Sparse Bayesian Logistic Regression with Hierarchical Prior and Variational Inference

Shunsuke Horii
Waseda University
s.horii@aoni.waseda.jp

Abstract

In this paper, we present a hierarchical model which assumes the logistic regression function as the observation model and assumes hierarchical priors which promote sparsity of the estimated parameters. We also develop an inference algorithm based on the variational method. The effectiveness of the proposed algorithm is validated through some experiments on both synthetic and real-world data.

1 Introduction

Learning from high-dimensional data is one of the emerging tasks in machine learning research. There are many researches that address the problem based on optimization methods such as LASSO [1]. LASSO can be considered as a method of finding the maximum a posterior (MAP) estimate of the parameter assuming that the prior distribution is the Laplace distribution. For some Bayesian decision-making problems, such as sequential experimental design, the full information of the posterior distribution is useful. There are some approaches to obtain an approximate posterior distribution for the sparse linear model. Relevance vector machine (RVM) assumes a Gaussian prior for the unknown parameters and try to find the maximum likelihood or MAP estimates for the scale parameters of the Gaussian [2]. Since some estimators for the scale parameters go to infinity, it results in a sparse estimator for the unknown parameters. Another approach is the Majorize Minimization (MM) approach. In this approach, sparsity inducing priors, such as the Laplace distribution and the Student's t distribution, are assumed for the unknown parameters and it finds a Gaussian approximation for the posterior [3]. Finally, there is a hierarchical modeling approach, which assumes a Gaussian prior for the unknown parameters and further assumes priors for the scale parameters. For the prior distributions of the scale parameters, the exponential distribution, the gamma distribution, or more generally, the generalized inverse Gaussian distribution are used [4][5][6]. Compared to the other approaches, the hierarchical modeling approach assumes one more deeper hierarchical structure. Figueiredo proposed to treat the scale parameters as hidden variables and developed an approximation algorithm based on the EM algorithm [4]. Park and Casella derived the Gibbs sampling for the hierarchical sparse linear model [5]. In general, sampling based methods such as Gibbs sampling require long time to converge and it is hard to apply to the large scale problems. As a deterministic approximation for the hierarchical sparse linear model, Babacan et al. proposed an approximation algorithm based on the variational Bayes (VB) method [6].

While there are a lot of works that address the problem of finding the approximate posterior distribution for the sparse linear model, limited efforts have been devoted to the sparse nonlinear models. In order to apply the methods to the classification problems, sparse nonlinear models are also important. Among the works listed above, RVM can also be applied to the classification problems [2]. Figueiredo also provided how to apply the EM algorithm based method for the probit classification model [4]. Very recently, Serra et al. proposed a Bayesian logistic regression with sparsity inducing priors [7]. The methods proposed in [7] is based on the MM approach. In this paper, we propose a Bayesian logistic regression with hierarchical modeling which induces sparsity for the unknown parameters. The proposed algorithm is based on the VB and the EM algorithms and it can be considered as an extension of the Bayesian logistic regression [8] and the VB algorithm for the hierarchical sparse linear model [6].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Hierarchical Model for Sparse Logistic Regression

Let $\boldsymbol{y} = (y_1, \ldots, y_n) \in \{0, 1\}^n$ be the observation label vector and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$ be the design matrix and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ be the unknown parameter vector. We consider the following logistic regression model,

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{j=1}^{n} p(y_j|\boldsymbol{\beta}) = \prod_{j=1}^{n} \sigma(\boldsymbol{x}_j^T \boldsymbol{\beta})^{y_j} (1 - \sigma(\boldsymbol{x}_j^T \boldsymbol{\beta}))^{1-y_j}, \tag{1}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

We assume an independent Gaussian prior for the unknown parameter $\boldsymbol{\beta}$,

$$p(\boldsymbol{\beta}|\boldsymbol{\tau}) = \mathcal{N}(\boldsymbol{\beta}|0, \boldsymbol{S}_{\boldsymbol{\tau}}), \tag{3}$$

where $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)$ and $\boldsymbol{S}_{\boldsymbol{\tau}} = \mathrm{diag}(\tau_1, \ldots, \tau_p)$. Further, we assume a prior distribution $p(\boldsymbol{\tau})$ for $\boldsymbol{\tau}$, which is called mixing distribution. As in [6], in this paper, for the mixing distribution, we consider the independent generalized inverse Gaussian (GIG) distribution,

$$p(\boldsymbol{\tau}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho}) = \prod_{i=1}^{p} p(\tau_i|a_i, b_i, \rho_i) = \prod_{i=1}^{p} \frac{(a_i/b_i)^{\rho_i/2}}{2K_{\rho_i}(\sqrt{a_i b_i})} \tau_i^{\rho_i - 1} \exp\left(-\frac{1}{2}(a_i \tau_i + b_i \tau_i^{-1})\right), \tag{4}$$

where $\boldsymbol{a} = (a_1, \ldots, a_p)$, $\boldsymbol{b} = (b_1, \ldots, b_p)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)$, and K_{ρ_i} is the modified Bessel function of the second kind. As special cases, the GIG distribution coincides with the exponential distribution when $b_i = 0, \rho_i = 1$ and the inverse gamma distribution when $a_i = 0, \rho_i < 0$.

In summary, the following joint distribution is obtained.

$$p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho}) = p(\boldsymbol{y}|\boldsymbol{\beta}) p(\boldsymbol{\beta}|\boldsymbol{\tau}) p(\boldsymbol{\tau}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho}) p(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\rho}). \tag{5}$$

For the sake of the simplicity, hereafter, we consider only the case where $\boldsymbol{\rho} = \mathbf{1}, \boldsymbol{b} = \mathbf{0}$ (in this case, $p(\tau_j|a_j)$ is the exponential distribution) and assume that the hyperprior $p(\boldsymbol{a})$ is the independent gamma distribution, that is,

$$p(\boldsymbol{a}) = \prod_{i=1}^{p} \mathrm{Gam}(a_i; k_a, \theta_a) = \prod_{i=1}^{p} \frac{\theta_a^{k_a}}{\Gamma(k_a)} a_i^{k_a - 1} \exp(-\theta_a a_i). \tag{6}$$

3 Variational Inference

Given the joint distribution (5), we want to calculate posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$, however, complex integral calculation is required to find the posterior and it is very hard. In this paper, we give an approximation algorithm based on the VB method [9]. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{a})$ and the VB method finds an approximation distribution $q(\boldsymbol{\theta})$ that approximates $p(\boldsymbol{\theta}|\boldsymbol{y})$. More specifically, the goal is to find $q(\boldsymbol{\theta})$ that minimizes the Kullback-Leibler divergence $\mathrm{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y}))$:

$$q^*(\boldsymbol{\theta}) = \operatorname*{argmin}_{q(\boldsymbol{\theta})} \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} = \operatorname*{argmin}_{q(\boldsymbol{\theta})} \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \boldsymbol{y})} d\boldsymbol{\theta}. \tag{7}$$

However, it is difficult to minimize (7) for arbitrary probability distributions. First, we limit the optimization distribution to $q(\boldsymbol{\theta})$ that can be factorized as follows.

$$q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{a}) = q(\boldsymbol{\beta}) q(\boldsymbol{\tau}) q(\boldsymbol{a}). \tag{8}$$

This is known as the mean-field approximation [9]. Using this factorization, we can minimize (7) by iteratively updating each component distribution $q(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in \boldsymbol{\theta}$ as [9]

$$\ln q^*(\boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}_k)}[\ln p(\boldsymbol{y}, \boldsymbol{\theta})] + \text{const.}, \tag{9}$$

where $\boldsymbol{\theta} \setminus \boldsymbol{\theta}_k$ is the set $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_k$ removed. However, it is still difficult to calculate (9) since there is no analytic solution for $\ln q^*(\boldsymbol{\beta}) = \mathbb{E}_{q(\boldsymbol{\theta} \setminus \boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \boldsymbol{\theta})] + \text{const.}$. Then, we use the Majorize-Minimization (MM) approach which is used for Bayesian logistic regression in [8]. The idea is

Algorithm 1 Sparse Bayesian Logistic Regression

Input: design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, label vector $\boldsymbol{y} \in \{0,1\}^n$, hyper parameters $k_a > 0, \theta_a > 0$, initial variational parameter $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^n$.
Output: $\bar{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.

Set $t = 0$, Initialize $\bar{a}_j^{(0)} = k_a/\theta_a$ for $j = 1, \ldots, p$, $\boldsymbol{S}^{(0)} = \boldsymbol{I}_p$ (p-dimensional identity matrix).
repeat
  if $p > n$ **then**
    $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{S}^{(t)} - \boldsymbol{S}^{(t)}\boldsymbol{X}^T \left(\frac{1}{2}\boldsymbol{\Lambda}_{\boldsymbol{\xi}^{(t)}}^{-1} + \boldsymbol{X}\boldsymbol{S}^{(t)}\boldsymbol{X}^T\right)^{-1} \boldsymbol{X}\boldsymbol{S}^{(t)}$,
  else
    $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{U}^{(t)} \left(2\boldsymbol{U}^{(t)}\boldsymbol{X}^T\boldsymbol{\Lambda}_{\boldsymbol{\xi}^{(t)}}\boldsymbol{X}\boldsymbol{U}^{(t)} + \boldsymbol{I}_p\right)^{-1} \boldsymbol{U}^{(t)}$,
  end if
  where $\boldsymbol{\Lambda}_{\boldsymbol{\xi}^{(t)}} = \mathrm{diag}\left(\lambda(\xi_1^{(t)}), \ldots, \lambda(\xi_n^{(t)})\right)$ and $\boldsymbol{U}^{(t)} = \mathrm{diag}\left(\sqrt{S_{(1,1)}^{(t)}}, \ldots, \sqrt{S_{(p,p)}^{(t)}}\right)$.
  $\bar{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(t+1)} \left(\sum_{j=1}^n \left(y_j - \frac{1}{2}\right)\boldsymbol{x}_j\right)$.
  $\bar{\tau}_i^{(t+1)} = \frac{1+\sqrt{\bar{a}_i^{(t)}}\sqrt{\left(\bar{\beta}_i^{(t+1)}\right)^2 + \left(\Sigma_{\boldsymbol{\beta},(i,i)}^{(t+1)}\right)}}{\bar{a}_i^{(t)}}$ for $i = 1, \ldots, p$.
  $\boldsymbol{S}^{(t+1)} = \mathrm{diag}\left(\frac{\sqrt{\left(\bar{\beta}_1^{(t+1)}\right)^2 + \left(\Sigma_{\boldsymbol{\beta},(1,1)}^{(t+1)}\right)}}{\sqrt{\bar{a}_1^{(t)}}}, \ldots, \frac{\sqrt{\left(\bar{\beta}_p^{(t+1)}\right)^2 + \left(\Sigma_{\boldsymbol{\beta},(p,p)}^{(t+1)}\right)}}{\sqrt{\bar{a}_p^{(t)}}}\right)$.
  $\bar{a}_i^{(t+1)} = (k_a + 1)\left(\theta_a + \frac{\bar{\tau}_i^{(t+1)}}{2}\right)$ for $i = 1, \ldots, p$.
  $\xi_j^{(t+1)} = \sqrt{\boldsymbol{x}_j^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{(t+1)}\boldsymbol{x}_j + \left(\boldsymbol{x}_j^T\bar{\boldsymbol{\beta}}^{(t+1)}\right)^2}$ for $j = 1, \ldots, n$.
  $t = t + 1$.
until Converges

to introduce a variational parameter $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ and to find a function $h(\boldsymbol{\beta}, \boldsymbol{\xi})$, which is a quadratic function of $\boldsymbol{\beta}$ and a lower bound of $p(\boldsymbol{y}|\boldsymbol{\beta})$. Then, $h(\boldsymbol{\beta}, \boldsymbol{\xi})$ is substituted to $p(\boldsymbol{y}|\boldsymbol{\beta})$. More specifically, it is known that the following inequality holds [8].

$$p(\boldsymbol{y}|\boldsymbol{\beta}) \geq h(\boldsymbol{\beta}, \boldsymbol{\xi}) = \prod_{j=1}^n \sigma(\xi_j) \exp\left(y_j\boldsymbol{x}_j^T\boldsymbol{\beta} - \frac{\boldsymbol{x}_j^T\boldsymbol{\beta} + \xi_j}{2} - \lambda(\xi_j)((\boldsymbol{x}_j^T\boldsymbol{\beta})^2 - \xi_j^2)\right), \tag{10}$$

where

$$\lambda(\xi) = \frac{1}{2\xi}\left(\sigma(\xi) - \frac{1}{2}\right). \tag{11}$$

In this paper, we use the following strategy to find an approximate posterior distribution. First, fix the variational parameter $\boldsymbol{\xi}$ and update $q(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k \in \boldsymbol{\theta}$ as

$$\ln q^*(\boldsymbol{\theta}_k) = \mathbb{E}_{q(\boldsymbol{\theta}\setminus\boldsymbol{\theta}_k)}[\ln h(\boldsymbol{\beta}, \boldsymbol{\xi})p(\boldsymbol{\theta})] + \text{const.}. \tag{12}$$

Then we use the EM algorithm to update $\boldsymbol{\xi}$. It is updated according to the following equation.

$$\boldsymbol{\xi}^* = \underset{\boldsymbol{\xi}}{\mathrm{argmax}} \int q(\boldsymbol{\beta}) \ln h(\boldsymbol{\beta}, \boldsymbol{\xi}) d\boldsymbol{\beta}. \tag{13}$$

Since it turns out that $q^*(\boldsymbol{\beta})$ is a Gaussian, we can use the result of [8] for solving (13). The resulting algorithm is summarized in Algorithm 1. Using $\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, we can construct an approximate predictive distribution as follows [9]

$$p(y = 1|\boldsymbol{x}, \boldsymbol{X}, \boldsymbol{y}) \approx \sigma\left(\frac{\boldsymbol{x}^T\bar{\boldsymbol{\beta}}}{\sqrt{1 + \frac{\pi}{8}\boldsymbol{x}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{x}}}\right). \tag{14}$$

Table 1: MSE and prediction accuracy for synthetic data.

	MSE	Accuracy
Proposed	$\mathbf{0.1589 \pm 0.1133}$	$\mathbf{0.8195 \pm 0.0477}$
L_1	0.3974 ± 0.2939	0.7750 ± 0.0456
L_2	0.4391 ± 0.2597	0.7112 ± 0.0242

Table 2: Prediction accuracy for real world data.

	a1a	w1a	covtype
Proposed	**0.8400**	0.9757	**0.7436**
L_1	**0.8400**	0.9702	0.7412
L_2	0.8386	**0.9779**	0.7398

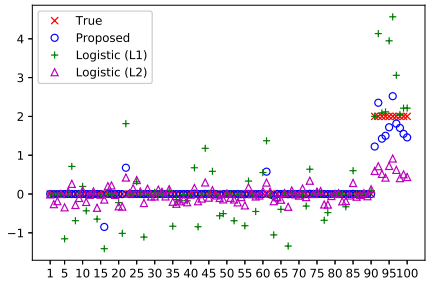

Figure 1: Estimation results on synthetic data.

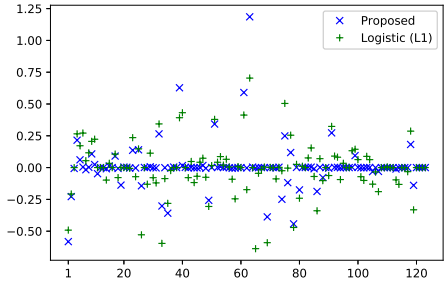

Figure 2: Estimation results on 'a1a' data.

4 Experiments

In this section, we illustrate the performance of our proposed method on synthetic data and real world data.

4.1 Experiments on Synthetic Data

We illustrate the behavior of the proposed method using a synthetic data. In this experiment, we fix the true sparse parameter $\boldsymbol{\beta}^* = (\mathbf{0}_{90}, \mathbf{2}_{10})$ and generate training samples $\boldsymbol{X}_{train}, \boldsymbol{y}_{train}$ and test samples $\boldsymbol{X}_{test}, \boldsymbol{y}_{test}$. The number of training sample is 100 and that of test sample is 1000. The elements of $\boldsymbol{X}_{train}, \boldsymbol{X}_{test}$ are generated from standard normal distribution and $\boldsymbol{y}_{train}, \boldsymbol{y}_{test}$ are generated according to the model (1) with $\boldsymbol{\beta}^*$. We examined the mean squared error (MSE) of the estimators for $\boldsymbol{\beta}^*$ and the prediction accuracy for the test data. For comparison, we compared the proposed method with the L_1 regularized logistic regression and L_2 regularized logistic regression with tuned regularization parameters through cross validation. For the proposed algorithm, we set $k_a = \theta_a = 10^{-6}$ so that the prior is almost non-informative and $\boldsymbol{\xi}^{(0)} = \mathbf{1}$. Table 1 shows the average and standard deviation of the MSE and the prediction accuracy for 1000 experiments. We can see that the proposed method shows better performance in terms of both the MSE and the prediction accuracy. Furthermore, Figure 1 is the plot of the true parameter and the estimators for one training data. We can see that the estimator of the proposed method is more sparse than the outputs of the regularized logistic regressions.

4.2 Experiments on Real World Data

We evaluate our proposed method on some real world classification tasks. As in the case of synthetic data, we compared the prediction accuracy of the proposed method with that of the L_1 regularized logistic regression and L_2 regularized logistic regression. The regularization parameters for the regularized logistic regressions are tuned through cross validation and $k_a = \theta_a = 10^{-6}$ and $\boldsymbol{\xi}^{(0)} = \mathbf{1}$ are used for the proposed algorithm. All data sets are from the LIBSVM archive. Table 2 shows the prediction accuracy for some data and it shows that the proposed method has competitive or even better performance compared with the regularized logistic regressions. Furthermore, from Figure 2, which shows the estimation results of the proposed method and L_1 regularized logistic regression on 'a1a' data, the proposed method outputs a sparse solution even for real world data. For 'a1a' data, the ratio of the number of the estimated coefficients whose absolute value is greater than 0.1 to $p = 123$ were 0.276 for the L_1 regularized logistic regression and 0.228 for the proposed method.

References

[1] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[2] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

[3] Matthias W Seeger and Hannes Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011.

[4] Mário AT Figueiredo. Adaptive sparseness for supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 25(9):1150–1159, 2003.

[5] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[6] S Derin Babacan, Shinichi Nakajima, and Minh N Do. Bayesian group-sparse modeling and variational inference. *IEEE transactions on signal processing*, 62(11):2906–2921, 2014.

[7] Juan G Serra, Pablo Ruiz, Rafael Molina, and Aggelos K Katsaggelos. Bayesian logistic regression with sparse general representation prior for multispectral image classification. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1893–1897. IEEE, 2016.

[8] T Jaakkola and M Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, volume 82, page 4, 1997.

[9] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.