Proceedings of Machine Learning Research 94:52–66, 2018 LIDTA 2018

Proper Losses for Learning with Example-Dependent Costs

Alexander Hepburn AH13558@BRISTOL.AC.UK
Ryan McConville RYAN.MCCONVILLE@BRISTOL.AC.UK
Raúl Santos-Rodríguez ENRSR@BRISTOL.AC.UK
University of Bristol
Bristol, UK
Jesús Cid-Sueiro JCID@TSC.UC3M.ES
Universidad Carlos III de Madrid
Madrid, Spain
Dario García-García DARIOGG@FB.COM
Facebook
New York, US

Editors: Luís Torgo, Stan Matwin, Nathalie Japkowicz, Bartosz Krawczyk, Nuno Moniz, and Paula Branco

Abstract

We study the design of cost-sensitive learning algorithms with example-dependent costs, when cost matrices for each example are given both during training and test. The approach is based on the empirical risk minimization framework, where we replace the standard loss function by a combination of surrogate losses belonging to the family of *proper losses*. The actual contribution of each example to the risk is then given by a loss that depends on the cost matrix for the specific example. We then evaluate the use of such example-dependent loss functions in real-world binary and multiclass problems, namely credit risk assessment and musical genre classification. Using different neural network architectures, we show that with the appropriate choice of the example-dependent losses, we can outperform conventional cost-sensitive methods in terms of total cost, making a more efficient use of cost information during training and test as compared to existing discriminative approaches.

Keywords: Cost-sensitive, Proper losses, Bregman divergences

1. Introduction

Cost-sensitive learning aims at making use of the cost information associated with choosing a category for a specific sample. The cost usually weights mistakes according their importance and is also used as a proxy when dealing with imbalanced data (the less representatives a class has, the higher the cost (Thai-Nghe et al., 2010)). Interestingly, the cost associated with selecting a label for a given example depends not only on the actual and estimated labels, but also on the example itself. Let us introduce the task of credit risk classification. The problem in its simplest form involves categorizing borrowers as either high-risk or low-risk customers, depending on how likely they are to return a requested loan. The cost of each decision is determined by a number of factors that vary for different customers, e.g., amount of the loan and its length in time or the existing relationship between the costumer and the company. Consider also the problem of musical genre classification. The task consists in classifying a given song as belonging to several predefined categories (Tzanetakis

and Essl, 2001). In this case, the cost of each decision is influenced by a number factors that vary from song to song, including the popularity, the context or the past experience of the listener.

However, the literature in machine learning is generally focused on cost-insensitive learning, assuming that minimizing the error probability suffices. Even when the problem is addressed as cost-sensitive, most existing approaches rely on the assumption that there exists a class-dependent cost matrix to penalize incorrect decisions and to reward correct ones (Elkan, 2001). In the credit risk example above, this is equivalent to, for instance, assume that the cost of deciding that a high-risk customer is reliable is higher than the cost of deciding that low-risk customer is not suitable for a loan. However, in both scenarios above this is neither natural nor optimal as this is just a special case of the more general example-dependent cost framework and, as such, considers only a part of the whole picture. For example, as a music content provider one might argue that decisions affecting songs that are accessed by a large audience might potentially be more damaging to the business that those with more limited visibility.

In this paper we frame the task as an Empirical Risk Minimization (ERM) problem (Devroye et al., 1996) and present a general approach for constructing loss functions that are tailored to naturally incorporate example-dependent costs. For this, we use the family of proper losses (Reid and Williamson, 2011) (also referred to as Bregman Divergences (Bregman, 1967) or proper scoring rules (Buja et al., 2005)). In our proposal, each example contributes to the loss function depending not only on the distance to the boundary but also through its cost. This way we are able to obtain loss functions which accurately approximate the posterior probability in the areas of interest from a classification perspective, that is to say, near the optimal classification boundary of the example-dependent cost problem. We show theoretically and empirically that minimizing these loss functions is asymptotically equivalent to minimizing the total cost.

2. Empirical risk minimization with example-dependent costs

In this section we present our approach for binary problems for the sake of simplicity but the extension to multiclass is straightforward. Let $\mathcal{X}$ be a sample space, $\mathcal{Y}$ be a label space and $\mathcal{C}$ the space of 2×2 cost matrices. Let (X, Y, C) be a triple of random variables taking values in $\mathcal{X} \times \{0, 1\} \times \mathcal{C}$, according to a joint probability distribution $P(X, Y, C)$. The element c_{iy} of matrix $\mathbf{C} \in \mathcal{C}$ denotes the cost of predicting class i when the actual class is y. The objective of a cost-sensitive classifier is then to predict the correct label y when both the $\mathbf{x}$ and $\mathbf{C}$ are observed, by minimizing the expected risk or cost

$$R = \mathbb{E}_{(\mathbf{x},y,\mathbf{C})\sim P}\{c_{iy}\}. \tag{1}$$

For every example $\mathbf{x}$, the Bayes optimal classifier selects class i^* such that

$$i^*(\mathbf{x}, \mathbf{C}) \in \arg\min_i \{(1 - \eta(\mathbf{x}))c_{i0} + \eta(\mathbf{x})c_{i1}\}, \tag{2}$$

where $\eta(\mathbf{x}) = P(Y = 1 \mid X = \mathbf{x})$ is the *posterior probability*. Assuming non-negative regrets $(c_{10} - c_{11})$ and $(c_{01} - c_{00})$, it is easy to verify that R is minimized by the assignment

$$i^*(\mathbf{x}, \mathbf{C}) = I_{\eta(\mathbf{x})\geq q(\mathbf{C})}, \tag{3}$$

where I denotes the indicator function and $q(\mathbf{C})$ is the *normalized regret*

$$q(\mathbf{C}) = \frac{c_{10} - c_{00}}{c_{10} - c_{11} + c_{01} - c_{00}}. \tag{4}$$

Eq. (4) illustrates that the classification depends just on the regrets and not on the absolute costs.

We assume that $P(X, Y, C)$ is unknown and a training set of i.i.d. pairs drawn from P and their corresponding cost matrices $\mathbf{C}^{(k)} = \begin{pmatrix} c_{00}^{(k)} & c_{01}^{(k)} \\ c_{10}^{(k)} & c_{11}^{(k)} \end{pmatrix}$ is given, $\mathcal{S} = \left\{(\mathbf{x}^{(k)}, y^{(k)}, \mathbf{C}^{(k)}), k = 1, \ldots, K\right\}$. A standard approach is based on estimating a posterior probability map $\hat{\eta}(\mathbf{x})$ using the training set $\mathcal{S}$ following the Empirical Risk Minimization (ERM) principle (Devroye et al., 1996),

$$R_{emp} = \frac{1}{K} \sum_{k=1}^{K} c_{i^{(k)}, y^{(k)}}^{(k)} \tag{5}$$

where $i^{(k)} = I_{\hat{\eta}(\mathbf{x}^{(k)}) \geq q(\mathbf{x}^{(k)})}$ is the classification for the k-th example following $\hat{\eta}$.

3. Proper losses and example-dependent costs

As the expression in Eq. (1) is neither convex nor differentiable, we suggest the use of other *surrogate* losses. Due to their well-studied properties, the family of *proper losses* (Reid and Williamson, 2011) is a natural choice. We begin by defining Bregman divergences as they have a one-to-one correspondence with the set of proper or Fisher consistent losses.

Definition 1 *(Bregman Divergence) Given a differentiable strictly convex function (Bregman generator) $\phi : \mathcal{A} \to \mathbb{R}$ defined in the convex set $\mathcal{A}$, and two points $p, z \in \mathcal{A}$, the Bregman divergence $D : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ relative to ϕ is defined as*

$$D_\phi(p, z) = \phi(p) - \phi(z) - \langle p - z, \nabla\phi(z)\rangle \tag{6}$$

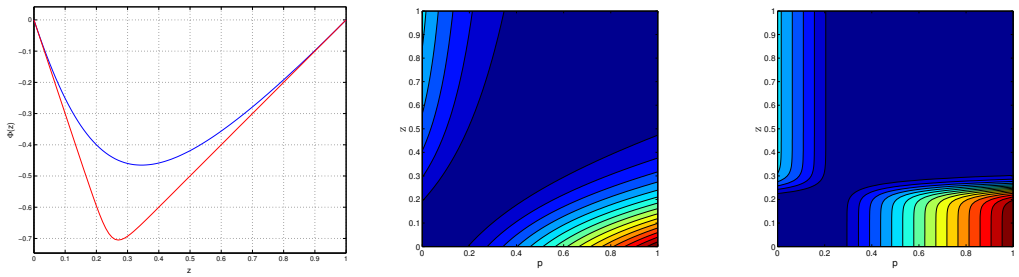
For a random variable $Y \in \{0, 1\}$ with $P\{Y = 1\} = \eta$, the loss $D_\phi(y, \hat{\eta})$ satisfies

$$\arg \min_{\hat{\eta} \in [0,1]} \{\mathbb{E}_{y \sim \eta}\{D_\phi(y, \hat{\eta})\}\} = \eta \tag{7}$$

According to this, given a class $\mathcal{M}$ of functions $\hat{\eta} : \mathcal{X} \to [0, 1]$, if the true posterior probability η is in $\mathcal{M}$, then, for any (strictly convex) ϕ,

$$\arg \min_{\hat{\eta} \in \mathcal{M}} \{\mathbb{E}_{(\mathbf{x}, y) \sim P}\{D_\phi(y, \hat{\eta})|\mathbf{x}\}\} = \eta \tag{8}$$

(both η and $\hat{\eta}$ are scalar variables in Eq. (7) and functions in Eq. (8)). Therefore, if the true posterior belongs to the function class, the choice of ϕ is not critical, and any estimator, e.g., the Maximum Likelihood estimate (which is equivalent to $\phi(p) = p\log(p) + (1-p\log(1-p))$),



(a) (b) (c)

Figure 1: Cost-sensitive proper loss ($c_{10} = 1, c_{01} = 3$). (a) Bregman generator (ϕ) for $n = 2$ (blue) and $n = 10$ (red), (b) Loss, $n = 2$, (c) Loss, $n = 10$.

is equally efficient (Dmochowski et al., 2010). However, if $\eta \notin \mathcal{M}$, Eq. (8) no longer holds, and different choices of ϕ lead to different results.

The second derivative of $\phi(\eta)$ relates to the sensitivity of the loss function D_ϕ to deviations of $\hat{\eta}$ from the true posterior, η (Santos-Rodriguez et al., 2009b). Then, in a cost-sensitive setting, since accurate posterior probability estimates are critical in the vicinity of the normalized regret, q, given by Eq. (4), generators with higher sensitivity at q may be more efficient for classification purposes.

For example, consider the Bregman generator given by

$$\phi_{n,\mathbf{C}}(z) = ((c_{10}z)^n + (c_{01}(1-z))^n)^{1/n} - c_{10}z - c_{01}(1-z) \tag{9}$$

where $n \in \mathbb{N}$ is a smoothness parameter. It naturally defines a cost-sensitive proper loss $D_{\phi_{n,\mathbf{C}}}(y, z)$. Figure 1 shows the generator and its associated loss for different values of n. Note that the highest curvature region varies with $\mathbf{C}$, achieving greater sensitivity in areas close to $q(\mathbf{C})$. As n grows larger, the sensitivity around the boundary increases, but the loss becomes less well-behaved from a numerical optimization point of view.

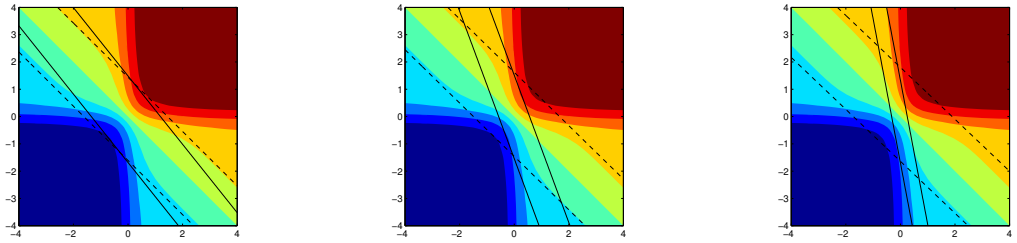
In our case, the threshold q is sample dependent and, therefore, there is no single choice of the Bregman generator that is appropriate for every example. Our approach is then to use a different generator for each example. Making use of the properties of proper losses, we note that

$$\mathbb{E}_{(\mathbf{x},y,C)\sim P}\{D_{\phi_C}(y, \hat{\eta})\} = \mathbb{E}_{\mathbf{C}\sim P}\{\mathbb{E}_{(\mathbf{x},y)\sim P}\{D_{\phi_\mathbf{C}}(y, \hat{\eta})|\mathbf{C}\}\} \tag{10}$$

Since the inner expectation is minimized by $\hat{\eta} = \eta$ for any $\mathbf{C}$, the whole expectation is also minimized by η. We then pose the task as minizing the empirical risk in Eq. (10),

$$\hat{R}(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K} D^{(k)}(y^{(k)}, \hat{\eta}^{(k)}) \tag{11}$$

where $\hat{\eta}^{(k)} = \hat{\eta}(\mathbf{x}^{(k)})$, $D^{(k)}$ is the proper loss given by generator $\phi_{\mathbf{C}^{(k)}}$, and $\mathbf{w}$ is a parameter vector specifying $\hat{\eta}$ in $\mathcal{M}$. Note that each example is associated with its very own loss. $\hat{R}$ represents the average of these losses evaluated for their corresponding examples.



(a) (b) (c)

Figure 2: Probability map as defined in Eq. (12). Dashed line: boundary minimizing the loss given by the average costs; solid line: boundary minimizing EDBD, (a) $n = 2$, (b) $n = 4$, (c) $n = 8$.

3.1. An example

This synthetic example tries to illustrate the difference between minimizing the empirical risk from Eq. (11) (from now on, example-dependent Bregman divergence or EDBD) and example-independent proper losses in a scenario where the capacity of the learning machine is limited. Consider the two-class problem with classes "0" and "1" and the probability map given by

$$\eta(\mathbf{x}) = \frac{1}{3}\left(\Phi(\mathbf{w_0}^T\mathbf{x}) + \Phi(\mathbf{w_1}^T\mathbf{x}) + \Phi(\mathbf{w_2}^T\mathbf{x})\right) \tag{12}$$

where $\mathbf{x} \in \mathbb{R}^2$, $\mathbf{w_0} = (4, 0)$, $\mathbf{w_1} = (0, 4)$ and $\mathbf{w_2} = (1, 1)$. The inverse link function $\Phi : \mathbb{R} \to [0, 1]$ is the logistic function given by $\Phi(z) = 1/\left(1 + \exp(z)\right)$ The contour-plot of this probabilistic map is represented in Fig. 2. Colder colours correspond to higher values of $1 - \eta(\mathbf{x})$, the posterior probability of class "0". We generated 8000 training samples uniformly distributed in the square $[-4, 4] \times [-4, 4]]$. The label of every sample was assigned stochastically according to the probability map. A single layer perceptron (SLP) with soft decisions given by

$$\hat{\eta}(\mathbf{x}) = \Phi(\mathbf{w}^T\mathbf{x}) \tag{13}$$

is used to estimate this map. Since the SLP has not enough capacity to do it exactly, different proper loss functions provide different approximations. We estimate $\mathbf{w}$ by minimizing the loss using a quasi-Newton method (Broyden-Fletcher-Goldfarb-Shanno (BFGS)).

Two cost policies are assigned to the samples depending on their position in the input space: for points satisfying that their ordinate is greater or equal than their abscissa, $\mathbf{C} = \begin{pmatrix} 0 & 3 \\ 7 & 0 \end{pmatrix}$; in any other case, $\mathbf{C} = \begin{pmatrix} 0 & 7 \\ 3 & 0 \end{pmatrix}$. Here, for the purpose of illustrating the advantages of using the proposed approach, D_{ϕ_n} is the loss given by the convex function in Eq. (9) (but it could have been replaced with any other cost-sensitive proper loss). The example-independent proper loss is also based on in Eq. (9) but makes use of the average costs to estimate the posterior probability.

The result of any comparison between proper losses depends on how we measure the quality of a probability estimate. Figures 2(a), 2(b) and 2(c) show the probability map and

the decision boundaries for both cost matrices and $n = \{2, 4, 8\}$ respectively. It becomes clear that, as n increases, the boundary obtained from EDBD varies its direction towards the Bayes solution, while the boundary corresponding to example-independent divergence remains practically unchanged. In this scenario, our method clearly improves locally the accuracy of the probability estimates. The importance of n is highlighted in the next section.

3.2. Convergence to the minimum total cost

In this section we show that the minimization of an empirical risk of the form of Eq. (11) with an adequately chosen Bregman generator leads, asymptotically, to the minimization of the overall cost regret (or the minimum number of errors in a cost-insensitive scenario). Note that the following results are asymptotical with respect to a parameter of the Bregman generator and *do not depend on the number of samples*.

Consider the cost-weighted classification loss given by

$$\hat{L}_c(y, \hat{\eta}) = c(1 - y)\hat{y} + (1 - c)y(1 - \hat{y}) \tag{14}$$

where $\hat{y} = I_{\hat{\eta} \geq c}$. For $c = 1/2$ this becomes half the zero-one loss. For any other constant, the loss is a cost indicator: for any cost-sensitive problem with normalized costs c and $1-c$, $\hat{L}_c$ computes the cost of decision rule $\hat{y} = I_{\hat{\eta} \geq c}$ (which is optimal if $\hat{\eta} = \eta$). Given a sample set $\mathcal{S} = \{(\mathbf{x}^{(k)}, y^{(k)}, c^{(k)}), k = 1, \ldots, K\}$, the corresponding risk is

$$\hat{R}(\hat{\eta}) = \sum_{k=1}^{K} \hat{L}_c^{(k)}(y^{(k)}, \hat{\eta}^{(k)})) = \sum_{k=1}^{K} (c^{(k)} - y^{(k)})(\hat{y}^{(k)} - y^{(k)}) \tag{15}$$

We present a general condition on a sequence of cost-sensitive proper losses to converge to the optimal risk.

Theorem 2 *Consider the sequence of proper losses* $\{L_n^{(k)} = D_{\phi_n^{(k)}}(y, \hat{\eta}), n = 0, 1, 2, ...\}$ *with corresponding weighting functions* $\{g_n^{(k)}(z) = \frac{\partial^2 \phi_n^{(k)}}{\partial z^2}, n = 0, 1, 2, \ldots\}$, *and let* $R_n(\hat{\eta})$ *be the corresponding sequence of empirical risks given by* $R_n(\hat{\eta}) = \sum_{k=1}^{K} L_n^{(k)}(y^{(k)}, \hat{\eta}^{(k)})$, *where* $y^{(k)} \in \{0, 1\}$ *and* $\hat{\eta}^{(k)} \in [0, 1]$. *If* $g_n^{(k)}$ *converges to a delta distribution shifted to* $c^{(k)}$, *then*

$$\lim_{n \to \infty} R_n(\hat{\eta}) = \hat{R}(\hat{\eta}) \tag{16}$$

Proof

The proof is straightforward. Since $L_n^{(k)}$ is a proper loss, we can express the empirical risk as $R_n(f_{\mathbf{w}}) = \sum_{k=1}^{K} \int_{y^{(k)}}^{\hat{\eta}^{(k)}} g_n^{(k)}(\alpha)(\alpha - y^{(k)}) d\alpha$ (see e.g. (Miller et al., 1991)). Taking the

limit,

$$\begin{aligned}\lim_{n\to\infty} R_n(\hat{\eta}) &= \sum_{k=1}^{K} \lim_{n\to\infty} \int_{y^{(k)}}^{\hat{\eta}^{(k)}} g_n^{(k)}(\alpha)(\alpha - y^{(k)})d\alpha \\ &= \sum_{k=1}^{K} \lim_{n\to\infty} \int_{0}^{1} g_n^{(k)}(\alpha)(\alpha - y^{(k)}) \cdot \left(I_{\hat{\eta}^{(k)}\geq\alpha} - I_{y^{(k)}\geq\alpha}\right) d\alpha \\ &= \sum_{k=1}^{K} (\hat{\eta}^{(k)} - y^{(k)}) \left(I_{\hat{\eta}^{(k)}\geq c^{(k)}} - I_{y^{(k)}\geq c^{(k)}}\right) = \sum_{k=1}^{K} (c^{(k)} - y^{(k)})(\hat{y}^{(k)} - y^{(k)}),\end{aligned} \tag{17}$$

where the third equality follows from the condition on the convergence to a delta distribution, which implies $\lim_{n\to\infty} \int_0^1 f(z) g_n^{(k)}(z) dz = f(c^{(k)})$ for any regular f. ■

In summary, if the individual generators $g_n^{(k)}$ behave asymptotically as delta distribution centered in $c^{(k)}$, as n goes to ∞, R_n converges to the minimum total cost. The proper loss defined by Eq. (9) was shown in (Santos-Rodriguez et al., 2009b) to satisfy the conditions of the above theorem. Therefore, for large n, the empirical risk in (11) converges to the total empirical cost. On the other hand, smaller values of n provide smoother approximations to the total cost. This way, n provides a "knob" to adjust the behaviour of the empirical risk between a generalized 0-1 loss and numerically and analytically better-behaved versions.

4. Related Work

The example-dependent cost setting has been rarely explored, mainly due to the difficulty of accurately obtaining reliable costs, being first introduced in (Lenarcik and Piasta, 1998) and also discussed by Provost and Fawcett in (Provost and Fawcett, 2001). We build upon (Zadrozny and Elkan, 2001), where costs are different for different examples in the same way that class membership probabilities are example-dependent. Zadrozny and Elkan depict domains under unknown costs and probabilities for test examples, so both cost and probability estimators are learned. Assuming that costs are given, this setting has been studied through cost-sensitive logistic regression for credit scoring in (Bahnsen et al., 2014). Similar principles are applied in (Bahnsen et al., 2015) making use of decision trees. The main limitation in these approaches is that they only consider binary classification. In (Brefeld et al., 2003) the standard support vector machine was adapted to account example-dependent costs.

Proper losses have recently become very popular (Reid and Williamson, 2011). They have also been translated to the cost-sensitive scenario, by defining Bregman divergences which are specially sensitive to changes near the classification boundaries (Santos-Rodriguez et al., 2009b,a; Santos-Rodríguez and Cid-Sueiro, 2012). The choice of loss function becomes more relevant when the knowledge about the problem is limited or the available training data is particularly demanding (for instance, applications where the high-cost examples are scarce). There, different loss functions lead to very different posterior probabilities estimates and *tailoring* of the loss becomes key (Buja et al., 2005). From a theoretical per-

spective, (Scott, 2011) explores the conditions for the existence of surrogate regret bounds for example-dependent surrogate losses.

5. Experiments

In this section we will detail the experiments performed using two datasets to explore the effect of the choice of loss in standard neural network architectures. First we present a binary classification task with low-dimensional data and then we focus on a more complex multiclass classification problem with high-dimensional inputs. For the later, we briefly introduce the multiclass extension of the loss functions discussed earlier. Let $\mathbf{y}^{(k)}$ be a one-hot encoding of the target class of the k-th example, $\hat{\mathbf{y}}^{(k)}$ the predicted output and $\mathbf{C}^{(k)}$ the cost matrix, with elements $c_{ij}^{(k)}$ being the cost of deciding in favour of i when the true label is j. We will explore the use of the multiclass version of example-dependent proper loss defined in Eq. (11) (EDBD),

$$L_n(\mathbf{y}^{(k)}, \hat{\mathbf{y}}^{(k)}) = \|\mathbf{z}(\mathbf{y}^{(k)})\|_n - \|\mathbf{z}(\hat{\mathbf{y}}^{(k)})\|_n^{1-n}(\mathbf{z}^{n-1}(\hat{\mathbf{y}}^{(k)}))^T\mathbf{z}(\mathbf{y}^{(k)}) \tag{18}$$

where $\|.\|_n$ denotes the n-norm and $\mathbf{z}(\mathbf{y})$ is as follows,

$$\mathbf{z}(\mathbf{y}) = \mathbf{s} - \mathbf{C}^{(k)}\mathbf{y}. \tag{19}$$

where $\mathbf{s} = \max_j\{c_{ij}\}$. The version of this loss that only uses the class-dependent cost matrix is referred to as BD in the experiments. We use as baseline the ever so popular cross entropy (CE). We also explore the class-dependent cross entropy (CDCE) and example-dependent cross entropy (EDCE) usually available in most neural network toolboxes defined as

$$L_{EDCE}(\mathbf{y}^{(k)}, \hat{\mathbf{y}}^{(k)}) = -\sum_i c_{ij}^{(k)}\left(y_i^{(k)}\log\hat{y}_i^{(k)}\right), \tag{20}$$

where i is an index that goes over the number of classes in the classification task, $j = \text{argmax}_i\{\hat{y}_i^{(k)}\}$. For CDCE $\mathbf{C}^{(k)}$ is deterministic.

For the binary classification task, these are also compared with the current state of the art in example-dependent cost-sensitive algorithms (EDDT)(Bahnsen et al., 2015). An implementation of this is available in the `CostCla` package [1]. As this implementation does not extend to the multiclass classification setting, it is not used in the second case study.

The main performance measure that we use is the total cost for the test dataset. For all experiments the total cost is evaluated using the example-dependent costs matrices. We will also look at the classification performance on a subset of the data with large example-dependent costs, namely the examples with example-dependent costs in the top 10%.

5.1. Case study: Credit Risk Assessment

5.1.1. Data

The German credit dataset[2] consists of 1000 instances, 700 of which belong to creditworthy applicants and the remaining 300 belong to applicants to whom credit should not be

1. https://pypi.python.org/pypi/costcla/0.5
2. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

extended. Each applicant is described by 24 attributes, depicting the status of existing accounts, credit history records, loan amount and purpose, employment status and an assortment of personal information. We use the numerical version of the dataset, provided by Strathclyde University. We use some information from the symbolic attributes to compute the costs (exact quantity and period of the loan). The task consists of classifying customers in two classes, namely bad or good clients.

5.1.2. Costs for Credit Risk

Let us consider a simple cost policy. Classifying a good credit customer as bad incurs a loss of $c_{10}^{(k)} = Q^{(k)} \cdot ((1+i)^{t^{(k)}} - 1)$. where i represent a quantity related to the interest rate, $Q^{(k)}$ is the amount of the loan and $t^{(k)}$ is the period of the load for sample k. Equivalently, classifying a bad credit customer as good incurs a loss of $c_{01}^{k} = Q^{(k)}$. We chose $i = 0.05$ but any other value behaves similarly. The costs are then normalized to lie on the range $(0, 1]$. A separate class-dependent cost matrix is provided with the dataset where $c_{10} = 5$ and $c_{01} = 1$ to account for the imbalance of the dataset and is used in CDCE and BD.

5.1.3. Architecture

A multilayer perceptron with 4 layers was used in the experiments. Each layer has a ReLU activation function except for the last layer where a Softmax function is applied. This results in the output being a probability vector. The details of the architecture can be found in the supplementary material. The network was trained using stochastic gradient descent with a learning rate of 1e$-$6 and batch size 100. All experiments were stopped after 200 epochs.

5.1.4. Results

Each experiment was performed 5 times on random training/test splits. Although random sampling was used to create splits we found that the distributions of Q were similar between splits. This becomes important when evaluating the models with respect to the total test cost. The distribution of Q for the training/test splits can be found in the supplementary material. For both EDBD and BD a basic search was performed for the value of n and it was found that $n = 2.1$ provided a good trade-off of performance and ease of optimization.

The results are summarized in Table 1. For both losses, including the example-dependent cost increased the average accuracy on examples with a value of Q in the top 10% for the test dataset. The total test cost is also significantly reduced with the inclusion of example-dependent costs. It is interesting to note that considering class-dependent costs also reduced the total test cost for CE. This is due to the unbalanced nature of the dataset and the cost matrix defined with the dataset has $c_{10} > c_{01}$. The example-dependent cost matrix defined in the previous section also satisfies the inequality. By training a neural network on the average of individual loss functions (EDBD), the total test loss is reduced. This agrees with the properties described in Sec. 3 in that minimizing the individual loss functions results in the global loss function also being minimized.

Table 1: Experimental results for the Credit Risk Assessment case study. Values are described in the form $mean \pm std$ across 5 experiments with different training/test splits of the dataset. In the table, NN denotes that a neural network was used. The loss functions are: CE - cross entropy; CDCE - class-dependent cross entropy; EDCE - example-dependent cross entropy; EDBD - example-dependent proper loss; BD - class-dependent proper loss; EDDT - example-dependent decision trees.

Algorithm	**Accuracy (%)**	**Top 10% of $Q^{(k)}$ Accuracy (%)**	**Total Test Example-Dependent Cost**
NN-CE	71.46 ± 4.53	61.99 ± 15.84	6.85 ± 2.25
NN-CDCE	72.40 ± 1.65	57.15 ± 10.99	3.41 ± 1.13
NN-EDCE	72.26 ± 1.73	64.64 ± 7.12	3.95 ± 0.69
NN-EDBD	74.82 ± 4.69	80.51 ± 11.26	2.06 ± 0.35
NN-BD	70.68 ± 2.88	49.74 ± 10.99	4.19 ± 0.73
EDDT	47.42 ± 5.80	66.14 ± 9.58	3.89 ± 0.89

5.2. Case study: Musical Genre Classification

5.2.1. Dataset

Musical Genre Classification is the task of finding a map between songs and a set of predifined musical genres (Tzanetakis and Essl, 2001). The musical features are a combination of pitch and timbre temporal data from the Million Song Dataset (MSD)(Bertin-Mahieux et al., 2011). The songs are segmented and at each segment the pitch and timbre calculated for 12 bins of frequencies resulting in matrices of size $t \times 12$, where t is the number of time-steps. Each song has a different number of segments and as such the pitch and timbre data is interpolated to a length of 500. The pitch and timbre data are then stacked to create a feature matrix of size 1000×12. These values are precomputed in the MSD.

The genre labels were gathered from the MSD Allmusic Top Genre Dataset (TopMAGD) and includes 13 popular genres. Roughly 4000 samples were taken from each genre to create a balanced dataset. We gather the number of listens is gathered from the Taste Profile Subset of the MSD. This subset from `lastFM` includes anonymized user IDs, song IDs corresponding to the MSD and the number of times that a user listened to a song. We aggregate this to calculate the total number of listens per song.

In order to ensure that the training/testing data splits have similar distributions of song-dependent cost information, the samples are placed into bins of equal size corresponding to the popularity measure. These bins are then randomly sampled using a ratio of 3:1 to create the training/testing split. The distribution of the song-dependent cost information used in the experiments detailed in this paper can be seen in the supplementary material.

5.2.2. Costs in musical genre classification

Until now, we have assumed that both types of costs were given. However, in practice, it is difficult to precisely quantify the costs. In this section, we outline a procedure to use data to estimate the costs.

Song-dependent costs There are a myriad of possible sources for song-dependent costs. These could include information about each specific song, artist or any other metadata that can be gathered. We propose that the use of *popularity* as a measure of significance. Popularity is inherently difficult to define Ni et al. (2011). As a proxy, we assume that songs which are accessed or queried more regularly are more important and thus the classifier has to be particularly careful with those. Websites such as ReverbNation[3] track the amount of interaction between bands and their fans and calculates a score that can also be used as a measure of popularity. Due to the limited scope of this type of sites, it is not feasible to collect this data for a large number of artists. In this paper, we suggest that the number of times a song has been listened to is a good substitute for its popularity.

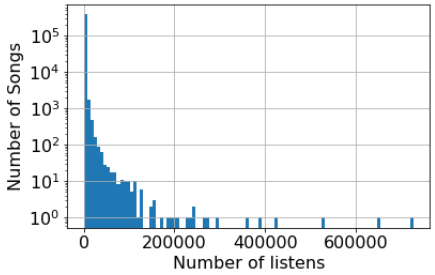


Figure 3: Histogram showing the distribution of the number of listens per song on.

As with popularity, the number of listens for a set of songs is well modelled as a long tail distribution. In Fig. 3 we show the number of listens for songs in the Million Song Dataset for which `Last.fm` data is available. This suggests that if number of listens was used as a song-dependent cost some songs would be several orders of magnitude more important than others. Scaling the number of listens to lie in the range $[0, 1]$ would cause most of the values to be almost zero which would result in difficulties when optimising the tailored loss functions. In order to avoid this we take the log of the number listens. The final popularity (example-dependent cost) measure *pop* is defined as

$$pop^{(k)} = \frac{\log(l^{(k)} + 1)}{\max\{pop\}}, \tag{21}$$

where the k-th the song that is being evaluated and l is the number of listens. This measure will be used for EDCE and EDBD as an importance weighting of the cost matrix described below.

Genre-dependent costs Various taxonomies for music genres have been defined in the literature, both in the form of general Directed Acyclic Graphs or trees. In order to simplify the task of constructing a taxonomy, we will be using a tree adapted from (Li and Ogihara, 2005) to only include genres that are in the data used in the empirical analysis. This tree can be seen in Fig. 4. With c_{ij} being the cost associated with labelling the sample as genre i when the true label is j, we define this cost as a shortest path between genre i and j in

3. https://www.reverbnation.com/

the tree defined above, following (Sordo et al., 2008). These distances are then scaled so that they lie in the range $[0, 1]$. These distances will be used for CDCE and BD.

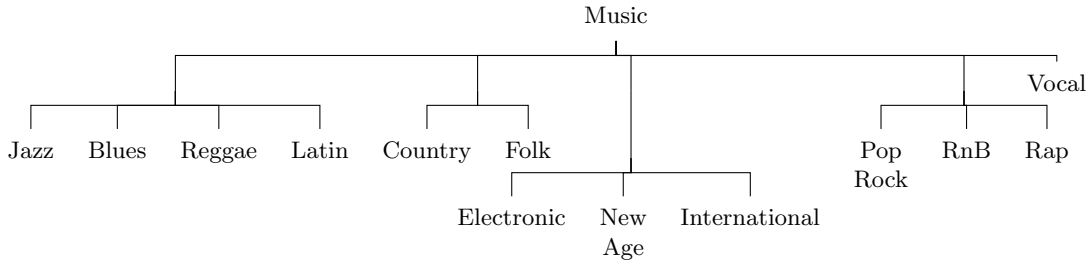


Figure 4: Tree describing the taxonomy of genres used to compute the distances among genres.

5.2.3. Neural Network Architecture

Due to the focus of this paper being on exploring loss function that include example-dependent costs and not achieving state of the art results in the field of genre classification, a simple 3 layer convolutional neural network architecture was used. The first layer has stride 1 and the next two have stride 2 across only one dimension. Batch-normalization (BN) is used in every layer and the ReLU activation applied. There are then two fully connected layers with a Softmax activation being applied to the last layer, resulting in the output becoming a probability vector of the input belongs to each of the 13 classes in the dataset. Exact details for the network can be found in the supplementary material. For all experiments the network was trained using the ADAM optimizer with learning rate $1e-5$ and an ℓ_2-norm weight-decay penalty of $5e-4$ to prevent overfitting. All experiments are stopped after 20 epochs.

5.2.4. Results

The experimental results for the musical genre classification case study can be found in Table 2. Utilizing the prior information given via the class-dependent (CDCE) cost shows an improvement of the overall accuracy of the classifier for cross entropy (CE). This is due to the penalization of misclassifying classes that are further away in the taxonomy defined above. Including the example-dependent costs further increases accuracy of the examples within the top 10% of *pop* values although the reduction of the total test cost is negligible due to the large standard deviation among experiments.

A search was performed over the parameter n for both the proper loss (BD) and example-dependent proper loss (EDBD) and it was found the $n = 2.15$ was a suitable trade-off (Table 3). EDBD outperforms BD in relation to the accuracy of the top 10% of examples and the total test cost. The standard deviation of the of the total test cost is also greatly reduced due to the large weighting that more popular songs have. For both CE and BD the inclusion of an example-dependent cost, popularity, within the loss function helps specialize

the classifier to a smaller subset of more popular songs. BD exhibits this behaviour to a greater degree with a larger decrease in total test cost.

In summary, using the song-dependent and genre-dependent costs, the accuracy does sacrifice accuracy for the whole dataset but it could be argued that not missclassifying the genre of popular songs is more critical.

Table 2: Experimental results for the Musical Genre Classification case study. Values are described in the form $mean \pm std$ across 3 experiments with different training/test splits of the dataset. The loss functions are: CE - cross entropy; CDCE - class-dependent cross entropy; EDCE - example-dependent cross entropy; EDBD - example-dependent proper loss; BD - class-dependent proper loss.

Algorithm	**Accuracy (%)**	**Top 10% of $pop^{(k)}$ Accuracy (%)**	**Total Test Example-Dependent Cost**
NN-CE	38.23 ± 0.11	38.4 ± 1.87	586.14 ± 8.53
NN-CDCE	40.32 ± 1.16	39.25 ± 1.13	572.29 ± 5.28
NN-EDCE	35.81 ± 0.86	42.11 ± 2.51	575.74 ± 13.08
NN-EDBD	37.23 ± 0.70	43.03 ± 1.56	556.19 ± 3.02
NN-BD	36.75 ± 3.51	37.56 ± 2.56	581.46 ± 17.03

Table 3: Experimental results for varying n in the example-dependent proper loss (EDBD) for the musical genre classification case study.

n	**Accuracy (%)**	**Top 10% of $pop^{(k)}$ Accuracy (%)**	**Total Test Example-Dependent Cost**
$n = 2$	33.74 ± 1.78	39.78 ± 2.86	581.19 ± 18.80
$n = 2.15$	37.23 ± 0.70	43.03 ± 1.56	556.19 ± 3.02
$n = 3$	35.32 ± 1.28	41.85 ± 1.09	570.20 ± 6.911
$n = 4$	34.64 ± 0.84	39.89 ± 0.99	592.85 ± 8.61

6. Conclusions

Example-dependent cost scenarios are pervasive and generalize standard minimum-error and (class-dependent) cost-sensitive classification. We have explored the use of linear combinations of proper losses to learning in example-dependent cost scenarios. The key idea is to associate each sample in the training set with its very own divergence, which intrinsically reflects the cost structure of the corresponding sample. Then, a global divergence is constructed by averaging the individual divergences. Optimizing this global divergence leads to posterior probability estimates which are specially accurate near the optimal boundary of the example-dependent cost-sensitive classification problem. We have explored the application of example-dependent cost functions in two separate case studies. In credit risk classification, including costs in the example-dependent loss decreased the total test cost

by up to 60% when compared to cost insensitive loss functions. In the multiclass setting of musical genre classification, example-dependent losses were constructed based off a proxy for the popularity of the song. We found that EDBD and EDCE outperformed the more conventional class-dependent cost-sensitive losses in terms of the test accuracy on the top 10% of the example-dependent costs.

Acknowledgements

This work was supported in part by EurValve (Personalised Decision Support for Heart Valve Disease), Project Number: H2020 PHC-30-2015. Additionally this work was supported by the Spanish Ministry of Economy and Competitiveness under Grants TEC2016-81900-REDT/AEI, and TEC2017-83838-R.

References

Alejandro Correa Bahnsen, Djamila Aouada, and Björn Ottersten. Example-dependent cost-sensitive logistic regression for credit scoring. In *International Conference on Machine Learning and Applications*, pages 263–269, 2014.

Alejandro Correa Bahnsen, Djamila Aouada, and Björn Ottersten. Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, 42(19):6609–6619, 2015.

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. *International Society for Music Information Retrieval Conference*, 2011.

Ulf Brefeld, Peter Geibel, and Fritz Wysotzki. Support vector machines with example dependent costs. *European Conference on Machine Learning*, pages 23–34, 2003.

Lev M Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(10):200–217, 1967.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, Department of of Statistics, University of Pennsylvania, 2005.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

Jacek P Dmochowski, Paul Sajda, and Lucas C Parra. Maximum likelihood in cost-sensitive learning: Model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11:3313–3332, 2010.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

Andrzej Lenarcik and Zdzisław Piasta. Rough classifiers sensitive to costs varying from object to object. In *International Conference on Rough Sets and Current Trends in Computing*, pages 222–230. Springer-Verlag, 1998.

Tao Li and Mitsunori Ogihara. Music genre classification with taxonomy. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

John W Miller, Rod Goodman, and P Smyth. Objective functions for probability estimation. In *International Conference on Neural Networks*, volume 1, pages 881–886, 1991.

Yizhao Ni, Raul Santos-Rodriguez, Matt McVicar, and Tijl De Bie. Hit song science once again a science? In *Proceedings of 4th international workshop on Machine learning and music*. ACM, 2011.

Foster Provost and Tom Fawcett. Robust classification systems for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

Raúl Santos-Rodríguez and Jesús Cid-Sueiro. Cost-sensitive sequences of bregman divergences. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1896–1904, 2012.

Raul Santos-Rodriguez, Darío García-García, and Jesús Cid-Sueiro. Cost-sensitive classification based on bregman divergences for medical diagnosis. In *Proceedings of the International Conference on Machine Learning and Applications*, pages 551–556, 2009a.

Raul Santos-Rodriguez, Alicia Guerrero-Curieses, Rocío Alaiz-Rodríguez, and Jesús Cid-Sueiro. Cost-sensitive learning based on bregman divergences. *Mach. Learn.*, 76:271–285, 2009b.

Clayton Scott. Surrogate losses and regret bounds for cost-sensitive classification with example-dependent costs. *28th International Conference on Machine Learning*, 2011.

Mohamed Sordo, Oscar Celma, Martin Blech, and Enric Guaus. The quest for musical genres: Do the experts and the wisdom of crowds agree? In *International Conference on Music Information Retrieval*, pages 255–260, 2008.

Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *International Joint Conference on Neural Networks*, 2010.

George Tzanetakis and Georg Essl. Automatic musical genre classification of audio signals. In *IEEE Transactions on Speech and Audio Processing*, pages 293–302, 2001.

Bianca Zadrozny and Charles Elkan. Learning and making decisions when costs and probabilities are both unknown. In *7th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213. ACM, 2001.