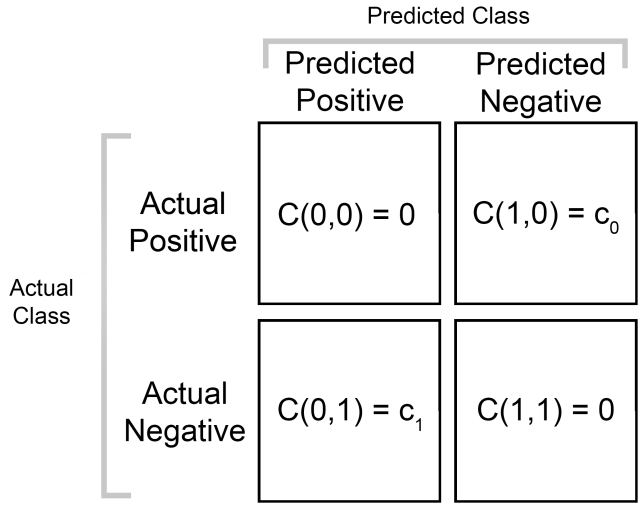

Figure 2. Cost matrix for binary classification. Each entry in a cell is a cost of prediction of certain instance.

Metrics for Imbalanced Classes In Figure 1, there is a confusion matrix, which is used to calculate most of common evaluation metrics. There are 4 possible cases for a single instance. If instance is predicted positive and is actually of positive class, then it is *true positive (TP)*. If instance is predicted negative and is actually positive, then it is *false negative (FN)*. Similarly, if instance is actually negative and is predicted positive or negative, it is *false positive (FP)* or *true negative (TN)*, respectively.

In such cases, *precision* and *recall* are pretty useful, since they focus on predicting positive class (the rare one) correctly. Precision, also called positive predicted value, is equal to $\frac{TP}{TP+FP}$, and recall, in also called true positive rate, is equal to $\frac{TP}{TP+FN}$ [Tharwat, 2018] [Fawcett, 2005] .

There are many more classification evaluation measurements, like, for example, F_1 *score*, which is harmonic mean between precision and recall, or *Informedness*, that takes into account true positive and true negative rates. Every one of them is useful and could be an ultimate way to measure adequacy of some models if classification is not complicated further [Fawcett, 2015] [Powers, 2011].

Loss Functions The previous paragraphs gave a lot of examples of evaluations where higher result (closer to 1) is better and lower result (closer to 0) is worse. However, that is not always the case, and there are a lot of measures for which 0 is the perfect result and more is bad result. Such measures are called loss measures, or, more commonly, loss functions.

A *loss function* (or cost function) for probabilistic classifiers is a non-negative function that takes as an input model's score of an instance and instance's actual label, and outputs a real number. Low value of loss function indicates better result, with 0 indicating perfect match between a prediction and instance's label. The goal for optimization is to minimize the loss function.

Error rate was already introduced earlier and it is one of the easiest loss functions, and is equal to the proportion of incorrectly predicted instances. Most of loss functions are more complex than that as they are based on probabilistic classifiers rather than labelling ones.

Bayes-Optimal Model The best possible model is called *Bayes-optimal model*, denoted by $\mathbf{p}^*(x)$, $\mathbf{p}^*(x) = \left(p_1^*(x), ..., p_k^*(x)\right)$, where $p_i^*(x) = P(Y = i|X = x)$. That is, for each instance, it produces probabilities of getting every label, given instance's features. This model is theoretical and cannot be learned.

For binary classification, we will have classes 0 and 1, and only one probability p will be used: $p = p_1^*(x)$ will mean probability of class 1 and $1 - p = p_0^*(x)$ will mean probability of class 0. Bayes-optimal is perfectly calibrated, which means it has threshold $t = 0.5$:

$$
\begin{aligned}
p_1^*(x) &> p_0^*(x) \\
p &> 1-p \\
2p &> 1 \\
p &> 0.5
\end{aligned}
$$

2.3 Cost-Sensitive Classification

As it was already discussed in Section 2.2, false negatives and false positive errors need to be taken into account separately. Moreover, it is common that one kind of error (usually false negative) is worse than another. For example, a bank needs to find from client's transactions unusual (possibly fraud) transactions. Let the usual transactions be of negative class and unusual ones be of positive class. Then, it is a much bigger problem of classifying fraud as usual payment (and missing theft) than it is to classify casualty as fraud (a client would simply confirm their transaction and payment will go through) [Sun et al., 2011]. To indicate such differences in importance, there is set of parameters called *costs*, which is a measure of how bad consequences of a certain prediction is.

The general assumption about cost-sensitive classification is that the cost does not depend on the instance itself but rather on its class. Therefore, costs are usually represented as cost matrices for simplicity, and an example of cost matrix for binary case is visualized in Figure 2.

Cost $C(i,j)$ stands for cost of predicting class i when the actual class is j. It is common for right predictions to not be penalized at all; therefore it will be assumed that $C(0,0) = 0$ and $C(1,1) = 0$. Negative values of costs will not be explored in this work. Conceptually, cost of misclassification should always be greater than predicting correctly. Mathematically, $C(0,1) > C(1,1) \Rightarrow C(0,1) > 0$ and $C(1,0) > C(0,0) \Rightarrow C(1,1) > 0$. These conditions are called the *reasonableness* conditions [Elkan, 2001].

For simpler representation, $C(0,1)$ (false positive) will be further denoted as c_1 and $C(1,0)$ (false negative) will be denoted as c_0, where indexes of c_0, c_1 denote the actual class.

Values (or distributions) of all costs are called *cost context*.

Cost-Sensitive Bayes-Optimal Model Bayes-Optimal model was already briefly discussed in Section 2.2. It was claimed that most of the times it uses threshold $t = 0.5$. However, for cost-sensitive learning there will be different threshold.

Probability of getting class 1 is $p_1 = p$ and probability of getting class 0 is $p_0 = 1 - p$. Then, classifier will predict class 1 if its probability is bigger when

probabilities are multiplied by respective costs, $p_1 c_1 > p_0 c_0$:

$$\begin{aligned} p_1 c_1 &> p_0 c_0 \\ p c_1 &> (1-p) c_0 \\ p c_1 &> c_0 - p c_0 \\ p c_1 + p c_0 &> c_0 \\ p &> \frac{c_0}{c_0 + c_1} \end{aligned}$$

Thus, classifier will predict class 1 if probability p is bigger than $\frac{c_0}{c_0+c_1}$. That means that cost-sensitive threshold is $t = \frac{c0}{c0+c1}$.

Probability Density Function, Cumulative Distribution Function According to the notation in Section 2.1, positive class is denoted by 0 and negative class is denoted by 1.

Probability density function (p.d.f.) for probabilistic scores s of class 0 is denoted by $f_0(s)$, and cumulative distribution function (c.d.f.) is denoted by $F_0(s)$.

Defined integral of p.d.f under limits $[a;b]$, $\int_a^b f_0(s)ds$ means probability of getting score s in range $[a;b]$, $a < s < b$.

And c.d.f at value t, $F_0(t) = \int_{-\infty}^{t} f_0(s)ds$, means probability of getting s in range $[-\infty;t]$, or simply less than t, $s < t$.

If t is a threshold, then $F_0(t)$ describes all the positive instances which scores are less than the threshold. Positive instances below the threshold will be classified as positive; that means that $F_0(t)$ is probability of positive instances to be predicted as positive, which is true positive rate, $F_0(t) = TP/Pos$, where TP is expected number of true positive instances and Pos is actual number of all positive instances. Following that, true negative rate is a complement of true positive rate, which means true negative rate is equal to $(1 - F_0(t))$.

Similarly, $F_1(t)$ describes all the negative instances which scores are less than the threshold t. Negative instances below the threshold will be classified as positive, which makes $F_1(t)$ false positive rate, $F_1(t) = FP/Neg$, where FP stands for expected number of false positives and Neg is actual number of all negative instances.

Lower scores are more likely to be of positive class 0, and greater scores are more likely to be of class 1. Let this relationship be described by two normal distributions, as shown in Figure 3. For positive class, $f_0 \sim N(-1,1)$ and for negative class $f_1 \sim N(1,1)$. The same distributions for scores of positive and negative classes will also be used in experiments in Section 5.

Two cases will be reviewed: in Figure 3 costs are of the same value, $c_0 = c_1$,

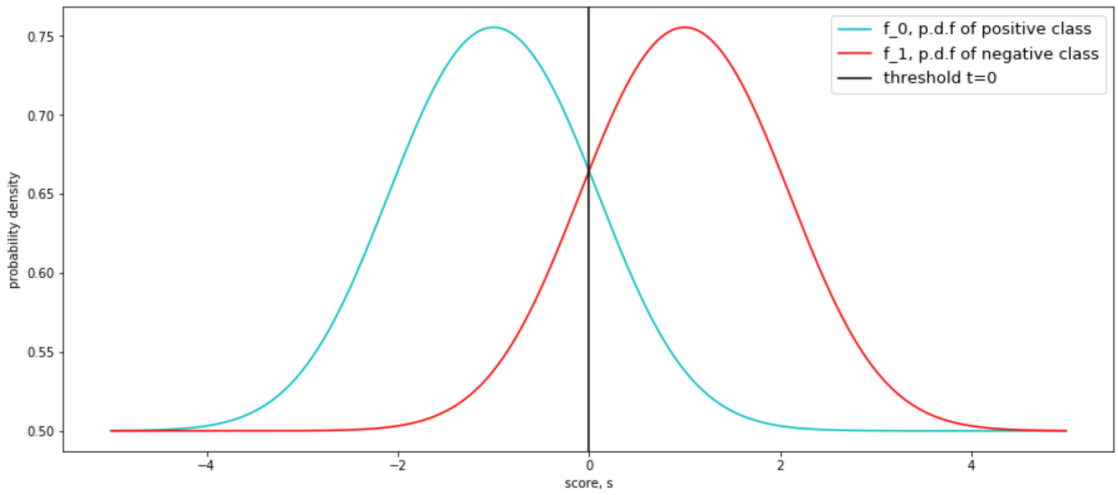

Figure 3. Probability density functions of scores s for positive and negative class f_0 and f_1, respectively. Threshold is $t = 0$ because classes are calibrated and have the same cost.

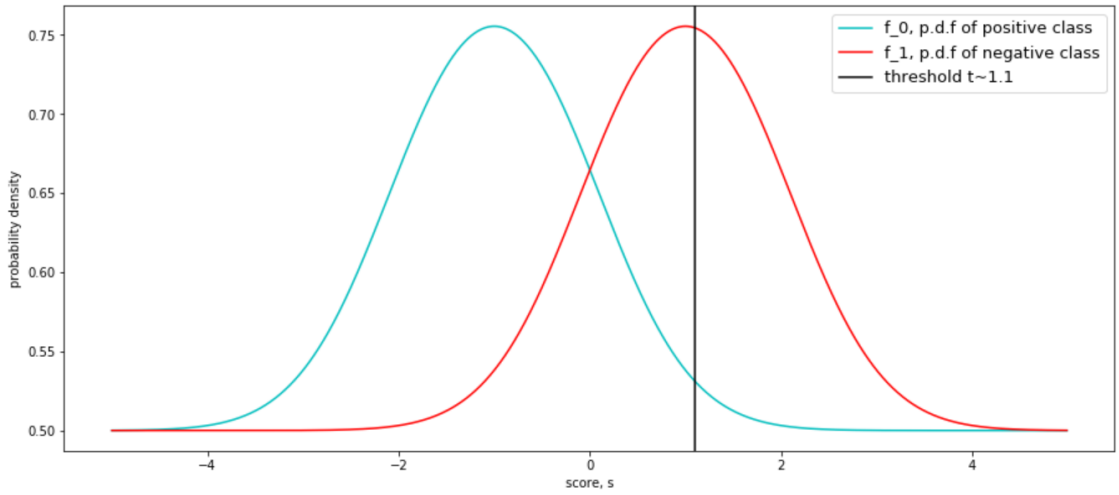

Figure 4. Probability density functions of scores s for positive and negative class f_0 and f_1, respectively. Threshold is $t = (\ln 9)/2 \approx 1.1$ because class errors are of different cost, $c_0 = 9c_1$.

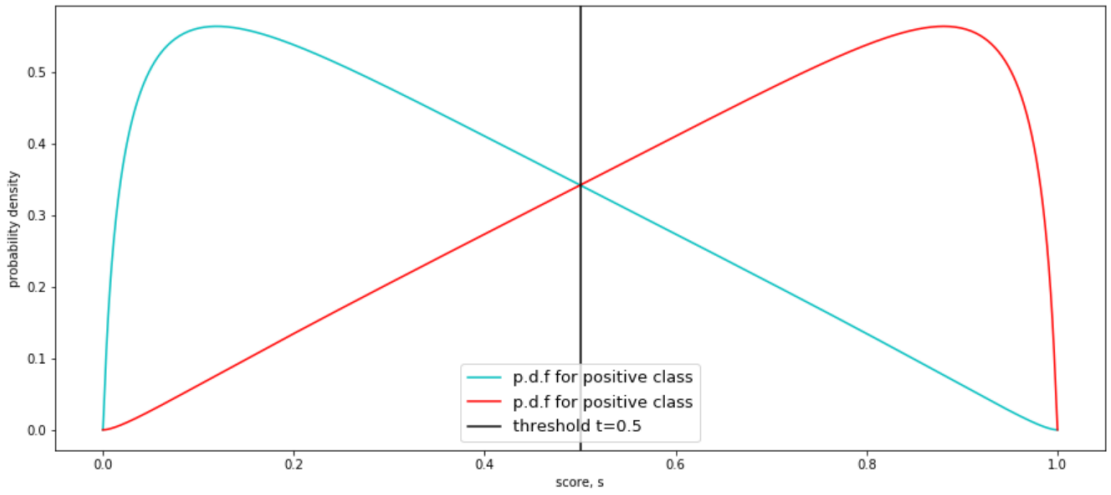

Figure 5. Probability density functions of probabilistic p for positive and negative classes. Threshold is $t = 0.5$ because classes are calibrated and have the same cost.

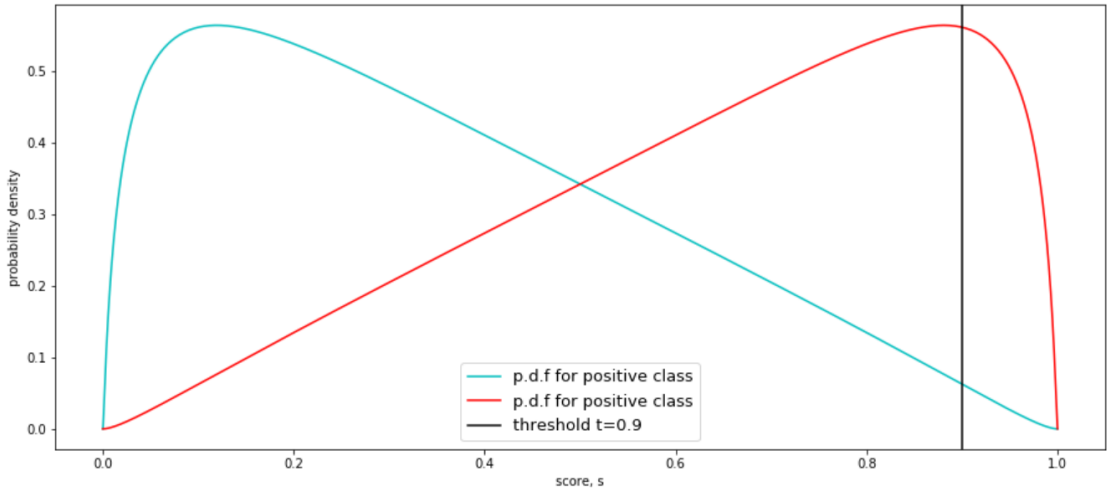

Figure 6. Probability density functions of probabilistic scores p for positive and negative classes. Threshold is $t = \frac{c_0}{c_0+c_1} = 0.9$ because class errors are of different cost, $c_0 = 9c_1$.

and in Figure 3 costs for false positives are nine times bigger than costs of false negatives, $c_0 = 9c_1$.

Threshold t can be calculated two ways, and the first is to scale scores s into probabilistic scores p in range $[0;1]$, using perfect calibration map:

$$p = \frac{1}{1+e^{-2s}}$$

Such threshold will be denoted by t_p. For first case where costs are of the same value, threshold for probabilistic scores s will be $t_p = 0.5$ because data is calibrated, as shown in Figure 5.

To avoid confusion, for uncalibrated scores s, threshold will be denoted by t_s. Threshold t_s would be such value that satisfies equality $f_0(t_s) = f_1(t_s)$. This threshold is $t_s = 0$, as shown in plot in Figure 3.

Now let us take a look at case where costs are different, $c_0 = 9c_1$. For probabilistic scores a threshold can be calculated using the following formula:

$$t_p = \frac{c_0}{c_0 + c_1} = \frac{9c_1}{9c_1 + c_1} = 0.9$$

And this result is shown in Figure 6. And for original scores s with distributions f_0 and f_1 it is such t_s that satisfies the equation:

$$\frac{f_0(t_s)}{f_0(t_s) + f_1(t_s)} = \frac{c_0}{c_0 + c_1}$$

In this particular case with cost context $c_0 = 9c_1$ and densities $f_0 \sim N(-1,1)$ for positive class and $f_1 \sim N(1,1)$ for negative class, threshold is equal to $t_s = \frac{\ln 9}{2}$.

To justify the need change of a threshold with change of cost proportion, it will be proven with the experiments.

Let $c_0 = 9$ and $c_1 = 1$. When threshold was $t_s = 0$ (or $t_p = 0.5$ for probabilistic scores) cost is ~ 0.78265. But with threshold $t_s = 1.1$ (or $t_p = 0.9$ for probabilistic scores) cost is reduced to ~ 0.34942, and it is a minimal cost in this cost context. The distributions and this threshold are shown in Figure 4.

In the experiments in Section 5, uncalibrated scores s will also be used. However, they will first be calibrated to probabilistic scores p and then a decision rule would be used. Further on threshold would be denoted simply as t because it would be calculated only for probabilistic scores.

To conclude, $F_0(t) = \int_{-\infty}^{t} f_0(s)ds = P(s \leq t|Y = 0)$ and it is true positive rate at threshold t. As false negative rate is complement of true positive rate, than false negative rate is equal to $1 - F_0(t)$. Correspondingly, $F_1(t) = \int_{-\infty}^{t} f_1(s)ds = P(s \leq t|Y = 1)$, which is equal to false positive rate at threshold t [Flach, 2015].

Because 1 represents negative class and 0 represents positive, true positive and false positive rates ($F_0(t)$ and $F_1(t)$, respectively) are non-decreasing functions with increase of $p(x)$.

Operating Condition There is an alternative parametrization of costs that will also be used here further. *Cost magnitude* b is equal to sum of costs $b = c_0 + c_1$, and *cost proportion* c is proportion of c_0 to sum of costs, $c = c_0/b$.

Moreover, π_i is proportion of number of instances of class i to the number of all instances [Hernández-Orallo et al., 2012]. So, $\pi_0 = Pos/Total$ and $\pi_1 = 1 - \pi_0 = Neg/Total$, where $Neg = TN + FP$ is number of all actual negatives, and, similarly, $Pos = TP + FN$ is number of all actual positives. Then $Total = Pos + Neg$ is number of all instances together.

Then, *operating condition* can be defined as tuple $\langle b, c, \pi_0 \rangle$, and set of all possible operating conditions is defined as Θ.

As it was mentioned earlier, $(1 - F_0(t))$ stands for false negative rate, which is proportion of false negatives to all positives. To calculate cost of all false negatives, one needs to multiply cost of one false negative by proportion of false negatives to all instances. Cost of one false negative is c_0, and proportion of false negatives at threshold t to all instances is $\pi_0(1 - F_0(t))$, because:

$$\pi_0 = \frac{Pos}{Total}$$
$$1 - F_0(t) = \frac{FN}{Pos}$$
$$\pi_0 \dot{(} 1 - F_0(t)) = \frac{Pos}{Total} \cdot \frac{FN}{Pos} = \frac{FN}{Total}$$

Likewise, proportion of false positives to all instances is $\pi_1 F_1(t)$, and to find out cost of all false positives, one needs to multiply proportion by c_1.

Note that proportion of actual classes, π_0 and π_0 are known values, and FN, FP, TN, TP are expected quantities.

Hence, total cost given threshold t and operating condition θ is:

$$Q(t;\theta) = Q(t;\langle b, c, \pi_0 \rangle) = c_0 \pi_0 (1 - F_0(t)) + c_1 \pi_1 F_1(t) \tag{1}$$

Since $c = c_0/b$, one can represent c_0 and c_1 through cost magnitude and cost proportion. Then, $c_0 = bc$, and $c_1 = b - c_0 = b - bc = b(1 - c)$. Then, repeating b can be put outside of the brackets (1) [Santos-Rodríguez et al., 2009]. With new notation, total cost Q will be calculated as:

$$Q(t;\theta) = b\Big[c\pi_0(1-F_0(t)) + (1-c)\pi_1 F_1(t)\Big] \tag{2}$$

It is sometimes necessary to evaluate classifier performance over an interval of operating conditions, rather than at a point estimate. If precise operating conditions are not known, a distribution of operating conditions can be defined as multivariate probability density function (p.d.f.) $w(\theta)$ over each component of θ, and loss will be integrated over components of θ. Therefore, loss over all operating conditions in Θ is integral of Q over θ:

$$L = \int_{\Theta} Q(T(\theta);\theta)w(\theta)d\theta \tag{3}$$

Let us look at more precise example on the basis of Eq.(1) and (3). Usually, b is a constant and equal to 2, because then loss has the advantage of being equal to error rate, that assumes $c_0 = c_1 = 1$ [Hernández-Orallo et al., 2011]. Moreover, it will be assumed that data is balanced, $\pi_0 = \pi_1 = 0.5$. In that case, the only integrable part of the operating condition $\theta = \langle b, c, \pi_0 \rangle$ is c, and c is also a threshold. Space of all possible c is $[0;+\infty]$ because costs are always non-negative. Then, the integral would be equal to:

$$\begin{aligned} L &= \int_0^{\infty} b\Big[c\pi_0(1-F_0(c)) + (1-c)\pi_1 F_1(c)\Big]w(c)dc \\ &= \int_0^{\infty} 2\Big[c0.5(1-F_0(c)) + (1-c)0.5F_1(c)\Big]w(c)dc \\ &\quad \int_0^{\infty} \Big[c(1-F_0(c)) + (1-c)F_1(c)\Big]w(c)dc \end{aligned} \tag{4}$$

3 Proper Scoring Rules

Proper scoring rules are loss functions which give the lowest losses to the ideal model outputting the actual class posterior probabilities $P(Y = i|X = x)$. The most widely known proper scoring rules are log-loss (also known as ignorance score) and Brier score (also known as squared loss).

Estimated probability vector p, was introduced earlier in Section 2.2 as probabilistic scores vector; p is a vector of estimated probabilities of getting each class, $p = (p_1, ..., p_k)$. Probabilities of all classed add up to one, $\sum_{i=1}^{k} p_i = 1$.

Actual class will be represented by vector $y = (y_1, ..., y_k)$, where $y_j = 1$ if j is an actual class of the instance and $y_j = 0$ otherwise.

A *scoring rule* $\phi(p, y)$ is a loss function that determines the goodness of a match between p and y, with 0 being a perfect match [Winkler and Murphy, 1968].

Scoring rules are meant to reward the probabilistic classifier for making careful assessments and for being honest (not being overconfident, etc.). They are also meant to measure the quality of the probabilistic predictions (goodness of a match with actual class) [Garthwaite et al., 2005] [Gneiting and Raftery, 2007].

Logarithmic Loss (or log-loss or ignorance score), denoted by $\phi_{LL}(p, y)$, is one of the simplest proper scoring rules [Good, 1952].

$$\phi_{LL}(p, y) := -\log p_y$$

Here, p_y means such p_j for which $y_j = 1$ [Kull and Flach, 2015].

As it could be seen from plot of log-loss on Figure 7, this scoring rule highly penalizes overconfident wrong predictions. For example, yellow line stands for negative (1) class. As scores approach 1, loss decreases to 0: since the actual class is 1, it is a good thing that scores are close to 1. But if for class 1 scores are close to 0, it means that prediction will be incorrect and loss function penalized it. For log-loss, loss of confident false positive (when score $s = 0$ while class is 1) is $+\infty$. Likewise, loss of confident false negative is also $+\infty$. Therefore it is advised not to use this scoring rule for models that can output all-or-nothing predictions [Brownlee, 2018].

Brier Score (or squared loss or quadratic score), denoted by $\phi_{BS}(p, y)$, is the most well-known proper scoring rule [Brier, 1950].

$$\phi_{BS}(p, y) := \sum_{i=1}^{k} (p_i - y_i)^2$$

Brier score plot could be seen in Figure 8. Brier score is less harsh in penalizing wrongly confident instances, as its loss for any instance has maximum of 2.

Both log-loss and Brier score are proper, in a sense that is defined further.

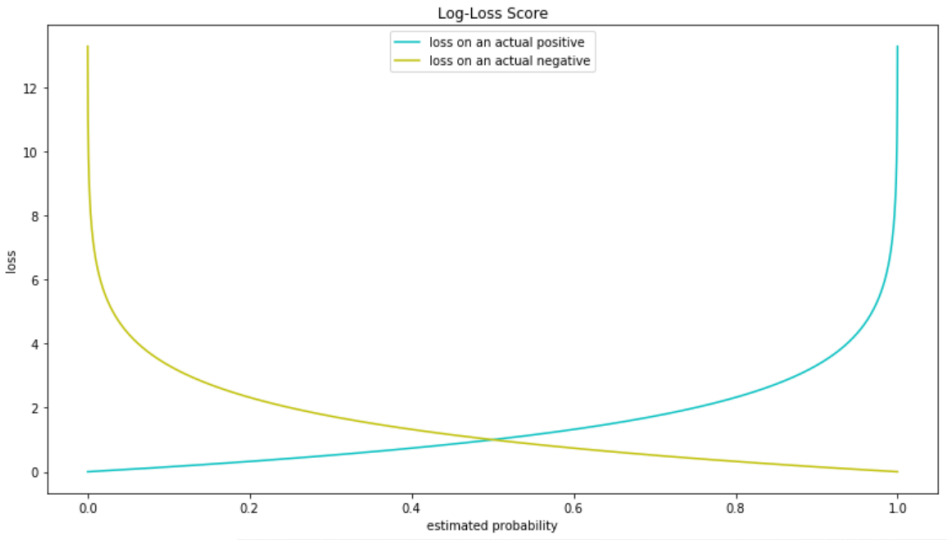

Figure 7. Plot of log-loss score for binary case. Blue line shows how log-loss changes over estimated probability when actual class is positive (0) and yellow line highlights log-loss over estimated probability when actual class is negative (1).

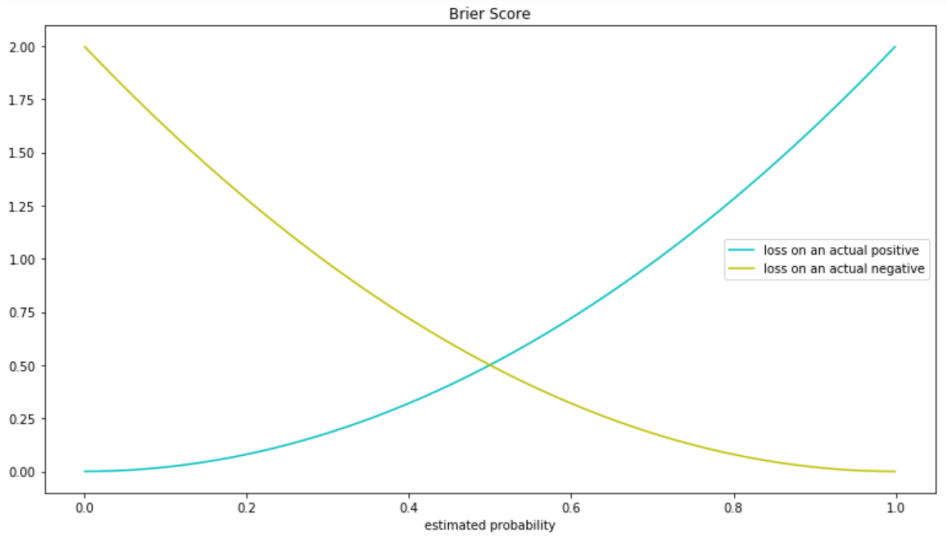

Figure 8. Plot of Brier score for binary case. Blue line shows how Brier changes over estimated probability when actual class is positive (0) and yellow line highlights Brier over estimated probability when actual class is negative (1).

3.1 Properness

Consider any instance x and suppose its probabilistic predictions for k classes is $\mathbf{p} = (p_1, ..., p_k)$. Actual probability to belong to classes $1..k$ is $\mathbf{q} = (q_1, ..., q_k)$, where $q_i = P(Y = i|X = x)$.

Assessment is a priori task, which means that assessment happens without knowledge of actual class. That is why actual scores (based on 'similarity' of prediction with actual prediction) are of secondary interest, and expected scores are widely used [Murphy and Winkler, 1970]. Then, *expected score*, also termed as expected loss, $s(\mathbf{p}, \mathbf{q})$ can be calculated as follows:

$$s(\mathbf{p}, \mathbf{q}) := \sum_{j=1}^{k} \phi(\mathbf{p}, \mathbf{e_j}) q_j$$

where $\mathbf{e_j}$ is vector of size k with 1 in j-th place and 0s everywhere else. Intuitively, expected score is weighted average of scoring rules with all the possible actual class vectors. For each j, $\mathbf{e_j}$ is a case when j is actual class, and q_j is probability of j being the actual class. Then, scoring rule ϕ of $\mathbf{e_j}$ and probability of $\mathbf{p}$ is calculated and is multiplied by q_j.

Scoring rule is called *proper* if for any $\mathbf{q}$, $s(\mathbf{p}, \mathbf{q}) \geq s(\mathbf{q}, \mathbf{q})$, which means that $\mathbf{q}$ minimizes expected score $s(\mathbf{p}, \mathbf{q})$ [Buja et al., 2005]:

$$\mathbf{q} : \arg\min_{\mathbf{q}} s(\mathbf{p}, \mathbf{q}) = \mathbf{q}$$

Scoring rule is called *strictly proper* when $s(\mathbf{p}, \mathbf{q}) = s(\mathbf{q}, \mathbf{q})$ if and only if $\mathbf{p} = \mathbf{q}$.

Expected score that takes as both arguments the same vector is called *entropy* e, $e(\mathbf{q}) = s(\mathbf{q}, \mathbf{q})$. For log-loss entropy is called *information entropy*, or *Shannon entropy*:

$$e_{LL}(\mathbf{q}) = -\sum_{i=1}^{k} q_i \log q_i$$

And for Brier score, entropy is called *Gini index*:

$$e_{BS}(\mathbf{q}) = \sum_{i=1}^{k} q_i (1 - q_i)$$

Divergence There is a parameter that directly corresponds to properness, which is divergence. *Divergence*, or *contrast function*, is a function that establishes 'distance' between one probability distribution ($\mathbf{p}$) and another ($\mathbf{q}$), and can be calculated from expected score:

$$d(\mathbf{p}, \mathbf{q}) := s(\mathbf{p}, \mathbf{q}) - s(\mathbf{q}, \mathbf{q})$$

Divergence function is less strict than distance function; unlike distance, it does not need to be symmetric nor does it need to satisfy triangle inequality.

Properness A scoring rule ϕ is called proper if its divergence is always non-negative, and strictly proper if $d(\mathbf{p},\mathbf{q}) = 0$ only in the case of $\mathbf{p} = \mathbf{q}$.

Lemma 1. *If vector of actual probabilities* $\mathbf{q}$ *is equal to actual class vector* $\mathbf{y}$*, divergence* d *is equal to the proper scoring rule* ϕ *itself:*

$$d(\mathbf{p},\mathbf{y}) = \phi(\mathbf{p},\mathbf{y})$$

Proof.

$$d(\mathbf{p},\mathbf{y}) = s(\mathbf{p},\mathbf{y}) - s(\mathbf{y},\mathbf{y}) \tag{5}$$

Let j be the actual class. Then, $y_j = 1$ and all other $y_i, i = 1..k, i \neq j$ are equal to 0:

$$s(\mathbf{p},\mathbf{y}) = \sum_{i=1}^{k} \phi(\mathbf{p},\mathbf{e_i}) y_i = \phi(\mathbf{p},\mathbf{e_j}) y_j = \phi(\mathbf{p},\mathbf{e_j}) \tag{6}$$

Similarly, for the second part:

$$s(\mathbf{y},\mathbf{y}) = \sum_{i=1}^{k} \phi(\mathbf{y},\mathbf{e_i}) y_i = \phi(\mathbf{y},\mathbf{e_j}) y_j = \phi(\mathbf{y},\mathbf{e_j}) \tag{7}$$

Vectors $\mathbf{e_j}$ and $\mathbf{y}$ are the same and have 1 in j-th place and 0s everywhere else; therefore, $\phi(\mathbf{y},\mathbf{e_j}) = 0$ because $\mathbf{y} = \mathbf{e_j}$ according to the definition of proper scoring rules.

Finally, after substitution of Eq. (6) and Eq. (7) in Eq. (5):

$$d(\mathbf{p},\mathbf{y}) = \phi(\mathbf{p},\mathbf{e_j}) - \phi(\mathbf{y},\mathbf{e_j}) = \phi(\mathbf{y},\mathbf{e_j}) - 0 = \phi(\mathbf{p},\mathbf{e_j}) = \phi(\mathbf{p},\mathbf{y}) \tag{8}$$

■

Brier Score and Properness Below are the formulas of entropy for Brier score (also called *Gini index*) and divergence for Brier score (also called *mean squared difference*):

$$e_{BS}(\mathbf{q}) = \sum_{i=1}^{k} q_i(1 - q_i)$$

$$d_{BS}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{k} (p_i - q_i)^2$$

Brier Score's divergence is a sum of squares; since squares are always non-negative, their sum is non-negative as well. Therefore, Brier Score is proper.

Log-Loss and Properness Below are the formulas of entropy for log-loss (also called *information entropy*) and divergence for log-loss (also called *Kullback–Leibler (KL) divergence*):

$$e_{LL}(\mathbf{q}) = -\sum_{i=1}^{k} q_i \log q_i$$
$$d_{LL}(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{k} q_i \log \frac{q_i}{p_i}$$

Lemma 2. *Log-loss is proper.*

Proof. To prove that log-loss is proper, we need to prove that its divergence $d_{LL}(\mathbf{p},\mathbf{q})$ is non-negative for all $\mathbf{p},\mathbf{q}$.

$$\log a \leq a - 1 \tag{9}$$

for any $a > 0$ (obvious from logarithmic plot). Then,

$$-d_{LL}(\mathbf{p},\mathbf{q}) = -\sum_{i=1}^{k} q_i \log \frac{q_i}{p_i} = \sum_{i=1}^{k} q_i \log \frac{p_i}{q_i} \tag{10}$$

Following logarithmic inequality in Eq. (9):

$$\sum_{i=1}^{k} q_i \log \frac{p_i}{q_i} \leq \sum_{i=1}^{k} q_i \left(\frac{p_i}{q_i} - 1\right) \tag{11}$$

$$\sum_{i=1}^{k} q_i \left(\frac{p_i}{q_i} - 1\right) = \sum_{i=1}^{k} (p_i - q_i) = \sum_{i=1}^{k} p_i - \sum_{i=1}^{k} q_i = 1 - 1 = 0 \tag{12}$$

It was proven $-d_{LL}(\mathbf{p},\mathbf{q}) \leq 0$, then $d_{LL}(\mathbf{p},\mathbf{q}) \geq 0$, i.e. log-loss's divergence is always non-negative.

Therefore, log-loss is proper. ■

3.2 Numerical Example

For example, consider a classification task with 4 classes, $k = 4$. For a fixed instance, let actual class be 2, then vector $\mathbf{y}$ has 1 in 2nd place and 0 everywhere else, $\mathbf{y} = (0,1,0,0)$.

Suppose a probabilistic classifier produces vector $\mathbf{p}$ that highlights probabilities of getting each class, let $\mathbf{p} = (0, 0.4, 0.3, 0.3)$. It can be noted that classifier works quite well; even though probability p_2 of getting actual class is not high, it is the biggest probability among other classes.

Brier score is equal to:

$$\phi_{BS}(\mathbf{p},\mathbf{y}) = \sum_{i=1}^{k}(p_i - y_i)^2 = (0-0)^2 + (0.4-1)^2 + (0.3-0)^2 + (0.3-0)^2 = 0.54$$

While $\mathbf{p}$ is an estimated probability of getting each class, there is an actual distribution over classes $\mathbf{q}$. Suppose $\mathbf{q} = (0.1, 0.1, 0.3, 0.5)$ and $\mathbf{e}$ corresponds to all possible vectors $\mathbf{y}$, $\mathbf{e} = ((1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1))$.

Expected score $s(\mathbf{p},\mathbf{q})$:

$$\begin{aligned} s(\mathbf{p},\mathbf{q}) &= \sum_{j=1}^{k}\phi(\mathbf{p},\mathbf{e_j})q_j = \\ &= \left[(0-1)^2 + (0.4-0)^2 + (0.3-0)^2 + (0.3-0)^2\right] * 0.1 \\ &+ \left[(0-0)^2 + (0.4-1)^2 + (0.3-0)^2 + (0.3-0)^2\right] * 0.1 \\ &+ \left[(0-0)^2 + (0.4-0)^2 + (0.3-1)^2 + (0.3-0)^2\right] * 0.3 \\ &+ \left[(0-0)^2 + (0.4-0)^2 + (0.3-0)^2 + (0.3-1)^2\right] * 0.5 \\ &= 0.78 \end{aligned}$$

,

Entropy $e(\mathbf{q})$:

$$\begin{aligned} e(\mathbf{q}) = s(\mathbf{q},\mathbf{q}) &= \sum_{j=1}^{k}\phi(\mathbf{q},\mathbf{e_j})q_j = \\ &= \left[(0.1-1)^2 + (0.1-0)^2 + (0.3-0)^2 + (0.5-0)^2\right] * 0.1 \\ &+ \left[(0.1-0)^2 + (0.1-1)^2 + (0.3-0)^2 + (0.5-0)^2\right] * 0.1 \\ &+ \left[(0.1-0)^2 + (0.1-0)^2 + (0.3-1)^2 + (0.5-0)^2\right] * 0.3 \\ &+ \left[(0.1-0)^2 + (0.1-0)^2 + (0.3-0)^2 + (0.5-1)^2\right] * 0.5 \\ &= 0.64 \end{aligned}$$

,

Divergence $d(\mathbf{p},\mathbf{q})$ calculated from expected score:

$$d(\mathbf{p},\mathbf{q}) = s(\mathbf{p},\mathbf{q}) - s(\mathbf{q},\mathbf{q}) = 0.78 - 0.64 = 0.14$$

Divergence for Brier score $d(\mathbf{p},\mathbf{q})$:

$$d_{BS}(\mathbf{p},\mathbf{q}) = \sum_{j=1}^{k}(p_i - q_i)^2 = (0-0.1)^2 + (0.4-0.1)^2 + (0.3-0.3)^2 + (0.3-0.5)^2 = 0.14$$

As expected, divergences are equal when calculated in different ways, and $d(\mathbf{p},\mathbf{q}) \geq 0$ for all cases and in this case $d(\mathbf{p},\mathbf{q}) > 0$ strictly since $\mathbf{p}$ and $\mathbf{q}$ are not equal.

3.3 Bregman Divergences

Convexity A set S is *convex* if for any $a, b \in S$ and any $\theta \in [0;1]$,

$$\theta a + (1-\theta)b \in S$$

Geometrically, it means that a set S is convex if the line segment between any two points in S lies in S.

A function f is *convex* if domain of f is a convex set and if for all a, b in domain of f and any $\theta \in [0;1]$,

$$f(\theta a + (1-\theta)b) \leq \theta f(a) + (1-\theta)f(b)$$

Geometrically, this inequality means that the line segment between $(a, f(a))$ and $(b, f(b))$ lies above the plot of f, as shown in Figure 9 [Boyd and Vandenberghe, 2004].

Bregman divergences Divergences of proper scoring rules form family of Bregman divergences, which have geometrical representation. If it might be hard to grasp the idea of scoring rules and their properness using divergence, Bregman divergences are much more intuitive because of their visualization.

Let $\phi : S \mapsto \mathbb{R}$ be a strictly convex function defined on a convex set $S \subseteq \mathbb{R}^k$ such that ϕ is differentiable on the relative interior of S, $ri(S)$. The *Bregman divergence*, or *Bregman loss functions (BLFs)*, $d_\phi : ri(S) \times S \mapsto [0, \infty)$ is defined as

$$d_\phi(a,b) = \phi(b) - \phi(a) - \left\langle b - a, \nabla\phi(a) \right\rangle \tag{13}$$

where $\nabla\phi(a)$ is a gradiant of ϕ in point a and angle brackets $\langle\rangle$ is notation for dot product [Bregman, 1967] [Banerjee et al., 2005]. In Figure 10, there is graphical representation of Bregman divergence.

Log Loss as Divergence It is called *Kullback-Leiber (KL) divergence* when used as a divergence measure, and *logarithmic loss (log loss)* when used as a loss measure [Kullback and Leibler, 1951]:

$$d_{LL}(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{d} b_i \log \frac{b_i}{a_i}$$

Graphical representation of KL-divergence is in Figure 11.

Brier Score as Divergence It is called *squared Euclidean distance* when used as a divergence measure, and *mean squared error* or *Brier score* when used as a loss measure.

$$d_{BS}(\mathbf{a},\mathbf{b}) = ||\mathbf{a}-\mathbf{b}||^2 = \sum_{i=1}^{d}(a_i - b_i)^2$$

Graphical representation of Euclidean distance is in Figure 12.

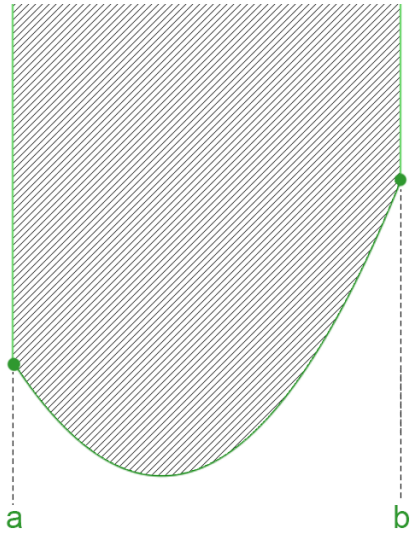

Figure 9. Convex function.

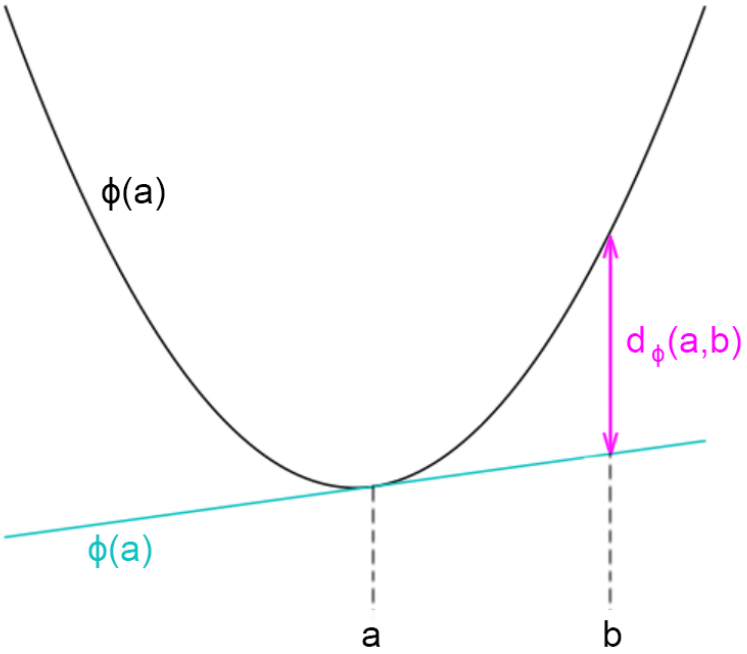

Figure 10. Visual representation of Bregman divergence. Divergence $d_\phi(a,b)$ is the difference between $\nabla\phi(a)$ in point b and $\phi(b)$.

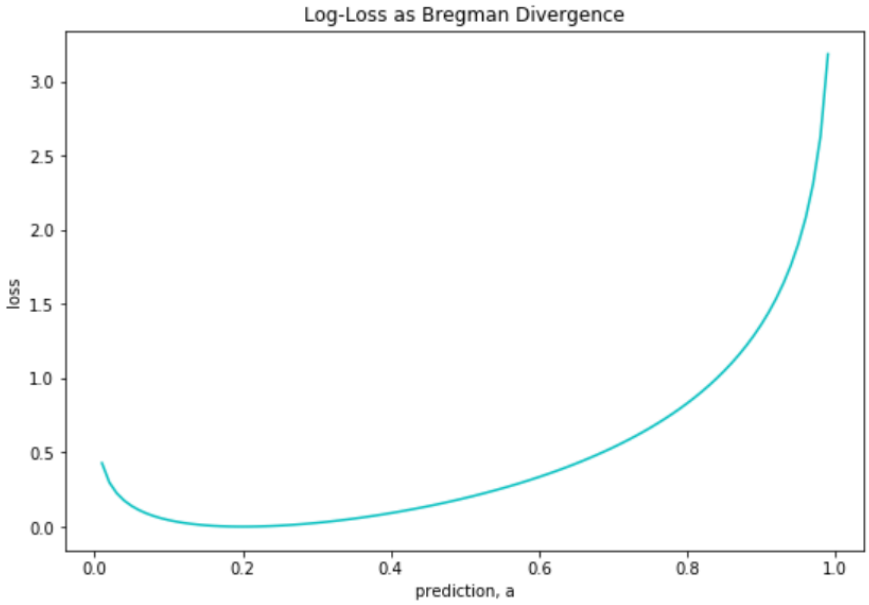

Figure 11. Log-loss as Bregman divergence, with $b = 0.2$.

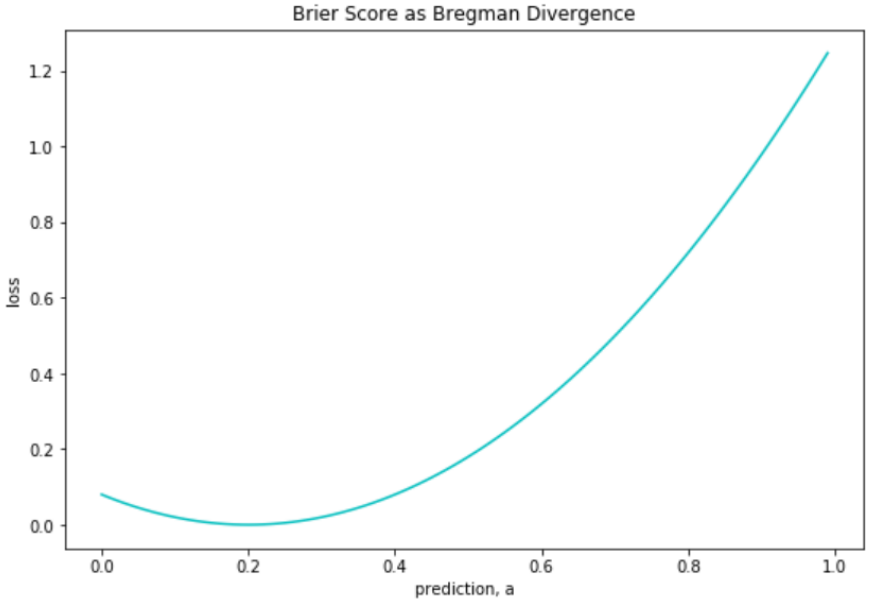

Figure 12. Brier score as Bregman divergence, with $b = 0.2$.

4 Derivation of Proper Scoring Rules

4.1 Brier Score

In this section, we will find and prove the relation between Brier score and additive cost context. This section mostly follows [Hernández-Orallo et al., 2012].

As it was discussed earlier, c_0 stands for cost of false negative and c_1 is cost of false positive. Further, cost magnitude b is equal to $b = c_0 + c_1$. Cost magnitude will be scaled to 2 to ensure commensurability with error rate, $b = 2$. Let us suppose that c_0 has uniform distribution in $[0;2]$, then $c_1 = 2 - c_0$. To make notation simpler, a new variable s will be introduced that has uniform distribution over $[0,1]$. Then, $c_0 = 2s$ and $c_1 = 2 - c_0 = 2 - 2s = 2(1-s)$. This cost context will be called *additive*.

According to the definitions in Section 2, expected proportion of false positives at threshold s is $\pi_0(1-F_0(s))$, and expected proportion of false negative at threshold s is $\pi_1 F_1(s)$.

To calculate cost of false negatives, one would multiply cost of one false negative $(2s)$ by number of false negatives $\pi_0(1-F_0(s))$. Similarly, cost of false positives is multiplication of $2(1-s)$ and $\pi_1 F_1(s)$. Clearly, the total cost at threshold s denoted by L is sum of false negative and false positive costs.

By default, integral has limits $[0;+\infty]$ and range of s is shown by p.d.f is shown by p.d.f. $w(s)$. Function $w(s)$ is equal to 1 only if $0 \le s \le 1$. Then total cost L at threshold s is:

$$L = \int_0^{\infty} \Big[2s\pi_0(1-F_0(s)) + 2(1-s)\pi_1 F_1(s) \Big] w(s) ds$$

Since $w(s)$ is equal to 1 only if $0 \le s \le 1$, integral's limits will be changed to $[0;1]$, $w(s)$ will cancel out because it is always 1 in new limits.

$$L = \int_0^{1} \Big[2s\pi_0(1-F_0(s)) + 2(1-s)\pi_1 F_1(s) \Big] ds \tag{14}$$

Theorem 1. *Let us assume additive cost context, $c_0 + c_1 = 2$, and probabilistic scores and a decision threshold of probabilistic scores equal to s. Then expected loss L under a uniform distribution of s is equal to expected Brier Score.*

Proof. Expected Brier score for actual class 1 will be denoted as BS_1. Assuming s is the predicted probability of getting class 1, loss of for class 1 will be equal to squared difference between probability s and actual class 1. Since value of s is not known, loss will be integrated over s in range $[0;1]$, as it is range for all possible values of s. Then, expected Brier score when actual class is 1:

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

Analogically, Brier score if actual class is 0 will be denoted as BS_0 and loss of class 0 will be equal to squared difference between s and 0 and it will be integrated over s in range $[0;1]$. Expected Brier score when actual class is 0:

$$BS_0 = \int_0^1 (0-s)^2 f_0(s) ds = \int_0^1 s^2 f_0(s) ds$$

Expected loss for both classes will be weighted average of Brier scores for each classes, weighted by proportion of those classes π_0 and π_1:

$$BS = \pi_0 \int_0^1 s^2 f_0(s) ds + \pi_1 \int_0^1 (1-s)^2 f_1(s) ds \tag{15}$$

Next step we will take expected loss for additive context from Eq.(14) and break it into 2 separate integrals:

$$\begin{aligned} L &= \int_0^1 \Big[2s\pi_0(1-F_0(s)) + 2(1-s)\pi_1 F_1(s) \Big] ds \\ &= \pi_0 \int_0^1 2s(1-F_0(s)) ds + \pi_1 \int_0^1 2(1-s)F_1(s) ds \end{aligned} \tag{16}$$

Clearly, first integral is expected loss when actual class is 0 and second integral is expected loss when actual class is 1.

$$\begin{aligned} L &= \pi_0 L_0 + \pi_1 L_1 \\ L_0 &= \int_0^1 2s(1-F_0(s)) ds \\ L_1 &= \int_0^1 2(1-s)F_1(s) ds \end{aligned} \tag{17}$$

Now let us integrate the expected losses using integration by parts, where $u = (1-F_0(s))$ and $dv = 2sds$, then $du = -f_0(s)ds$ and $v = s^2$:

$$L_0 = \int_0^1 2s\Big(1-F_0(s)\Big) ds = s^2\Big(1-F_0(s)\Big)\Big|_{s=0}^1 + \int_0^1 s^2 f_0(s) ds \tag{18}$$

Since $F_0(0) = 0$ and $F_0(1) = 1$:

$$s^2\Big(1-F_0(s)\Big)\Big|_{s=0}^1 = 1^2\Big(1-F_0(1)\Big) - 0^2\Big(1-F_0(0)\Big) = 1(1-1) - 0 = 0$$

Then expected loss for class 0:

$$L_0 = s^2\Big(1-F_0(s)\Big)\Big|_{s=0}^1 + \int_0^1 s^2 f_0(s) ds = \int_0^1 s^2 f_0(s) ds \tag{19}$$

Expected loss for class 1, using integration by parts, where $u = F_1(s)$ and $dv = 2(1-s)ds$, then $du = f_1(s)ds$ and $v = -(1-s)^2$:

$$L_1 = \int_0^1 2(1-s)F_1(s)ds = -(1-s)^2F_1(s)\Big|_{s=0}^1 + \int_0^1 (1-s)^2 f_1(s)ds \tag{20}$$

Since $F_1(0) = 0$:

$$(1-s)^2F_1(s)\Big|_{s=0}^1 = (1-1)^2F_1(1) - (1-0)^2F_1(0) = 0^2 \cdot 1 - 1^2 \cdot 0 = 0$$

Then expected loss for class 1:

$$L_1 = (1-s)^2F_1(s)\Big|_{s=0}^1 + \int_0^1 (1-s)^2 f_1(s)ds = \int_0^1 (1-s)^2 f_1(s)ds \tag{21}$$

Expected loss for both classes, following Eq. (17):

$$L = \pi_0 \int_0^1 s^2 f_0(s)ds + \pi_1 \int_0^1 (1-s)^2 f_1(s)ds$$

Which is equal to Brier score in Eq. (15). Thus, it is proven that Brier score is equal to expected loss under additive cost context, probabilistic scores and a decision threshold equal to s.

■

Brier score is proper for any number of classes; below is the proof of Brier score's properness for binary classification.

Theorem 2. *Assuming binary classification, Brier score is proper.*

Proof. In binary classification, y has two possible values, $Y = \{0,1\}$. If $y = 1$, expected Brier score is:

$$BS_1 = (y-s)^2 = (1-s)^2$$

And if $y = 0$, expected Brier score is:

$$BS_0 = (y-s)^2 = (0-s)^2 = s^2$$

Brier score is proper if for any q it is minimized at $s = q$. One can take derivative of Brier Score show that derivative is equal to 0 if and only if $s = q$. Then:

$$\begin{gathered} q\left((1-s)^2\right)' + (1-q)\left((s^2)\right)' = 0 \\ -2q(1-s) + (1-q)2s = 0 \\ -q + qs + s - qs = 0 \\ q = s \end{gathered} \tag{22}$$

When score is minimal, $q = s$. Thus, Brier score is proper. ■

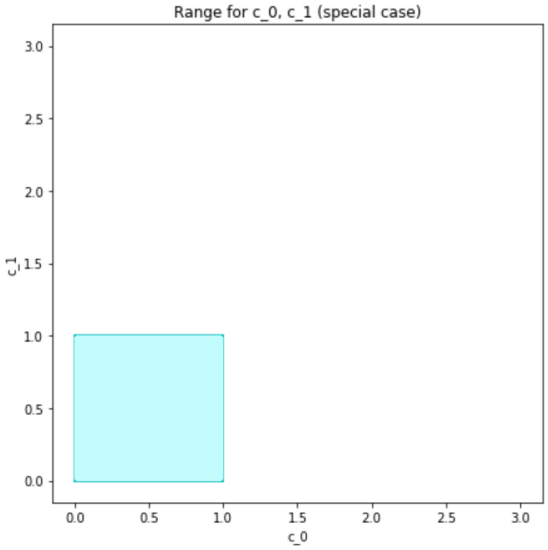

Figure 13. Restrictions to c_0 and c_1 for Inverse score.

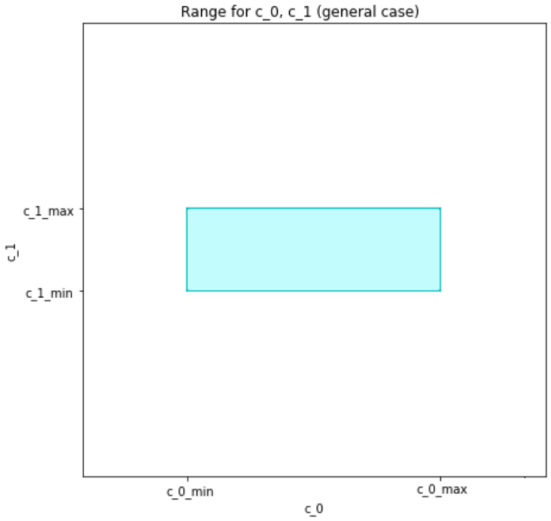

Figure 14. Restrictions to c_0 and c_1 for possible generalization of Inverse score.

4.2 Inverse Score

In the previous section, costs c_0, c_1 both had uniform distribution but were perfectly anti-correlated. Let us now introduce a new cost context, where both c_0, c_1 will have independent uniform distributions. This cost context will be called *independent uniform*. A new score will be introduced, based on independent uniform cost context.

At first, a special case will be introduced, where distributions of c_0, c_1 will be over $[0;1]$. These distributions are displayed in Figure 13.

Assuming calibrated threshold, $t = \frac{c_0}{c_0+c_1}$. At threshold t proportions of false positives and false negatives are denoted by $1-F_0(t)$ and $F_1(t)$, respectively.

Then expected loss is equal to:

$$L = \int_0^\infty \int_0^\infty \left[c_0\pi_0\Big(1-F_0\Big(\frac{c_0}{c_0+c_1}\Big) + c_1\pi_1 F_1\Big(\frac{c_0}{c_0+c_1}\Big)\right] w(c_0,c_1)dc_0dc_1 \tag{23}$$

where p.d.f. $w(c_0,c_1)$ in special case is equal to 1 when c_0 and c_1 fall into their defined distributions, that is $0 \le c_0 \le 1$ and $0 \le c_1 \le 1$, and $w(c_0,c_1)$ is equal to 0 otherwise.

Expected Inverse Score can be calculated as:

$$\begin{aligned} IS_0 &= \begin{cases} \frac{2p-1}{6(p-1)^2} + \frac{1}{6} & \text{if } p \in [0;\frac{1}{2}] \\ -\frac{1}{3p} + \frac{5}{6} & \text{if } p \in (\frac{1}{2};1] \end{cases} \\ IS_1 &= \begin{cases} \frac{1}{3(p-1)} + \frac{5}{6} & \text{if } p \in [0;\frac{1}{2}] \\ \frac{1-2p}{6p^2} + \frac{1}{6} & \text{if } p \in (\frac{1}{2};1] \end{cases} \end{aligned} \tag{24}$$

where p is estimated probability of predicting class 1.

Theorem 3. *Let us assume independent uniform cost context, c_0, c_1 have uniform distributions in $[0;1]$, and probabilistic scores and a decision threshold equal to $s = \frac{c_0}{c_0+c_1}$. Then expected loss L under a uniform distribution of c_0 and c_1 is equal to expected Inverse score.*

Proof. Inverse score:

$$\begin{aligned} IS = \frac{1}{3}\Bigg[&\int_0^{\frac{1}{2}} \Big(\frac{2s-1}{2(s-1)^2} + \frac{1}{2}\Big) f_0(s)ds + \int_{\frac{1}{2}}^1 \Big(-\frac{1}{s} + \frac{5}{2}\Big) f_0(s)ds \\ &+ \int_0^{\frac{1}{2}} \Big(\frac{1}{s-1} + \frac{5}{2}\Big) f_1(s)ds + \int_{\frac{1}{2}}^1 \Big(\frac{1-2s}{2s^2} + \frac{1}{2}\Big) f_1(s)ds \Bigg] \end{aligned}$$

In expected loss from Eq. (23), substitution $s = \frac{c_0}{c_0+c_1}$ can be made. Then, $c_0 = \frac{sc_1}{1-s}$ and $dc_0 = \Big(\frac{sc_1}{(1-s)}\Big)' ds = \frac{c_1}{(1-s)^2} ds$. Following that integral limits for c_0 were $c_0 \in$

$[0;+\infty]$, s is in range $\left[\frac{0}{0+c_1};\frac{+\infty}{+\infty+c_1}\right]$, which is equal to $s\in[0;1]$. New integration is now over s in range $[0;1]$ and over c_1 in range $[0;+\infty]$:

$$L=\int_0^{\infty}\int_0^1\left[\frac{sc_1}{1-s}\Big(1-F_0(s)\Big)+c_1F_1(s)\right]w\Big(\frac{sc_1}{1-s},c_1\Big)\frac{c_1}{(1-s)^2}dsdc_1 \tag{25}$$

Then, $w\big(\frac{sc_1}{1-s},c_1\big)$ will be equal to 1 in two cases:

1. $s\in\big(0;\frac{1}{2}\big)$, $c_1\in\big(0;1\big)$, because $\frac{sc_1}{1-s}<1\Rightarrow sc_1<1-s\Rightarrow s<\frac{1}{c_1+1}$
2. $s\in\big(\frac{1}{2};1\big)$, $c_1\in\big(0;\frac{1-s}{s}\big)$, because $\frac{sc_1}{1-s}<1\Rightarrow\frac{s}{1-s}c_1<1\Rightarrow c_1<\frac{1-s}{s}$

With new restrictions to s and c_1, Eq. (25) could be represented as the sum of four following integrals:

$$\begin{aligned}L=&\int_0^{\frac{1}{2}}\int_0^1\frac{sc_1}{1-s}\frac{c_1}{(1-s)^2}\Big(1-F_0(s)\Big)dc_1ds\\&+\int_{\frac{1}{2}}^1\int_0^{(1-s)/s}\frac{sc_1}{1-s}\frac{c_1}{(1-s)^2}\Big(1-F_0(s)\Big)dc_1ds\\&+\int_0^{\frac{1}{2}}\int_0^1\frac{c_1^2}{(1-s)^2}F_1(s)dc_1ds\\&+\int_{\frac{1}{2}}^1\int_0^{(1-s)/s}\frac{c_1^2}{(1-s)^2}F_1(s)dc_1ds\end{aligned} \tag{26}$$

Let us solve these 4 integrals separately.

Integration of 1st integral Integration over c_1:

$$\int_0^1\frac{sc_1}{1-s}\frac{c_1}{(1-s)^2}\Big(1-F_0(s)\Big)dc_1=\Big(1-F_0(s)\Big)\frac{s}{(1-s)^3}\int_0^1c_1^2dc_1=\Big(1-F_0(s)\Big)\frac{s}{(1-s)^3}\frac{1}{3}$$

Integration by parts over s, where $u=1-F_0(s)$ and $dv=\frac{s}{(1-s)^3}ds$, then $du=-f_0(s)ds$ and $v=\frac{2s-1}{2(s-1)^2}$:

$$\frac{1}{3}\int_0^{\frac{1}{2}}\frac{s}{(1-s)^3}\Big(1-F_0(s)\Big)ds=\frac{1}{3}\left[\frac{1}{2}-\frac{1}{2}F_0(0)+\int_0^{\frac{1}{2}}\frac{2s-1}{2(s-1)^2}f_0(s)ds\right] \tag{27}$$

Integration of 2nd integral Integration over c_1:

$$\int_0^{(1-s)/s}\frac{sc_1}{1-s}\frac{c_1}{(1-s)^2}\Big(1-F_0(s)\Big)dc_1=\Big(1-F_0(s)\Big)\frac{s}{(1-s)^3}\int_0^{(1-s)/s}c_1^2dc_1=\Big(1-F_0(s)\Big)\frac{1}{3s^2}$$

Integration by parts over s, where $u = 1 - F_0(s)$ and $dv = \frac{1}{s^2}ds$, then $du = -f_0(s)ds$ and $v = -\frac{1}{s}$:

$$\frac{1}{3}\int_{\frac{1}{2}}^{1}\frac{1}{s^2}\Big(1-F_0(s)\Big)ds = \frac{1}{3}\left[2-2F_0(\frac{1}{2})-1+F_0(1)\int_{\frac{1}{2}}^{1}\Big(-\frac{1}{s}\Big)f_0(s)ds\right] \tag{28}$$

Integration of 3rd integral Integration over c_1:

$$\int_0^1 \frac{c_1}{(1-s)^2}c_1F_1(s)dc_1 = F_1(s)\frac{1}{(1-s)^2}\int_0^1 c_1^2dc_1 = F_1(s)\frac{1}{(1-s)^2}\frac{1}{3}$$

Integration by parts over s, where $u = F_1(s)$ and $dv = \frac{1}{(1-s)^2}ds$, then $du = f_1(s)ds$ and $v = \frac{1}{1-s}$:

$$\frac{1}{3}\int_0^{\frac{1}{2}}\frac{1}{(1-s)^2}F_1(s)ds = \frac{1}{3}\left[2F_1\Big(\frac{1}{2}\Big)-F_1(0)+\int_0^{\frac{1}{2}}\frac{1}{s-1}f_1(s)ds\right] \tag{29}$$

Integration of 4th integral Integration over c_1:

$$\int_0^{(1-s)/s}\frac{c_1}{(1-s)^2}c_1F_1(s)dc_1 = F_1(s)\frac{1}{(1-s)^2}\int_0^{(1-s)/s}c_1^2dc_1 = F_1(s)\frac{1-s}{3s^3}$$

Integration by parts over s, where $u = F_1(s)$ and $dv = \frac{1-s}{s^3}ds$, then $du = f_1(s)ds$ and $v = \frac{2s-1}{2s^2}$:

$$\frac{1}{3}\int_{\frac{1}{2}}^{1}\frac{1-s}{s^3}F_1(s)ds = \frac{1}{3}\left[\frac{1}{2}F_1(1)+\int_{\frac{1}{2}}^{1}\frac{1-2s}{2s^2}f_1(s)ds\right] \tag{30}$$

Sum of results of integrals in Eq. (27 – 30) is equal to:

$$\begin{aligned} L = \frac{1}{3}\Bigg[&\frac{1}{2}-\frac{1}{2}F_0\Big(0\Big)-1+F_0\Big(1\Big)+2-2F_0\Big(\frac{1}{2}\Big)\\ &+2F_1\Big(\frac{1}{2}\Big)-F_1\Big(0\Big)+\frac{1}{2}F_1\Big(1\Big)\\ &+\int_0^{\frac{1}{2}}f_0(s)\Big(\frac{2s-1}{2(s-1)^2}\Big)ds+\int_{\frac{1}{2}}^{1}f_0(s)\Big(-\frac{1}{s}\Big)ds\\ &+\int_0^{\frac{1}{2}}f_1(s)\Big(\frac{1}{s-1}\Big)ds+\int_{\frac{1}{2}}^{1}f_1(s)\Big(\frac{1-2s}{2s^2}\Big)ds\Bigg] \end{aligned} \tag{31}$$

Some transformations need to be done, for $i = 0, 1$:

$$F_i\Big(0\Big) = 0, \quad F_i\Big(1\Big) = 1$$

Expected loss after transformations:

$$L = \frac{1}{3}\Big[\frac{1}{2} - 0 - 1 + 1 + 2 - 2F_0\Big(\frac{1}{2}\Big) + 2F_1\Big(\frac{1}{2}\Big) - 0 + \frac{1}{2}$$
$$+ \int_0^{\frac{1}{2}} f_0(s)\Big(\frac{2s-1}{2(s-1)^2}\Big)ds + \int_{\frac{1}{2}}^1 f_0(s)\Big(-\frac{1}{s}\Big)ds + \int_0^{\frac{1}{2}} f_1(s)\Big(\frac{1}{s-1}\Big)ds + \int_{\frac{1}{2}}^1 f_1(s)\Big(\frac{1-2s}{2s^2}\Big)ds\Big]$$
$$= \frac{1}{3}\Big[3 - 2F_0\Big(\frac{1}{2}\Big) + 2F_1\Big(\frac{1}{2}\Big) -$$
$$+ \int_0^{\frac{1}{2}} f_0(s)\Big(\frac{2s-1}{2(s-1)^2}\Big)ds + \int_{\frac{1}{2}}^1 f_0(s)\Big(-\frac{1}{s}\Big)ds + \int_0^{\frac{1}{2}} f_1(s)\Big(\frac{1}{s-1}\Big)ds + \int_{\frac{1}{2}}^1 f_1(s)\Big(\frac{1-2s}{2s^2}\Big)ds\Big]$$

Transformations for $2F_0\Big(\frac{1}{2}\Big)$ and $2F_1\Big(\frac{1}{2}\Big)$:

$$2F_1\Big(\frac{1}{2}\Big) = 2\int_0^{\frac{1}{2}} f_1(s)ds = \int_0^{\frac{1}{2}} f_1(s)ds + 1 - \int_{\frac{1}{2}}^1 f_1(s)ds$$

$$-2F_0\Big(\frac{1}{2}\Big) = -2\int_0^{\frac{1}{2}} f_0(s)ds = -\int_0^{\frac{1}{2}} f_0(s)ds - 1 + \int_{\frac{1}{2}}^1 f_0(s)ds$$

Which means that -1 will be added to coefficients of $\int_0^{\frac{1}{2}} f_0(s)ds$ and $\int_{\frac{1}{2}}^1 f_1(s)ds$, and 1 will be added to the other two coefficients. Expected loss after transformations:

$$L = \frac{1}{3}\Big[3 + \int_0^{\frac{1}{2}} f_0(s)\Big(\frac{2s-1}{2(s-1)^2} - 1\Big)ds + \int_{\frac{1}{2}}^1 f_0(s)\Big(-\frac{1}{s} + 1\Big)ds$$
$$+ \int_0^{\frac{1}{2}} f_1(s)\Big(\frac{1}{s-1} + 1\Big)ds + \int_{\frac{1}{2}}^1 f_1(s)\Big(\frac{1-2s}{2s^2} - 1\Big)ds\Big]$$

Transformations for 3:

$$3 = \frac{3}{2} + \frac{3}{2} = \frac{3}{2}\Big[\int_0^{\frac{1}{2}} f_0(s)ds + \int_{\frac{1}{2}}^1 f_0(s)ds\Big] + \frac{3}{2}\Big[\int_0^{\frac{1}{2}} f_1(s)ds + \int_{\frac{1}{2}}^1 f_1(s)ds\Big]$$
$$= \frac{3}{2}\Big[\int_0^{\frac{1}{2}} f_0(s)ds + \int_{\frac{1}{2}}^1 f_0(s)ds + \int_0^{\frac{1}{2}} f_1(s)ds + \int_{\frac{1}{2}}^1 f_1(s)ds\Big]$$

Which means that to all coefficients under integral $\frac{3}{2}$ will be added. Final form after transformations:

$$L = \frac{1}{3}\Big[\int_0^{\frac{1}{2}} f_0(s)\Big(\frac{2s-1}{2(s-1)^2}+\frac{1}{2}\Big)ds + \int_{\frac{1}{2}}^1 f_0(s)\Big(-\frac{1}{s}+\frac{5}{2}\Big)ds + \int_0^{\frac{1}{2}} f_1(s)\Big(\frac{1}{s-1}+\frac{5}{2}\Big)ds + \int_{\frac{1}{2}}^1 f_1(s)\Big(\frac{1-2s}{2s^2}+\frac{1}{2}\Big)ds\Big] \tag{32}$$

■

If we take coefficients of $f_0(s)$ from Eq. (32), we get loss function L_0 for actual positive label (0) that depends on probabilistic scores, and coefficients of $f_1(s)$ makes loss function L_1 for actual negative label (1) in the same way. Eq. (33) highlights final loss, functions in Figures 15 and 16 display their plots.

$$\begin{aligned} L_0 &= \begin{cases} \frac{2s-1}{6(s-1)^2}+\frac{1}{6} & \text{if } s \in [0;\frac{1}{2}] \\ -\frac{1}{3s}+\frac{5}{6} & \text{if } s \in (\frac{1}{2};1] \end{cases} \\ L_1 &= \begin{cases} \frac{1}{3(s-1)}+\frac{5}{6} & \text{if } s \in [0;\frac{1}{2}] \\ \frac{1-2s}{6s^2}+\frac{1}{6} & \text{if } s \in (\frac{1}{2};1] \end{cases} \end{aligned} \tag{33}$$

Theorem 4. *Inverse Score is proper.*

Proof. Similarly to Theorem (2) about Brier score's properness, Inverse Score is proper if for any q, it is minimized at $s = q$. One can take derivative of Inverse Score show that derivative is equal to 0 if and only if $s = q$. Then:

If $s \in [0;\frac{1}{2}]$:

$$\begin{aligned} q\Big(\frac{1}{s-1}\Big)' + (1-q)\Big(\frac{2s-1}{2(s-1)^2}\Big)' &= 0 \\ q\Big(-\frac{1}{(1-s)^2}\Big) + (1-q)\Big(\frac{s}{(1-s)^3}\Big) &= 0 \\ -q(1-s)+(1-q)s &= 0 \\ s &= q \end{aligned} \tag{34}$$

If $s \in (\frac{1}{2};1]$:

$$\begin{aligned} q\Big(\frac{1-2s}{2s^2}\Big)' + (1-q)\Big(-\frac{1}{s}\Big)' &= 0 \\ q\Big(\frac{s-1}{s^3}\Big) + (1-q)\Big(\frac{1}{s^2}\Big) &= 0 \\ q(s-1)+(1-q)s &= 0 \\ s &= q \end{aligned} \tag{35}$$

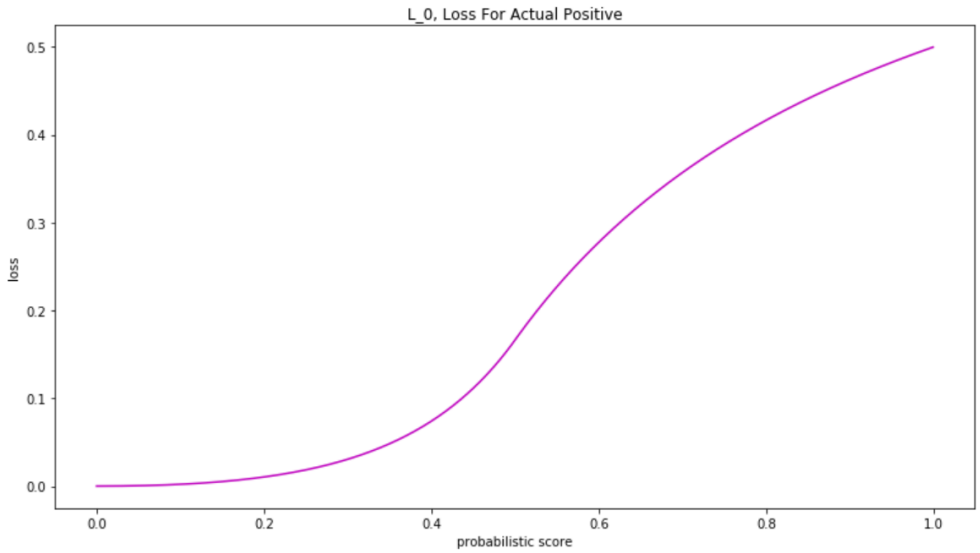

Figure 15. Plot of L_0, loss when actual class is positive, with respect to probabilistic score

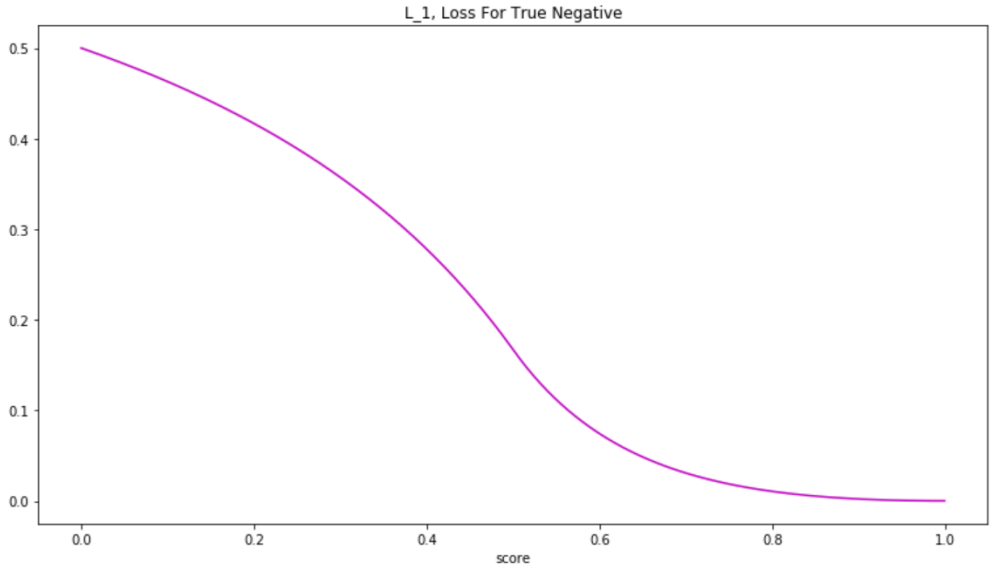

Figure 16. Plot of L_1, loss when actual class is negative, with respect to probabilistic score

When score is minimal, $q = s$. Thus, Inverse Score is proper. ■

4.3 Possible Generalization of Inverse Score

In general case, c_0 has uniform distribution in $[c_{0min}, c_{0max}]$ and c_1 has uniform distribution in $(c_{1min}, c_{1max}]$. These distributions are displayed in Figure 14.

Let $a = c_{0min}$, $b = c_{0max}$, $d = c_{0min}$, $e = c_{1max}$ for the easiness to comprehend. Then, following the formula in Eq. (25) on page 33, $w\left(\frac{sc_1}{1-s}, c_1\right)$ will be equal to 1 in two cases:

1. $s \in \left(\frac{e}{a+e}; \frac{b}{b+e}\right)$, $c_1 \in \left(d; e\right)$
2. $s \in \left(\frac{b}{b+e}; \frac{b}{b+d}\right)$, $c_1 \in \left(d; b\frac{1-s}{s}\right)$

Then the integral from Eq. (25) can be broken down to the sum of 4 integrals:

$$\begin{aligned} L = &\int_{a/(a+e)}^{b/(b+e)} \int_{d}^{e} \frac{sc_1}{1-s} \frac{c_1}{(1-s)^2} \Big(1 - F_0(s)\Big) dc_1 ds \\ &+ \int_{b/(b+e)}^{b/(d+b)} \int_{d}^{b(1-s)/s} \frac{sc_1}{1-s} \frac{c_1}{(1-s)^2} \Big(1 - F_0(s)\Big) dc_1 ds \\ &+ \int_{a/(a+e)}^{b/(b+e)} \int_{d}^{e} \frac{c_1}{(1-s)^2} F_1(s) dc_1 ds \\ &+ \int_{b/(b+e)}^{b/(d+b)} \int_{d}^{b(1-s)/s} \frac{c_1}{(1-s)^2} F_1(s) dc_1 ds \end{aligned} \tag{36}$$

Solving this integral would result in a family of proper scoring rules. This part remains to be done in the future work.

5 Experiments

In Section 4 Theorem 3 was proven, which states that for probabilistic scores, independent cost contexts and a threshold equal to $s = \frac{c_0}{c_0+c_1}$, expected loss under a uniform distribution of c_0 and c_1 is equal to Inverse Score.

In this section, it will be shown on the experiments.

5.1 Setup

The experiments began by creating toy dataset with 5,000 instances. Binary classification was used and dataset was balanced, labels $\{0,1\}$were assigned to instances randomly with probability 0.5. Scores of instances with label 1 had distribution $N(-\mu,1)$ and scores of instances with label 0 had distribution $N(\mu,1)$. From scores, probabilistic scores were calculated using formula $p = \frac{1}{1+e^{-2s}}$.

Cost Calculations To calculate total cost, independent cost context was used, which was introduced in Section 4.2. It means that false negative cost c_0 and false positive cost c_1 both have uniform distributions and do not depend on the values of each other. Threshold s for classification is $s = \frac{c_0}{c_0+c_1}$. So, instance will be predicted negative if its probabilistic score $p > s$ and positive otherwise.

In this case, distributions will be the same and equal to $[0;1]$. 5,000 iterations were made; in each iteration, number of false positives FP and proportion of false negatives FN were calculated. $Total$ is the number of all instances.For each iteration, cost was equal to:

$$C = \frac{FP \cdot c_0 + FN \cdot c_1}{Total}$$

To estimate expected total cost, costs C from all iterations were averaged.

Loss Calculations To calculate loss, loss function in Eq. (33) from page 36 was used:

If instance is actually of positive class:

$$L_0 = \begin{cases} \frac{2p-1}{6(p-1)^2} + \frac{1}{6} & \text{if } p \in [0;\frac{1}{2}] \\ -\frac{1}{3p} + \frac{5}{6} & \text{if } p \in (\frac{1}{2};1] \end{cases}$$

If instance is actually of negative class:

$$L_1 = \begin{cases} \frac{1}{3(p-1)} + \frac{5}{6} & \text{if } p \in [0;\frac{1}{2}] \\ \frac{1-2p}{6p^2} + \frac{1}{6} & \text{if } p \in (\frac{1}{2};1] \end{cases}$$

Expected loss is average of losses for all 5,000 instances.

5.2 Results

In Figure 17, total cost and expected loss are compared using different μ. As expected, expected loss generalized total cost very well with little to no bias, where pink line describes expected loss calculated from the experiments, and green line represents total cost. Loss and cost both decrease with growth of μ, because instances of different classes are separated further and it is harder to predict them incorrectly. It is possible to reduce bias further by increasing the number of iterations when calculating total cost.

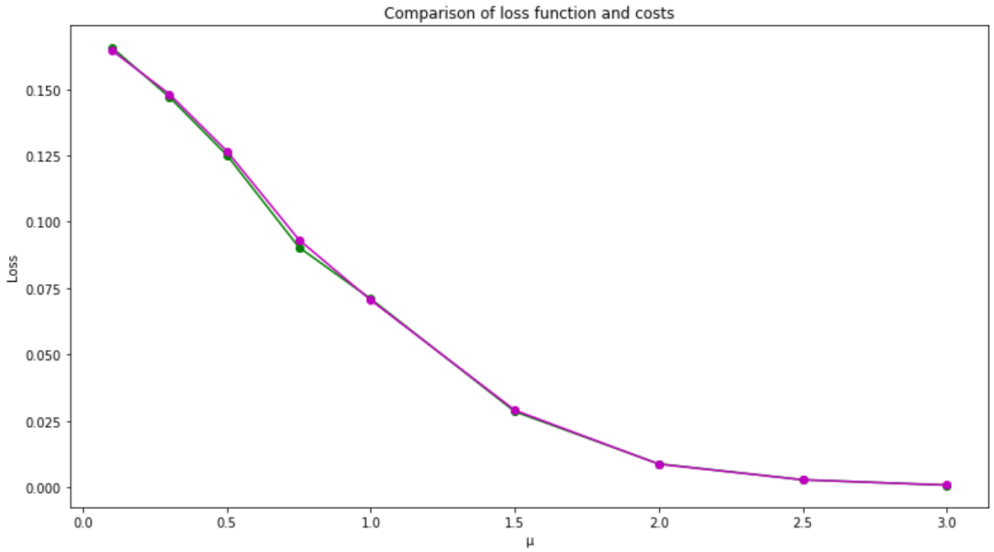

Figure 17. Comparison of loss and cost.

6 Conclusion

Classification is a fundamental task in supervised machine learning with an extremely wide range of applications. In case of cost-sensitive learning, evaluation of classifier performance becomes non-trivial due to operating conditions and cost contexts, which span ranges of cost values.

It has been previously shown that classifier evaluation under well-known operating conditions are handled with appropriate proper scoring rules.

In this thesis, a new operating condition for binary classification with uniformly distributed costs c_0 and c_1 was introduced, along with a corresponding new proper scoring rule. Its applicability and properness was proven theoretically and confirmed experimentally. The proof was performed for a special case of cost context with fixed distributions, where both costs had distributions $[0; 1]$.

Following the proof, new proper scoring rule can be used when cost context involves independent uniform distributions for both costs.

The future work may involve complete the proof for general case of distributions, which would result in a family of proper scoring rules.

References

[Banerjee et al., 2005] Banerjee, A., Guo, X., and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions of Information Theory*.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, 7 edition.

[Bregman, 1967] Bregman, L. M. (1967). A relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Journal of Computational Mathematics and Mathematical Physics*.

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

[Brownlee, 2018] Brownlee, J. (2018). A gentle introduction to probability scoring methods in python. `https://machinelearningmastery.com/how-to-score-probability-predictions-in-python/`.

[Buja et al., 2005] Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. `http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf`.

[Elkan, 2001] Elkan, C. (2001). The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*.

[Fawcett, 2005] Fawcett, T. (2005). An introduction to ROC analysis. *Pattern Recognition Letters*.

[Fawcett, 2015] Fawcett, T. (2015). The Basics of Classifier Evaluation: Part 1. `https://www.svds.com/the-basics-of-classifier-evaluation-part-1/`.

[Flach, 2015] Flach, P. A. (2015). Cost-sensitive classification meets proper scoring rules. `http://dmip.webs.upv.es/LMCE2015/Papers/LMCE_2015_submission_5.pdf`.

[Garthwaite et al., 2005] Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*.

[Gneiting and Raftery, 2007] Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*.

[Good, 1952] Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*.

[Hand, 2009] Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Springer Science+Business Media*.

[Hernández-Orallo et al., 2011] Hernández-Orallo, J., Flach, P. A., and Ferri, C. (2011). Brier Curves: A New Cost-Based Visualisation of Classifier Performance. *International Conference on Machine Learning*.

[Hernández-Orallo et al., 2012] Hernández-Orallo, J., Flach, P. A., and Ferri, C. (2012). A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. *Journal of Machine Learning Research*.

[Kull and Flach, 2015] Kull, M. and Flach, P. A. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*.

[Murphy and Winkler, 1970] Murphy, A. H. and Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*.

[Powers, 2011] Powers, D. M. W. (2011). Evaluation: from precision, recall and f-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*.

[Santos-Rodríguez et al., 2009] Santos-Rodríguez, R., Guerrero-Curieses, A., Alaiz-Rodríguez, R., and Cid-Sueir, J. (2009). Cost-sensitive learning based on bregman divergences. *Springer Science+Business Media*.

[Sun et al., 2011] Sun, Y., Wong, A., and Kamel, M. S. (2011). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*.

[Tharwat, 2018] Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*.

[Volovich, 2019] Volovich, K. (2019). What's a Good Clickthrough Rate? New Benchmark Data for Google AdWords. https://blog.hubspot.com/agency/google-adwords-benchmark-data.

[Winkler and Murphy, 1968] Winkler, R. L. and Murphy, A. H. (1968). 'Good' Probability Assessors. *Journal of Applied Meteorology and Climatology*.

[Wolpert, 1996] Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*.

[Yao, 2017] Yao, M. (2017). Chihuahua or muffin? My search for the best computer vision API. https://medium.freecodecamp.org/chihuahua-or-muffin-my-search-for-the-best-computer-vision-api-cbda4d6b425d.

Appendix

I. Code

Source code for this thesis is located in the following GitHub repository:
`https://github.com/sherlie/inverse-score`

II. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Diana Grygorian**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Classifier Evaluation With Proper Scoring Rules

   supervised by Meelis Kull

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Diana Grygorian
16.05.2019