Semi-supervised logistic discrimination via labeled data and unlabeled data from different sampling distributions

Shuichi Kawano

Department of Mathematical Sciences, Graduate School of Engineering,

Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka 599-8531, Japan.

skawano@ms.osakafu-u.ac.jp

Abstract: This article addresses the problem of classification method based on both labeled and unlabeled data, where we assume that a density function for labeled data is different from that for unlabeled data. We propose a semi-supervised logistic regression model for classification problem along with the technique of covariate shift adaptation. Unknown parameters involved in proposed models are estimated by regularization with EM algorithm. A crucial issue in the modeling process is the choices of tuning parameters in our semi-supervised logistic models. In order to select the parameters, a model selection criterion is derived from an information-theoretic approach. Some numerical studies show that our modeling procedure performs well in various cases.

Key Words and Phrases: Covariate shift; EM algorithm; Model selection; Regularization; Semi-supervised learning.

1 Introduction

In recent years, with the wide availability of fast and high-powered computers, high-throughput data of unexampled size and complexity have frequently been seen in the contemporary statistics and machine learning. Examples involve data from genomics, proteomics, natural language processing, and signal processing. For the huge amount of data, it is difficult to label data by a human operator, since its work requires vast times and efforts. Only small labeled data set may, therefore, be available, while an unlabeled data set can be more easily obtained. Under such a circumstance, a classification method that

combines both labeled and unlabeled data, called semi-supervised learning, has received an enormous amount of attention in the late machine learning and statistical literature (see, e.g., Chapelle *et al.*, 2006; Liang *et al.*, 2007). For overviews of semi-supervised learning methods, we refer to Zhu (2008), and references given therein.

Many classification techniques for semi-supervised learning have been proposed by various researchers, e.g., Amini and Gallinari (2002), Basu *et al.* (2004), Bennett and Demiriz (1998), Chen and Wang (2007), Dean *et al.* (2006), Kawano and Konishi (2011), Kawano *et al.* (2012), Lafferty and Wasserman (2007), and Zhou *et al.* (2004). Most of these semi-supervised methods implicitly assumes that a density function for labeled data is the same as that for unlabeled data. On the other hand, we, here, consider the case that the densities for labeled data and unlabeled data are different, since the densities are not always same in practical situations. In such a case, several semi-supervised methods have been presented, e.g., Jiang and Zhai (2007), Wu *et al.* (2009), and Zadrozny (2004). However, for these methods, there remains a problem of evaluating constructed semi-supervised models, which is a crucial issue in the model building process. Cross validation (CV) is often used in evaluating models constructed by semi-supervised procedures. An advantage of CV lies in its independence from probabilistic assumptions. The computational time of the procedures is, however, very large, and the high variability and tendency to undersmooth in CV are not negligible in the analysis of complex or high-dimensional data, since the selectors are repeatedly applied.

In this paper, we propose a logistic model for the semi-supervised classification problem by using statistical methods under covariate shift (Shimodaira, 2000) in the case that the density function for labeled data is different from that for unlabeled data. The unknown parameters in the model are estimated by the regularization method with the help of EM algorithm. A crucial issue in our modeling strategy is to choose values of some tuning parameters included in semi-supervised logistic models, which corresponds to evaluating models determined by our proposed procedures. In order to objectively select optimal values of tuning parameters, we then introduce a model selection criterion based on an information-theoretic approach (Konishi and Kitagawa, 1996) that evalu-

ates the semi-supervised logistic models estimated by the regularization method. Some numerical examples demonstrate that the proposed procedure works well and performs better than competing methods.

This paper is organized as follows. In Section 2, we present a semi-supervised logistic model for classification problem based on covariate shift adaptation and its estimation procedure by the regularization method. Section 3 provides a model selection criterion derived from an information-theoretic viewpoint to select some tuning parameters in semi-supervised logistic models. In Section 4, Monte Carlo simulations and benchmark data analysis are given to assess the performances of our proposed semi-supervised logistic discrimination. Some concluding remarks are given in Section 5.

2 Semi-supervised logistic modeling from different sampling distributions

2.1 Linear logistic modeling for semi-supervised learning

We review here semi-supervised linear logistic models developed by early researchers (e.g., Amini and Gallinari, 2002; Vittaut *et al.*, 2002). Suppose that we have an n_1 labeled data set $\{(\boldsymbol{x}_\alpha, y_\alpha); \alpha = 1, \ldots, n_1\}$ and an $(n - n_1)$ unlabeled data set $\{\boldsymbol{x}_\alpha; \alpha = n_1 + 1, \ldots, n\}$, where $\boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha p})^T$ denotes a p-dimensional explanatory variable and Y_α is a random variable taking values 0 or 1 with probabilities

$$\Pr(Y_\alpha = 1|\boldsymbol{x}_\alpha) = \pi(\boldsymbol{x}_\alpha), \qquad \Pr(Y_\alpha = 0|\boldsymbol{x}_\alpha) = 1 - \pi(\boldsymbol{x}_\alpha). \tag{1}$$

Note that logistic models are first constructed by only labeled data set, while the unlabeled data set is used in estimating the parameters involved in the logistic models.

Using conditional probabilities in Equation (1) and the labeled data set, a linear logistic model (see, e.g., Hastie *et al.*, 2009) is formulated by

$$\log\left\{\frac{\pi(\boldsymbol{x}_\alpha)}{1 - \pi(\boldsymbol{x}_\alpha)}\right\} = w_0 + \sum_{j=1}^{p} w_j x_{\alpha j} = \boldsymbol{w}^T \boldsymbol{x}_\alpha^*, \quad \alpha = 1, \ldots, n_1, \tag{2}$$

where $\boldsymbol{w} = (w_0, w_1, \ldots, w_p)^T$ is an unknown parameter vector and $\boldsymbol{x}_\alpha^* = (1, \boldsymbol{x}_\alpha^T)^T$. Hereafter, we denote conditional probabilities by $\pi(\boldsymbol{x}_\alpha; \boldsymbol{w})$, since the conditional probabilities

3

depend on the parameter vector $\boldsymbol{w}$. It follows from Equation (2) that conditional probabilities can be rewritten as

$$\pi(\boldsymbol{x}_\alpha; \boldsymbol{w}) = \frac{\exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)}{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)}. \tag{3}$$

Also, a probability function of the random variable Y_α is the Bernoulli distribution in the form

$$f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) = \pi(\boldsymbol{x}_\alpha; \boldsymbol{w})^{y_\alpha} \{1 - \pi(\boldsymbol{x}_\alpha; \boldsymbol{w})\}^{1-y_\alpha}, \qquad y_\alpha = 0, 1. \tag{4}$$

Under the linear logistic model, the log-likelihood function for y_α in terms of $\boldsymbol{w}$ is induced into

$$\begin{aligned}
\ell(\boldsymbol{w}) &= \sum_{\alpha=1}^{n_1} \log f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) \\
&= \sum_{\alpha=1}^{n_1} [y_\alpha \log \pi(\boldsymbol{x}_\alpha; \boldsymbol{w}) + (1 - y_\alpha) \log\{1 - \pi(\boldsymbol{x}_\alpha; \boldsymbol{w})\}] \\
&= \sum_{\alpha=1}^{n_1} \left[y_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\} \right].
\end{aligned} \tag{5}$$

The unknown parameter $\boldsymbol{w}$ included in the logistic model is usually estimated by maximizing the log-likelihood function with respect to the parameter. The procedure is known as the supervised learning, i.e., the parameter is determined by using only labeled data set. Since we have an additional unlabeled data set, the parameter should be estimated by both labeled and unlabeled data set, which is called the semi-supervised learning. Thereby, Amini and Gallinari (2002) proposed a log-likelihood function with additional unlabeled data given by

$$\begin{aligned}
\ell^*(\boldsymbol{w}) &= \sum_{\alpha=1}^{n_1} \left[y_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\} \right] \\
&\quad + \sum_{\alpha=n_1+1}^{n} \left[t_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\} \right],
\end{aligned} \tag{6}$$

where t_α $(\alpha = n_1 + 1, \ldots, n)$ is a latent variable coded as 0 or 1. Amini and Gallinari (2002) estimated the parameter by maximizing the Equation (6) with the technique of EM algorithm, while Kawano and Konishi (2011) employed the Equation (6) with a

regularization term in estimating the parameter in the context of nonlinear logistic models based on basis expansions.

Given the estimate $\hat{\boldsymbol{w}}$, we assign a future observation $\boldsymbol{x}_{\mathrm{f}}$ into class j $(j = 0, 1)$ that has the maximum conditional probability in the Equation (3).

2.2   Semi-supervised logistic model for different distributions

Logistic models for semi-supervised learning described in Section 2.1 usually assumes that a density function for the labeled data set is the same as that for the unlabeled data set, i.e., when we denote that $q_{\mathrm{label}}(\boldsymbol{x})$ is a probability density function of explanatory variables for the labeled data and $q_{\mathrm{unlabel}}(\boldsymbol{x})$ is that for the unlabeled data, $q_{\mathrm{label}}(\boldsymbol{x}) = q_{\mathrm{unlabel}}(\boldsymbol{x})$. Our aim in this section is to construct logistic models under the situation that a density for the labeled data set is different from that for the unlabeled data set, i.e., $q_{\mathrm{label}}(\boldsymbol{x}) \neq q_{\mathrm{unlabel}}(\boldsymbol{x})$.

We recall the log-likelihood function for logistic models with unlabeled data in Equation (6). For the log-likelihood function, we propose a weighted log-likelihood function with unlabeled data in the form

$$\ell^*(\boldsymbol{w}; \gamma_1, \gamma_2) = \sum_{\alpha=1}^{n_1} \left\{ \frac{q_{\mathrm{unlabel}}(\boldsymbol{x}_\alpha)}{q_{\mathrm{label}}(\boldsymbol{x}_\alpha)} \right\}^{\gamma_1} \left[y_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\} \right]$$
$$+ \sum_{\alpha=n_1+1}^{n} \left\{ \frac{q_{\mathrm{label}}(\boldsymbol{x}_\alpha)}{q_{\mathrm{unlabel}}(\boldsymbol{x}_\alpha)} \right\}^{\gamma_2} \left[t_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\} \right], \quad (7)$$

where $\gamma_1, \gamma_2 \in [0, 1]$ are tuning parameters. If both γ_1 and γ_2 are 0, the log-likelihood function in Equation (7) coincides with that in Equation (6). Note that the weight on the first term, $q_{\mathrm{unlabel}}(\boldsymbol{x})/q_{\mathrm{label}}(\boldsymbol{x})$, is bigger near high densities of unlabeled data compared to those of labeled data, while that on the second term, $q_{\mathrm{label}}(\boldsymbol{x})/q_{\mathrm{unlabel}}(\boldsymbol{x})$, is strengthen near high densities of labeled data compared to those of unlabeled data. Hence, the log-likelihood function on the first term is highly weighted near high densities of unlabeled data compared to those of labeled data, while that on the second term has high weighting near high densities of labeled data compared to those of unlabeled data. An idea of the weight, the ratio of $q_{\mathrm{label}}(\boldsymbol{x})$ and $q_{\mathrm{unlabel}}(\boldsymbol{x})$, arises from a statistical inference under covariate shift (Shimodaira, 2000). In the semi-supervised learning, employing a ratio of

densities in log-likelihood functions is not new. For example, Kawakita and Kanamori (2012), Sokolovska *et al.* (2008), and Zou *et al.* (2007) use a ratio of densities in the semi-supervised inference. However, the Equation (7) is a novel formulation in the semi-supervised context.

The Equation (7) includes unknown values of ratios, $q_{\text{unlabel}}(\boldsymbol{x})/q_{\text{label}}(\boldsymbol{x})$ and $q_{\text{label}}(\boldsymbol{x})/q_{\text{unlabel}}(\boldsymbol{x})$, which are to be estimated. Various researchers address the problem of estimating the ratios by using several methods of statistics or machine learning (Bickel *et al.*, 2009; Huang *et al.*, 2007; Kanamori *et al.*, 2009; Sugiyama *et al.*, 2008; Sugiyama and Kawanabe, 2012; Sugiyama *et al.*, 2012). In this paper, we employ a uLSIF method proposed by Kanamori *et al.* (2009) in determining values of the ratios, where the determination is performed before estimating the parameter $\boldsymbol{w}$. Also, a source code of the method uLSIF is available in *http://www.math.cm.is.nagoya-u.ac.jp/~kanamori/software/LSIF*. We do not follow details of the density ratio estimation procedure by the uLSIF method, since these are not our focus in this paper. For readers that are interested in the topics, we refer to Kanamori *et al.* (2009), and Sugiyama and Kawanabe (2012).

2.3 Parameter estimation via regularization

In estimating parameters in logistic models, the log-likelihood function often diverges to infinity when the maximum likelihood method is applied (Konishi and Kitagawa, 2008). Hence, the parameter vector $\boldsymbol{w}$ in Equation (7) is estimated by the regularization method. The regularization method is to maximize a following regularized log-likelihood function

$$\ell_\lambda^*(\boldsymbol{w}; \gamma_1, \gamma_2) = \ell^*(\boldsymbol{w}; \gamma_1, \gamma_2) - \frac{n_1 \lambda}{2} \boldsymbol{w}^T K \boldsymbol{w}, \tag{8}$$

where λ is a regularization parameter that has positive values and $K = \text{diag}(0, I_p)$ is a $(p+1) \times (p+1)$ matrix. Here, the matrix I_p is a p-dimensional identity matrix.

It is not easy to optimize the parameter involved in Equation (8), since the latent variables t_α ($\alpha = n_1+1, \ldots, n$) are unobserved. Hence, we employ an EM-based algorithm developed by Kawano and Konishi (2011) as follows:

Step1 Estimate the parameter vector $\boldsymbol{w}$ by maximizing the regularized log-likelihood

function using only labeled data set $\{(\boldsymbol{x}_\alpha, y_\alpha); \alpha = 1, \ldots, n_1\}$ along with the technique of Newton-Raphson method.

Step2 Construct a classification rule $\pi(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}})$.

Step3 (E-step) According to the classification rule in Step2, compute the conditional probabilities $\pi(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}})$ for unlabeled data $\boldsymbol{x}_\alpha$ ($\alpha = n_1 + 1, \ldots, n$). By using the conditional probabilities, estimate t_α in the form $\hat{t}_\alpha = \pi(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}})$.

Step4 (M-step) Replace t_α into $\hat{t}_\alpha$ in the regularized log-likelihood function (8), and then determine the parameter vector $\boldsymbol{w}$ through the maximization of the log-likelihood function in Equation (8) with the help of Newton-Raphson method.

Step5 Repeat the Step2 to the Step4 until the following condition

$$|\ell_\lambda^*(\hat{\boldsymbol{w}}^{(k+1)}; \gamma_1, \gamma_2) - \ell_\lambda^*(\hat{\boldsymbol{w}}^{(k)}; \gamma_1, \gamma_2)| < \varepsilon \tag{9}$$

is satisfied, where $\hat{\boldsymbol{w}}^{(k)}$ is the value of $\boldsymbol{w}$ after the k-th EM iteration and ε is an arbitrary small number (e.g., 10^{-5}).

It follows from these procedures that we obtain a statistical model in the form

$$f(y|\boldsymbol{x}; \hat{\boldsymbol{w}}) = \pi(\boldsymbol{x}; \hat{\boldsymbol{w}})^y \{1 - \pi(\boldsymbol{x}; \hat{\boldsymbol{w}})\}^{1-y}. \tag{10}$$

Note that the statistical model is constructed by using both labeled data and unlabeled data.

3   Model selection criterion

The statistical model in Equation (10) contains some adjusted parameters including two tuning parameters γ_1, γ_2 in the weighted log-likelihood function and the regularization parameter λ. Regarding the selection of these adjusted parameters as that of candidate models, we introduce a model selection criterion from an information-theoretic approach.

Let $y_1, \ldots, y_{n_1}$ be n_1 observations drawn randomly from an unknown probability distribution function $G(y|x)$ having a density function $g(y|x)$. On the other hand, we assume that n_1 observations for explanatory variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}$ are non-random; i.e.,

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}$ are fixed (for details of this assumption, we refer to Konishi and Kitagawa, 2008). Under these settings, we derive a model selection criterion from the viewpoint of information theory.

Suppose that $\boldsymbol{z} = (z_1, \ldots, z_{n_1})^T$ are future observations for the response variable generated from $g(y|x)$. Let $f(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{w}}_G)^{\eta(\boldsymbol{x})} = \prod_{\alpha=1}^{n_1} f(z_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}_G)^{\eta(\boldsymbol{x}_\alpha)}$ and $g(\boldsymbol{z}|\boldsymbol{x}) = \prod_{\alpha=1}^{n_1} g(z_\alpha|\boldsymbol{x}_\alpha)$, where $\hat{\boldsymbol{w}}_G$ is an estimator of the parameter by any estimation procedures, $\eta(\boldsymbol{x}) = \eta(\boldsymbol{x}_1) + \cdots + \eta(\boldsymbol{x}_{n_1})$, and $\eta(\boldsymbol{x}_\alpha)$ ($\alpha = 1, \ldots, n_1$) are weights that depend on explanatory variables $\boldsymbol{x}_\alpha$, which satisfy $\eta(\boldsymbol{x}_\alpha) > 0$. Note that the weights $\eta(\boldsymbol{x}_\alpha)$ ($\alpha = 1, \ldots, n_1$) are fixed, since we assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_1}$ are non-random. Then Irizarry (2001) implicitly proposes a following Kullback–Leibler information in order to measure the divergence of the statistical model with weights from the true distribution:

$$
\begin{aligned}
I\{g; f\} &= E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{g(\boldsymbol{z}|\boldsymbol{x})}{f(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{w}}_G)^{\eta(\boldsymbol{x})}}\right] \\
&= E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\log g(\boldsymbol{z}|\boldsymbol{x})\right] - E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\log f(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{w}}_G)^{\eta(\boldsymbol{x})}\right] \\
&= E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\log g(\boldsymbol{z}|\boldsymbol{x})\right] - E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\eta(\boldsymbol{x}) \log f(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{w}}_G)\right]. \quad (11)
\end{aligned}
$$

The best model can be regarded as the best minimizer of the Kullback–Leibler information (Irizarry, 2001). Since the first term of Equation (11) does not depend on the models with the estimator $\hat{\boldsymbol{w}}_G$, we have only to consider the second term of Equation (11). Therefore, we focus on maximizing the second term of Equation (11) which leads to the minimization of the Kullback–Leibler information.

By introducing an estimator of the second term of Equation (11), a model selection criterion is, generally, given by

$$
\text{IC} = -2 \sum_{\alpha=1}^{n_1} \eta(\boldsymbol{x}_\alpha) \log f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}_G) + 2\hat{b}(G), \quad (12)
$$

where IC stands for information criterion and $\hat{b}(G)$ is an estimator of the bias $b(G)$ in the following:

$$
b(G) = E_{G(\boldsymbol{y}|\boldsymbol{x})}\left[\sum_{\alpha=1}^{n_1} \eta(\boldsymbol{x}_\alpha) \log f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}_G) - E_{G(\boldsymbol{z}|\boldsymbol{x})}\left[\eta(\boldsymbol{x}) \log f(\boldsymbol{z}|\boldsymbol{x}; \hat{\boldsymbol{w}}_G)\right]\right]. \quad (13)
$$

Suppose that the estimator $\hat{\boldsymbol{w}}_M$ of the parameter is an M-estimator defined as the

8

solution of the following implicit equation:

$$\sum_{\alpha=1}^{n_1} \boldsymbol{\psi}(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}_M) = \boldsymbol{0} \tag{14}$$

with $\boldsymbol{\psi}$ being referred to as $\boldsymbol{\psi}$–function (e.g., see, Huber, 2004). Using the idea of Konishi and Kitagawa (1996), we derive a model selection criterion for the statistical models with the M-estimator $\hat{\boldsymbol{w}}_M$ in the form

$$\text{IC}_M = -2\sum_{\alpha=1}^{n_1} \eta(\boldsymbol{x}_\alpha) \log f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}_M) + 2\text{tr}\left\{Q(\hat{\boldsymbol{w}}_M)R^{-1}(\hat{\boldsymbol{w}}_M)\right\}, \tag{15}$$

where $Q(\hat{\boldsymbol{w}}_M)$ and $R(\hat{\boldsymbol{w}}_M)$ are given by

$$Q(\hat{\boldsymbol{w}}_M) = \frac{1}{n_1}\sum_{\alpha=1}^{n_1} \boldsymbol{\psi}(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w})\frac{\eta(\boldsymbol{x}_\alpha)\partial \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w})}{\partial \boldsymbol{w}^T}\bigg|_{\boldsymbol{w}=\hat{\boldsymbol{w}}_M}, \tag{16}$$

$$R(\hat{\boldsymbol{w}}_M) = -\frac{1}{n_1}\sum_{\alpha=1}^{n_1} \frac{\partial \boldsymbol{\psi}(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w})^T}{\partial \boldsymbol{w}}\bigg|_{\boldsymbol{w}=\hat{\boldsymbol{w}}_M}. \tag{17}$$

In our models, the estimator $\hat{\boldsymbol{w}}$, which maximizes the regularized log-likelihood function in Equation (8), can be regarded as an M-estimator. Here, we set the $\boldsymbol{\psi}$–function of the estimator into

$$\boldsymbol{\psi}(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}}\left[\left\{\frac{q_{\text{unlabel}}(\boldsymbol{x}_\alpha)}{q_{\text{label}}(\boldsymbol{x}_\alpha)}\right\}^{\gamma_1}\left[y_\alpha \boldsymbol{w}^T \boldsymbol{x}_\alpha^* - \log\{1 + \exp(\boldsymbol{w}^T \boldsymbol{x}_\alpha^*)\}\right] - \frac{\lambda}{2}\boldsymbol{w}^T K \boldsymbol{w}\right]. \tag{18}$$

Note that the $\boldsymbol{\psi}$–function in Equation (18) is actually incorrect since the estimator $\hat{\boldsymbol{w}}$ is obtained by maximizing the Equation (8) with respect to the parameter; i.e., the estimator are constructed by using both labeled and unlabeled data. However, $\boldsymbol{\psi}$–functions in the context of model selection criteria must be given by a regularized or non-regularized log-likelihood function with incomplete data; i.e., the functions does not include latent variables (for details, see, Hirose *et al.*, 2008). Hence, we employ the $\boldsymbol{\psi}$–function in Equation (18) in order to derive a model selection criterion.

By using the $\boldsymbol{\psi}$–function in Equation (18) and substituting $\{q_{\text{unlabel}}(\boldsymbol{x}_\alpha)/q_{\text{label}}(\boldsymbol{x}_\alpha)\}^{\gamma_1}$ for the weights $\eta(\boldsymbol{x}_\alpha)$ ($\alpha = 1, \ldots, n_1$), we introduce a generalized information criterion (GIC) for evaluating our proposed semi-supervised logistic models estimated by the regularization method. The model selection criterion is given by

$$\text{GIC} = -2\sum_{\alpha=1}^{n_1} \left\{\frac{q_{\text{unlabel}}(\boldsymbol{x}_\alpha)}{q_{\text{label}}(\boldsymbol{x}_\alpha)}\right\}^{\gamma_1} \log f(y_\alpha|\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}}) + 2\text{tr}\left\{Q(\hat{\boldsymbol{w}})R^{-1}(\hat{\boldsymbol{w}})\right\}, \tag{19}$$

where the matrices $Q(\hat{\boldsymbol{w}})$ and $R(\hat{\boldsymbol{w}})$ are

$$Q(\hat{\boldsymbol{w}}) = \frac{1}{n_1} \left\{ X^T \hat{W}^2 \hat{\Lambda}^2 X - \lambda K \hat{\boldsymbol{w}} \mathbf{1}_{n_1}^T \hat{W} \hat{\Lambda} X \right\}, \tag{20}$$

$$R(\hat{\boldsymbol{w}}) = \frac{1}{n_1} X \hat{\Pi} \hat{W} (I_{n_1} - \hat{\Pi}) X + \lambda K. \tag{21}$$

Here, $\mathbf{1}_{n_1}$ is an n_1-dimensional vector of which the elements are all one, and I_{n_1} is an n_1-dimensional identity matrix. Also, $X, \hat{W}, \hat{\Lambda}$, and $\hat{\Pi}$ are, respectively, given by

$$X = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{n_1}^*)^T,$$

$$\hat{W} = \mathrm{diag} \left[\left\{ \frac{q_{\mathrm{unlabel}}(\boldsymbol{x}_1)}{q_{\mathrm{label}}(\boldsymbol{x}_1)} \right\}^{\gamma_1}, \ldots, \left\{ \frac{q_{\mathrm{unlabel}}(\boldsymbol{x}_{n_1})}{q_{\mathrm{label}}(\boldsymbol{x}_{n_1})} \right\}^{\gamma_1} \right],$$

$$\hat{\Lambda} = \mathrm{diag} \left[y_1 - \pi(\boldsymbol{x}_1; \hat{\boldsymbol{w}}), \ldots, y_{n_1} - \pi(\boldsymbol{x}_{n_1}; \hat{\boldsymbol{w}}) \right],$$

$$\hat{\Pi} = \mathrm{diag} \left[\pi(\boldsymbol{x}_1; \hat{\boldsymbol{w}}), \ldots, \pi(\boldsymbol{x}_{n_1}; \hat{\boldsymbol{w}}) \right].$$

Note that the GIC in Equation (19) seemingly appears not to depend on all adjusted parameters (in particular, γ_2). However, the GIC implicitly includes the adjusted parameters $(\lambda, \gamma_1, \gamma_2)$, since the estimator $\hat{\boldsymbol{w}}$ depends on all adjusted parameters.

We choose the adjusted parameters from the minimizer of the GIC in Equation (19).

4 Numerical studies

We studied some numerical examples to show the efficiency of our proposed modeling strategy. Two types of Monte Carlo simulations and benchmark data analysis are given to illustrate the proposed semi-supervised logistic discrimination.

4.1 Simulation 1

We investigated the effectiveness of the proposed modeling procedures through Monte Carlo simulations. In this simulation study, we generated data sets $\{(x_{1\alpha}, x_{2\alpha}, y_\alpha); \alpha = 1, \ldots, n\}$ as labeled data and $\{(x_{1\alpha}, x_{2\alpha}); \alpha = 1, \ldots, 500\}$ as unlabeled data. In labeled data, $(x_{1\alpha}, x_{2\alpha})$ were generated by a normal distribution $N((-0.9, 1-\sin(\sin(0.9^2\pi)))^T, \mathrm{diag}(0.0015, 2))$, and y_α was generated according to a following conditional probability

$$\Pr(Y = 1|x_1, x_2) = 1/\left[1 + \exp\left\{-\sin(2\pi x_1^2) - x_2 + 1\right\}\right]. \tag{22}$$

Table 1: Comparison of prediction error rates (%) and values of selected parameters for several number of labeled data points.

Method \ # of labeled data		25	50	100	150	200	250
SSLRCS	PE	33.3	33.3	33.9	34.8	35.5	35.0
	$\log_{10}(\lambda)$	–2.20	–3.00	–3.18	–3.54	–3.80	–3.72
	γ_1	0.10	0.10	0.10	0.10	0.10	0.10
	γ_2	0.61	0.71	0.74	0.82	0.86	0.82
LSSLR	PE	34.3	34.4	34.2	35.3	35.9	35.6
	$\log_{10}(\xi_1)$	–2.72	–3.36	–3.38	–3.72	–3.88	–3.92
SLR	PE	35.6	34.3	34.3	35.2	35.8	35.6
	$\log_{10}(\xi_2)$	–2.06	–2.32	–2.80	–3.10	–3.50	–3.68

Meanwhile, unlabeled data $(x_{1\alpha}, x_{2\alpha})$ were obtained by a normal distribution $N((-0.4, 1- \sin(\sin(0.4^2\pi)))^T, \mathrm{diag}(0.05, 1))$. Test data $\{(x_{1\alpha}, x_{2\alpha}, y_\alpha); \alpha = 1, \ldots, 1000\}$ were generated as follows. First, $(x_{1\alpha}, x_{2\alpha})$ were derived by a mixture of labeled and unlabeled data, where the mixing rate is equal (that is, 0.5). Second, for the $(x_{1\alpha}, x_{2\alpha})$, y_α was obtained according to the conditional probability in Equation (22). We assumed that labeled data sizes (n) were 25, 50, 100, 150, 200, and 250.

We fitted our semi-supervised logistic regression model to the data sets. Note that the density ratio estimation procedure by the uLSIF method described in Section 2.2 is not performed in these simulation trials, since the density ratio is exactly calculated. The simulation results were obtained by averaging over 50 repeated Monte Carlo trials. For each data set, we computed averages of prediction error rates (PE) for 50 iterations. The tuning parameters in our models were selected by using the GIC in Equation (19). For 50 trials, we computed averages of selected adjusted parameters. The results are summarized in Table 1. From the table, in the selection of adjusted parameters, the values of the tuning parameter γ_1 are 0.10 in all cases, while those of the parameter γ_2 increase with the increasing numbers of labeled data. The regularization parameter λ takes smaller values according to the increasing numbers of labeled data.

We compared the performances of the proposed semi-supervised methodologies (SSLRCS: semi-supervised logistic regression under covariate shift) with those of semi-supervised method proposed by Amini and Gallinari (2002) (LSSLR: linear semi-supervised logistic regression), which is developed under the condition that density functions for labeled and unlabeled data are same, and supervised linear logistic discriminant analysis (SLR: supervised logistic regression). Note that the SLR is constructed by using only labeled data. Semi-supervised and supervised logistic modeling strategies were applied into the data sets. The LSSLR and the SLR include a tuning parameter, respectively, where we denote the tuning parameters as ξ_1 and ξ_2, respectively. The parameter is determined by the GIC, where the GIC for LSSLR is obtained by setting $q_{\text{unlabel}}(\boldsymbol{x}_\alpha)/q_{\text{label}}(\boldsymbol{x}_\alpha) = 1$ ($\alpha = 1, \ldots, n_1$) in Equation (19) and that for SLR is given by Ando *et al.* (2008). For these methods, we also computed averages of prediction error rates and selected tuning parameters. It may be seen from Table 1 that SSLRCS is superior to other methods (LSSLR and SLR) in all cases in the sense that the proposed method gives smaller prediction error rates.

4.2   Simulation 2

We simulated three data sets given in Chakraborty (2011) to examine the performances of our proposed modeling strategy. For each of the simulation cases, we generated 100 data points in the labeled data set, 1000 data points in the unlabeled data set, and 1000 data points in the test data set. Using the data sets, we constructed the SSLRCS, the LSSLR, and the SLR. We repeated the procedure 50 times. Our simulation settings are given as follows (for details, see, Chakraborty (2011, p. 76)):

- Case 1 : In the labeled data set, generate $\boldsymbol{x} = (x_1, x_2)^T$ given by $x_i \sim N(2, 1)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(-2, 1)$ ($i = 1, 2$) for Class 2. In the unlabeled data set, $x_i \sim N(2, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim N(-2, 2)$ ($i = 1, 2$) for Class 2. In the test data set, $x_i \sim 0.5N(2, 1) + 0.5N(2, 2)$ ($i = 1, 2$) for Class 1 and $x_i \sim 0.5N(-2, 1) + 0.5N(-2, 2)$ ($i = 1, 2$) for Class 2.

- Case 2 : Generate $\boldsymbol{x} = (x_1, \ldots, x_{10})^T$ given by $x_i \sim N(1, 3)$ ($i = 1, \ldots, 10$) for Class 1 and $x_i \sim N(-1, 3)$ ($i = 1, \ldots, 10$) for Class 2.

Table 2: Comparison of prediction error rates (%) and values of selected parameters for several cases.

Method \ Data sets		Case 1	Case 2	Case 3
SSLRCS	PE	1.28	3.65	9.72
	$\log_{10}(\lambda)$	–2.50	–1.98	–1.98
	γ_1	1.00	1.00	1.00
	γ_2	0.102	0.106	0.106
LSSLR	PE	1.36	4.19	11.6
	$\log_{10}(\xi_1)$	–2.50	–2.00	–3.00
SLR	PE	1.43	5.05	11.7
	$\log_{10}(\xi_2)$	–2.50	–1.96	–2.18

- Case 3 : Generate $\boldsymbol{x} = (x_1, x_2)^T$ given by $x_i \sim N(5, 2)$ $(i = 1, 2)$ for Class 1 and $x_i \sim N(8, 2)$ $(i = 1, 2)$ for Class 2 in the labeled data set. In the unlabeled data set, $x_i \sim N(6, 2)$ $(i = 1, 2)$ for Class 1 and $x_i \sim N(9, 2)$ $(i = 1, 2)$ for Class 2. In the test data set, $x_i \sim 0.5N(5, 2) + 0.5N(6, 2)$ $(i = 1, 2)$ for Class 1 and $x_i \sim 0.5N(8, 2) + 0.5N(9, 2)$ $(i = 1, 2)$ for Class 2.

The results from the simulation studies are in Table 2. We obtained the values in the table by averaging over 50 trials. The optimal tuning parameters selected by our model selection criterion were 1.00 for γ_1 in all situations, 0.102 and 0.106 for γ_2 in Case 1 and Case 2, 3, respectively, and $10^{-2.50}$ and $10^{-1.98}$ for λ Case 1 and Case 2, 3, respectively. From the simulation results, we observe that our proposed procedure performs well in all cases with respect to minimizing prediction error rates even though Case 2 is an ordinary setting of semi-supervised learning, i.e., the density function for labeled data is same as that for unlabeled data. Hence, we conclude that our proposed method may be useful even if the densities for labeled and unlabeled data are same.

4.3 Benchmark data analysis

Thorough analyzing the g10 data set (Chapelle and Zien, 2005), the ionosphere data set (Sigillito *et al.*, 1989), and the pima data set (Ripley, 1996), we illustrated the effectiveness of the proposed semi-supervised methodology. The g10 data set includes 550 data points with 10 predictors, and we prepared 250 training data points and 300 test data points. The ionosphere data set consists of 356 data points with 33 predictors, and we split the whole 356 data points into 150 training data points and 206 test data points. The pima data set, which consists of 300 training data points and 232 test data points, is a binary classification with 7 predictors. In order to implement the semi-supervised procedure, the training data points were randomly split into two halves with labeled data points and unlabeled data points, where labeled data points were assigned as 5%, 10%, 20%, 30%, 40%, and 50% for training data points, respectively. We repeated the random splitting 50 times. We also compared our proposed method (SSLRCS) with the LSSLR and the SLR, which are described in Section 4.1.

Table 3 shows the summary of the prediction errors and selected adjusted parameters for the benchmark data sets. The values in the table were averaged over 50 repetitions. From the results, we observe that the tuning parameter γ_1 provides the largest values (i.e., 1.00) in almost all cases, while the parameter γ_2 gives relatively smaller values (i.e., from 0.10 to 0.40). We also find that our proposed procedure outperforms the previously proposed methods in almost all situations, although it is unclear that whether densities for labeled and unlabeled data are different. In particular, the proposed method seems to work well when the number of labeled data points is small.

5 Concluding remarks

We proposed a semi-supervised logistic classification methodology for different density functions of labeled and unlabeled data along with the technique of covariate shift adaptation and regularization. A crucial point for our semi-supervised modeling processes includes the choices of some tuning parameters in our proposed models. We introduced a

Table 3: Comparison of prediction error rates (%) and values of selected parameters for some data sets.

Method \ %		5	10	20	30	40	50
g10							
SSLRCS	PE	3.40	3.47	3.85	4.06	4.66	5.42
	$\log_{10}(\lambda)$	–3.20	–2.97	–2.99	–3.00	–3.00	–3.00
	γ_1	1.00	1.00	1.00	1.00	1.00	1.00
	γ_2	0.15	0.10	0.10	0.10	0.10	0.10
LSSLR	PE	26.6	16.2	9.94	7.04	5.66	4.77
	$\log_{10}(\xi_1)$	–3.50	–3.00	–3.00	–3.00	–3.00	–3.00
SLR	PE	26.4	16.4	9.30	6.85	5.45	4.62
	$\log_{10}(\xi_2)$	–3.50	–3.00	–3.00	–3.00	–3.00	–3.00
Ionosphere							
SSLRCS	PE	18.2	17.3	16.9	16.4	17.3	16.8
	$\log_{10}(\lambda)$	–2.89	–2.86	–2.70	–2.44	–2.61	–2.66
	γ_1	0.99	0.99	1.00	1.00	1.00	1.00
	γ_2	0.50	0.46	0.37	0.27	0.37	0.35
LSSLR	PE	29.0	22.8	18.9	17.4	16.2	15.4
	$\log_{10}(\xi_1)$	–3.92	–3.50	–3.50	–3.00	–3.00	–3.00
SLR	PE	28.9	23.1	19.5	18.0	16.7	15.7
	$\log_{10}(\xi_2)$	–3.92	–3.50	–3.50	–3.00	–3.00	–3.00
Pima							
SSLRCS	PE	26.6	26.9	26.6	26.8	26.7	26.7
	$\log_{10}(\lambda)$	1.41	1.53	1.35	1.30	1.27	1.36
	γ_1	1.00	1.00	1.00	1.00	1.00	1.00
	γ_2	0.30	0.28	0.26	0.23	0.24	0.23
LSSLR	PE	30.1	27.0	27.0	27.0	26.9	26.7
	$\log_{10}(\xi_1)$	1.27	1.41	1.53	1.72	1.71	1.61
SLR	PE	29.3	26.9	26.9	27.0	26.8	26.7
	$\log_{10}(\xi_2)$	2.46	2.37	2.34	2.23	2.16	2.10

model selection criterion from the viewpoint of information theory in order to select the values of the adjusted parameters. Through Monte Carlo simulations and the benchmark data analysis, we showed that our modeling strategy is effectiveness in practical situations in the viewpoints of yielding relatively lower prediction errors than previously developed methods. Our modeling procedure may be applied into the problem of constructing a nonlinear semi-supervised classification method based on basis expansions, which will be discussed in another paper.

Acknowledgement

References

[1] Amini, M-R. and Gallinari, P. (2002). Semi-supervised logistic regression. *Proc. 15th Eur. Conf. Artif. Intell.* 390–394.

[2] Ando, T., Konishi, S. and Imoto, S. (2008). Nonlinear regression modeling via regularized radial basis function networks. *J. Statist. Plann. Inference* **138**, 3616–3633.

[3] Basu, S., Bilenko, M. and Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discov. Data. Min.* ACM Press, 59–68.

[4] Bennett, K. P. and Demiriz, A. (1998). Semi-supervised support vector machines. *Adv. Neural Inform. Process. Syst.* **11**, 368–374.

[5] Bickel, S., Brückner, M. and Scheffer, T. (2009). Discriminative learning under covariate shift. *J. Mach. Learn. Res.* **10**, 2137–2155.

[6] Chakraborty, S. (2011). Bayesian semi-supervised learning with support vector machine. *Statist. Methodol.* **8**, 68–82.

[7] Chapelle, O., Schölkopf, B. and Zien, A. (2006). *Semi-Supervised Learning.* Cambridge, MA: MIT Press.

[8] Chapelle, O. and Zien, A. (2005). Semi-supervised classification by low density separation. *Proc. 10th Int. Workshop Artif. Intell. Stat.* 57–64.

[9] Chen, K. and Wang, S. (2007). Regularized boost for semi-supervised learning. *Adv. Neural Inform. Process. Syst.* **20**, 281–288.

[10] Dean, N., Murphy, T. B. and Downey, G. (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *J. Roy. Statist. Soc. Ser. C* **55**, 1–14.

[11] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning.* 2nd ed. New York: Springer.

[12] Hirose, K., Kawano, S. and Konishi, S. (2008). Bayesian factor analysis and information criterion. *Bull. Inform. Cybernet.* **40**, 75–87.

[13] Huang, J., Smola, A., Gretton, A., Borgwardt, K. M. and Schölkopf, B. (2007). Correcting sample selection bias by unlabeled data. *Adv. Neural Inform. Process. Syst.* **19**, 601–608.

[14] Huber, P. (2004). *Robust Statistics.* New York: Wiley.

[15] Irizarry, R. A. (2001). Information and posterior probability criteria for model selection in local likelihood estimation. *J. Am. Stat. Assoc.* **96**, 303–315.

[16] Jiang, J. and Zhai C-X. (2007). Instance weighting for domain adaptation in NLP. *Proc. 45th Annu. Meet. Assoc. Comput. Linguist.* 264–271.

[17] Kanamori, T., Hido, S. and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* **10**, 1391–1445.

[18] Kawakita, M. and Kanamori, T. (2012). Semi-supervised learning with density-ratio estimation. Preprint, arXiv:1204.3965.

[19] Kawano, S. and Konishi, S. (2011). Semi-supervised logistic discrimination via regularized Gaussian basis expansions. *Comm. Statist. Theory Methods* **40**, 2412–2423.

[20] Kawano, S. Misumi, T. and Konishi, S. (2012). Semi-supervised logistic discrimination via graph-based regularization. To appear in *Neural Process. Lett.*.

[21] Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

[22] Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York: Springer.

[23] Lafferty, J. and Wasserman, L. (2007). Statistical analysis of semi-supervised regression. *Adv. Neural Inform. Process. Syst.* **21**, 801–808.

[24] Liang, F., Mukherjee, S. and West, M. (2007). The use of unlabeled data in predictive modeling. *Statist. Sci.* **22**, 189–205.

[25] Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

[26] Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90**, 227–244.

[27] Sigillito, V. G., Wing, S. P., Hutton, L. V. and Baker, K. B. (1989). Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Tech. Digest* **10**, 262–266.

[28] Sokolovska, N., Cappé, O. and Yvon, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. *Proc. 25th Int. Conf. Mach. Learn.*, 984–991.

[29] Sugiyama, M. and Kawanabe, M. (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. Cambridge, MA: MIT Press.

[30] Sugiyama, M., Suzuki, T. and Kanamori, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge: Cambridge University Press.

[31] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P. and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Ann. Inst. Statist. Math.* **60**, 699–746.

[32] Vittaut, J-N., Amini, M-R. and Gallinari, P. (2002). Learning classification with both labeled and unlabeled data. *Proc. 13th Eur. Conf. Mach. Learn.* 468–479.

[33] Wu, D., Lee, W. S. and Ye, N. (2009). Domain adaptive bootstrapping for names entity recognition. *Proc. 2009 Conf. Empir. Methods Nat. Lang. Process.* 1523–1532.

[34] Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proc. 21th Int. Conf. Mach. Learn.*, 114–121.

[35] Zhou, D., Bousquet, O., Lal, T. N., Weston, J. and Schölkopf, B. (2004). Learning with local and global consistency. *Adv. Neural Inform. Process. Syst.* **16**, 321–328.

[36] Zhu, X. (2008). Semi-supervied learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin-Madison.

[37] Zou, H., Zhu, J., Rosset, S. and Hastie, T. (2007). Automatic bias correction methods in semi-supervised learning. *Contemp. Math.* **443**, 165–175.