PubH7440 - Project proposal

Zhaoliang Zhou

due 4/12/2022

1 Introduction

The Breast Cancer Wisconsin (Diagnostic) data set has been a very popular data set on the Kaggle website. For this data set, many researchers have applied many frequentist methods, machine learning algorithms, as well as deep learning methods to predict the binary diagnosis outcome of the breast cancer (benign vs. malignant). However, there has not been many Bayesian approach for this data set. Therefore, I plan to apply Bayesian binary regression for both inference and prediction using this data set and compare the predictive performance to those traditional frequentist methods.

All computation will be done in R and JAGS/WinBugs.

2 Data

The data used for this analysis can be obtained from Kaggle data set website, and the original data can be accessed via UW CS ftp server as well as the UCI Machine Learning Repository. + This data set contains unique patient ID as well as the binary outcome variable diagnosis, which indicates the status of the breast cancer of the patient (malignant vs. benign). There are total 357 patients with benign diagnosis, and 212 patients with malignant diagnosis. This data set contains features computed and extracted from digital image of a fine needle aspirate (FNA) of a breast mass. Those features describe the characteristics of the cell nuclei shown in the image. For each cell nucleus, 10 features have been computed, and those features are: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension. In addition, for each image, the mean, standard error, and the worst/largest of those features have been computed. Therefore, this data set contains total 30 features. There are no missing values for this data set.

3 Analysis and statistical methods

There are several goals for this project. First, I plan to apply Bayesian variable selection methods to extract the most important features for the diagnosis outcome. Then, I plan to use the selected features for inference for those features, and make predictions using sample splitting method to assess prediction accuracy.

3.1 Bayesian GLM - logistic regression

Since our diagnostic outcome is binary, it is natural to consider using Bayesian logistic regression.

First, for a general case, denote $y_i = (y_1, ..., y_n)^T$ to be the vector of independent binary response, and the p_i to be the probability of event (malignant). Then, we denote X_n to be the $n \times p$ design matrix, and β which is a

$p \times 1$ vector of regression coefficients. If we consider logit link, then we have the following:

$$Y_i|p_i \overset{i.i.d}{\sim} Bernoulli(p_i)\ i = 1, ..., n$$
$$logit(p_i) = log(\frac{p_i}{1-p_i}) = X\beta$$
$$p_i = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

Then, we can specify a non-informative or a weakly informative prior such as t distribution as suggested in BDA textbook. Then, we proceed with the MCMC methods to make posterior inference on this model.

3.2 Bayesian variable selection

From the data section, we can observe that we have a large number of features. Therefore, in order to reduce computational burden when fitting the model later on, some variable selection methods should be considered.

There have been many proposed Bayesian methods for variable selection. However, most of those methods require a local prior for regression coefficients in the true model. This is equivalent to have a prior on the regression coefficient that has a positive prior density function at 0. This is especially difficult in Bayesian framework as it could become difficult to differentiate models with regression coefficients close to 0 and those who do not. Therefore, to overcome this problem, Johnson and Rossell (2012) proposed using non-local prior densities, then use Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution.

Now, we make some change in notation from section 3.1 above. Here, we further denote x_i to be the i^{th} row of X_n, and model $k = (k_1, ..., k_j)$. Then, for model k, the design matrix is X_k, and the i^{th} row of X_k is then X_{ik}. Therefore, the our logistic regression model is then:

$$y_i|\beta_k \sim Bernoulli(\frac{exp(X_{ik}\beta_k)}{1 + exp(X_{ik}\beta)})$$

Then, for the non-local prior, Johnson and Rossell (2012) proposed a prior which is the product of piMOM densities, and which can be represented as follow:

$$\pi(\beta_k|\tau, r) = \frac{\tau^{rk/2}}{\Gamma(r/2)^k} \prod_{i=1}^{k} |\beta_i|^{-(r+1)} exp(-\frac{\tau}{\beta_i^2})$$

where $\tau > 0$ is a hyperparameter, and r is a shape parameter similar to the one in the Inverse Gamma distribution. Nikooienejad, Wang et al. (2016) proposed an algorithm for choosing the hyperparameter r and τ, which can be found in the Algorithm 1 of the paper. Lastly, the model k with the highest posterior probability for the data will be selected as the optimal model.

3.3 Prediction

For predictive density for future observation $\hat{y_i}$, we have the following:

$$f(\hat{Y_i} = \hat{y_i}) = \int \pi(\beta|y) f(\hat{y_i}, \beta) d\beta$$

where $\pi(\beta|y)$ is the posterior density of β, and $f(\hat{y_i}, \beta)$ is the sampling binomial density of $\hat{y_i}$ given regression coefficient β vector.

For this analysis, we plan to use sample splitting method using 20% total data as test set and the remaining for training set to calculate test set misclassification rate. In addition, ROC curve will be computed to asses predictive performance.

Reference and links

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Amir Nikooienejad, Wenyi Wang, Valen E. Johnson, Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors, Bioinformatics, Volume 32, Issue 9, 1 May 2016, Pages 1338–1345, https://doi.org/10.1093/bioinformatics/btv764

Johnson V.E. Rossell D. (2012) Bayesian model selection in high-dimensional settings. J. Am. Stat. Assoc., 107, 649–660. https://hannig.cloudapps.unc.edu/STOR757Bayes/handouts/JohnsonRussell2012.pdf