CS&SS/STAT/SOC 536: Bayesian Logistic Regression

Adrian Dobra
adobra@uw.edu

1 Bayesian Inference in Univariate Logistic Regression

We assume to have observed the n independent samples

$$\mathcal{D} = \{(y_1, x_1), \ldots, (y_n, x_n)\}.$$

The response variable Y is binary, i.e. $y_i \in \{0, 1\}$ for $i = 1, 2, \ldots, n$. The explanatory variable X can be continuous or discrete. We consider the univariate logistic regression model

$$\log \frac{P(y = 1|x)}{P(y = 0|x)} = \beta_0 + \beta_1 x. \quad (1)$$

Our model assumptions say that each y_i follows a Bernoulli distribution with probability of success $\pi_i = P(y_i = 1|x_i)$:

$$y_i \sim Ber(\pi_i).$$

Since the samples are assumed to be independent, the likelihood is:

$$L(\beta_0, \beta_1|\mathcal{D}) = \prod_{i=1}^{n} [P(y_i = 1|x_i)]^{y_i} [1 - P(y_i = 1|x_i)]^{1-y_i},$$

where

$$\pi_i = P(y_i = 1|x_i) = logit^{-1}(\beta_0 + \beta_1 x_i) \in (0, 1).$$

We assume that the logistic regression coefficients follow independent $N(0, 1)$ priors. The joint posterior distribution of β_0 and β_1 is therefore given by

$$P(\beta_0, \beta_1|\mathcal{D}) = \frac{1}{P(\mathcal{D})} \exp\left(l^*(\beta_0, \beta_1)\right), \quad (2)$$

where

$$l^*(\beta_0, \beta_1) = -\log(2\pi) - \frac{1}{2}\left(\beta_0^2 + \beta_1^2\right) + l\left(\beta_0, \beta_1|\mathcal{D}\right),$$

and

$$P(\mathcal{D}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(l^*(\beta_0, \beta_1)\right) d\beta_0 d\beta_1. \quad (3)$$

We call $P(\mathcal{D})$ the marginal likelihood associated with the univariate logistic regression (1).

The gradient of $l^*(\beta_0, \beta_1)$ is

$$\nabla l^*(\beta_0, \beta_1) = \begin{pmatrix} \frac{\partial l^*(\beta_0,\beta_1)}{\partial \beta_0} \\ \frac{\partial l^*(\beta_0,\beta_1)}{\partial \beta_1} \end{pmatrix}.$$

The Hessian matrix associated with $l^*(\beta_0, \beta_1)$ is

$$D^2 l^*(\beta_0, \beta_1) = \begin{bmatrix} \frac{\partial^2 l^*(\beta_0,\beta_1)}{\partial \beta_0^2} & \frac{\partial^2 l^*(\beta_0,\beta_1)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 l^*(\beta_0,\beta_1)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l^*(\beta_0,\beta_1)}{\partial \beta_1^2} \end{bmatrix}.$$

1.1 The Newton-Raphson Algorithm

We determine the mode of the posterior distribution (2), i.e.

$$\left(\widehat{\beta}_0, \widehat{\beta}_1\right) = \text{argmax}_{(\beta_0,\beta_1)\in\Re^2} l^*(\beta_0, \beta_1),$$

by employing the Newton-Raphson algorithm presented on slide 2 of "534montecarlo.pdf". The procedure starts with the initial values $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = (0, 0)$. At iteration k, we update our current estimate $\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$ of the mode $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ to a new estimate $\left(\beta_0^{(k)}, \beta_1^{(k)}\right)$ as follows:

$$\begin{pmatrix} \beta_0^{(k)} \\ \beta_1^{(k)} \end{pmatrix} = \begin{pmatrix} \beta_0^{(k-1)} \\ \beta_1^{(k-1)} \end{pmatrix} - \left[D^2 l^*(\beta_0^{(k-1)}, \beta_1^{(k-1)})\right]^{-1} \nabla l^*(\beta_0^{(k-1)}, \beta_1^{(k-1)}).$$

The procedure stops when the estimates of the mode do not change after performing a new update, i.e. $|\beta_0^{(k)} - \beta_0^{(k-1)}| < \epsilon$ and $|\beta_1^{(k)} - \beta_1^{(k-1)}| < \epsilon$. Here ϵ is some small positive number, e.g. 0.0001.

1.2 The Laplace Approximation

Since the integral (3) cannot be explicitly calculated, we need to approximate it numerically. We calculate the marginal likelihood $P(\mathcal{D})$ using the Laplace approximation described on slide 11 of "534montecarlo.pdf", i.e.

$$\widehat{P(\mathcal{D})} = 2\pi \exp\left(l^*(\widehat{\beta}_0, \widehat{\beta}_1)\right) \left[\det D^2 l^*(\widehat{\beta}_0, \widehat{\beta}_1)\right]^{-1/2}, \tag{4}$$

where $\left(\widehat{\beta}_0, \widehat{\beta}_1\right)$ is the mode of the posterior distribution (2). We note that you should not actually calculate $\widehat{P(\mathcal{D})}$. Instead, you should calculate the logarithm of the marginal likelihood $\log \widehat{P(\mathcal{D})}$.

1.3 The Metropolis-Hastings Algorithm

Sampling from the posterior distribution (2) can be done using the Metropolis-Hastings algorithm. The procedure starts with the initial values $\left(\beta_0^{(0)}, \beta_1^{(0)}\right) = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)$, i.e. we start right at the mode of the distribution (2). We update the current state $\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$ of the Markov chain to its next state $\left(\beta_0^{(k)}, \beta_1^{(k)}\right)$ as follows.

We generate a candidate state $\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right)$ by sampling from the bivariate normal distribution

$$N_2\left(\begin{pmatrix} \beta_0^{(k-1)} \\ \beta_1^{(k-1)} \end{pmatrix}, -\left[D^2 l^*\left(\widehat{\beta}_0, \widehat{\beta}_1\right)\right]^{-1}\right). \tag{5}$$

Note that the covariance matrix of the proposal (5) is the negative of the inverse of the Hessian matrix evaluated at the mode of (2).

We accept the move to the proposed state, i.e. we set $\left(\beta_0^{(k)}, \beta_1^{(k)}\right) = \left(\widetilde{\beta}_0, \widetilde{\beta}_1\right)$ with probability

$$\min\left\{1, \exp\left[l^*\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right) - l^*\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)\right]\right\}. \tag{6}$$

Otherwise the Markov chain stays at its current state, i.e. we set $\left(\beta_0^{(k)}, \beta_1^{(k)}\right) = \left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$. We see that the proposal distribution (5) is symmetric, i.e. the probability of proposing $\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right)$ if the chain is currently in $\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$ is equal with the probability of proposing $\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$ if the chain is currently in $\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right)$. As such, the proposal distribution (5) cancels when we calculate the acceptance probability (6).

The implementation of an iteration of the Metropolis-Hastings algorithm proceeds as follows. If the proposed state leads to an increase of l^*, i.e.

$$l^*\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right) \geq l^*\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right),$$

we accept the move to the proposed state. Otherwise, if the proposed state leads to a decrease in l^*, i.e.

$$l^*\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right) < l^*\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right),$$

we sample u from a Uniform$(0, 1)$ distribution. If

$$\log(u) \leq l^*\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right) - l^*\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right),$$

we accept the move to the proposed state. If

$$\log(u) > l^*\left(\widetilde{\beta}_0, \widetilde{\beta}_1\right) - l^*\left(\beta_0^{(k-1)}, \beta_1^{(k-1)}\right)$$

we reject the move and the chain stays at the current state.

2 Example: Selecting SNPs Using Logistic Regressions

Our running example involves 3000 SNPs (categorical variables with levels 1, 2 and 3) and a binary outcome (disease status). The dataset is split in 2982 training samples and 1988 test samples. We use the training samples to identify logistic regressions with high posterior probabilities. Since we consider that a priori all regressions are assumed to be equally likely, the posterior probability of a regression is proportional with its marginal likelihood which is calculated as described in Section 1. The most relevant SNPs will be present in the logistic regressions with the highest marginal likelihoods.

We start by identifying logistic regressions with at most three predictors. The two logistic regressions from Table 1 dominate the set of possible regressions: any other logistic regression with at most three predictors has a posterior probability very close to zero with respect to the highest posterior probability regression that involves SNP 1, SNP 2, SNP 3 and SNP 2838. This leads to the following posterior inclusion probabilities of the four SNPs that appear in these two regressions – see Table 2. The fact that SNP 1, SNP 2 and SNP 3 appear in these regressions is not surprising: these SNPs have the highest absolute correlations with the disease status! SNP 2838 appears because the *combination* of variables is what is being captured by logistic regression. SNP 2838 would not have been picked up by any one-at-a-time variable selection method. The importance of each SNP is quantified through its posterior inclusion probability which is defined as the sum of the posterior probabilities of all regressions in which this particular SNP appears as a predictor. Table 2 gives the posterior inclusion probabilities associated with the regressions from Table 2. The other SNPs have negligible posterior inclusion probabilities since they do not appear in the top logistic regressions.

		Predictor		
Regression	Posterior Probability	First	Second	Third
1	0.947	1	2	2838
2	0.053	1	2	3

Table 1: Logistic regressions with at most three predictors. The SNPs are identified by their indices ranging from 1 to 3000.

Posterior Inclusion Probability	SNP
1	1
1	2
0.947	2838
0.053	3

Table 2: Posterior inclusion probabilities of the SNPs that appear in the highest posterior probability logistic regressions with at most three variables.

Bayesian model averaging is key for constructing a classifier based on the logistic regressions from Table 2. The two regressions are weighted with respect to their posterior inclusion probabilities. In the training set, the disease status of 1983 (66.5%) samples is correctly predicted by this Bayesian classifier. The corresponding Brier score is 612.44 with a standard error of 0.778. In the test set, 1289 (64.8%) of the samples are correctly predicted. The corresponding Brier score is 414.11 with a standard error of 1.64.

What happens if we allow up to four predictors in the logistic regressions? The highest posterior probability regressions are shown in Table 3 while the corresponding posterior inclusion probabilities of the SNPs present in these regressions are shown in Table 4. The same four SNPs (1, 2, 3 and 2838) have the highest posterior inclusion probabilities, but other SNPs (4, 5, 7, 9, 10) start to appear in the top regressions. The Bayesian model averaging classifier formed with the regressions in Table 3 has the following performance. In the training data, 1983 (66.5%) samples are correctly predicted. The corresponding Brier score is 612.44 with a standard error of 0.778. In the test data, 1289 (64.8%) samples are correctly predicted. The corresponding Brier score is 414.11 with a standard error of 1.64.

The logistic regressions seem to indicate that there are three relevant SNPs: 1, 2 and 2838. SNP 3 could also be relevant although it appears in only one regression. The same SNPs appear in simpler regressions and in slightly richer regressions. Three of these SNPs could have been identified using univariate methods. The fourth SNP (2838) comes up only in combination with SNPs 1 and 2. Although the prediction accuracy is not impressive in itself, it is the same in the training and the test data. This indicates that the SNPs selected are actually statistically relevant. There does not seem to be any need to explore logistic regression models involving more regressors. It appears that the inclusion of additional covariates will not lead to an improvement in the prediction accuracy. Just for illustrative purposes, we give the posterior inclusion probabilities of the SNPs that appear in the highest posterior logistic regressions with at most five and at most six predictors — see Tables 5 and 6.

		Predictor			
Regression	Posterior Probability	First	Second	Third	Fourth
1	0.9888	1	2	3	2838
2	0.0033	1	2	7	2838
3	0.0023	1	2	10	2838
4	0.002	1	2	9	2838
5	0.0016	1	2	4	2838
6	0.0016	1	2	5	2838
7	0.0011	1	2	11	2838

Table 3: Logistic regressions with at most four predictors.

Posterior Inclusion Probability	SNP
1	1
1	2
1	2838
0.988	3

Table 4: Posterior inclusion probabilities of the SNPs that appear in the highest posterior probability logistic regressions with at most four variables. The other SNPs have negligible posterior inclusion probabilities.

Posterior Inclusion Probability	SNP
1	1
1	2
1	2838
0.983	3
0.677	10
0.295	8

Table 5: Posterior inclusion probabilities of the SNPs that appear in the highest posterior probability logistic regressions with at most five variables. The other SNPs have negligible posterior inclusion probabilities.

Posterior Inclusion Probability	SNP
1	1
1	2
1	2838
0.987	10
0.980	3
0.931	8

Table 6: Posterior inclusion probabilities of the SNPs that appear in the highest posterior probability logistic regressions with at most six variables. The other SNPs have negligible posterior inclusion probabilities.