JMLR: Workshop and Conference Proceedings 45:221–236, 2015 ACML 2015

Class-prior Estimation for Learning from Positive and Unlabeled Data

Marthinus Christoffel du Plessis CHRISTO@MS.K.U-TOKYO.AC.JP
Gang Niu GANG@MS.K.U-TOKYO.AC.JP
Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
The University of Tokyo, Tokyo, 113-0033, Japan.

Editor: Geoffrey Holmes and Tie-Yan Liu

Abstract

We consider the problem of estimating the *class prior* in an unlabeled dataset. Under the assumption that an additional labeled dataset is available, the class prior can be estimated by fitting a mixture of class-wise data distributions to the unlabeled data distribution. However, in practice, such an additional labeled dataset is often not available. In this paper, we show that, with additional samples coming only from the positive class, the class prior of the unlabeled dataset can be estimated correctly. Our key idea is to use properly penalized divergences for model fitting to cancel the error caused by the absence of negative samples. We further show that the use of the penalized L_1-distance gives a computationally efficient algorithm with an analytic solution, and establish its uniform deviation bound and estimation error bound. Finally, we experimentally demonstrate the usefulness of the proposed method.

Keywords: Class-prior estimation, positive and unlabeled data

1. Introduction

Suppose that we have two datasets $\mathcal{X}$ and $\mathcal{X}'$, which are i.i.d. samples from probability distributions with density $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x})$, respectively:

$$\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}|y=1), \quad \mathcal{X}' = \{\boldsymbol{x}'_j\}_{j=1}^{n'} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}).$$

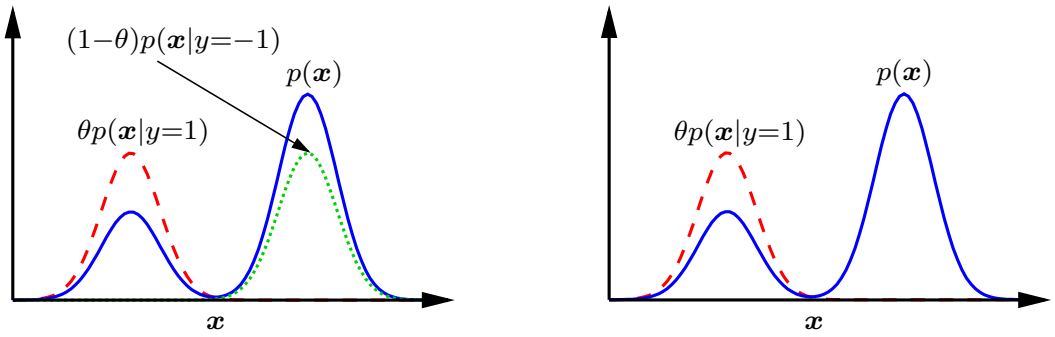
That is, $\mathcal{X}$ is a set of samples from the positive class and $\mathcal{X}'$ is a set of unlabeled samples (consisting of both the positive and negative samples). Our goal is to estimate the class prior

$$\pi = p(y=1),$$

in the unlabeled dataset $\mathcal{X}'$. Estimation of the class prior from positive and unlabeled data is of great practical importance, since it allows a classifier to be trained only from these datasets (Scott and Blanchard, 2009; du Plessis et al., 2014), in the absence of negative data.

If a mixture of class-wise input data densities,

$$q'(\boldsymbol{x};\theta) = \theta p(\boldsymbol{x}|y=1) + (1-\theta)p(\boldsymbol{x}|y=-1),$$



(a) Full matching with $q(\boldsymbol{x};\theta) = \theta p(\boldsymbol{x}|y=1) + (1-\theta)p(\boldsymbol{x}|y=-1)$

(b) Partial matching with $q(\boldsymbol{x};\theta) = \theta p(\boldsymbol{x}|y=1)$

Figure 1: Class-prior estimation by matching model $q(\boldsymbol{x};\theta)$ to unlabeled input data density $p(\boldsymbol{x})$.

is fitted to the unlabeled input data density $p(\boldsymbol{x})$, the true class prior π can be obtained (Saerens et al., 2002; du Plessis and Sugiyama, 2012), as illustrated in Figure 1(a). In practice, fitting may be performed under the f-divergence (Ali and Silvey, 1966; Csiszár, 1967):

$$\theta := \operatorname*{arg\,min}_{0\leq\theta\leq 1} \int f\left(\frac{q'(\boldsymbol{x};\theta)}{p(\boldsymbol{x})}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{1}$$

where $f(t)$ is a convex function with $f(1) = 0$. So far, class-prior estimation methods based on the Kullback-Leibler divergence (Saerens et al., 2002), and the Pearson divergence (du Plessis and Sugiyama, 2012) have been developed (Table 1). Additionally, class-prior estimation has been performed by L_2-distance minimization (Sugiyama et al., 2012).

However, since these methods require labeled samples from both positive and negative classes, they cannot be directly employed in the current setup. To cope with problem, a partial model,

$$q(\boldsymbol{x};\theta) = \theta p(\boldsymbol{x}|y=1),$$

was used in Elkan and Noto (2008) and du Plessis and Sugiyama (2014) to estimate the class prior in the absence of negative samples (Figure 1(b)):

$$\theta := \operatorname*{arg\,min}_{0\leq\theta\leq 1} \mathrm{Div}_f(\theta), \tag{2}$$

where

$$\mathrm{Div}_f(\theta) := \int f\left(\frac{q(\boldsymbol{x};\theta)}{p(\boldsymbol{x})}\right) p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

In this paper, we first show that the above partial matching approach consistently overestimates the true class prior. We then show that, by appropriately penalizing f-divergences, the class prior can be correctly obtained. We further show that the use of the

Table 1: Common f-divergences. $f^*(z)$ is the conjugate of $f(t)$ and $\tilde{f}^*(z)$ is the conjugate of the penalized function $\tilde{f}(t) = f(t)$ for $0 \leq t \leq 1$ and ∞ otherwise.

Divergence	Function $f(t)$	Conjugate $f^*(z)$	Penalized Conjugate $\tilde{f}^*(z)$
Kullback-Leibler divergence (Kullback and Leibler, 1951)	$-\log(t)$	$-\log(-z)-1$	$\begin{cases} -1-\log(-z) & z \leq -1 \\ z & z > -1 \end{cases}$
Pearson divergence (Pearson, 1900)	$\frac{1}{2}(t-1)^2$	$\frac{1}{2}z^2+z$	$\begin{cases} -\frac{1}{2} & z < -1 \\ \frac{1}{2}z^2+z & -1 \leq z \leq 0 \\ z & z > 0 \end{cases}$
L_1-distance	$\lvert t-1 \rvert$	$\begin{cases} z & -1 \leq z \leq 1 \\ \infty & \text{otherwise} \end{cases}$	$\max(z,-1)$

penalized L_1-distance drastically simplifies the estimation procedure, resulting in an analytic estimator that can be computed efficiently. We also establish a uniform deviation bound and an estimation error bound for the penalized L_1-distance estimator. Finally, through experiments, we demonstrate the usefulness of the proposed method in classification from positive and unlabeled data.

2. Class-prior estimation via penalized f-divergences

First, we investigate the behavior of the partial matching method (2), which can be regarded as an extension of the existing analysis for the Pearson divergence (du Plessis and Sugiyama, 2014) to more general divergences.

We show that naively using general divergences may result in an overestimate of the class prior, and show how this situation can be avoided by penalization.

2.1. Over-estimation of the class prior

For f-divergences, we focus on $f(t)$ such that its minimum is attained at $t \geq 1$. We also assume that it is differentiable and the derivative of $f(t)$ is $\partial f(t) < 0$, when $t < 1$, and $\partial f(t) \leq 0$ when $t = 1$. This condition is satisfied for divergences such as the Kullback-Leibler divergence and the Pearson divergence. Because of the divergence matching formulation, we expect that the objective function (2) is minimized at $\theta = \pi$. That is, based on the first-order optimality condition, we expect that the derivative of $\mathrm{Div}_f(\theta)$ w.r.t. θ, given by

$$\partial \mathrm{Div}_f(\theta) = \int \partial f\left(\frac{\theta p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x},$$

satisfies $\partial \mathrm{Div}_f(\pi) = 0$. Since

$$\frac{\pi p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})} = p(y=1|\boldsymbol{x}) \leq 1 \quad \Longrightarrow \quad \partial f\left(\frac{\pi p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x})}\right) \leq 0,$$

we have

$$\partial \mathrm{Div}_f(\pi) = \int \underbrace{\partial f\Big(p(y=1|\boldsymbol{x})\Big)}_{\leq 0} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} \leq 0.$$

The domain of the above integral, where $p(\boldsymbol{x}|y=1) > 0$, can be expressed as:

$$\mathcal{D}_1 = \{\boldsymbol{x} : p(y=1|\boldsymbol{x}) = 1 \wedge p(\boldsymbol{x}|y=1) > 0\}, \mathcal{D}_2 = \{\boldsymbol{x} : p(y=1|\boldsymbol{x}) < 1 \wedge p(\boldsymbol{x}|y=1) > 0\}.$$

The derivative is then expressed as

$$\partial \mathrm{Div}_f(\pi) = \int_{\mathcal{D}_1} \underbrace{\partial f(p(y=1|\boldsymbol{x}))}_{\leq 0} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} + \int_{\mathcal{D}_2} \underbrace{\partial f(p(y=1|\boldsymbol{x}))}_{<0} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}. \quad (3)$$

The posterior is $p(y=1|\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=1)/p(\boldsymbol{x})$, where $p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=1) + (1-\pi)p(\boldsymbol{x}|y=-1)$. $\mathcal{D}_1$ is the part of the domain where the two classes do not overlap, because $p(y=1|\boldsymbol{x}) = 1$ implies that $\pi p(\boldsymbol{x}|y=1) = p(\boldsymbol{x})$ and $(1-\pi)p(\boldsymbol{x}|y=-1) = 0$. Conversely, $\mathcal{D}_2$ is the part of the domain where the classes overlap because $p(y=1|\boldsymbol{x}) < 1$ implies that $\pi p(\boldsymbol{x}|y=1) < p(\boldsymbol{x})$, and $(1-\pi)p(\boldsymbol{x}|y=-1) > 0$.

Since the first term in (3) is non-positive, the derivative can be zero only if $\mathcal{D}_2$ is empty (i.e., there is no class overlap). If $\mathcal{D}_2$ is not empty (i.e., there is class overlap) the derivative will be negative. Since the objective function $\mathrm{Div}_f(\theta)$ is convex, the derivative $\partial \mathrm{Div}_f(\theta)$ is a monotone non-decreasing function. Therefore, if the function $\mathrm{Div}_f(\theta)$ has a minimizer, it will be larger than the true class prior π.

2.2. Partial distribution matching via penalized f-divergences

In this section, we consider a function

$$f(t) = \begin{cases} -(t-1) & t < 1, \\ c(t-1) & t \geq 1. \end{cases}$$

This function coincides with the L_1 distance when $c = 1$. The analysis here is slightly more involved, since the subderivative should be taken at $t = 1$. This gives the following:

$$\partial \mathrm{Div}_f(\pi) = \int_{\mathcal{D}_1} \partial f(1) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} + \int_{\mathcal{D}_2} \underbrace{\partial f(p(y=1|\boldsymbol{x}))}_{<0} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}, \quad (4)$$

where the subderivative at $t = 1$ is $\partial f(1) = [-1, c]$ and the derivative for $t < 1$ is $\partial f(t) = -1$. We can therefore write the subderivative of the first term in (4) as

$$\int_{\mathcal{D}_1} \partial f(1) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} = \left[-\int_{\mathcal{D}_1} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}, c\int_{\mathcal{D}_1} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} \right].$$

The derivative for the second term in (4) is

$$\int_{\mathcal{D}_2} \partial f(p(y=1|\boldsymbol{x}) p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} = -\int_{\mathcal{D}_2} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}.$$

To achieve a minimum at π, we should have

$$0 \in \left[-\int_{\mathcal{D}_1} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} - \int_{\mathcal{D}_2} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}, c\int_{\mathcal{D}_1} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} - \int_{\mathcal{D}_2} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} \right].$$

However, depending on $\mathcal{D}_1$ and $\mathcal{D}_2$, we may have

$$c\int_{\mathcal{D}_1} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} - \int_{\mathcal{D}_2} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} < 0,$$

which means that $0 \notin \partial \mathrm{Div}_f(\pi)$ The solution is to take $c = \infty$ to ensure that $0 \in \mathrm{Div}_f(\pi)$ is always satisfied.

Using the same reasoning as above, we can also rectify the overestimation problem for other divergences specified by $f(t)$, by replacing $f(t)$ with a penalized function $\widetilde{f}(t)$:

$$\widetilde{f}(t) = \begin{cases} f(t) & 0 \le t \le 1, \\ \infty & \text{otherwise.} \end{cases}$$

In the next section, estimation of penalized f-divergences is discussed.

2.3. Direct evaluation of penalized f-divergences

Here, we show how distribution matching can be performed without density estimation under penalized f-divergences. We use the *Fenchel duality bounding technique* for f-divergences (Keziou, 2003), which is based on *Fenchel's inequality*:

$$f(t) \ge tz - f^*(z), \tag{5}$$

where $f^*(z)$ is the *Fenchel dual* or *convex conjugate* defined as

$$f^*(z) = \sup_{t'} t'z - f(t').$$

Applying the bound (5) in a pointwise manner, we obtain

$$f\left(\frac{\theta p(\boldsymbol{x} \mid y=1)}{p(\boldsymbol{x})}\right) \ge r(\boldsymbol{x})\left(\frac{\theta p(\boldsymbol{x} \mid y=1)}{p(\boldsymbol{x})}\right) - f^*(r(\boldsymbol{x})),$$

where $r(\boldsymbol{x})$ fulfills the role of z in (5). Multiplying both sides with $p(\boldsymbol{x})$ gives

$$f\left(\frac{\theta p(\boldsymbol{x} \mid y=1)}{p(\boldsymbol{x})}\right) p(\boldsymbol{x}) \ge \theta r(\boldsymbol{x}) p(\boldsymbol{x} \mid y=1) - f^*(r(\boldsymbol{x}))p(\boldsymbol{x}). \tag{6}$$

Integrating and then selecting the tightest bound gives

$$\mathrm{Div}_f(p\|q) \ge \sup_r \theta \int r(\boldsymbol{x})p(\boldsymbol{x}\,|\,y=1)\mathrm{d}\boldsymbol{x} - \int f^*(r(\boldsymbol{x}))p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{7}$$

Replacing expectations with sample averages gives

$$\widehat{\mathrm{Div}}_f(p\|q) \ge \sup_r \theta \frac{1}{n}\sum_{i=1}^{n} r(\boldsymbol{x}_i) - \frac{1}{n'}\sum_{j=1}^{n'} f^*(r(\boldsymbol{x}'_j)), \tag{8}$$

where $\widehat{\mathrm{Div}}_f(p\|q)$ denotes $\mathrm{Div}_f(p\|q)$ estimated from sample averages. Note that the conjugate $f^*(z)$ of any function $f(t)$ is convex. Therefore, if $r(\boldsymbol{x})$ is linear in parameters, the above maximization problem is convex and thus can be easily solved.

The conjugates for selected penalized f-divergences are given in Table 1.

2.4. Penalized L_1-distance estimation

Here we focus on penalized L_1-distance $f(t) = |t-1|$ as a specific example of penalized f-divergences, and show that an analytic and computationally efficient solution can be obtained.

The conjugate for the penalized L_1-distance is $\tilde{f}^*(z) = \max(z, -1)$. Then we can see that the lower-bound in (7) will be met with equality when

$$r(\boldsymbol{x}) = \begin{cases} -1 & \theta p(\boldsymbol{x}|y=1) \leq p(\boldsymbol{x}), \\ \infty & \text{otherwise.} \end{cases} \tag{9}$$

For this reason, we use the following linear model as $r(\boldsymbol{x})$:

$$r(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}) - 1, \tag{10}$$

where $\{\varphi_\ell(\boldsymbol{x})\}_{\ell=1}^b$ is the set of non-negative basis functions[1]. Then the empirical estimate in the right-hand side of (8) can be expressed as

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_b) = \underset{(\alpha_1,\ldots,\alpha_b)}{\arg\min} \frac{1}{n'} \sum_{j=1}^{n'} \max\left(\sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}'_j), 0\right) - \frac{\theta}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}_i) + \theta + \frac{\lambda}{2} \sum_{\ell=1}^{b} \alpha_\ell^2, \tag{11}$$

where the regularization term $\frac{\lambda}{2}\sum_{\ell=1}^{b} \alpha_\ell^2$ for $\lambda > 0$ is included to avoid overfitting.

The optimal value for the lower-bound (7) occurs when $r(\boldsymbol{x}) \geq -1$ (see (9)). This fact can be incorporated by constraining the parameters of the model (10), so that $\alpha_\ell \geq 0, \forall \ell = 1, \ldots, b$. If the basis functions, $\varphi_\ell(\boldsymbol{x}), \ell = 1, \ldots, b$, are non-negative, the term inside the max in (11) is always positive and the max operation becomes superfluous. This allows us to obtain the parameter vector $(\alpha_1, \ldots, \alpha_b)$ as the solution to the following constrained optimization problem:

$$\begin{aligned} (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_b) = \underset{(\alpha_1,\ldots,\alpha_b)}{\arg\min} \quad & \sum_{\ell=1}^{b} \frac{\lambda}{2} \alpha_\ell^2 - \sum_{\ell=1}^{b} \alpha_\ell \beta_\ell, \\ \text{s.t.} \quad & \alpha_\ell \geq 0, \qquad \ell = 1, \ldots, b, \end{aligned}$$

where

$$\beta_\ell = \frac{\theta}{n} \sum_{i=1}^{n} \varphi_\ell(\boldsymbol{x}_i) - \frac{1}{n'} \sum_{j=1}^{n'} \varphi_\ell(\boldsymbol{x}'_j).$$

The above optimization problem decouples for all α_ℓ values and can be solved separately as

$$\widehat{\alpha}_\ell = \frac{1}{\lambda} \max(0, \beta_\ell).$$

1. In practice, we use Gaussian kernels centered at all sample points as the basis functions: $\varphi_\ell(\boldsymbol{x}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{c}_\ell\|^2/(2\sigma^2)\right)$, where $\sigma > 0$, and $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n, \boldsymbol{c}_{n+1}, \ldots, \boldsymbol{c}_{n+n'}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots \boldsymbol{x}'_{n'})$

Since $\widehat{\boldsymbol{\alpha}}$ can be just calculated with a max operation, the above solution is extremely fast to calculate. All hyper-parameters including the Gaussian width σ and the regularization parameter λ are selected for each θ via straightforward cross-validation.

Finally, our estimate of the penalized L_1-distance (i.e., the maximizer of the empirical estimate in the right-hand side of (8)) is obtained as

$$\widehat{\mathrm{penL}}_1(\theta) = \frac{1}{\lambda}\sum_{\ell=1}^{b} \max\left(0, \beta_\ell\right)\beta_\ell - \theta + 1.$$

The class prior is then selected so as to minimize the above estimator.

3. Stability analysis

Regarding the estimation stability of $\widehat{\mathrm{penL}}_1(\theta)$ for fixed θ, we have the following deviation bound. Without loss of generality, we assume that the basis functions are upper bounded by one, i.e., $\forall \boldsymbol{x}, \varphi_\ell(\boldsymbol{x}) \leq 1$ for $\ell = 1, \ldots, b$.

Theorem 1 (Deviation bound) *Fix θ, then, for any $0 < \delta < 1$, with probability at least $1-\delta$ over the repeated sampling of $\mathcal{D} = \mathcal{X} \cup \mathcal{X}'$ for estimating $\widehat{\mathrm{penL}}_1(\theta; \mathcal{D})$, we have*

$$\left|\widehat{\mathrm{penL}}_1(\theta; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta; \mathcal{D})]\right| \leq \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2 + \frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

Proof 1 *We prove the theorem based on a technique known as* the method of bounded difference. *Let $f_\ell(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'}) = \max\left(0, \beta_\ell\right)\beta_\ell$, so that*

$$\widehat{\mathrm{penL}}_1(\theta; \mathcal{D}) = \frac{1}{\lambda}\sum_{\ell=1}^{b} f_\ell(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'}) - \theta + 1.$$

Next, we replace $\boldsymbol{x}_i$ with $\bar{\boldsymbol{x}}_i$ and bound the difference between $f_\ell(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$ and $f_\ell(\boldsymbol{x}_1, \ldots, \bar{\boldsymbol{x}}_i, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$. Let $t = \varphi_\ell(\boldsymbol{x}_i)$, $t' = \varphi_\ell(\bar{\boldsymbol{x}}_i)$, and $\xi_\ell = \beta_\ell - t$. Since the basis functions are bounded, we know that $0 \leq t, t' \leq 1$, as well as $-1 \leq \xi_\ell \leq \theta$. Then the maximum difference is,

$$c_\ell = \sup_{0\leq t\leq 1, 0\leq t'\leq 1} \max\left(0, \frac{\theta}{n}t' - \xi_\ell\right)\left(\frac{\theta}{n}t' - \xi_\ell\right) - \max\left(0, \frac{\theta}{n}t - \xi_\ell\right)\left(\frac{\theta}{n}t - \xi_\ell\right).$$

By analyzing the above for different cases where the constraints are active, the the maximum difference for replacing $\boldsymbol{x}_i$ with is $\bar{\boldsymbol{x}}_i$ is $(\theta/n)(2+\theta/n)$.

We can use the same argument for replacing $\boldsymbol{x}'_j$ with $\bar{\boldsymbol{x}}'_j$ in $f_\ell(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{n'})$, resulting in a maximum difference of $(1/n)(2+1/n)$. Note that this holds for all $f_\ell(\cdot)$ simultaneously, and thus the change of $\widehat{\mathrm{penL}}_1(\theta; \mathcal{D})$ is no more than $(b/\lambda)(\theta/n)(2+\theta/n)$ if $\boldsymbol{x}_i$ is replaced with $\bar{\boldsymbol{x}}_i$ or $(b/\lambda)(1/n)(2+1/n)$ if $\boldsymbol{x}'_j$ is replaced with $\bar{\boldsymbol{x}}'_j$. We can therefore apply McDiarmid's inequality *to obtain, with probability at least $1-\delta/2$,*

$$\widehat{\mathrm{penL}}_1(\theta; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[\widehat{\mathrm{penL}}_1(\theta; \mathcal{D})\right] \leq \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2 + \frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

Applying McDiarmid's inequality again for $\mathbb{E}_{\mathcal{D}}\left[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})\right] - \widehat{\mathrm{penL}}_1(\theta;\mathcal{D})$ *proves the result.* ■

Theorem 1 shows that the deviation from our estimate to its expectation is small with high probability. Nevertheless, θ must be fixed before seeing the data, so we cannot use the estimate to choose θ.

This motivates us to derive a uniform deviation bound. Let us define the constants

$$C_x = \sup_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left(\sum_{\ell=1}^{b}\varphi_\ell^2(\boldsymbol{x})\right)^{1/2} \leq \sqrt{b},$$
$$C_\alpha = \sup_{0\leq\theta\leq 1}\sup_{\mathcal{X}\sim p^n(\boldsymbol{x}|y=1),\mathcal{X}'\sim p^{n'}(\boldsymbol{x})}\left(\sum_{\ell=1}^{b}\widehat{\alpha}_\ell^2(\theta,\mathcal{D})\right)^{1/2},$$

where we write $\widehat{\alpha}_\ell(\theta,\mathcal{D})$ to emphasize that $\widehat{\alpha}_\ell$ depends on θ and $\mathcal{D}$.

Theorem 2 (Uniform deviation bound) *For any* $0<\delta<1$*, with probability at least* $1-\delta$ *over the repeated sampling of* $\mathcal{D}$*, the following holds for all* $0\leq\theta\leq 1$*,*

$$\left|\widehat{\mathrm{penL}}_1(\theta;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})]\right| \leq \left(\frac{2}{\sqrt{n}}+\frac{2}{\sqrt{n'}}\right)C_\alpha C_x + \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2+\frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

Proof 2 *We denote* $g(\theta;\mathcal{D}) = \sum_{\ell=1}^{b}\widehat{\alpha}_\ell\beta_\ell$*, and* $g(\theta) = \mathbb{E}_{\mathcal{D}}[g(\theta;\mathcal{D})]$ *so that*

$$\widehat{\mathrm{penL}}_1(\theta;\mathcal{D}) = g(\theta;\mathcal{D}) - \theta + 1, \quad \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})] = g(\theta) - \theta + 1,$$

and

$$\widehat{\mathrm{penL}}_1(\theta;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})] = g(\theta;\mathcal{D}) - g(\theta).$$

Step 1: *We first consider one direction. By definition,* $\forall\theta$*,*

$$g(\theta;\mathcal{D}) - g(\theta) \leq \sup_\theta\left\{g(\theta;\mathcal{D}) - g(\theta)\right\}.$$

We cannot simply apply Theorem 1 to bound the right-hand side since θ *in the right-hand side is not fixed. Nevertheless, according to the proof of Theorem 1, if we replace a single point* $\boldsymbol{x}_i$ *or* $\boldsymbol{x}'_j$ *in* $\mathcal{D}$*, the change of* $\sup_\theta\{g(\theta;\mathcal{D}) - g(\theta)\}$ *is also bounded by* $\frac{b}{\lambda n}(2+\frac{1}{n})$ *or* $\frac{b}{\lambda n'}(2+\frac{1}{n'})$*. We then have, with probability at least* $1-\delta/2$*,*

$$\sup_\theta\{g(\theta;\mathcal{D})-g(\theta)\} \leq \mathbb{E}_{\mathcal{D}}\left[\sup_\theta\{g(\theta;\mathcal{D})-g(\theta)\}\right] + \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2+\frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

Step 2: *Next we bound* $\mathbb{E}_{\mathcal{D}}[\sup_\theta\{g(\theta;\mathcal{D})-g(\theta)\}]$ *based on a technique known as symmetrization. Note that the function* $g(\theta;\mathcal{D})=\sum_{\ell=1}^{b}\widehat{\alpha}_\ell\beta_\ell$ *can be rewritten in a point-wise manner other than a base-wise manner,*

$$\begin{aligned} g(\theta;\mathcal{D}) &= \sum_{i=1}^{n}\sum_{\ell=1}^{b}\left(\frac{\widehat{\alpha}_\ell\theta}{n}\right)\varphi_\ell(\boldsymbol{x}_i)-\sum_{j=1}^{n'}\sum_{\ell=1}^{b}\left(\frac{\widehat{\alpha}_\ell}{n'}\right)\varphi_\ell(\boldsymbol{x}'_j)\\ &= \sum_{i=1}^{n}\omega(\boldsymbol{x}_i)-\sum_{j=1}^{n'}\omega'(\boldsymbol{x}'_j), \end{aligned}$$

where for simplicity we define

$$\omega(\boldsymbol{x})=\sum_{\ell=1}^{b}\left(\frac{\widehat{\alpha}_\ell\theta}{n}\right)\varphi_\ell(\boldsymbol{x}),\quad \textit{and},\quad \omega'(\boldsymbol{x})=\sum_{\ell=1}^{b}\left(\frac{\widehat{\alpha}_\ell}{n'}\right)\varphi_\ell(\boldsymbol{x}).$$

Let $\mathcal{D}'=\{\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_n,\bar{\boldsymbol{x}}'_1,\ldots,\bar{\boldsymbol{x}}'_{n'}\}$ *be a ghost sample,*

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}\left[\sup_\theta\{g(\theta;\mathcal{D})-g(\theta)\}\right] &= \mathbb{E}_{\mathcal{D}}\left[\sup_\theta\{g(\theta;\mathcal{D})-\mathbb{E}_{\mathcal{D}'}[g(\theta;\mathcal{D}')]\}\right],\\ &= \mathbb{E}_{\mathcal{D}}\left[\sup_\theta\{\mathbb{E}_{\mathcal{D}'}[g(\theta;\mathcal{D})-g(\theta;\mathcal{D}')]\}\right],\\ &\le \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_\theta\{g(\theta;\mathcal{D})-g(\theta;\mathcal{D}')\}\right], \end{aligned}$$

where we apply Jensen's inequality with the fact that the supremum is a convex function. Moreover, let $\boldsymbol{\sigma}=\{\sigma_1,\ldots,\sigma_n,\sigma'_1,\ldots,\sigma'_{n'}\}$ *be a set of Rademacher variables of size* $n+n'$*,*

$$\begin{aligned} &\mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_\theta\{g(\theta;\mathcal{D})-g(\theta;\mathcal{D}')\}\right]\\ &\quad= \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_\theta\left\{\left(\sum_{i=1}^{n}\omega(\boldsymbol{x}_i)-\sum_{j=1}^{n'}\omega'(\boldsymbol{x}'_j)\right)-\left(\sum_{i=1}^{n}\omega(\bar{\boldsymbol{x}}_i)-\sum_{j=1}^{n'}\omega'(\bar{\boldsymbol{x}}'_j)\right)\right\}\right]\\ &\quad= \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_\theta\left\{\sum_{i=1}^{n}(\omega(\boldsymbol{x}_i)-\omega(\bar{\boldsymbol{x}}_i))-\sum_{j=1}^{n'}(\omega'(\boldsymbol{x}'_j)-\omega'(\bar{\boldsymbol{x}}'_j))\right\}\right]\\ &\quad= \mathbb{E}_{\boldsymbol{\sigma},\mathcal{D},\mathcal{D}'}\left[\sup_\theta\left\{\sum_{i=1}^{n}\sigma_i(\omega(\boldsymbol{x}_i)-\omega(\bar{\boldsymbol{x}}_i))-\sum_{j=1}^{n'}\sigma'_j(\omega'(\boldsymbol{x}'_j)-\omega'(\bar{\boldsymbol{x}}'_j))\right\}\right], \end{aligned}$$

since the original and ghost samples are symmetric and each $(\omega(\boldsymbol{x}_i)-\omega(\bar{\boldsymbol{x}}_i))$ *shares the same distribution with* $\sigma_i(\omega(\boldsymbol{x}_i)-\omega(\bar{\boldsymbol{x}}_i))$ *and each* $(\omega'(\boldsymbol{x}'_j)-\omega'(\bar{\boldsymbol{x}}'_j))$ *shares the same distribution*

with $\sigma'_j(\omega'(\boldsymbol{x}'_j) - \omega'(\bar{\boldsymbol{x}}'_j))$. *Subsequently,*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma},\mathcal{D},\mathcal{D}'}\left[\sup_\theta\left\{\left(\sum_{i=1}^n \sigma_i\omega(\boldsymbol{x}_i) - \sum_{j=1}^{n'}\sigma'_j\omega'(\boldsymbol{x}'_j)\right) + \left(\sum_{i=1}^n(-\sigma_i)\omega(\bar{\boldsymbol{x}}_i) - \sum_{j=1}^{n'}(-\sigma'_j)\omega'(\bar{\boldsymbol{x}}'_j)\right)\right\}\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},\mathcal{D}}\left[\sup_\theta\left\{\sum_{i=1}^n \sigma_i\omega(\boldsymbol{x}_i) - \sum_{j=1}^{n'}\sigma'_j\omega'(\boldsymbol{x}'_j)\right\}\right] + \mathbb{E}_{\boldsymbol{\sigma},\mathcal{D}'}\left[\sup_\theta\left\{\sum_{i=1}^n(-\sigma_i)\omega(\bar{\boldsymbol{x}}_i) - \sum_{j=1}^{n'}(-\sigma'_j)\omega'(\bar{\boldsymbol{x}}'_j)\right\}\right] \\
&= 2\mathbb{E}_{\boldsymbol{\sigma},\mathcal{D}}\left[\sup_\theta\left\{\sum_{i=1}^n \sigma_i\omega(\boldsymbol{x}_i) - \sum_{j=1}^{n'}\sigma'_j\omega'(\boldsymbol{x}'_j)\right\}\right],
\end{aligned}$$

where we first apply the triangle inequality, and then make use of that the original and ghost samples have the same distribution and all Rademacher variables have the same distribution.

Step 3: *The Rademacher complexity still remains to be bound. To this end, we decompose the Rademacher complexity into two,*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_\theta\left\{\sum_{i=1}^n \sigma_i\omega(\boldsymbol{x}_i) - \sum_{j=1}^{n'}\sigma'_j\omega'(\boldsymbol{x}'_j)\right\}\right] \\
&\quad= \mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_\theta\left\{\sum_{i=1}^n \sigma_i\sum_{\ell=1}^b\left(\frac{\widehat{\alpha}_\ell\theta}{n}\right)\varphi_\ell(\boldsymbol{x}_i) - \sum_{j=1}^{n'}\sigma'_j\sum_{\ell=1}^b\left(\frac{\widehat{\alpha}_\ell}{n'}\right)\varphi_\ell(\boldsymbol{x}'_j)\right\}\right] \\
&\quad\leq \frac{1}{n}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_\theta\sum_{\ell=1}^b(\widehat{\alpha}_\ell\theta)\sum_{i=1}^n\sigma_i\varphi_\ell(\boldsymbol{x}_i)\right] + \frac{1}{n'}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_\theta\sum_{\ell=1}^b\widehat{\alpha}_\ell\sum_{j=1}^{n'}\sigma'_j\varphi_\ell(\boldsymbol{x}'_j)\right].
\end{aligned}$$

Applying the Cauchy-Schwarz inequality followed by Jensen's inequality to the first Rademacher average gives

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_\theta\sum_{\ell=1}^b(\widehat{\alpha}_\ell\theta)\sum_{i=1}^n\sigma_i\varphi_\ell(\boldsymbol{x}_i)\right] &\leq \frac{C_\alpha}{n}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\left(\sum_{\ell=1}^b\left(\sum_{i=1}^n\sigma_i\varphi_\ell(\boldsymbol{x}_i)\right)^2\right)^{1/2}\right] \\
&\leq \frac{C_\alpha}{n}\left(\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sum_{\ell=1}^b\left(\sum_{i=1}^n\sigma_i\varphi_\ell(\boldsymbol{x}_i)\right)^2\right]\right)^{1/2} \\
&= \frac{C_\alpha}{n}\left(\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sum_{\ell=1}^b\sum_{i,i'=1}^n\sigma_i\sigma_{i'}\varphi_\ell(\boldsymbol{x}_i)\varphi_\ell(\boldsymbol{x}_{i'})\right]\right)^{1/2}.
\end{aligned}$$

Since $\sigma_1, \ldots, \sigma_n$ *are Rademacher variables,*

$$\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sum_{\ell=1}^b\sum_{i,i'=1}^n\sigma_i\sigma_{i'}\varphi_\ell(\boldsymbol{x}_i)\varphi_\ell(\boldsymbol{x}_{i'})\right] = \mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^n\sum_{\ell=1}^b\varphi_\ell^2(\boldsymbol{x}_i)\right] \leq nC_x^2.$$

Consequently, we have

$$\frac{1}{n}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_{\theta}\sum_{\ell=1}^{b}(\widehat{\alpha}_\ell\theta)\sum_{i=1}^{n}\sigma_i\varphi_\ell(\boldsymbol{x}_i)\right] \leq \frac{C_\alpha C_x}{\sqrt{n}},$$

$$\frac{1}{n'}\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_{\theta}\sum_{\ell=1}^{b}\widehat{\alpha}_\ell\sum_{j=1}^{n'}\sigma'_j\varphi_\ell(\boldsymbol{x}'_j)\right] \leq \frac{C_\alpha C_x}{\sqrt{n'}},$$

and

$$\mathbb{E}_{\mathcal{D},\boldsymbol{\sigma}}\left[\sup_{\theta}\left\{\sum_{i=1}^{n}\sigma_i\omega(\boldsymbol{x}_i)-\sum_{j=1}^{n'}\sigma'_j\omega'(\boldsymbol{x}'_j)\right\}\right] \leq \left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{n'}}\right)C_\alpha C_x.$$

3.0.1. Step 4

Combining the three steps together, we obtain that with probability at least $1-\delta/2$, $\forall\theta$,

$$g(\theta;\mathcal{D})-g(\theta) \leq \left(\frac{2}{\sqrt{n}}+\frac{2}{\sqrt{n'}}\right)C_\alpha C_x + \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2+\frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

The same argument can be used to bound $g(\theta)-g(\theta;\mathcal{D})$. *Combining these two tail inequalities proves the theorem.* ■

Theorem 2 shows that the uniform deviation bound is of order $O(1/\sqrt{n}+1/\sqrt{n'})$, whereas the deviation bound for fixed θ is of order $O(\sqrt{1/n+1/n'})$ as shown in Theorem 1. For this special estimation problem, the convergence rate of the uniform deviation bound is clearly worse than the convergence rate of the deviation bound for fixed θ.

However, after obtaining the uniform deviation bound, we are able to bound the estimation error, that is, the gap between the expectation of our estimate and the best possible estimate within the model. To do so, we need to constrain the parameters of the best estimate via C_α since (10) takes (9) as the target, while (9) is an unbounded function. Furthermore, we assume that the regularization is weak enough so that it would not affect the solution $\widehat{\boldsymbol{\alpha}}$ too much. Specifically, given $\mathcal{D}$, let $\widetilde{\boldsymbol{\alpha}}$ be the minimizer of the objective function in (11) but without the regularization term $\frac{\lambda}{2}\sum_{\ell=1}^{b}\alpha_\ell^2$, subjecting to $\|\widetilde{\boldsymbol{\alpha}}\|_2 \leq C_\alpha$. Let $\widetilde{\mathrm{penL}_1}(\theta;\mathcal{D})$ be an estimator of the penalized L_1-distance corresponding to $\widetilde{\boldsymbol{\alpha}}$, and we assume that there exists $\Delta_\alpha > 0$ such that $\forall\theta$, $\forall\mathcal{D}$, $\widetilde{\mathrm{penL}_1}(\theta;\mathcal{D})-\widehat{\mathrm{penL}_1}(\theta;\mathcal{D}) \leq \Delta_\alpha$. Then we have the following theorem.

Theorem 3 (Estimation error bound) *Let* $\mathrm{penL}_1(\theta)$ *be the maximizer of the estimate in the right-hand side of* (7) *based on* (10) *with the best possible* $\boldsymbol{\alpha}^*$ *where* $\|\boldsymbol{\alpha}^*\|_2 \leq C_\alpha$. *For any* $0<\delta<1$, *with probability at least* $1-\delta$ *over the repeated sampling of* $\mathcal{D}$, *the following holds for all* $0 \leq \theta \leq 1$,

$$\mathrm{penL}_1(\theta)-\mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}_1}(\theta;\mathcal{D})] \leq \Delta_\alpha + \left(\frac{4}{\sqrt{n}}+\frac{4}{\sqrt{n'}}\right)C_\alpha C_x + \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2+\frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)}.$$

Proof 3 *Since $\boldsymbol{\alpha}^*$ is fixed, we have $\mathbb{E}_{\mathcal{D}}[\mathrm{penL}_1(\theta;\mathcal{D})] = \mathrm{penL}_1(\theta)$. Then,*

$$\begin{aligned}\mathrm{penL}_1(\theta) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})] &= \mathbb{E}_{\mathcal{D}}[\mathrm{penL}_1(\theta;\mathcal{D})] - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})] \\ &= (\mathbb{E}_{\mathcal{D}}[\mathrm{penL}_1(\theta;\mathcal{D})] - \mathrm{penL}_1(\theta;\mathcal{D})) + \left(\widehat{\mathrm{penL}}_1(\theta;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})]\right) \\ &\quad + \left(\mathrm{penL}_1(\theta;\mathcal{D}) - \widetilde{\mathrm{penL}}_1(\theta;\mathcal{D})\right) + \left(\widetilde{\mathrm{penL}}_1(\theta;\mathcal{D}) - \widehat{\mathrm{penL}}_1(\theta;\mathcal{D})\right).\end{aligned}$$

We bound each of the four terms separately. According to the proof of Theorem 2, with probability at least $1-\delta/2$, $\forall\theta$,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[\mathrm{penL}_1(\theta;\mathcal{D})] - \mathrm{penL}_1(\theta;\mathcal{D}) \\ &\quad\le \left(\frac{2}{\sqrt{n}} + \frac{2}{\sqrt{n'}}\right) C_\alpha C_x + \sqrt{\frac{\ln(2/\delta)}{2\lambda^2/b^2}\left(\frac{1}{n}\left(2+\frac{1}{n}\right)^2 + \frac{1}{n'}\left(2+\frac{1}{n'}\right)^2\right)},\end{aligned}$$

The same can be proven for $\widehat{\mathrm{penL}}_1(\theta;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathrm{penL}}_1(\theta;\mathcal{D})]$. The third term must be non-positive since $\widetilde{\mathrm{penL}}_1(\theta;\mathcal{D})$ is the maximizer of the empirical estimate. Finally, the four term is upper bounded by Δ_α, which completes the proof. ■

Theorem 3 shows that the deviation from the expectation of our estimate from the optimal value in the model is small with high probability.

4. Related work

In Scott and Blanchard (2009) and Blanchard et al. (2010), it was proposed to reduce the problem of estimating the class prior to Neyman-Pearson classification[2]. A Neyman-Pearson classifier f minimizes the *false-negative rate* $R_1(f)$, while keeping the *false-positive rate* $R_{-1}(f)$ constrained under a user-specified threshold (Scott and Nowak, 2005):

$$R_1(f) = P_1(f(\boldsymbol{x}) \neq 1), \quad R_{-1}(f) = P_{-1}(f(\boldsymbol{x}) \neq -1),$$

where P_1 and P_{-1} denote the probabilities for the positive-class and negative-class conditional densities, respectively. The false-negative rate on the unlabeled dataset is defined and expressed as

$$\begin{aligned}R_X(f) &= P_X(f(\boldsymbol{x}) = 1) \\ &= \pi(1 - R_1(f)) + (1-\pi)R_{-1}(f),\end{aligned}$$

where P_X denotes the probability for unlabeled input data density.

The Neyman-Pearson classifier between P_1 and P_X is defined as

$$R^*_{X,\alpha} = \inf_f \; R_X(f) \qquad \text{s.t.} \quad R_1(f) \le \alpha.$$

2. The papers (Scott and Blanchard, 2009; Blanchard et al., 2010) considered the nominal class as $y = 0$, and the novel class as $y = 1$. The aim was to estimate $p(y = 1)$. We use a different notation with the nominal class as $y = 1$ and the novel class as $y = -1$ and estimate $\pi = p(y = 1)$. To simplify the exposition, we use the same notation here as in the rest of the paper.

Then the minimum false-negative rate for the unlabeled dataset given false positive rate α is expressed as

$$R^*_{X,\alpha} = \theta(1-\alpha) + (1-\theta)R^*_{-1,\alpha}. \tag{12}$$

Theorem 1 in Scott and Blanchard (2009) says that if the supports for P_1 and P_{-1} are different, there exists α such that $R^*_{-1,\alpha} = 0$. Therefore, the class prior can be determined as

$$\theta = -\left.\frac{\mathrm{d}R^*_{X,\alpha}}{\mathrm{d}\alpha}\right|_{\alpha=1^-}, \tag{13}$$

where $\alpha \to 1^-$ is the limit from the left-hand side. Note that this limit is necessary since the first term in (12) will be zero when $\alpha = 1$.

However, estimating the derivative when $\alpha \to 1^-$ is not straightforward in practice. The curve of $1 - R^*_X$ vs. R^*_1 can be interpreted as an ROC curve (with a suitable change in class notation), but the empirical ROC curve is often unstable at the right endpoint when the input dimensionality is high (Sanderson and Scott, 2014). One approach to overcome this problem is to fit a curve to the right endpoint of the ROC curve in order to enable the estimation (as in Sanderson and Scott (2014)). However, it is not clear how the estimated class-prior is affected by this curve-fitting.

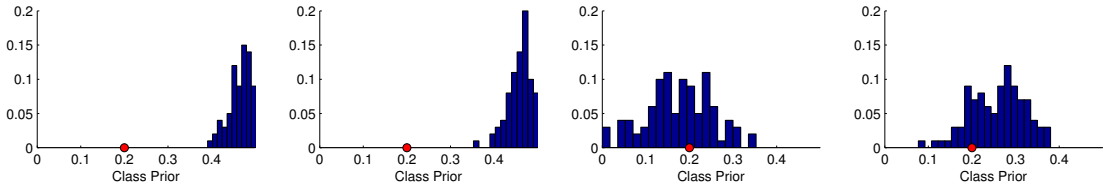
5. Experiments

In this section, we experimentally compare the performance of the proposed method and alternative methods for estimating the class prior. We compared the following methods:

- **EN**: The method of Elkan and Noto (2008) with the classifier as a squared-loss variant of logistic regression classifier (Sugiyama, 2010).
- **PE**: The direct Pearson-divergence matching method proposed in du Plessis and Sugiyama (2014).
- **SB** The method of Blanchard et al. (2010). The Neyman-Pearson classifier was implemented as a the thresholded ratio of two kernel density estimates, each with a bandwidth parameter. As in Blanchard et al. (2010), the bandwidth parameters were jointly optimized by maximizing the cross-validated estimate of the AUC. The prior was obtained by estimating (13) from the empirical ROC curve.
- **pen-L_1** (proposed): The penalized L_1-distance method with an analytic solution. The basis functions were selected as Gaussians centered at all training samples. All hyper-parameters were determined by cross-validation.

First, we illustrate the systematic overestimation of the class prior by two previously proposed methods, EN (Elkan and Noto, 2008) and PE (du Plessis and Sugiyama, 2014) when the classes significantly overlap. The class-conditional densities are

$$p(\boldsymbol{x}|y=1) = \mathcal{N}_x\left(0, 1^2\right) \quad \text{and} \quad p(\boldsymbol{x}|y=-1) = \mathcal{N}_x\left(2, 1^2\right),$$



(a) EN (b) PE (c) SB (d) pen-L_1

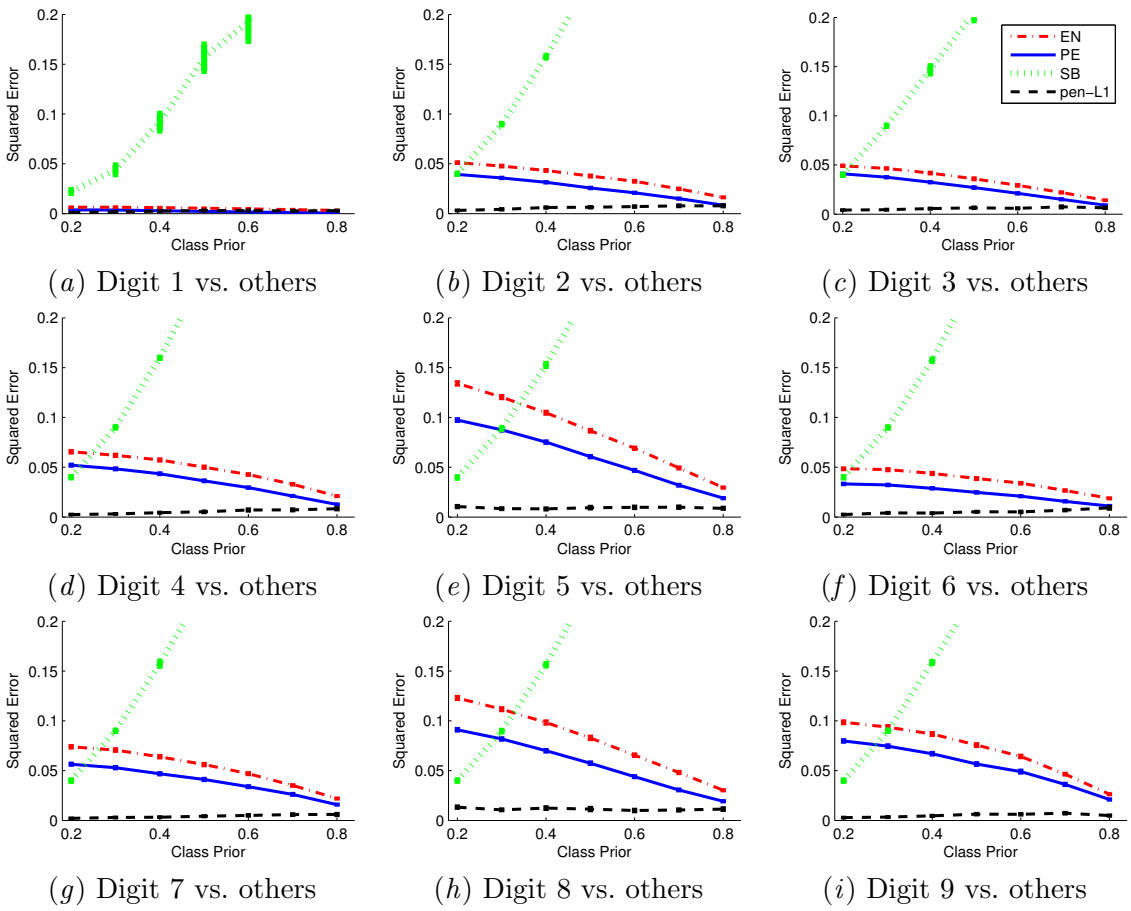
Figure 2: Histograms of class-prior estimates when the true class prior is $p(y = 1) = 0.2$. The EN method and the PE method have an intrinsic bias that does not decrease even when the number of samples is increased, while the SB method and the proposed pen-L_1 method work reasonably well.

where $\mathcal{N}_x\left(\mu, \sigma^2\right)$ denotes the Gaussian density with mean μ and variance σ^2 with respect to x. The true class prior is set at $\pi = p(y = 1) = 0.2$. The sizes of the unlabeled dataset and the labeled dataset were both set at $n = n' = 300$. The histograms of class prior estimates are plotted in Figure 2, showing that the EN and PE methods clearly overestimate the true class prior. We also confirmed that this overestimation does not decrease even when the number of samples is increased. On the other hand, the SB and pen-L_1 methods work reasonably well.

Finally, we use the MNIST hand-written digit dataset. For each digit, all the other digits were assumed to be in the opposite class (i.e., one-versus-rest). The dataset was reduced to 4-dimensions using principal component analysis. The squared error of class-prior estimation is given in Figure 3, showing that the proposed pen-L_1 method overall gives accurate estimates of the class prior, while the EN and PE methods tend to give less accurate estimates for low class priors and more accurate estimates for higher class priors, which agrees with the observation in du Plessis and Sugiyama (2014). On the other hand, the SB method tends to perform poorly, which is caused by the instability of the empirical ROC curve at the right endpoint when the input dimensionality is larger, as pointed out in Section 4.

6. Conclusion

In this paper, we discussed the problem of class-prior estimation from positive and unlabeled data. We first showed that class-prior estimation from positive and unlabeled data by partial distribution matching under f-divergences yields systematic overestimation of the class prior. We then proposed to use penalized f-divergences to rectify this problem. We further showed that the use of L_1-distance as an example of f-divergences yields a computationally efficient algorithm with an analytic solution. We provided its uniform deviation bound and estimation error bound, which theoretically supports the usefulness of the proposed method. Finally, through experiments, we demonstrated that the proposed method compares favorably with existing approaches.



(*a*) Digit 1 vs. others (*b*) Digit 2 vs. others (*c*) Digit 3 vs. others

(*d*) Digit 4 vs. others (*e*) Digit 5 vs. others (*f*) Digit 6 vs. others

(*g*) Digit 7 vs. others (*h*) Digit 8 vs. others (*i*) Digit 9 vs. others

Figure 3: Class-prior estimation accuracy for the MNIST dataset. The EN method and the PE method behave similarly, and the proposed pen-L_1 method consistently outperforms them. The SB method performed poorly due to the instability of the empirical ROC curve at the right endpoint.

Acknowledgments

MCdP and GN were supported by the JST CREST program and MS was supported by KAKENHI 25700022.

References

S. M. Ali and S. D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *The Journal of Machine Learning Research*, 9999:2973–3009, 2010.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

M. C. du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *ICML 2012*, pages 823–830, Jun. 26–Jul. 1 2012.

M. C. du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, E97-D, 2014.

M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 703–711, 2014.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD 14*, pages 213–220, 2008.

A. Keziou. Dual representation of ϕ-divergences and applications. *Comptes Rendus Mathématique*, 336(10):857–862, 2003.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

C. Scott and G. Blanchard. Novelty detection: Unlabeled data definitely help. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS2009)*, Clearwater Beach, Florida USA, Apr. 16-18 2009.

C. Scott and R. Nowak. A Neyman-Pearson approach to statistical learning. *Information Theory, IEEE Transactions on*, 51(11):3806–3819, 2005.

M. Sugiyama. Superfast-trainable multi-class probabilistic classifier by least-squares posterior fitting. *IEICE Transactions on Information and Systems*, E93-D(10):2690–2701, 2010.

M. Sugiyama, T. Suzuki, T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. In *Advances in Neural Information Processing Systems 25*, pages 692–700, 2012.