GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare

Arpit Agarwal ARPIT.AGARWAL@CSA.IISC.ERNET.IN
Harikrishna Narasimhan HARIKRISHNA@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore 560012, India

Shivaram Kalyanakrishnan SHIVARAM@YAHOO-INC.COM
Yahoo Labs Bangalore, Bangalore 560071, India

Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore 560012, India

Abstract

We consider the problem of binary class probability estimation (CPE) when one class is rare compared to the other. It is well known that standard algorithms such as logistic regression do not perform well in this setting as they tend to underestimate the probability of the rare class. Common fixes include under-sampling and weighting, together with various correction schemes. Recently, Wang & Dey (2010) suggested the use of a parametrized family of asymmetric link functions based on the generalized extreme value (GEV) distribution, which has been used for modeling rare events in statistics. The approach showed promising initial results, but combined with the logarithmic CPE loss implicitly used in their work, it results in a non-convex composite loss that is difficult to optimize. In this paper, we use tools from the theory of proper composite losses (Buja et al., 2005; Reid & Williamson, 2010) to construct a canonical underlying CPE loss corresponding to the GEV link, which yields a convex proper composite loss that we call the *GEV-canonical loss*; this loss can be tailored to CPE settings where one class is rare, and is easily minimized using an IRLS-type algorithm similar to that used for logistic regression. Our experiments on both synthetic and real data suggest that the resulting algorithm – which we term *GEV-canonical regression* – performs well compared to common approaches such as under-sampling and weights-correction for this problem.

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

1. Introduction

The problem of estimating class probabilities from data with binary labels is a fundamental one in machine learning, and arises in several applications in practice, including for example medical diagnosis, fraud prediction, click-rate prediction in web advertising, etc. In many of these applications, one class is rare compared to the other: in medical diagnosis, only a few patients develop a given disease; in fraud prediction, only a few transactions turn out to be fraudulent; in web advertising, only a few ad impressions result in clicks, and so on. Yet in all these applications, it is important to accurately estimate the *probability* of the rare class occurring: in medical diagnosis, these probabilities help in deciding the right course of treatment; in fraud prediction, these probabilities help in estimating the risk of various actions; in web advertising, these probabilities help in deciding how to rank or display various ads, and so on.

It is well known that classical approaches such as logistic regression do not perform well in such settings, especially when the probability of the rare class is very small and the number of training examples is limited (Czado & Santner, 1992; King & Zeng, 2001; Zhang, 2004). The reason for this is that the logistic loss used in logistic regression is symmetric in nature, i.e. it assigns equal penalty for the losses on positive and negative examples. Common fixes used in practice include under-sampling the majority class to balance the two classes before training or weighting losses on positive and negative examples differently, and then applying some form of correction scheme when estimating probabilities from the learned model (King & Zeng, 2001; Wallace & Dahabreh, 2012).

The logistic loss can be viewed as a proper composite loss that combines the well-known logarithmic loss for binary class probability estimation (CPE) with the symmetric logit link, which is the 'canonical' link for the loga-

rithmic loss (Buja et al., 2005; Reid & Williamson, 2010). An alternative approach is to use an asymmetric link function that helps penalize mispredictions on positive examples differently from those on negative examples. Recently, Wang & Dey (2010) suggested the use of a parametrized family of asymmetric link functions based on the generalized extreme value (GEV) distribution, which has been used for modeling rare events in statistics (Kotz & Nadarajah, 2000; Embrechts et al., 1997); a similar approach was used in (Calabrese & Osmetti, 2011). These works use a probabilistic model together with maximum likelihood/maximum a posteriori estimation, which effectively composes the GEV links with the same logarithmic CPE loss as that used in logistic regression; unfortunately, this results in a non-convex optimization problem.

In this paper, we use tools from the theory of proper composite losses, which are natural choices for CPE problems in general (Buja et al., 2005; Reid & Williamson, 2010), to construct a proper composite loss with desirable properties for CPE problems when one class is rare. Specifically, we derive a family of underlying CPE losses for which the GEV links form the 'canonical' link. The resulting proper composite loss family, which we call the *GEV-canonical loss* family, can be used to adapt to the degree of rarity in the data and accordingly penalize wrong predictions on positive examples more heavily than those on negative examples. In addition, due to properties of proper composite losses formed using canonical links, each loss in the GEV-canonical loss family is convex, allowing us to use an iterative reweighted least squares (IRLS) algorithm for its minimization, similar to that used in common logistic regression implementations. The resulting algorithm, which we term *GEV-canonical regression*, outperforms various baselines in experiments with both synthetic and real data, particularly when the number of training examples is limited.

Related Work. In addition to the work mentioned above, there has been much interest in learning in class imbalance settings in general, with several workshops, survey articles, and editorials devoted to the topic over the years (Provost, 2000; Japkowicz, 2000; Chawla et al., 2004; Van Hulse et al., 2007; He & Garcia, 2009). Much of this work focuses on *classification* in class imbalance settings, where again approaches such as weighting the two classes differently, subsampling the majority class or over-sampling the minority class before training, and using calibration to correct predictions afterwards are widely used (Chawla et al., 2002; Drummond & Holte, 2003; Van Hulse et al., 2007; Lee et al., 2012). Such techniques are also used for cost-sensitive learning (Elkan, 2001; Zadrozny et al., 2003; Masnadi-Shirazi & Vasconcelos, 2010), wherein different misclassification errors are penalized differently. A different approach to cost-sensitive learning, which bears some relation to our work, is that of Guerrero-Curieses et al. (2004), who designed loss functions to predict class probabilities accurately around a given classification threshold in order to minimize classification errors. In this paper, our interest is primarily in CPE problems in class imbalance settings, where we wish to estimate reliably the class probabilities, particularly in the region of small probabilities but in other regions as well, despite the imbalance in the data.

Organization. We start with preliminaries and background on proper loss functions, link functions, the GEV link family, and proper composite losses in Section 2. Section 3 derives the proper loss for which the GEV link forms the 'canonical' link, constructs the GEV-canonical proper composite loss, and describes the resulting GEV-canonical regression algorithm. Section 4 gives our experimental results comparing the GEV-canonical regression algorithm with several baselines on both synthetic and real data. We conclude with a brief discussion in Section 5.

2. Preliminaries and Background

Notation. We denote $\mathbb{R} = (-\infty, \infty)$, $\overline{\mathbb{R}} = [-\infty, \infty]$, $\mathbb{R}_+ = [0, \infty)$, and $\overline{\mathbb{R}}_+ = [0, \infty]$. For $z \in \mathbb{R}$, we denote $z_+ = \max(0, z)$.

Problem Setup. We consider binary CPE problems where there is an instance space $\mathcal{X}$, binary label space $\mathcal{Y} = \{\pm 1\}$, and an underlying (unknown) probability distribution D on $\mathcal{X} \times \{\pm 1\}$ from which both training examples and future test examples are assumed to be drawn i.i.d. Let (X, Y) denote a random example drawn from D. Let $p = \mathbf{P}(Y = 1)$ denote the overall probability of the positive class under D, and let $\eta : \mathcal{X} \to [0, 1]$ denote the associated class probability function: $\eta(x) = \mathbf{P}(Y = 1 \,|\, X = x)$. Given a training sample $S = ((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)) \sim D^n$, the goal is to learn from S a CPE model $\widehat{\eta}_S : \mathcal{X} \to [0, 1]$ that accurately estimates the true class probability function η. We are interested in settings where one class (say the positive class) is rare, so that $p \ll 0.5$.

CPE Loss Functions and Proper Losses. A CPE loss function is any loss function $c : \{\pm 1\} \times [0, 1] \to \overline{\mathbb{R}}_+$ that assigns a penalty $c(y, \widehat{\eta})$ for predicting $\widehat{\eta} \in [0, 1]$ as the probability of a positive label when the true label is $y \in \{\pm 1\}$. A CPE loss c can equivalently be defined through its partial losses $c_1 : [0, 1] \to \overline{\mathbb{R}}_+$ and $c_{-1} : [0, 1] \to \overline{\mathbb{R}}_+$, given by $c_y(\widehat{\eta}) = c(y, \widehat{\eta})$. A popular CPE loss is the *logarithmic loss* $c^{\log} : \{\pm 1\} \times [0, 1] \to \overline{\mathbb{R}}_+$, with partial losses given by

$$c_1^{\log}(\widehat{\eta}) = -\ln(\widehat{\eta}) \,; \tag{1}$$

$$c_{-1}^{\log}(\widehat{\eta}) = -\ln(1 - \widehat{\eta}) \,. \tag{2}$$

For any CPE loss function c, define the point-wise c-risk $L_c : [0, 1] \times [0, 1] \to \overline{\mathbb{R}}_+$ as follows:[1]

[1] Note that we overload notation and use η to denote both the class probability function and a number in $[0, 1]$; the usage should be clear from context.

$$L_c(\eta, \widehat{\eta}) \;=\; \eta\, c_1(\widehat{\eta}) + (1-\eta)\, c_{-1}(\widehat{\eta})\,.$$

A CPE loss c is said to be *proper* if the point-wise c-risk $L_c(\eta, \widehat{\eta})$ is minimized by $\widehat{\eta} = \eta$ for all $\eta \in [0,1]$, i.e. if

$$\eta \in \operatorname*{arg\,min}_{\widehat{\eta} \in [0,1]} L_c(\eta, \widehat{\eta}) \qquad \forall \eta \in [0,1]\,,$$

and *strictly proper* if in addition this minimizer is unique. This is a desirable property for any CPE loss as it ensures that minimizing the loss yields the correct probability estimates. Proper losses are related to proper scoring rules that have been used in the probability forecasting literature (Savage, 1971; Hendrickson & Buehler, 1971; Schervish, 1989; Gneiting & Raftery, 2007), and have received significant interest in the machine learning community recently (Buja et al., 2005; Reid & Williamson, 2009; 2010; Agarwal, 2013). It can be verified that the logarithmic loss $c^{\log}$ defined above is strictly proper.

Link Functions and the GEV Link Family. In practice, when learning a CPE model $\widehat{\eta}_S : \mathcal{X} \to [0,1]$, one usually learns a real-valued scoring model $f_S : \mathcal{X} \to \mathcal{V}$ for some $\mathcal{V} \subseteq \overline{\mathbb{R}}$, and then maps the real-valued scores to probability estimates in $[0,1]$ via a *link function*. A link function is any strictly increasing function $\psi : [0,1] \to \mathcal{V}$; one then uses the *inverse link function* $\psi^{-1} : \mathcal{V} \to [0,1]$ to map scores $f_S(x) \in \mathcal{V}$ to probability estimates $\widehat{\eta}_S(x) = \psi^{-1}(f_S(x))$. One of the most widely used link functions is the *logit link* $\psi_{\text{logit}} : [0,1] \to \overline{\mathbb{R}}$, defined as

$$\psi_{\text{logit}}(\widehat{\eta}) = \ln\left(\frac{\widehat{\eta}}{1-\widehat{\eta}}\right).$$

Other common links include the *probit link* $\psi_{\text{probit}} : [0,1] \to \overline{\mathbb{R}}$, defined as

$$\psi_{\text{probit}}(\widehat{\eta}) = \Phi^{-1}(\widehat{\eta})\,,$$

where Φ denotes the standard normal CDF, and the *complementary log-log (cloglog) link* $\psi_{\text{cloglog}} : [0,1] \to \overline{\mathbb{R}}$, defined as

$$\psi_{\text{cloglog}}(\widehat{\eta}) = \ln(-\ln(1-\widehat{\eta}))\,.$$

The logit and probit links are both symmetric, in that they satisfy $\psi(\widehat{\eta}) = -\psi(1-\widehat{\eta})$; the cloglog link is asymmetric. A general method for constructing a link function is to use the inverse CDF of a continuous real-valued random variable, just as the probit link uses the inverse standard normal CDF. Recently, Wang & Dey (2010) proposed the use of the CDF of the generalized extreme value (GEV) distribution, used in statistics for modeling rare events (Kotz & Nadarajah, 2000; Embrechts et al., 1997), for constructing a parametric family of asymmetric links. Specifically, the CDF of the GEV distribution with location parameter $\mu = 0$, scale parameter $\sigma = 1$, and shape parameter $\xi \in \mathbb{R}$, which we shall denote as $F_\xi : \overline{\mathbb{R}} \to [0,1]$, is defined as

$$F_\xi(v) \;=\; \exp(-(1+\xi v)_+^{-1/\xi})\,.$$

Clearly, this distribution has support $[-\frac{1}{\xi}, \infty)$ for $\xi > 0$; $(-\infty, -\frac{1}{\xi}]$ for $\xi < 0$; and $\mathbb{R}$ for $\xi = 0$; taking the limit in the above as $\xi \to 0$, one recovers the Gumbel distribution: $F_0(v) = \exp(-\exp(-v))$. Denote the extension of the above support by $\overline{\mathbb{R}}_\xi$:

$$\overline{\mathbb{R}}_\xi \;=\; \begin{cases} [-\frac{1}{\xi}, \infty] & \text{if } \xi > 0 \\ [-\infty, -\frac{1}{\xi}] & \text{if } \xi < 0 \\ \overline{\mathbb{R}} & \text{if } \xi = 0. \end{cases}$$

The corresponding GEV link, parametrized by $\xi \in \mathbb{R}$ and which we denote as $\psi_{\text{GEV}(\xi)} : [0,1] \to \overline{\mathbb{R}}_\xi$, is then defined as

$$\psi_{\text{GEV}(\xi)}(\widehat{\eta}) \;=\; \frac{1}{\xi}\left(\frac{1}{\big(-\ln(\widehat{\eta})\big)^\xi} - 1\right) \qquad \forall \xi \in \mathbb{R}\backslash\{0\}\,.$$

In the limit $\xi \to 0$, it leads to the standard log-log link: $\psi_{\text{GEV}(0)}(\widehat{\eta}) = -\ln(-\ln(\widehat{\eta}))$. The parameter ξ can be adjusted to yield different degrees of asymmetry in the above link, which in turn can be used to fit different degrees of rarity in the data. This is similar to how the GEV distribution is used traditionally, where one selects the most appropriate distribution in the GEV family to model the underlying data by adjusting the parameter ξ .

Proper Composite Losses and Canonical Links. A common way to learn a real-valued scoring function $f_S : \mathcal{X} \to \mathcal{V}$ (for $\mathcal{V} \subseteq \overline{\mathbb{R}}$ as above) is to minimize a loss function $\ell : \{\pm 1\} \times \mathcal{V} \to \overline{\mathbb{R}}_+$ on the training sample S, i.e. to minimize $\sum_{i=1}^n \ell(y_i, f(x_i))$ over some suitable class of functions f, where $\ell(y,v)$ can be viewed as the penalty assigned by ℓ for predicting a score $v \in \mathcal{V}$ when the true label is $y \in \{\pm 1\}$. Again, any such loss ℓ can equivalently be defined through its partial losses $\ell_1 : \mathcal{V} \to \overline{\mathbb{R}}_+$ and $\ell_{-1} : \mathcal{V} \to \overline{\mathbb{R}}_+$, given by $\ell_y(v) = \ell(y,v)$. A popular loss operating on scores in $\overline{\mathbb{R}}$ is the *logistic loss* $\ell^{\text{logistic}} : \{\pm 1\} \times \overline{\mathbb{R}} \to \overline{\mathbb{R}}_+$ used in logistic regression, defined as

$$\ell^{\text{logistic}}(y,v) = \ln(1 + e^{-yv})\,.$$

A loss function $\ell : \{\pm 1\} \times \mathcal{V} \to \overline{\mathbb{R}}_+$ is said to be *proper composite* (Buja et al., 2005; Reid & Williamson, 2010) if it can be written as a composition of a proper CPE loss $c : \{\pm 1\} \times [0,1] \to \overline{\mathbb{R}}_+$ and a link $\psi : [0,1] \to \mathcal{V}$, so that

$$\ell(y,v) = c(y, \psi^{-1}(v)) \qquad \forall y \in \{\pm 1\}, v \in \mathcal{V}\,,$$

and *strictly proper composite* if in addition c is strictly proper. It is easy to verify that the popular logistic loss is a strictly proper composite loss, composed of the strictly proper logarithmic CPE loss and the logit link. It is common to compose the logarithmic CPE loss with other link functions as well, such as the probit or cloglog links. The approach in (Wang & Dey, 2010; Calabrese & Osmetti, 2011) uses the GEV link in a probabilistic model and performs maximum likelihood or maximum a posteriori estimation under a suitable prior, which also amounts to effectively composing the GEV link with the logarithmic

Table 1. Summary of proper composite losses considered in this paper.

	COMPOSITE LOSS $\ell(y,v)$	INVERSE LINK $\psi^{-1}(v)$	UNDERLYING CPE LOSS $c(y,\widehat{\eta})$	FLEXIBLE LINK?	CONVEX?
LOGISTIC	$\ln(1-e^{-yv})$	$1/(1+e^{-v})$	SEE EQS. (1-3)	$\times$	$\surd$
PROBIT	$c(y,\psi^{-1}(v))$	$\Phi(v)$	SEE EQS. (1-3)	$\times$	$\surd$
CLOGLOG	$c(y,\psi^{-1}(v))$	$1-e^{-e^v}$	SEE EQS. (1-3)	$\times$	$\surd$
GEV-LOG(ξ)	$c(y,\psi^{-1}(v))$	$\exp(-(1+\xi v)^{-1/\xi})$	SEE EQS. (1-3)	$\surd$	$\times$
GEV-CANONICAL(ξ)	$c(y,\psi^{-1}(v))$	$\exp(-(1+\xi v)^{-1/\xi})$	SEE EQS. (3-4)	$\surd$	$\surd$

CPE loss (and possibly adding regularization to the resulting minimization problem). Unfortunately, this results in a non-convex optimization problem.

For any strictly proper loss, there is a unique 'canonical' link for which the resulting composite loss satisfies various desirable properties, including convexity; conversely, for any link function, there is a unique 'canonical' strictly proper loss (Buja et al., 2005; Reid & Williamson, 2010). The logit link and logarithmic loss form a canonical pair. Below we construct a proper composite loss using the GEV link and its corresponding canonical proper loss.

3. GEV-Canonical Regression

In this section we examine the GEV link family more closely through the lens of proper composite losses. Using results of (Buja et al., 2005; Reid & Williamson, 2010), we derive a parametric proper CPE loss for which the GEV link forms the canonical link. This allows us to maintain the attractive properties of the GEV link for CPE settings when one class is rare, namely flexibility of the link function to adapt to varying degrees of rarity in the data, as well as obtain desirable properties for the overall composite loss, such as convexity in the second argument. We term the resulting proper composite loss the *GEV-canonical loss* (see Table 1 and Figure 1 for a summary). This loss can be minimized efficiently using an IRLS algorithm similar to that used in logistic regression implementations; we term the resulting algorithm *GEV-canonical regression*.

3.1. GEV-Canonical Loss

As described in (Buja et al., 2005; Reid & Williamson, 2010), for any link function $\psi : [0,1]\to\mathcal{V}$, the strictly proper CPE loss $c : \{\pm 1\}\times[0,1]\to\overline{\mathbb{R}}_+$ that yields a canonical pair with ψ is given by

$$c_1(\widehat{\eta}) = \int_{\widehat{\eta}}^{1} (1-q)\,\omega(q)\,dq\,;$$

$$c_{-1}(\widehat{\eta}) = \int_{0}^{\widehat{\eta}} q\,\omega(q)\,dq\,,$$

where $\omega : (0,1)\to\mathbb{R}_+$ is a weight function given by

$$\omega(q) = \psi'(q) \qquad \forall q\in(0,1)\,.$$

Applying this result to the parametric GEV link, we get that for any $\xi\in\mathbb{R}$, the following is the unique strictly proper CPE loss for which the GEV link with parameter ξ forms the canonical link:

$$c_1^{\text{GEV-can}(\xi)}(\widehat{\eta}) = \int_{\widehat{\eta}}^{1} \frac{1-q}{q(-\ln q)^{1+\xi}}\,dq\,; \tag{3}$$

$$c_{-1}^{\text{GEV-can}(\xi)}(\widehat{\eta}) = \int_{0}^{\widehat{\eta}} \frac{1}{(-\ln q)^{1+\xi}}\,dq\,. \tag{4}$$

The resulting proper composite loss, which we refer to as the *GEV-canonical loss*, is given by

$$\ell_y^{\text{GEV-can}(\xi)}(v) = c_y^{\text{GEV-can}(\xi)}\big(\psi^{-1}_{\text{GEV}(\xi)}(v)\big) \qquad \forall v\in\overline{\mathbb{R}}_\xi\,.$$

This loss is not available in closed form, but is guaranteed to be convex on $\overline{\mathbb{R}}_\xi$ for all ξ, and moreover, as we describe below, can be minimized efficiently using an IRLS algorithm. Plots of the GEV-canonical loss for various values of ξ (obtained using numerical integration) are shown in Figure 1. As can be seen, different values of ξ yield different forms of asymmetry; for larger values of ξ, the loss effectively penalizes mispredictions on positive examples more heavily than those on negative examples.

For comparison, Figure 1 also shows plots of the logistic, probit and cloglog losses, as well as the GEV-log loss effectively used in (Wang & Dey, 2010; Calabrese & Osmetti, 2011), which is composed of the GEV link together with the logarithmic CPE loss:

$$\ell_y^{\text{GEV-}\log(\xi)}(v) = c_y^{\log}\big(\psi^{-1}_{\text{GEV}(\xi)}(v)\big) \qquad \forall v\in\overline{\mathbb{R}}_\xi\,.$$

The GEV-log loss is non-convex for $\xi\notin[-1,0.1)$, making its minimization prone to local minima.

3.2. IRLS Algorithm for GEV-Canonical Regression

In the following, fix any $\xi\in\mathbb{R}$. (In practice, ξ will be selected based on the training data, by cross-validation or by using a validation set.) For Euclidean instance spaces, we show how the GEV-canonical loss for any fixed $\xi\in\mathbb{R}$ can be minimized over linear functions using an IRLS algorithm; extension to non-linear functions or to non-Euclidean instance spaces via kernels is straightforward.

Let $\mathcal{X}=\mathbb{R}^k$ for some $k\in\mathbb{Z}_+$, and let $S=((\mathbf{x}_i,y_i))_{i=1}^n\in(\mathbb{R}^k\times\{\pm 1\})^n$. Since the GEV-canonical loss is defined

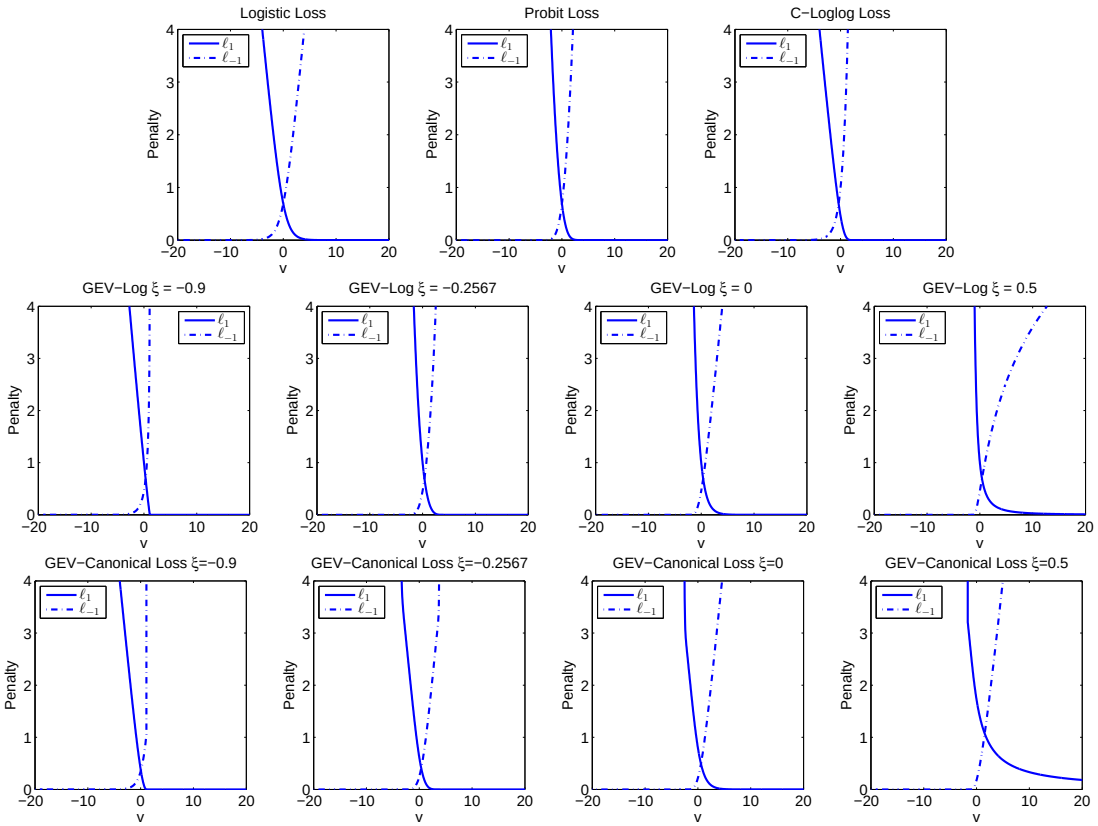

Figure 1. Partial losses for various proper composite losses considered in this paper: logistic, probit, cloglog, GEV-log for various values of ξ, and GEV-canonical for various values of ξ. The logistic, probit, cloglog, and GEV-log losses all use the symmetric underlying logarithmic CPE loss (with different link functions); the GEV-canonical loss uses an asymmetric underlying CPE loss for which the GEV link forms the canonical link. GEV-log and GEV-canonical losses with with $\xi = -0.2567$ can be used when one wants the composite loss to be (close to) symmetric. More interestingly, for larger values of ξ, GEV-log and and GEV-canonical losses penalize mispredictions on positive examples more heavily than those on negative examples, which is desirable in CPE settings when one class is rare. For $\xi \notin [-1, 0.1)$, the GEV-log loss is non-convex. On the contrary, the GEV-canonical loss is guaranteed to be convex for all ξ.

only for scores in $\mathcal{V} = \overline{\mathbb{R}}_\xi$, we consider learning a 'clipped' linear function $f_S : \mathcal{X} \to \overline{\mathbb{R}}_\xi$ of the form

$$f_S(\mathbf{x}) = \text{clip}_\xi(\boldsymbol{\beta}^\top \mathbf{x})$$

for some $\boldsymbol{\beta} \in \mathbb{R}^k$, where $\text{clip}_\xi : \overline{\mathbb{R}} \to \overline{\mathbb{R}}_\xi$ clips values outside the interval $\overline{\mathbb{R}}_\xi$ to the closest endpoint of the interval:

$$\text{clip}_\xi(v) \;=\; \begin{cases} \max(-\frac{1}{\xi}, v) & \text{if } \xi > 0 \\ \min(-\frac{1}{\xi}, v) & \text{if } \xi < 0 \\ v & \text{if } \xi = 0. \end{cases}$$

Thus we would like to minimize the following objective over $\boldsymbol{\beta}$:

$$\widehat{L}_\xi(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n \ell_{y_i}^{\text{GEV}(\xi)\text{-can}}\big(\text{clip}_\xi(\boldsymbol{\beta}^\top \mathbf{x}_i)\big)\,.$$

Now, while $\ell_1^{\text{GEV}(\xi)\text{-can}}(v)$ and $\ell_{-1}^{\text{GEV}(\xi)\text{-can}}(v)$ are both convex in v over $\overline{\mathbb{R}}_\xi$, the losses in the above sum are not always convex in $\boldsymbol{\beta}$. In particular, when $\xi > 0$, we have that for any $\mathbf{x} \in \mathbb{R}^k$, $\ell_{-1}^{\text{GEV}(\xi)\text{-can}}\big(\text{clip}_\xi(\boldsymbol{\beta}^\top \mathbf{x})\big)$ is convex in $\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^k$, but $\ell_1^{\text{GEV}(\xi)\text{-can}}\big(\text{clip}_\xi(\boldsymbol{\beta}^\top \mathbf{x})\big)$ is convex in $\boldsymbol{\beta}$ only for $\boldsymbol{\beta} : \boldsymbol{\beta}^\top \mathbf{x} \in \overline{\mathbb{R}}_\xi$; the reverse is true when $\xi < 0$. Therefore we would like to solve the following convex optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathcal{C}_\xi} \widehat{L}_\xi(\boldsymbol{\beta})\,,$$

where

$$\mathcal{C}_\xi = \begin{cases} \big\{\boldsymbol{\beta} \in \mathbb{R}^k \,\big|\, \boldsymbol{\beta}^\top \mathbf{x}_i \in \overline{\mathbb{R}}_\xi \;\forall i \in [n] : y_i = 1\big\} & \text{if } \xi > 0 \\ \big\{\boldsymbol{\beta} \in \mathbb{R}^k \,\big|\, \boldsymbol{\beta}^\top \mathbf{x}_i \in \overline{\mathbb{R}}_\xi \;\forall i \in [n] : y_i = -1\big\} & \text{if } \xi < 0 \\ \mathbb{R}^k & \text{if } \xi = 0\,. \end{cases}$$

While the objective above is not available in closed form, its gradient and Hessian in $\mathcal{C}_\xi$ can be expressed in closed form:

$$\nabla_{\boldsymbol{\beta}} \widehat{L}_\xi(\boldsymbol{\beta}) \;=\; -\sum_{i=1}^n \big(\mathbf{1}(y_i = 1) - \eta_i\big)\,\mathbf{x}_i \quad \forall \boldsymbol{\beta} \in \mathcal{C}_\xi\,;$$

$$\nabla^2_{\boldsymbol{\beta}} \widehat{L}_\xi(\boldsymbol{\beta}) \;=\; \sum_{i=1}^n \eta_i\big(-\ln(\eta_i)\big)^{\xi+1} \mathbf{x}_i \mathbf{x}_i^\top \quad \forall \boldsymbol{\beta} \in \mathcal{C}_\xi\,;$$

where $\eta_i = \psi^{-1}_{\text{GEV}(\xi)}\big(\text{clip}_\xi(\boldsymbol{\beta}^\top \mathbf{x}_i)\big)\;\forall i \in [n]$.

Algorithm 1 GEV-Canonical Regression (using IRLS)

Input: Data $S = ((\mathbf{x}_i, y_i))_{i=1}^n \in (\mathbb{R}^k \times \{\pm 1\})^n$

Initialize: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times k}$

$$\eta_i^{(1)} = \begin{cases} 0.75 & \text{if } y_i = 1 \\ 0.25 & \text{if } y_i = -1 \end{cases} \quad \forall i \in [n]$$

$$v_i^{(1)} = \psi_{\text{GEV}(\xi)}(\eta_i^{(1)}) \quad \forall i \in [n]$$

$t = 1$

repeat

  for $i = 1$ **to** n **do**

    $w_i^{(t)} = \eta_i^{(t)}(-\ln(\eta_i^{(t)}))^{\xi+1}$

    choose a suitable step size $\gamma^{(t)}$

    $z_i^{(t)} = v_i^{(t)} + \gamma^{(t)} \cdot \left(\mathbf{1}(y_i = 1) - \eta_i^{(t)}\right) \cdot \psi'_{\text{GEV}(\xi)}(\eta_i^{(t)})$

  end for

  $\mathbf{W}^{(\mathbf{t})} = \text{diag}(w_1^{(t)}, \cdots, w_n^{(t)})$

  $\boldsymbol{\beta}^{(t)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}$

    // compute $\boldsymbol{\beta}^{(t)}$ via weighted least squares (WLS)

  for $i = 1$ **to** n **do**

    $v_i^{(t+1)} = (\boldsymbol{\beta}^{(t)})^\top \mathbf{x}_i$

    $\eta_i^{(t+1)} = \psi^{-1}_{\text{GEV}(\xi)}(\text{clip}_\xi(v_i^{(t+1)}))$

  end for

  $t \leftarrow t + 1$

until convergence

Output: Coefficient vector $\boldsymbol{\beta}^{(t-1)} \in \mathbb{R}^k$

Given the gradient and the Hessian, one can use Newton's method iteratively:

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} - \gamma \cdot \left(\nabla^2_{\boldsymbol{\beta}} \widehat{L}_\xi(\boldsymbol{\beta})\right)^{-1} \left(\nabla_{\boldsymbol{\beta}} \widehat{L}_\xi(\boldsymbol{\beta})\right)\Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_{old}},$$

where γ is a suitable step size. It can be verified that if $\boldsymbol{\beta}_{old} \in \mathcal{C}_\xi$, then Newton's update with an appropriate step size will result in $\boldsymbol{\beta}_{new} \in \mathcal{C}_\xi$. For simplicity, in our experiments, we fix $\gamma = 1$, although it may be worth exploring variable step size schedules in future work.

Implementing Newton's method directly as above requires inverting the Hessian, which turns out to be a costly operation. Therefore we use a variant of the iterative reweighed least squares (IRLS) algorithm (Green, 1984) to implement Newton's method (see Algorithm 1). To avoid overfitting, one can add L_2-norm regularization by simply replacing the WLS step in the algorithm with the following:

$$\boldsymbol{\beta}^{(t)} = (\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)},$$

where λ is the regularization parameter, $\mathbf{X}$, $\mathbf{W}$ are as defined in Algorithm 1, and $\mathbf{I}$ is the $k \times k$ identity matrix.

4. Experiments

We conducted experiments with both synthetic and real data. In the case of synthetic data (Section 4.1), we model settings where $\eta(x)$ is small for most x; this might be the case, for example, in a web advertising application, where most ad impressions have fairly small probabilities of resulting in a click. Since here we know the true class probability function η, we measure performance of the learned CPE model $\widehat{\eta}_S$ relative to η on a test set. In the case of real data (Section 4.2), for which we use 12 data sets from the UCI repository (Frank & Asuncion, 2010), we do not have the true class probability function η, but only observed labels y, and so we measure performance of the learned CPE model $\widehat{\eta}_S$ relative to the observed labels y on a test set.

In both sets of experiments, we compare our proposed GEV-canonical regression algorithm with the following algorithms as baselines (all implemented to learn a linear function): logistic regression, probit regression, C-loglog regression, under-sampled logistic regression with King & Zeng's correction to the learned $\boldsymbol{\beta}$ coefficients (King & Zeng, 2001), weighted logistic regression with correction to the learned $\boldsymbol{\beta}$ coefficients (Buja et al., 2005), and minimization of the GEV-log loss used in (Wang & Dey, 2010; Calabrese & Osmetti, 2011).[2] In the case of weighted logistic regression, the weights are the inverse empirical class probabilities, i.e. if $\widehat{p}$ is the fraction of positives in the training sample, then the losses on positives are multiplied by $1/\widehat{p}$ and those on negatives by $1/(1-\widehat{p})$. In (Wang & Dey, 2010; Calabrese & Osmetti, 2011), the parameter ξ in the GEV-log loss is incorporated as a variable in the optimization problem, which adds an additional layer of non-convexity. In our experiments with both GEV-log and GEV-canonical losses, we select ξ by validation from the set $\{-1, -0.9, \ldots, 0, 0.1, \ldots, 1.5\} \cup \{-0.2567\}$. This range of ξ values covers a wide range of shapes for the GEV link, and in our experiments, is sufficient to accurately estimate class probabilities for various degrees of rarity.

4.1. Experiments with Synthetic Data

We generated synthetic data in $\mathcal{X} = \mathbb{R}^k$ (for $k = 100$) from three distributions for which $\eta(\mathbf{x})$ is small for most $\mathbf{x} \in \mathcal{X}$, and consequently, p is small. Details of the distributions can be found in Appendix A; for the specific distributions generated, we had $p = 0.0158$, $p = 0.0312$, and $p = 0.095$.

For each of the three distributions, we generated training sets of increasing sizes, and tested the learned CPE models on a test set containing 5000 examples drawn independently from the same distribution. Since we know the true class probability function in this case, we used the *root mean squared error* (RMSE) of the learned model relative to the true class probability function as the performance measure; for a test sample containing n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the RMSE of a CPE model $\widehat{\eta} : \mathcal{X} \to [0, 1]$ is defined as[3]

[2] We also tried under-sampling without any correction, but it consistently gave worse performance than with the correction step, so we do not report the results here.

[3] Again, we overload notation for $\widehat{\eta}$ ($\widehat{\eta}$ was earlier used to denote a number in $[0, 1]$); the usage should be clear from context.

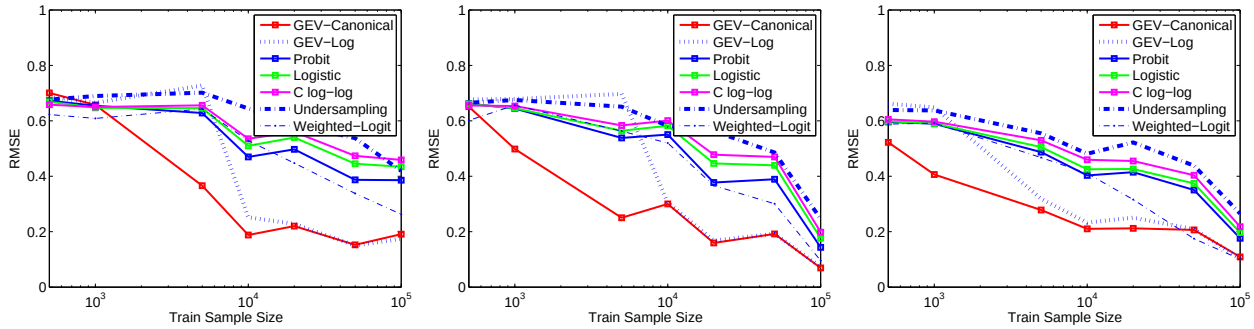

Figure 2. Results on synthetic data generated from various distributions with $p \ll 0.5$. **Left:** 'Sine' distribution with $r = 64$ ($p = 0.0158$). **Middle:** 'Sine' distribution with $r = 32$ ($p = 0.0312$). **Right:** 'Step' distribution ($p = 0.095$). GEV-canonical regression outperforms most baselines, especially for small training sample sizes. See Appendix A for details of the distributions used.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}(\mathbf{x}_i) - \eta(\mathbf{x}_i)\right)^2}.$$

We also used the RMSE on a seperately generated validation set containing 500 examples (drawn from the same distribution as training set) to select the GEV parameter ξ.

The results, averaged over 10 random generations of training samples for each distribution, are shown in Figure 2. As can be seen, the proposed GEV-canonical regression algorithm has better RMSE performance than most baseline algorithms, especially for small training sample sizes.

4.2. Experiments with Real Data

We conducted experiments with 12 real-world data sets from the UCI repository (Frank & Asuncion, 2010). Properties of these data sets are summarized in Table 2. As can be seen, the data sets have varying degrees of rarity: 4 data sets have $p \in (0, 0.1]$; the next 4 data sets have (roughly) $p \in (0.1, 0.25]$, and the remaining 4 data sets have $p \in (0.25, 0.35]$. We randomly split each data set into 70% for training and 30% for testing, and report average results over 10 such random splits. We used L_2-norm regularization in all the algorithms, and for each train/test split, further used 30% of the training set as a validation set to select the GEV parameter ξ and regularization parameter λ; the latter is chosen from the set $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$.

As noted earlier, in the case of real-world data, we do not know the true class probabilities, and therefore we cannot directly evaluate the learned CPE models relative to the true class probability function η. Instead, we use the squared error with respect to the observed binary labels y, more commonly referred to as the *Brier score* in the probability forecasting literature (Brier, 1950), as one performance measure; for a test sample containing n examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, the Brier score of a CPE model $\widehat{\eta} : \mathcal{X} \to [0, 1]$ is defined as

$$\text{BS} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}(\mathbf{x}_i) - \mathbf{1}(y_i = 1)\right)^2.$$

Table 2. Characteristics of UCI data sets used in our experiments.

DATASET	# EXAMPLES	# FEATURES	p
NURSERY	12960	27	.025
LETTER-A	20000	16	.039
CAR	1728	21	.040
GLASS	214	9	.079
ECOLI	336	9	.104
LETTER-VOWEL	20000	16	.194
CMC	1473	24	.226
VEHICLE	846	18	.251
HABERMAN	306	3	.265
YEAST	1484	9	.289
GERMAN	1000	61	.300
PIMA	768	8	.349

We also use the Brier score on the validation set to select select the GEV parameter ξ and regularization parameter λ. We note that Wallace & Dahabreh (2012) recently proposed measuring the Brier score on only the positive examples, without regard to the performance on negative examples; we do not consider this here as our goal is to estimate the full class probability function well.

We also evaluate the performance of all the algorithms in terms of the *calibration loss*, which provides a more fine-grained analysis of the class probability estimates (Hernández-Orallo et al., 2012). Specifically, given n test examples as above, we assign each example to one of 10 bins based on whether its predicted class probability $\widehat{\eta}(\mathbf{x}_i)$ falls in the range $[0, 0.1]$, $(0.1, 0.2]$, $\ldots$, or $(0.9, 1]$. For each of these 10 bins, we count the fraction of examples assigned to that bin that are positive, and treat this as a 'proxy' true class probability for each example $\mathbf{x}_i$ in that bin, denoted $\widetilde{\eta}(\mathbf{x}_i)$. The calibration loss of the CPE model $\widehat{\eta}$ w.r.t. these bins is then defined as

$$\text{CL} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}(\mathbf{x}_i) - \widetilde{\eta}(\mathbf{x}_i)\right)^2.$$

The results in terms of the Brier score and in terms of the calibration loss, both averaged over 10 random train-test splits for each data set, are shown in Table 3 and Table 4,

Table 3. Results on UCI data sets in terms of Brier score. The symbols * and ** against a method indicate that it is statistically significantly different than the best method in the row (using the two-sided Wilcoxon test) with 90% and 95% confidence, respectively.

DATASET	LOGISTIC REGRESSION	PROBIT REGRESSION	CLOGLOG REGRESSION	UNDERSAMPLING + KING-ZENG CORRECTION	WEIGHTED LOGISTIC + CORRECTION	GEV-LOG REGRESSION	GEV-CANONICAL REGRESSION
NURSERY	**0.0084**	**0.0084**	0.0088 **	0.0124 **	0.0090 **	0.0172 **	**0.0084**
LETTER-A	0.0079 **	0.0084 **	**0.0074**	0.0111 **	0.0112 **	0.0313 **	0.0080 **
CAR	0.0266 **	0.0262	0.0267 **	0.0320 **	0.0271 *	0.0296 **	**0.0259**
GLASS	0.0670	0.0671	0.0744 **	0.0623	0.0637	**0.0614**	0.0649
ECOLI	0.0646	0.0644	0.0689 **	0.0756 **	**0.0635**	0.0641	0.0641
LETTER-VOWEL	0.1392 **	0.1392 **	0.1400 **	0.1416 **	0.1414 **	0.1405 **	**0.1367**
CMC	0.1617	0.1617	0.1621	0.1642 **	**0.1615**	0.1626	0.1622
VEHICLE	0.1399	0.1395	0.1422	0.1501 **	0.1408	0.1497 **	**0.1394**
HABERMAN	0.1828 *	0.1812	0.1907 **	0.1823 *	0.1814 **	**0.1761**	0.1769
YEAST	0.1634 **	0.1635 **	0.1666 **	0.1646 **	0.1635 **	0.1621	**0.1616**
GERMAN	0.1721	0.1731	0.1737	0.1754 **	**0.1714**	0.1787 **	0.1727
PIMA	0.1617 **	0.1623 **	0.1652 **	0.1662 *	0.1626 **	0.1616	**0.1603**

Table 4. Results on UCI data sets in terms of calibration loss. The symbols * and ** against a method indicate that it is statistically significantly different than the best method in the row (using the two-sided Wilcoxon test) with 90% and 95% confidence, respectively.

DATASET	LOGISTIC REGRESSION	PROBIT REGRESSION	CLOGLOG REGRESSION	UNDERSAMPLING + KING-ZENG CORRECTION	WEIGHTED LOGISTIC + CORRECTION	GEV-LOG REGRESSION	GEV-CANONICAL REGRESSION
NURSERY	0.0007	**0.0006**	**0.0006**	0.0016 **	0.0009	0.0010 **	0.0008 *
LETTER-A	**0.0005**	0.0006 **	**0.0005**	0.0009 *	0.0006	0.0038 **	0.0006 *
CAR	0.0052 *	0.0049 *	0.0053 **	0.0062	0.0068 **	**0.0033**	0.0037
GLASS	0.0222	0.0235	0.0309 **	**0.0109**	0.0266	0.0252 **	0.0238
ECOLI	0.0230	0.0245	0.0264	0.0260	0.0266	0.0229	**0.0202**
LETTER-VOWEL	0.0059 **	0.0061 **	0.0047 **	0.0090 **	0.0088 **	0.0042	**0.0038**
CMC	0.0056 **	0.0054	0.0063	0.0057	**0.0046**	0.0047	0.0057 *
VEHICLE	0.0112	0.0116	0.0111	0.0195 **	0.0127 **	0.0097	**0.0083**
HABERMAN	0.0295 **	0.0245	0.0331 **	0.0291	0.0309 **	**0.0216**	0.0236
YEAST	0.0083	0.0078	0.0090 **	0.0087 *	0.0091 **	**0.0060**	0.0064
GERMAN	0.0089 **	0.0088	0.0101 **	0.0067	**0.0065**	0.0105	0.0084
PIMA	0.0090	0.0106 **	0.0112 *	0.0116	0.0107 *	0.0091	**0.0081**

respectively. As can be seen, GEV-canonical regression performs well overall compared to other approaches. In particular, even when it is not the best performer itself, it is rarely significantly worse than the best performer, indicating that its performance is generally close to that of the best approach. In comparison to GEV-log loss, the GEV-canonical loss is easier to optimize due to its convexity; in our experiments, the optimization for GEV-canonical regression also converged faster than that for GEV-log regression (see Appendix B for run-time comparisons).

5. Conclusion

The problem of estimating class probabilities from data with binary labels, where one class is rare compared to the other, arises in several applications. We have considered a principled approach to this problem, based on the notion of proper composite losses that have received significant interest recently in the context of class probability estimation problems, and have applied tools from the theory of these losses to construct a flexible parametric family of convex, proper composite losses based on the GEV distribution that can be used to adapt to the degree of rarity in the data. Experiments with the resulting GEV-canonical regression algorithm on both synthetic and real data demonstrate improved class probability estimation performance as compared to a wide variety of baseline algorithms.

Future directions include developing large-scale extensions of our method, and studying statistical convergence rates for GEV-canonical regression in comparison to other CPE algorithms for distributions where one class is rare.

Acknowledgments

This work was supported in part by a Yahoo Labs FREP grant to SA. SA also thanks DST for its support under a Ramanujan Fellowship. AA thanks Google for a travel grant. HN is supported by a Google India PhD Fellowship.

References

Agarwal, Shivani. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *COLT*, 2013.

Boyd, Stephen P and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

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

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

Calabrese, Raffaella and Osmetti, Silvia Angela. Generalized extreme value regression for binary rare events data: an application to credit defaults. *Bulletin of the International Statistical Institute LXII, 58th Session of the International Statistical Institute*, pp. 5631–5634, 2011.

Chawla, Nitesh V., Japkowicz, Nathalie, and Kotcz, Aleksander. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explorations*, 6(1):1–6, 2004.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Czado, Claudia and Santner, Thomas J. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33(2):213–231, 1992.

Drummond, C. and Holte, R.C. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *ICML 2003 Workshop on Learning from Imbalanced Datasets*, 2003.

Elkan, Charles. The foundations of cost-sensitive learning. In *IJCAI*, 2001.

Embrechts, Paul, Klüppelberg, Claudia, and Mikosch, Thomas. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

Frank, A. and Asuncion, A. UCI machine learning repository, 2010. URL http://archive.ics.uci.edu/ml.

Gneiting, Tilmann and Raftery, Adrian E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Green, P. J. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 149–192, 1984.

Guerrero-Curieses, Alicia, Cid-Sueiro, Jesús, Alaiz-Rodríguez, Rocío, and Figueiras-Vidal, Aníbal R. Local estimation of posterior class probabilities to minimize classification errors. *IEEE Transactions on Neural Networks*, 15(2):309–317, 2004.

He, H. and Garcia, E.A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284, 2009.

Hendrickson, A. D. and Buehler, R. J. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42: 1916–1921, 1971.

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

Japkowicz, N. The class imbalance problem: Significance and strategies. In *ICAI*, 2000.

King, G. and Zeng, L. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Kotz, S. and Nadarajah, S. (eds.). *Extreme Value Distributions: Theory and Applications*. Imperial College Press, London, 2000.

Lee, Kuang-chih, Orten, Burkay, Dasdan, Ali, and Li, Wentong. Estimating conversion rate in display advertising from past performance data. In *SIGKDD*, 2012.

Masnadi-Shirazi, H. and Vasconcelos, N. Risk minimization, probability elicitation, and cost-sensitive SVMs. In *ICML*, 2010.

Provost, F. Machine learning from imbalanced data sets 101. In *AAAI-2000 Workshop on Imbalanced Data Sets*, 2000.

Reid, Mark D. and Williamson, Robert C. Surrogate regret bounds for proper losses. In *ICML*, 2009.

Reid, Mark D. and Williamson, Robert C. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Savage, Leonard J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66 (336):783–801, 1971.

Schervish, M. J. A general method for comparing probability assessors. *The Annals of Statistics*, 17:1856–1879, 1989.

Van Hulse, J., Khoshgoftaar, T.M., and Napolitano, A. Experimental perspectives on learning from imbalanced data. In *ICML*, 2007.

Wallace, B. C. and Dahabreh, I. J. Class probability estimates are unreliable for imbalanced data (and how to fix them). In *ICDM*, 2012.

Wang, Xia and Dey, Dipak K. Generalized extreme value regression for binary response data: an application to b2b electronic payments system adoption. *The Annals of Applied Statistics*, 4 (4):2000–2023, 2010.

Zadrozny, Bianca, Langford, John, and Abe, Naoki. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, 2003.

Zhang, Tong. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004.

Supplementary Material

A. Distributions for Synthetic Data

The first two distributions, which we call 'sine' distributions, can be described as follows (we used two different settings for the 'rarity' parameter r):

1. Fix $\boldsymbol{\mu} \in \mathbb{R}^k$ and a symmetric positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$. (In our experiments, $\boldsymbol{\mu}$ was sampled from a k-variate standard normal distribution, and $\boldsymbol{\Sigma}$ was sampled from a Wishart distribution with parameters $(\mathbf{I}, k)$, where $\mathbf{I}$ is the $k \times k$ identity matrix.)
2. Fix $\boldsymbol{\beta}^* \in \mathbb{R}^{k+1}$. (In our experiments, $\boldsymbol{\beta}^*$ was sampled from a $(k+1)$-variate standard normal distribution.)
3. The distribution D on $\mathcal{X} \times \{\pm 1\}$ is defined as follows: To generate an example $(\mathbf{x}, y)$, sample $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\widetilde{\mathbf{x}} = [\mathbf{x}\,; 1]$; then sample $y \in \{\pm 1\}$ according to the conditional probability
$$\eta(\mathbf{x}) = \frac{1}{r}\left(\sin\left(\frac{\pi}{4}(\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{x}}\right) + 1\right),$$
where $r > 0$ is a 'rarity' parameter that controls p: the higher the value of r, the more rare the positive class.

In our experiments, we used two 'sine' distributions with rarity parameters $r = 64$ (which for the specific distribution generated yielded $p = 0.0158$) and $r = 32$ (which for the specific distribution generated yielded $p = 0.0312$).

The third distribution, which we call the 'step' distribution, was defined similarly; the only difference was in the form of the class probability function:

1. Same as for 'sine' distribution above.
2. Same as for 'sine' distribution above.
3. The distribution D on $\mathcal{X} \times \{\pm 1\}$ is defined as follows: To generate an example $(\mathbf{x}, y)$, sample $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\widetilde{\mathbf{x}} = [\mathbf{x}\,; 1]$; then sample $y \in \{\pm 1\}$ according to the conditional probability
$$\eta(\mathbf{x}) = \begin{cases} 0.10 & \text{if } (\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{x}} < -30 \\ 0.03 & \text{if } -30 \le (\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{x}} \le 30 \\ 0.13 & \text{if } (\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{x}} > 30. \end{cases}$$

The 'step' distribution in our experiments had $p = 0.095$.

B. Run-Time Comparisons for Real Data

Table 5 shows the time (in seconds; rounded off to the nearest integer) it takes for the GEV-log and the GEV-canonical method to run for the data sets listed in Table 2. This includes the time for training as well as for validation of parameters. The results were averaged over 10 runs.

Table 5. Average training time (in seconds, including validation time) for GEV-log & GEV-canonical regression on UCI data sets.

DATASET	GEV-LOG	GEV-CANONICAL
NURSERY	21947	7242
LETTER-A	84037	39138
CAR	197	40
GLASS	7	1
ECOLI	7	2
LETTER-VOWEL	72628	6503
CMC	179	19
VEHICLE	39	5
HABERMAN	6	1
YEAST	200	27
GERMAN	114	13
PIMA	20	4

We observed that the optimization for GEV-canonical loss converges faster than for GEV-log on all the data sets. This is likely due to the Hessian of the GEV-canonical objective being better conditioned (the eigenvalues of the Hessian in this case are easily bounded, which is not the case for the Hessian of the GEV-log objective); the Newton method converges faster when the Hessian of the objective is well conditioned (Boyd & Vandenberghe, 2004).