

Sparse Bayesian Logistic Regression with Hierarchies

June 15, 2005

1 Introduction

Suppose the data we observe are broken down into groups; canonical examples are students in different schools or patients in different hospitals. To take advantage of this hopefully relevant information, we want to build separate models for groups. However these models should not be completely independent of each other; we want their estimates to "borrow statistical strength" from each other. A Bayesian model for this situation would be hierarchical: second level prior will squeeze together similar parameters for different groups.

2 Model

Binary regression model with group dependence looks like this:

$$P(y = 1 | \mathbf{x}, g, B) = \phi(\mathbf{x}^T \boldsymbol{\beta}_g). \quad (1)$$

where $\phi(z) = (1 + \exp(-z))^{-1}$ is logistic link function; g is the group index that the observation belongs to, $g = 1, \dots, G$; and B is the matrix of model parameters:

$$B = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_G], \quad (2)$$

where each column vector defines a model for one group.

This can be formulated slightly differently using dummy variables, which are essentially group indicators: each dummy variable takes value 1 for a case that belongs to a particular group and 0 otherwise. Now let $\mathbf{d} = (d_1, \dots, d_G)$ be the vector of dummy variables values for an observation at hand. We can rewrite (1) equivalently:

$$P(y = 1 | \mathbf{x}, \mathbf{d}, B) = \phi(\mathbf{x}^T B \mathbf{d}). \quad (3)$$

This new formulation is not limited to the case where each observation belongs to exactly one group: we can have several groupings, overlapping groups, or observations that do not belong to any group at all, just need to represent it properly with dummy variables. However there is a difference in the interpretation if we allow this to happen: we have to assume the additive effects of groups on model parameters.

3 Priors

We want to put two-level priors on model parameters: first level priors are similar in structure to priors for the non-hierarchical case; second level priors "squeeze" together parameter values of different groups corresponding to the same feature. We shall consider Gaussian and Laplace first level priors:

$$p(b_j|\sigma_1) = N(a, \sigma_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(b_j - a)^2}{2\sigma_1^2}\right), \quad (4)$$

$$p(b_j|\lambda_1) = \frac{\lambda_1}{2} \exp(-\lambda_1|b_j - a|), \quad (5)$$

and also Gaussian and Laplace second level priors:

$$p(\beta_{j,g}|\sigma_2) = N(b_j, \sigma_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(\beta_{j,g} - b_j)^2}{2\sigma_2^2}\right), \quad (6)$$

$$p(\beta_{j,g}|\lambda_2) = \frac{\lambda_2}{2} \exp(-\lambda_2|\beta_{j,g} - b_j|). \quad (7)$$

This gives rise to four possible combinations of priors on two levels. With both Gaussian priors we have a direct generalization of ridge logistic regression, and this will end up with a dense model. With both Laplace priors – a direct generalization of lasso logistic regression – the model will likely be sparse. Gaussian on the first level and Laplace on the second will yield the model where for each feature j many of $\beta_{j,1}, \dots, \beta_{j,G}$ will likely coincide with each other. It's hard at this point to characterize the model with Laplace prior on the first level and Gaussian on the second or interpret the prior belief that it expresses, but that's not to disallow this structure of priors.

4 MAP Estimation

For the time being we are considering only Gaussian-Gaussian and Laplace-Laplace prior combinations. Expressions for the log-posterior that we need to minimize are:

$$L^{ridge}(B, b_1, \dots, b_J) = l(X, Y, D, B) + 1/\sigma_1^2 \sum_j (b_j - a)^2 + 1/\sigma_2^2 \sum_{j,g} (\beta_{j,g} - b_j)^2, \quad (8)$$

$$L^{lasso}(B, b_1, \dots, b_J) = l(X, Y, D, B) + \lambda_1 \sum_j |b_j - a| + \lambda_2 \sum_{j,g} |\beta_{j,g} - b_j|, \quad (9)$$

where $l(X, Y, D, B)$ is negated loglikelihood. We need somehow to get rid of parameters b_1, \dots, b_J .

Note that the last two terms on the r.h.s. of both equations, penalty terms, do not depend on data: X , Y , or D . This allows us in both cases to minimize

the sum of those terms regarding b_j for any given value of model parameter matrix B . Namely, for each j :

$$b_j^{ridge}(B) = \arg \min_b [(b-a)^2/\sigma_1^2 + 1/\sigma_2^2 \sum_g (\beta_{j,g} - b)^2], \quad (10)$$

$$b_j^{lasso}(B) = \arg \min_b [\lambda_1 |b-a| + \lambda_2 \sum_g |\beta_{j,g} - b|]. \quad (11)$$

Looking at the equation (10), in the absence of the first term under minimum this would define the mean of $\beta_{j,1}, \dots, \beta_{j,G}$. The first term however shrinks it towards a , so taking all terms into account we can characterize b_j^{ridge} as shrunk mean. Analogously with (11), without the first term under minimum it would define the median, so we characterize b_j^{lasso} as shrunk median.

For (10) it's easy to obtain the explicit formula:

$$b_j^{ridge}(B) = \frac{a \frac{\sigma_2^2}{\sigma_1^2} + \sum_g \beta_{j,g}}{\frac{\sigma_2^2}{\sigma_1^2} + G}, \quad (12)$$

and substitute it back into (8):

$$L^{ridge}(B) \propto l(X, Y, D, B) + \frac{1}{\sigma_2^2} \sum_{j,g} \beta_{j,g}^2 - \frac{\sigma_1^2}{\sigma_2^2(\sigma_1^2 + \sigma_2^2 G)} \sum_j (\sum_g \beta_{j,g})^2 - \frac{2a}{\sigma_2^2 + \sigma_1^2 G} \sum_{j,g} \beta_{j,g}.$$

Here are the first and second derivatives for the purposes of Zhang and Oles' algorithm:

$$\frac{\partial L^{ridge}(B)}{\partial \beta_{j,g}} = \frac{\partial l(X, Y, D, B)}{\partial \beta_{j,g}} + \frac{2\beta_{j,g}}{\sigma_2^2} - \frac{2(a\sigma_2^2 + \sigma_1^2 \sum_g \beta_{j,g})}{\sigma_2^2(\sigma_2^2 + \sigma_1^2 G)} \quad (13)$$

$$\frac{\partial^2 L^{ridge}(B)}{\partial \beta_{j,g}^2} = \frac{\partial^2 l(X, Y, D, B)}{\partial \beta_{j,g}^2} + 2 \frac{\sigma_2^2 + \sigma_1^2 G - \sigma_1^2}{\sigma_2^2(\sigma_2^2 + \sigma_1^2 G)} \quad (14)$$

For (11) the same idea of eliminating b_j works, but instead of explicit formulas its implementation requires some combinatorial algorithms.