Logistic regression with spike-and-slab priors

Wei Lu

Theme

Bayesian classification models. This project compares the logistic regression with normal priors and with spike-and-slab priors in terms of accuracy and uncertainty of the posterior distributions. The frequentist's hypothesis test is employed during the comparison. The main challenges include the Rao-Blackwellization of the latent discrete variables and the diagnostics of the binary-response model.

1 Introduction

Pumpkin seeds are rich in nutrients, and are widely consumed around the world. Effective Statistical models for pumpkin seeds will benefit agricultural industries and botanical studies. The dataset contains two types of pumpkins seeds (Çerçevelik, Ürgüp Sivrisi) and their morphological features. Previous studies applied machine learning methods and achieved accuracy rates about 87% (Koklu et al., 2021). In this project, the classification problem is approached with Bayesian methods. The aim is to investigate the prediction performance and the feature selection ability of the spike-and-slab prior.

George and McCulloch (1993) studied the selection of variables with hierarchical mixture models. The idea is to assign large variance to the non-zero coefficients but set small variance to zero coefficients. Consequently, the zero coefficients have "spike-like" posterior distributions and non-zero ones have widely spread distributions. Ishwaran and Rao (2005) further improved the model with continuous priors and the rescaling method to perform variable selections on high-dimensional problems. Since the pumpkin seeds dataset has 12 features, this project employs the discrete prior model. In addition, a logistic link function is applied to address this classification problem.

2 Analysis

2.1 Exploratory Data Analysis

The prior information is important in Bayesian inference, so the exploratory analysis is conducted on the raw data. The pumpkin seed classes are nearly balanced (13:12), so

the baseline accuracy is 52% with a dummy classifier. Most distributions are unimodal and symmetric, except for "Eccentricity", "Solidity", "Extent". Hence, normal priors on the coefficients are proper choices. From the preliminary analysis, "Major_Axis_Length", "Eccentricity", "Roundness", "Aspect_Ration", "Compactness" might be key features to classify the pumpkin seeds (appx. Figure 3).

Before modeling, data preprocessing is performed. The dataset is split into training data ($n = 2000$) and test data ($n = 500$) for the purpose of prediction. The prediction accuracy on the test data and its uncertainty are the evaluation metrics. In addition, we standardized the morphological features based on the assumptions (Ishwaran and Rao, 2005). The different value ranges caused slow-mixing issues in our model building attempts.

2.2 Bayesian Models

The normal-prior model is the basic Bayesian logistic regression model as the follows:

$$\begin{aligned}
&\beta_i \overset{i.i.d}{\sim} \mathcal{N}(0, 10), \quad (1 \le i \le 12) \\
&\theta_j = \text{Logistic}(X_j \underline{\beta}), \quad (1 \le j \le 2000) \\
&y_j \overset{i.i.d}{\sim} \text{Bern}(\theta_j)
\end{aligned}$$

where X_j is the features of the j^{th} data (y_j), and $\underline{\beta} = (\beta_1, \cdots, \beta_{12})$. The standard deviation of priors is set manually to 10. According to George and McCulloch (1993) and Ishwaran and Rao (2005), the spike-and-slab hierarchical model is as the follows:

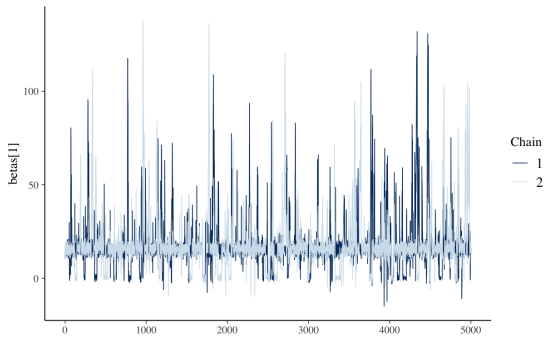
$$\begin{gathered}
\sigma^2 \sim \text{InverseGamma}(\frac{1}{2}, \frac{1}{2}), \quad \tau_i^2 \overset{i.i.d}{\sim} \text{Gamma}(2, 1), \quad p_i \overset{i.i.d}{\sim} \text{Beta}(1, 1), \\
\gamma_i | p_i \overset{i.i.d}{\sim} \text{Bern}(p_i), \quad a_i | \gamma_i, \sigma^2, \tau_i^2 = \gamma_i(v_1 \sigma \tau_i) + (1 - \gamma_i)\sigma \\
\beta_i | a_i \overset{i.i.d}{\sim} \mathcal{N}(0, a_i), \quad y_j \overset{i.i.d}{\sim} \text{Bern}(\text{logistic}(X_j \underline{\beta}))
\end{gathered}$$

The a_i serves as a "spike-and-slab" component, which provides a small variance for potential zero-valued β_i, and a large variance for non-zero β_i. The v_1 is set manually to control the "width of the slab". The zero-valued β_i's are expected to shrink to 0, while the non-zero β_i's are spread with large value. Note that γ_i acts as an indicator with binary supports, Rao-Blackwellization is needed for the model implementation in Stan. Given the independency between variables, this step is simplified as the follows:

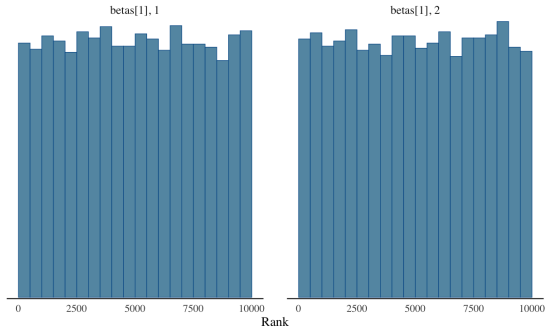
$$\begin{aligned}
\gamma(\beta_i, p_i, \tau_i, \sigma) &= \sum_{\gamma_i=0}^{1} \gamma(\beta_i, \gamma_i, p_i, \tau_i, \sigma), \quad (1 \le i \le 12) \\
&= f_{IG}(\sigma^2; \frac{1}{2}, \frac{1}{2}) f_{Gamma}(\tau_i^2; 2, 1) f_{Beta}(p_i; 1, 1) \\
&\quad \{p_i \gamma_{Norm}(\beta_i; 0, v_1 \sigma \tau_i) + (1 - p_i)\gamma_{Norm}(\beta_i; 0, \sigma)\}
\end{aligned}$$

2.3 Model Diagnostics

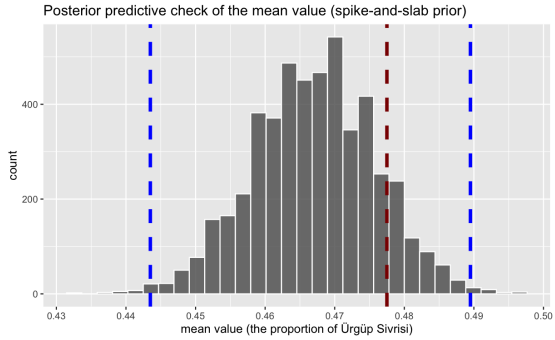
The MCMC methods are fast-mixing on both models. There is no obvious difference between chains in the trace plots (Figure 1a, appx.4a), and the rank plots are nearly uniform (Figure 1b, appx.4b). However, it is difficult to perform posterior predictive check on the binary data, since the credible set only contains 0 and 1. Gelman et al. (2000) provided some posterior check methods for discrete regressions. The simple checks for mean and standard deviation are performed. The predictions are generated with the MCMC coefficients and features of the training data. The mean and the standard deviation of the true data (brown lines) fall within the 99% credible intervals (blue lines) of the posterior statistics (Figure 1, appx.4). Therefore, both models are approximately well-specified.



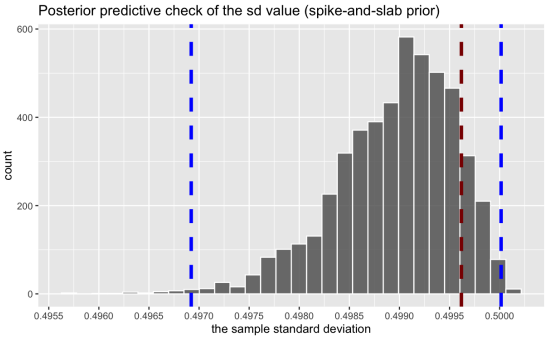
(a) trace plot of β_1



(b) rank plot of β_1



(c) predictive check for mean



(d) predictive check for sd

Figure 1: The diagnostics for the spike-and-slab model (partial)

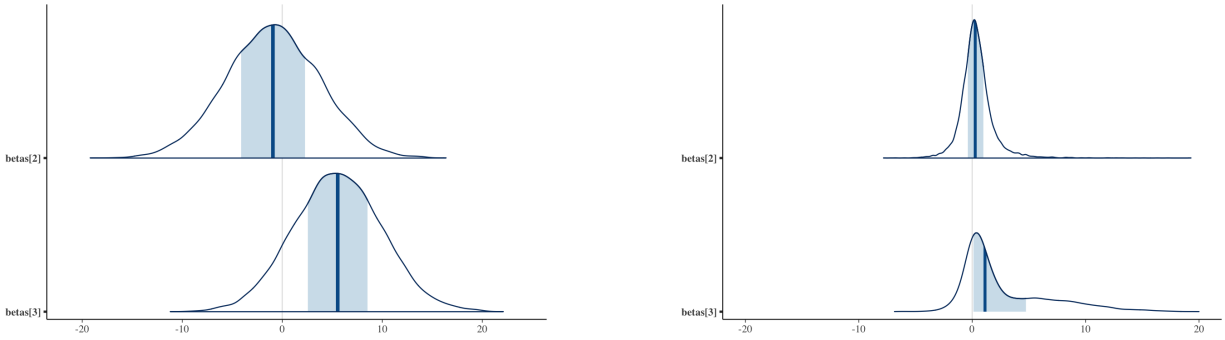
3 Results

The posterior distributions of β_2, β_3 are different between two models. In particular, the distributions of the spike-and-slab model shrink to 0 with less variance, while the distributions of the normal model have larger variance and shift from 0 (Figure 2). This discrep-

ancy implies that "Perimeter" and "Major_Axis_Length" may be irrelevant morphological features to classify the pumpkin seeds. Therefore, a "reduced" model is modified from the spike-and-slab model, which excludes "Perimeter" and "Major_Axis_Length". Furthermore, the coefficients ($\underline{\beta}$) of the three models are extracted from the MCMC to simulate predictions on the test data for each iteration, resulting in accuracy distributions. With the Kolmogorov-Smirnov test, the difference of distributions is not significant between the normal and the spike-and-slab model, but the accuracy of the reduced model is significantly improved from the spike-and-slab model with less uncertainty (Table 1).

Model	**95% CI of accuracy**	**KS p-value**
Normal	[0.7975 0.8280]	0.8367
Spike-and-slab	[0.7975 0.8285]	baseline
Reduced	[0.7980 0.8285]	0.0017

Table 1: Credible intervals for accuracy and p-values



(a) normal model

(b) spike-and-slab model

Figure 2: The posterior distributions of $\beta_{2,3}$

4 Conclusion

Both normal and spike-and-slab models have about 80% prediction accuracy with similar uncertainty on this seed classification problem. The spike-and-slab model shows the ability to select features, resulting in a simpler reduced model with similar prediction performance, but less uncertainty. Hence, the spike-and-slab model can be further explored to study the relationships between features and give guidance to the agricultural industries.

However, the involved features are relatively low-dimensional and non-sparse enough to show the power the the spike-and-slab prior. In addition, the spike-and-slab prior might be "washed away" during the MCMC iterations. Further studies might work on a higher dimensional sparse data with the rescaling methods (Ishwaran and Rao, 2005).

5 Appendix

For more visualizations and details, see github.com/weilu6/STAT447_Bayesian.

5.1 Stan code

The normal model and the spike-and-slab model

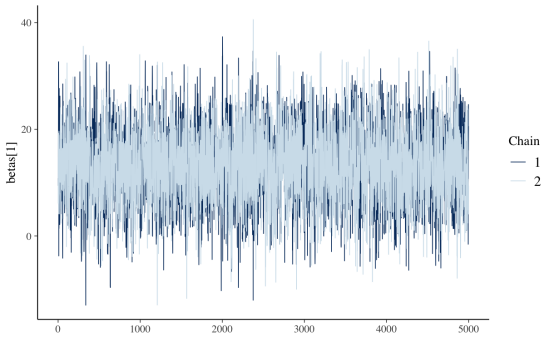
```
{stan output.var = "normal"}

data {
  int N; # number of train data
  int K; # number of features
  array[N] int y; # label of train
  matrix[N, K] X; # train features

}

parameters {
  vector[K] betas; # coefficients for features
}

model {
  betas ~ normal(0, 10);
  y ~ bernoulli_logit(X * betas);
}
```

```
{stan output.var = "sas"}
data {
  int N; # number of train data
  int K; # number of features
  array[N] int y; # label of train
  matrix[N, K] X; # train features
  real v_1; #control slab
}


parameters {
```

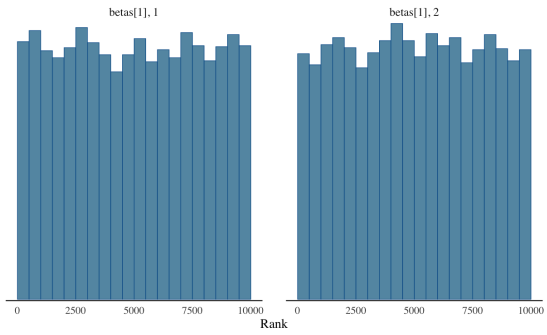
```
  real<lower=0> sigma_squared; # normal variance
  vector<lower=0>[K] tau_squared; #v ariance for each beta
  vector<lower=0.01, upper=1>[K] ps; # probability to generate 0 or 1 for indicators
  #vector[K] as; spike and slab component, no more needed since Blackwellization
  vector[K] betas; # coefficients
}

transformed parameters {
  real<lower=0> sigma = sqrt(sigma_squared); # for normal implementation, need std
  vector<lower=0>[K] tau = sqrt(tau_squared);
}

model {
  ps ~ beta(1, 1); # prior for p_i
  sigma_squared ~ inv_gamma(0.5,0.5); # prior for sigma
  tau_squared ~ gamma(2,1); # prior for tau
  #Blackwellization over Bernoulli
  for (i in 1:K) {
    target += log_sum_exp(
    log(ps[i])+gamma_lpdf(tau_squared[i]|2,1)+normal_lpdf(betas[i]|0,v_1*tau[i]*sigma),
    log(1-ps[i])+normal_lpdf(betas[i]|0,sigma));
}
  y ~ bernoulli_logit(X * betas);
}
```

5.2 Visualization

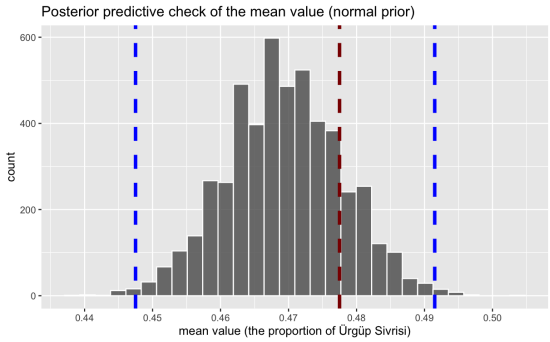
Figure 3: The histograms for morphological features (partial)



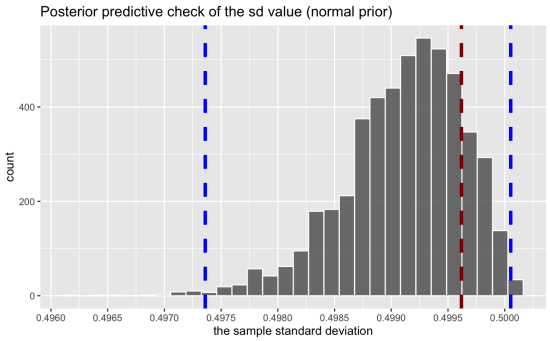
(a) trace plot of β_1



(b) rank plot of β_1



(c) predictive check for mean



(d) predictive check for sd

Figure 4: The diagnostics for the normal model (partial)

References

A. Gelman, Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen. Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied statistics*, 49(2):247–268, 2000.

Edward I. George and Robert E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Hemant Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of statistics*, 33(2):730–773, 2005.

Murat Koklu, Seyma Sarigil, and Osman Ozbek. The use of machine learning methods in classification of pumpkin seeds (cucurbita pepo l.). *Genetic resources and crop evolution*, 68(7):2713–2726, 2021.