# Semi-supervised logistic discrimination via regularized Gaussian basis expansions

Kawano, Shuichi
Graduate School of Mathematics, Kyushu University

Konishi, Sadanori
Faculty of Mathematics, Kyushu University



出版情報：MI Preprint Series. 2009-20, 2011-04-13. Taylor & Francis Group, LLC
バージョン：

MI Preprint Series
Kyushu University
The Global COE Program
Math-for-Industry Education & Research Hub

# Semi-supervised logistic discrimination via regularized Gaussian basis expansions

S. Kawano & S. Konishi

MI 2009-20

( Received June 17, 2009 )

Faculty of Mathematics
Kyushu University
Fukuoka, JAPAN

# Semi-supervised logistic discrimination via regularized Gaussian basis expansions

Shuichi Kawano[1] and Sadanori Konishi[2]

[1] *Graduate School of Mathematics, Kyushu University,*
*6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.*

[2] *Faculty of Mathematics, Kyushu University,*
*6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan.*

s-kawano@math.kyushu-u.ac.jp konishi@math.kyushu-u.ac.jp

**Abstract:** The problem of constructing classification methods based on both classified and unclassified data sets is considered for analyzing data with complex structures. We introduce a semi-supervised logistic discriminant model with Gaussian basis expansions. Unknown parameters included in the logistic model are estimated by regularization method along with the technique of EM algorithm. For selection of adjusted parameters, we derive a model selection criterion from Bayesian viewpoints. Numerical studies are conducted to investigate the effectiveness of our proposed modeling procedures.

**Key Words and Phrases:** Bayesian approach, EM algorithm, Logistic regression, Regularization, Semi-supervised learning.

## 1 Introduction

The classification or discrimination method is one of the most useful statistical tools in various fields of research, including engineering, artificial intelligence and life science (see, e.g., Bishop, 2006; Hastie *et al.*, 2009). In practical situations such as medical diagnosis, classifying data sets may require expensive tests or human efforts, and hence only small classified data sets may be available, whereas unclassified data sets can be easily obtained. Also, for the problem of prediction of protein function, we have known functions of some proteins through several biological experiments, while those of the others are unknown because of the demanding experimental cost and effort. Under these circum-

stances, a classification method that combines both classified and unclassified samples, called semi-supervised learning, has received considerable attention in the recent statistical and machine learning literature (Chapelle *et al.*, 2006).

Various model approaches have been taken to exploit information from the sets of classified and unclassified data; e.g., a mixture model approach (Miller and Uyer, 1997; Dean *et al.*, 2006), a logistic discriminant model approach (Amini and Gallinari, 2002; Vittaut *et al.*, 2002), a graphical model approach (Kai *et al.*, 2004; Zhou *et al.*, 2004), a support vector machine approach (Bennett and Demiriz, 1998; Vapnik, 1998), a boosting approach (Bennett *et al.*, 2002; Chen and Wang, 2007) and so on. A logistic discriminant model approach constructs models by extending linear logistic discriminant models to cope with additional unclassified data, and unknown parameters in the model are estimated by the maximum likelihood method. This method, however, has some drawbacks. First, the estimated models cannot capture complex structures with the nonlinear decision boundaries, since the models produce only the linear decision boundaries. Second, a large number of predictors leads to unstable or infinite maximum likelihood parameter estimates and, consequently, may result in incorrect classification results.

In this article, we develop a semi-supervised nonlinear logistic model based on Gaussian basis expansions. The unknown parameters are estimated by the regularization method with the help of EM algorithm. The crucial points for model building process are the choice of the number of basis functions and the values of the regularization parameter and hyperparameter included in Gaussian basis functions. In order to select the adjusted parameters, we introduce a Bayesian type criterion for evaluating models estimated by the method of regularization according to the basic idea of Konishi *et al.* (2004). The numerical examples are conducted to investigate the effectiveness of our modeling strategies. We also applied our proposed model to a high-dimensional data set with small sample size, which is a increasing feature in many areas of contemporary statistics.

The remainder of this article is organized as follows: Section 2 describes a nonlinear logistic discrimination using Gaussian basis functions. In this section, we also provide an estimation procedure based on a regularized log-likelihood function, constructed by both

classified and unclassified samples, along with the technique of EM algorithm. Section 3 presents a model selection criterion to choose adjusted parameters in the logistic models from a Bayesian perspective. In Section 4, numerical studies are illustrated to assess the performances of proposed semi-supervised logistic discriminant models. Concluding remarks are given in Section 5.

# 2 Semi-supervised logistic discrimination with basis expansions

## 2.1 Nonlinear logistic model using Gaussian basis functions

Suppose we have $n_1$ classified observations $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n_1\}$ and $(n - n_1)$ unclassified observations $\{\boldsymbol{x}_\alpha; \alpha = n_1 + 1, \ldots, n\}$, where $\boldsymbol{x}_\alpha$ are $p$-dimensional vector of observations and $g_\alpha \in \{1, 2, \ldots, L\}$ indicates the class label to which $\boldsymbol{x}_\alpha$ belongs. Let $\Pr(g_\alpha = k|\boldsymbol{x}_\alpha)$ $(k = 1, \ldots, L)$ be posterior probabilities that $\boldsymbol{x}_\alpha$ belongs to the class $k$. Using the posterior probabilities and $n_1$ classified observations, we construct a nonlinear logistic model in the following:

$$\log\left\{\frac{\Pr(g_\alpha = k|\boldsymbol{x}_\alpha)}{\Pr(g_\alpha = L|\boldsymbol{x}_\alpha)}\right\} = w_{k0} + \sum_{j=1}^{m} w_{kj}\phi_j(\boldsymbol{x}_\alpha) = \boldsymbol{w}_k^T\boldsymbol{\phi}(\boldsymbol{x}_\alpha), \quad k = 1, \ldots, L-1, \qquad (1)$$

where $\boldsymbol{w}_k = (w_{k0}, w_{k1}, \ldots, w_{km})^T$ is an unknown parameter vector for class $k$ and $\boldsymbol{\phi}(\boldsymbol{x}) = (1, \phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x}))^T$ is a vector of basis functions. For basis functions $\phi_j(\boldsymbol{x})$, we use Gaussian basis functions with hyperparameter given by

$$\phi_j(\boldsymbol{x}; \boldsymbol{\mu}_j, h_j^2, \nu) = \exp\left(-\frac{||\boldsymbol{x} - \boldsymbol{\mu}_j||^2}{2\nu h_j^2}\right), \quad (j = 1, \ldots, m), \qquad (2)$$

where $\boldsymbol{\mu}_j$ is a $p$-dimensional vector that determines the center of the basis function, $h_j^2$ is the width parameter and $\nu$ $(> 0)$ is hyperparameter. The hyperparameter $\nu$ plays a key role in adjusting the smoothness of the decision boundary (for details, Ando and Konishi, 2009).

The centers $\boldsymbol{\mu}_j$ and width parameters $h_j^2$ included in Gaussian basis functions in Equation (2) are generally determined by using the $k$-means clustering algorithm (Moody and Darken, 1989). Using this algorithm, we assign a set of observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ into

$m$ clusters $\{C_1, \ldots, C_m\}$ corresponding to the number of basis functions. The centers $\boldsymbol{\mu}_j$ and the width parameters $h_j^2$ are, respectively, determined by $\hat{\boldsymbol{\mu}}_j = \sum_{\boldsymbol{x}_\alpha \in C_j} \boldsymbol{x}_\alpha / n_j$ and $\hat{h}_j^2 = \sum_{\boldsymbol{x}_\alpha \in C_j} ||\boldsymbol{x}_\alpha - \hat{\boldsymbol{\mu}}_j||^2 / n_j$, where $n_j$ is the number of observations that belongs to the $j$-th cluster $C_j$. Replacing $\boldsymbol{\mu}_j$ with $\hat{\boldsymbol{\mu}}_j$ and $h_j^2$ with $\hat{h}_j^2$, we obtain a set of $m$ basis functions given by

$$\phi_j(\boldsymbol{x}; \hat{\boldsymbol{\mu}}_j, \hat{h}_j^2, \nu) = \exp\left(-\frac{||\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j||^2}{2\nu \hat{h}_j^2}\right), \quad j = 1, \ldots, m. \tag{3}$$

The hyperparameter $\nu$ is determined by a model selection criterion given in Section 3.

From Equation (1) the posterior probability can be rewritten as

$$\Pr(g_\alpha = k | \boldsymbol{x}_\alpha) = \frac{\exp\{\boldsymbol{w}_k^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}{1 + \sum_{j=1}^{L-1} \exp\{\boldsymbol{w}_j^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}, \quad k = 1, \ldots, L-1, \tag{4}$$

$$\Pr(g_\alpha = L | \boldsymbol{x}_\alpha) = 1 - \sum_{k=1}^{L-1} \Pr(g_\alpha = k | \boldsymbol{x}_\alpha) = \frac{1}{1 + \sum_{j=1}^{L-1} \exp\{\boldsymbol{w}_j^T \boldsymbol{\phi}(\boldsymbol{x}_\alpha)\}}. \tag{5}$$

Henceforth, we set $\Pr(g_\alpha = k | \boldsymbol{x}_\alpha)$ as $\pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})$, since the posterior probabilities depend on the parameter vector $\boldsymbol{w} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_{L-1}^T)^T$.

For $n_1$ classified observations $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n_1\}$, we introduce an $(L-1)$-dimensional response variable $\boldsymbol{y} = (y_1, \ldots, y_{L-1})^T$, the components of which are either 0 or 1. The $k$-th element of $\boldsymbol{y}_\alpha$ is set to 1 if the corresponding $\boldsymbol{x}_\alpha$ belongs to the $k$-th class, i.e.,

$$\boldsymbol{y}_\alpha = (y_1^{(\alpha)}, \ldots, y_{L-1}^{(\alpha)})^T = \begin{cases} (0, \ldots, 0, \overset{(k)}{1}, 0, \ldots, 0)^T & \text{if } \ g_\alpha = k, \quad (k = 1, \ldots, L-1), \\ (0, \ldots, 0)^T & \text{if } \ g_\alpha = L. \end{cases}$$

The response vector $\boldsymbol{y}_\alpha$ is then distributed as a multinomial distribution with the posterior probabilities $\pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})$ given by

$$f(\boldsymbol{y}_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{w}) = \prod_{k=1}^{L-1} \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})^{y_k^{(\alpha)}} \{\pi_L(\boldsymbol{x}_\alpha; \boldsymbol{w})\}^{1 - \sum_{l=1}^{L-1} y_l^{(\alpha)}}. \tag{6}$$

For $(n - n_1)$ unclassified observations $\{\boldsymbol{x}_\alpha; \alpha = n_1 + 1, \ldots, n\}$, we define a missing indicator vector as follows:

$$\boldsymbol{t}_\alpha = (t_1^{(\alpha)}, \ldots, t_{L-1}^{(\alpha)})^T = \begin{cases} (0, \ldots, 0, \overset{(k)}{1}, 0, \ldots, 0)^T & \text{if } \ \boldsymbol{x}_\alpha \text{ belongs to } k\text{-th class}, \\ (0, \ldots, 0)^T & \text{if } \ \boldsymbol{x}_\alpha \text{ belongs to } L\text{-th class}. \end{cases}$$

It is assumed that $\boldsymbol{t}_\alpha$ is distributed as the same multinomial distribution with the posterior probabilities $\pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w})$ as in (6). Hence, we have the log-likelihood function based on the $n_1$ classified and $(n - n_1)$ unclassified observations in the form

$$\ell(\boldsymbol{w}) = \sum_{\alpha=1}^{n_1} \left[ \sum_{k=1}^{L-1} y_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w}) + \left(1 - \sum_{l=1}^{L-1} y_l^{(\alpha)}\right) \log \pi_L(\boldsymbol{x}_\alpha; \boldsymbol{w}) \right] + \sum_{\alpha=n_1+1}^{n} \left[ \sum_{k=1}^{L-1} t_k^{(\alpha)} \log \pi_k(\boldsymbol{x}_\alpha; \boldsymbol{w}) + \left(1 - \sum_{l=1}^{L-1} t_l^{(\alpha)}\right) \log \pi_L(\boldsymbol{x}_\alpha; \boldsymbol{w}) \right]. \tag{7}$$

## 2.2 Estimation

The maximum likelihood estimator of an unknown parameter $\boldsymbol{w}$ can be obtained by maximizing the log-likelihood function in (7). However, the maximum likelihood method often yields unstable estimates of parameters. We thus maximize a regularized or penalized log-likelihood function as follows:

$$\ell_\lambda(\boldsymbol{w}) = \ell(\boldsymbol{w}) - \frac{n_1 \lambda}{2} \sum_{k=1}^{L-1} \boldsymbol{w}_k^T K \boldsymbol{w}_k, \tag{8}$$

where $\lambda$ $(> 0)$ is a regularization parameter that reduces the variances of the parameter estimates, $K$ is an $(m+1) \times (m+1)$ matrix given by

$$K = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & K^* \end{pmatrix}. \tag{9}$$

Here $\mathbf{0}$ is an $m$-dimensional 0 vector and $K^*$ is an $m \times m$ positive semi-definite matrix (for details, Konishi and Kitagawa, 2008). In our numerical examples in Section 4, we use an identity matrix $I_m$ as the positive semi-definite matrix $K^*$.

Amini and Gallinari (2002) proposed a log-likelihood function based on classified and unclassified data sets for linear logistic models in the context of binary classification problem. Vittaut *et al.* (2002) also provided an extension of the log-likelihood function for semi-supervised multi-class classification problem. It is noted, however, that these log-likelihood functions have been proposed in the framework of linear logistic models and the regularization method has not been applied to the log-likelihood functions.

The maximum penalized likelihood estimator $\hat{\boldsymbol{w}}$ in Equation (8) is the solution of $\partial \ell_\lambda(\boldsymbol{w}) / \partial \boldsymbol{w} = \mathbf{0}$. It is difficult to maximize the regularized log-likelihood function given

in (8), since $\boldsymbol{t}_\alpha$ is treated as an unknown missing vector. In order to overcome these problems, we employ an EM algorithm (Dempster *et al.*, 1977) with Fisher's scoring method (Green and Silverman, 1994) given as below:

**Step1** Estimate the parameter vector $\boldsymbol{w}$ through the maximization of the penalized log-likelihood function based on only classified data set $\{(\boldsymbol{x}_\alpha, g_\alpha); \alpha = 1, \ldots, n_1\}$ along with the technique of Fisher's scoring method.

**Step2** Construct a classification rule $\hat{\Pr}(g_\alpha = k|\boldsymbol{x}_\alpha) = \pi_k(\boldsymbol{x}_\alpha; \hat{\boldsymbol{w}})$.

**Step3** Using the classification rule given by Step2, calculate the posterior probabilities $\hat{\Pr}(g_\alpha = k|\boldsymbol{x}_\alpha)$ $(k = 1, \ldots, L)$ for unclassified data set $\boldsymbol{x}_\alpha$ $(\alpha = n_1 + 1, \ldots, n)$. According to the posterior probabilities, estimate $\boldsymbol{t}_\alpha$ as follows:

$$\hat{\boldsymbol{t}}_\alpha = (\hat{t}_1^{(\alpha)}, \ldots, \hat{t}_{L-1}^{(\alpha)})^T = (\hat{\Pr}(g_\alpha = 1|\boldsymbol{x}_\alpha), \ldots, \hat{\Pr}(g_\alpha = L-1|\boldsymbol{x}_\alpha))^T. \tag{10}$$

**Step4** Replace $t_k^{(\alpha)}$ into $\hat{t}_k^{(\alpha)}$ in the regularized log-likelihood function (8), and estimate the parameter vector $\boldsymbol{w}$ by maximizing the function (8) with the help of Fisher's scoring method.

**Step5** Repeat the Srep2 to the Step4 until the condition

$$|\ell_\lambda(\hat{\boldsymbol{w}}^{(k+1)}) - \ell_\lambda(\hat{\boldsymbol{w}}^{(k)})| < \varepsilon \tag{11}$$

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

Given the estimate $\hat{\boldsymbol{w}}$, a future observation $\boldsymbol{x}$ is assigned to class $k$ that has the maximum posterior probability $\pi_k(\boldsymbol{x}; \hat{\boldsymbol{w}})$ among $L$ classes, where

$$\pi_k(\boldsymbol{x}; \hat{\boldsymbol{w}}) = \frac{\exp\{\hat{\boldsymbol{w}}_k^T \boldsymbol{\phi}(\boldsymbol{x})\}}{1 + \sum_{j=1}^{L-1} \exp\{\hat{\boldsymbol{w}}_j^T \boldsymbol{\phi}(\boldsymbol{x})\}}, \quad k = 1, \ldots, L-1, \tag{12}$$

$$\pi_L(\boldsymbol{x}; \hat{\boldsymbol{w}}) = \frac{1}{1 + \sum_{j=1}^{L-1} \exp\{\hat{\boldsymbol{w}}_j^T \boldsymbol{\phi}(\boldsymbol{x})\}}. \tag{13}$$

The estimates $\hat{\boldsymbol{w}}$ depend on the number of basis functions $m$ and the values of the regularization parameter $\lambda$ and hyperparameter $\nu$. In order to select the values of these adjusted parameters, we introduce a model selection criterion for evaluating the statistical model

$$f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) = \prod_{k=1}^{L-1}\{\pi_k(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}})\}^{y_k^{(\alpha)}}\{\pi_L(\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}})\}^{1-\sum_{l=1}^{L-1}y_l^{(\alpha)}}, \tag{14}$$

which is constructed based on the classified and unclassified observations.

# 3 Model selection criterion

The Bayesian information criterion (BIC) has been proposed by Schwarz (1978) from a Bayesian viewpoint. The basic idea of the BIC is to select the model maximizing the posterior probability of candidate models. However, the BIC only covers models estimated by the maximum likelihood method. Konishi *et al.* (2004) extended the BIC such that it could be used for evaluating models estimated by the maximum penalized likelihood method, thus deriving GBIC.

Using the result given in Konishi *et al.* (2004), we obtain a criterion for evaluating the statistical model in (14) as follows:

$$\begin{aligned}\mathrm{GBIC} = &-2\sum_{\alpha=1}^{n_1}\log f(\boldsymbol{y}_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{w}}) + n_1\lambda\sum_{k=1}^{L-1}\hat{\boldsymbol{w}}_k^T K\hat{\boldsymbol{w}}_k + \log|R| \\ &- (L-1)\log|K|_+ - (L-1)(m+1-d)\log\lambda - (L-1)d\log\left(\frac{2\pi}{n_1}\right),\end{aligned} \tag{15}$$

where $|K|_+$ is the product of the positive eigenvalues of $K$ with rank $d$ and $R$ are an $(L-1)(m+1)\times(L-1)(m+1)$ matrix given by

$$R = -\frac{1}{n_1}(G\odot E)^T(G\odot E) + \frac{1}{n_1}H + \lambda I \tag{16}$$

with $E = (\Phi,\ldots,\Phi)$, $G = (\boldsymbol{\pi}_{(1)}\mathbf{1}_{m+1}^T,\ldots,\boldsymbol{\pi}_{(L-1)}\mathbf{1}_{m+1}^T)$, $H = \mathrm{diag}\{\Phi^T\mathrm{diag}\{\boldsymbol{\pi}_{(1)}\}\Phi,\ldots,$ $\Phi^T\mathrm{diag}\{\boldsymbol{\pi}_{(L-1)}\}\Phi\}$, $I = \mathrm{diag}\{K,\ldots,K\}$, $\boldsymbol{y}_{(k)} = (y_k^{(1)},\ldots,y_k^{(n_1)})^T$, $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1),\ldots,\boldsymbol{\phi}(\boldsymbol{x}_{n_1}))^T$ and $\boldsymbol{\pi}_{(k)} = (\pi_k(\boldsymbol{x}_1;\hat{\boldsymbol{w}}),\ldots,\pi_k(\boldsymbol{x}_{n_1};\hat{\boldsymbol{w}}))^T$. Here the operator $\odot$ indicates the Hadamard product, which means the elementwise product of matrices.

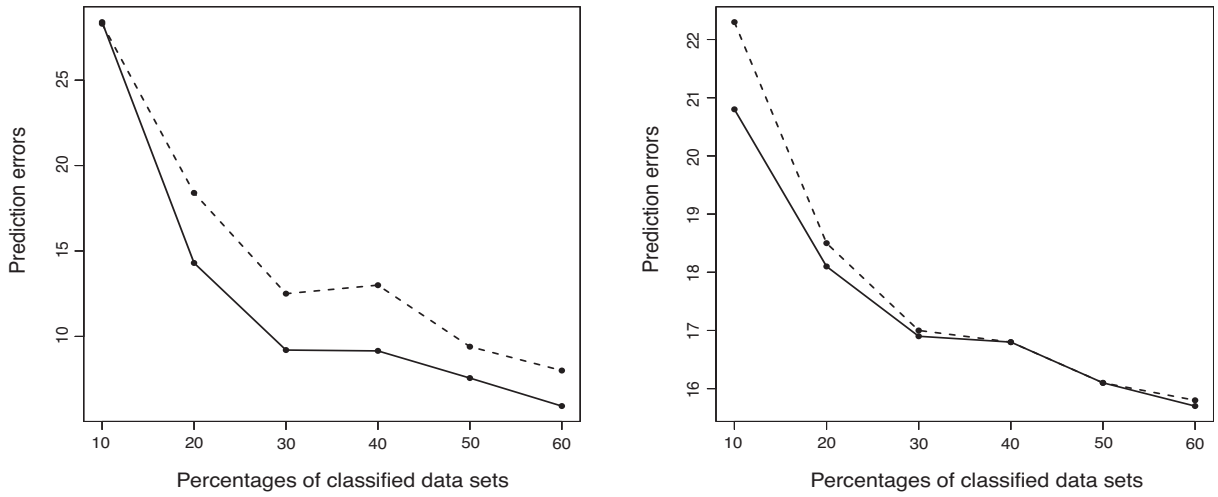


Figure 1: Performances of SSLDA (solid line) and SLDA (dashed line) for different percentages of classified data in the training sets. The left panel shows the result for ionosphere data, while the right panel shows that for waveform data.

We select the adjusted parameters including the number of basis functions and the values of the regularization parameter and the hyperparameter by minimizing the GBIC in Equation (15).

# 4 Numerical studies

In this section, our proposed semi-supervised logistic discrimination is applied to several data sets including high-dimensional and low-sample size data. These data sets are available from the UCI machine learning repository: *http://archive.ics.uci.edu/ml*.

## 4.1 Benchmark data sets

We investigate the performance of our proposed modeling procedure by analyzing ionosphere data (Sigillito *et al.*, 1989) and waveform data (Hastie *et al.,* 2009). The ionosphere data set consists of two classes with 33 predictors, and we prepared 150 sets of training data for each class and 201 sets of test data.

The waveform data consist of three classes with 21-dimensional predictors, and were

Table 1: Comparisons of prediction error rates with different percentages of classified data in the training sets for ionosphere data. Figures in parentheses indicate the values of tuning parameters.

| Method ＼ % | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| SSLDA | 28.4 | 14.3 | 9.20 | 9.15 | 7.56 | 5.92 |
| LLGC (0.5) | 26.8 | 17.5 | 20.6 | 18.1 | 17.3 | 16.5 |
| LLGC (0.3) | 26.8 | 17.2 | 18.0 | 17.0 | 14.8 | 15.4 |
| LLGC (0.1) | 26.8 | 17.4 | 17.1 | 16.1 | 13.7 | 14.2 |
| ILLGC (1) | 33.7 | 24.0 | 21.9 | 16.6 | 15.9 | 12.1 |
| ILLGC (0.1) | 25.1 | 16.1 | 11.5 | 8.15 | 8.60 | 6.86 |
| ILLGC (0.01) | 21.9 | 17.0 | 12.0 | 8.30 | 8.40 | 6.81 |

generated from the following functions:

$$x_k = \begin{cases} uH_1(k) + (1-u)H_2(k) + \varepsilon_k & \text{if } g = 1 \\ uH_1(k) + (1-u)H_3(k) + \varepsilon_k & \text{if } g = 2 \\ uH_2(k) + (1-u)H_3(k) + \varepsilon_k & \text{if } g = 3 \end{cases} \qquad k = 1, \ldots, 21, \tag{17}$$

where $u$ is uniform on [0,1], $\varepsilon_k$ are the standard normal variates and $H_i$ are the shifted triangular waveforms, $H_1(k) = \max\{6-|k-11|, 0\}$, $H_2(k) = H_1(k-4)$, $H_3(k) = H_1(k+4)$. We generated 300 sets of training data with equal prior probability for each class and 500 sets of test data. In order to implement a semi-supervised learning, the training data set was randomly divided into two halves with classified data sets and unclassified data sets, where classified data sets were assigned as training data sets of 10%, 20%, 30%, 40%, 50%, 60%, respectively.

We compared the performances of our proposed methodology (SSLDA: Semi-Supervised Logistic Discriminant Analysis) with those of several procedures. As for other semi-supervised learning, semi-supervised methods using graphical model approaches proposed by Kai *et al.* (2004) (ILLGC: Inductive Learning with Local and Global Consistency) and Zhou *et al.* (2004) (LLGC: Learning with Local and Global Consistency) were used. We also employed a nonlinear logistic discrimination (SLDA: Supervised Logistic Discrimi-

Table 2: Comparisons of prediction error rates with different percentages of classified data in the training sets for waveform data. Figures in parentheses indicate the values of tuning parameters.

| Method ＼ % | 10 | 20 | 30 | 40 | 50 | 60 |
|---|---|---|---|---|---|---|
| SSLDA | 20.8 | 18.1 | 16.9 | 16.8 | 16.1 | 15.7 |
| LLGC (0.5) | 33.7 | 34.3 | 28.1 | 28.1 | 29.7 | 26.3 |
| LLGC (0.3) | 31.8 | 31.8 | 26.1 | 26.4 | 27.7 | 25.0 |
| LLGC (0.1) | 29.5 | 29.8 | 25.0 | 24.6 | 26.3 | 23.6 |
| ILLGC (1) | 40.5 | 28.7 | 22.2 | 20.7 | 19.1 | 19.0 |
| ILLGC (0.1) | 28.0 | 20.5 | 19.4 | 18.6 | 17.2 | 17.9 |
| ILLGC (0.01) | 31.2 | 20.9 | 19.9 | 18.4 | 17.2 | 18.0 |

nant Analysis), which is introduced by Ando and Konishi (2009). It is noted that the SLDA method is estimated by using only classified data sets. Since the LLGC and ILLGC have a tuning parameter, we set the values of the parameter into 0.5, 0.3, 0.1 for LLGC and 1, 0.1, 0.01 for ILLGC, respectively. Results were averaged over 10 repetitions for random splits of classified data sets.

Figure 1 represents the prediction errors of SSLDA and SLDA for different ratio of classified-unclassified data in the training sets. Compared to the supervised learning (SLDA), our semi-supervised methods (SSLDA) improve the predictive accuracy in classification. Table 1 shows a summary of the prediction error rates for ionosphere data set, while Table 2 shows that for waveform data set. From these tables, our proposed models using the GBIC give lower prediction errors than other semi-supervised methods in almost situations.

## 4.2 High-dimensional data set with small sample size

We applied our proposed modeling procedure to the Gisette data set (LeCun *et al.*, 1998). This data set consists of two classes with 5000 dimensional explanatory variables, and we obtain 500 training sets and 1000 test sets. Such a situation where the dimension of

Table 3: Comparison of prediction error rates for Gisette data set. Figures in parentheses indicate the values of tuning parameters.

| Method | Prediction error rate (%) |
|---|---|
| SSLDA | 11.6 |
| ILLGC (1) | 23.5 |
| ILLGC (0.1) | 10.6 |
| ILLGC (0.01) | 11.9 |
| SLDA | 19.3 |

predictors is larger than the sample size has often arisen in recent statistical settings: e.g., microarray data analysis or image processing. To perform the semi-supervised learning, we randomly assigned 500 training data sets into 50 classified data sets and 450 unclassified data sets.

Our semi-supervised logistic model was applied to the data set with the help of regularization. Some adjusted parameters were selected by the model evaluation criterion GBIC given in Section 3. We compared the performance of our procedure with that of other methods described in Section 4.1 except for the LLGC method.

Summaries of the results are given in Table 3. We observe that the ILLGC method is superior to other methods when the value of the tuning parameter is 0.1, while the nonlinear semi-supervised logistic model based on Gaussian bases provides second higher predictive accuracy. However, in general, the values of tuning parameters for the ILLGC method should be objectively determined. In this point, our proposed semi-supervised discrimination seems to perform well in practical situations, since the values of tuning parameters in our models are automatically selected by the model selection criterion GBIC and the proposed models give a relatively lower prediction error rate.

# 5 Concluding remarks

In this article, we presented a nonlinear semi-supervised logistic model based on Gaussian

basis functions in the framework of multi-class classification problem. In order to select the values of adjusted parameters, we introduced a model selection criterion from Bayesian approaches. An advantage of the use of Gaussian bases is that models based on the basis functions are easily applied to analyze complex or high-dimensional data. Some numerical examples including the high-dimension data analysis illustrated that our modeling strategies yield lower prediction error rates than previously developed models. We believe that our semi-supervised logistic discrimination has the potential to be useful in variety fields of research: e.g., bioinformatics, text classification and webpage classification.

## References


Amini, M-R. and Gallinari, P. (2002): Semi-supervised logistic regression. *Proceedings of the 15th European Conference on Artificial Intelligence*, 390–394.

Ando, T. and Konishi, S. (2009): Nonlinear logistic discrimination via regularized radial basis functions for classifying high-dimensional data. *Annals of the Institute of Statistical Mathematics*, **61**, 331–353.

Bennett, K.P. and Demiriz, A. (1998): Semi-supervised support vector machines. *Advances in Neural Information Processing Systems*, **11**, 368–374.

Bennett, K.P., Demiriz, A. and Maclin, R. (2002): Exploiting unlabeled data in ensemble methods. *Proceedings of ACM International Conference on Knowledge Discovery and Data Mining*, 289–296.

Bishop, C.M. (2006): *Pattern Recognition and Machine Learning*. Springer, New York.

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

Chen, K. and Wang, S. (2007): Regularized boost for semi-supervised learning. *Advances in Neural Information Processing Systems*, **20**, 281–288.

Dean, N., Murphy, T.B. and Downey, G. (2006): Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society C*, **55**, 1–14.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977): Maximum likelihood from incom-

plete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

Green, P.J. and Silverman, B.W. (1994): *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, London.

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

Kai, Y., Tresp, V. and Zhou, D. (2004): Semi-supervised induction with basis functions. Technical Report. Department of Empirical Inference, Max Planck Institute for Biological Cybernetics, Tuebingen, Germany.

Konishi, S., Ando, T. and Imoto, S. (2004): Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, **91**, 27–43.

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

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998): Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278–2324.

Miller, D. and Uyar, H.S. (1997): A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in Neural Information Processing Systems*, **9**, 571–577.

Moody, J. and Darken, C.J. (1989): Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281–294.

Schwarz, G. (1978): Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

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 Technical Digest*, **10**, 262–266.

Vapnik, V. (1998): *Statistical Learning Theory*. Wiley, New York.

Vittaut, J-N., Amini, M-R. and Gallinari, P. (2002): Learning classification with both labeled and unlabeled data. *Proceedings of the 13th European Conference on Machine Learning*, 468–479.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J. and Schölkopf, B. (2004): Learning with local and global consistency. *Advances in Neural Information Processing Systems*, **16**, 321–328.

# List of MI Preprint Series, Kyushu University

**The Global COE Program**
**Math-for-Industry Education & Research Hub**

MI

MI2008-1 Takahiro ITO, Shuichi INOKUCHI & Yoshihiro MIZOGUCHI
Abstract collision systems simulated by cellular automata

MI2008-2 Eiji ONODERA
The intial value problem for a third-order dispersive flow into compact almost Hermitian manifolds

MI2008-3 Hiroaki KIDO
On isosceles sets in the 4-dimensional Euclidean space

MI2008-4 Hirofumi NOTSU
Numerical computations of cavity flow problems by a pressure stabilized characteristic-curve nite element scheme

MI2008-5 Yoshiyasu OZEKI
Torsion points of abelian varieties with values in n nite extensions over a p-adic eld

MI2008-6 Yoshiyuki TOMIYAMA
Lifting Galois representations over arbitrary number elds

MI2008-7 Takehiro HIROTSU & Setsuo TANIGUCHI
The random walk model revisited

MI2008-8 Silvia GANDY, Masaaki KANNO, Hirokazu ANAI & Kazuhiro YOKOYAMA
Optimizing a particular real root of a polynomial by a special cylindrical algebraic decomposition

MI2008-9 Kazufumi KIMOTO, Sho MATSUMOTO & Masato WAKAYAMA
Alpha-determinant cyclic modules and Jacobi polynomials

MI2008-10 Sangyeol LEE & Hiroki MASUDA
Jarque-Bera Normality Test for the Driving Lévy Process of a Discretely Observed Univariate SDE

MI2008-11 Hiroyuki CHIHARA & Eiji ONODERA
A third order dispersive flow for closed curves into almost Hermitian manifolds

MI2008-12 Takehiko KINOSHITA, Kouji HASHIMOTO and Mitsuhiro T. NAKAO
On the $L^2$ a priori error estimates to the nite element solution of elliptic problems with singular adjoint operator

MI2008-13 Jacques FARAUT and Masato WAKAYAMA
Hermitian symmetric spaces of tube type and multivariate Meixner-Pollaczek polynomials

MI2008-14 Takashi NAKAMURA
Riemann zeta-values, Euler polynomials and the best constant of Sobolev inequality

MI2008-15 Takashi NAKAMURA
Some topics related to Hurwitz-Lerch zeta functions

MI2009-1 Yasuhide FUKUMOTO
Global time evolution of viscous vortex rings

MI2009-2 Hidetoshi MATSUI & Sadanori KONISHI
Regularized functional regression modeling for functional response and predictors

MI2009-3 Hidetoshi MATSUI & Sadanori KONISHI
Variable selection for functional regression model via the $L_1$ regularization

MI2009-4 Shuichi KAWANO & Sadanori KONISHI
Nonlinear logistic discrimination via regularized Gaussian basis expansions

MI2009-5 Toshiro HIRANOUCHI & Yuichiro TAGUCHII
Flat modules and Groebner bases over truncated discrete valuation rings

MI2009-6 Kenji KAJIWARA & Yasuhiro OHTA
Bilinearization and Casorati determinant solutions to non-autonomous 1+1 dimensional discrete soliton equations

MI2009-7 Yoshiyuki KAGEI
Asymptotic behavior of solutions of the compressible Navier-Stokes equation around the plane Couette flow

MI2009-8 Shohei TATEISHI, Hidetoshi MATSUI & Sadanori KONISHI
Nonlinear regression modeling via the lasso-type regularization

MI2009-9 Takeshi TAKAISHI & Masato KIMURA
Phase eld model for mode III crack growth in two dimensional elasticity

MI2009-10 Shingo SAITO
Generalisation of Mack's formula for claims reserving with arbitrary exponents for the variance assumption

MI2009-11 Kenji KAJIWARA, Masanobu KANEKO, Atsushi NOBE & Teruhisa TSUDA
Ultradiscretization of a solvable two-dimensional chaotic map associated with the Hesse cubic curve

MI2009-12 Tetsu MASUDA
Hypergeometric $\tau$-functions of the q-Painlevé system of type $E_8^{(1)}$

MI2009-13 Hidenao IWANE, Hitoshi YANAMI, Hirokazu ANAI & Kazuhiro YOKOYAMA
A Practical Implementation of a Symbolic-Numeric Cylindrical Algebraic Decomposition for Quanti er Elimination

MI2009-14 Yasunori MAEKAWA
On Gaussian decay estimates of solutions to some linear elliptic equations and its applications

MI2009-15 Yuya ISHIHARA & Yoshiyuki KAGEI
Large time behavior of the semigroup on $L^p$ spaces associated with the linearized compressible Navier-Stokes equation in a cylindrical domain

MI2009-16 Chikashi ARITA, Atsuo KUNIBA, Kazumitsu SAKAI & Tsuyoshi SAWABE
Spectrum in multi-species asymmetric simple exclusion process on a ring

MI2009-17 Masato WAKAYAMA & Keitaro YAMAMOTO
Non-linear algebraic di erential equations satis ed by certain family of elliptic functions

MI2009-18 Me Me NAING & Yasuhide FUKUMOTO
Local Instability of an Elliptical Flow Subjected to a Coriolis Force

MI2009-19 Mitsunori KAYANO & Sadanori KONISHI
Sparse functional principal component analysis via regularized basis expansions and its application

MI2009-20 Shuichi KAWANO & Sadanori KONISHI
Semi-supervised logistic discrimination via regularized Gaussian basis expansions