Bayesian methods for the Wisconsin Breast Cancer data set

Zhaoliang Zhou

Due: Friday 5/6/2022

1 Introduction

1.1 Motivation and background

The Breast Cancer Wisconsin (Diagnostic) data set is one of the most popular data sets 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 breast cancer (benign vs. malignant). For instance, K.P. Bennett et al[1] used this data set to evaluate and compare the performance of traditional K-means clustering and a proposed constraint version of the K-means clustering algorithm. Gavin Brown (2004)[2] used this data set to evaluate the prediction performance of using different hyperparameters for the deep learning algorithm neural network (NN). In addition, many papers and studies such as Kristin P. Bennett and Ayhan Demiriz, and Richard Maclin[3] and Nikunj C. Oza and Stuart J. Russell[4] applied ensembles methods such as bagging and boosting and used this data to evaluate predictive performance against other methods.

From the past publications and literature listed above, we could see most of the methods that have been applied to this data set are complex machine learning methods that specifically prioritize predictions and predictive performance. However, since this is a data set with clinical outcomes, clinical applicability and easy interpretation should also be considered when conducting statistical inference. Under those considerations, a Bayesian approach seems more attractive as the Bayesian framework has several advantages. For example, clinicians or physicians could incorporate domain knowledge into the modeling processing by specifying a prior, the inference is conditional on the data and is robust even when the sample size is small and without any distribution assumptions, and the results are easier to interpret as we could obtain posterior distribution for each coefficient and a probabilistic interpretation of intervals for each coefficient from the posterior credible interval.

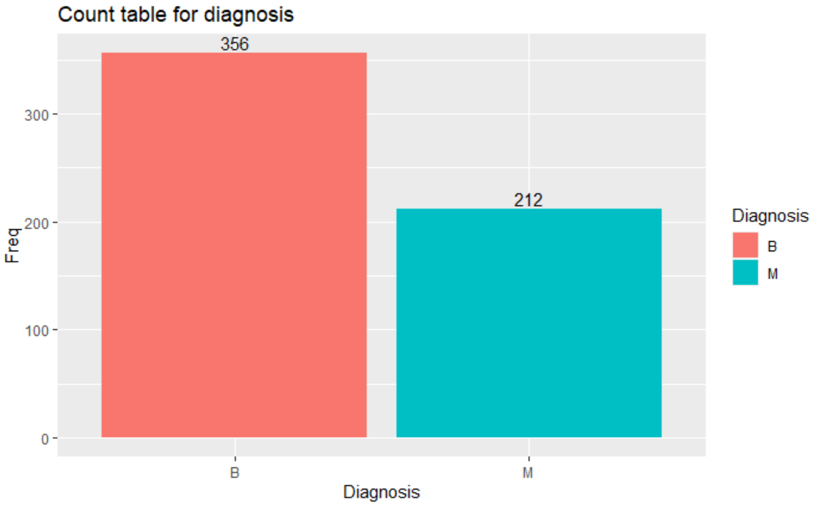
This project was initially motivated by the lack of a Bayesian approach and evaluation of performance using this dataset. In addition, this project aims to provide a Bayesian solution to the problem by first conducting Bayesian variable selection, fitting Bayesian logistic regression and inference on selected variables, making Bayesian predictions, and calculating prediction accuracy. In addition, we would compare the predictive performance of the Bayesian approach with a more traditional approach, the penalized logistic regression with LASSO.

1.2 Data

The data used for this analysis can be obtained from Kaggle data set website directly, and the original data can be accessed via UW CS ftp server as well as the UCI Machine Learning Repository. A link to this data set is included in the reference section [10].

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 569 patients in this data set. 357 patients (63%) have benign diagnosis, and 212 patients (37%) have malignant diagnosis. Figure 1 below shows graphically the distribution for each of the diagnosis outcome.

Figure 1: Distribution for the diagnosis outcome



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 [10].

1.3 Methods overview

For a brief summary of the methods, we first apply Bayesian variable selection using a non-local prior (NLP) density on the coefficients. After, we fit a Bayesian logistic regression using the diagnosis status as outcome and the selected variables as covariates. Then, we make predictions, and we will use several measurements to assess the predictive performance. Those include: training accuracy, receiver operating characteristic (ROC), and leave-one-out cross validation (loocv) to calculate classification accuracy. In addition, we will also calculate leave-one-out information criterion (looic) and expected log predictive density (ELPD). All the computations will be performed using R version 4.1.3.

2 Methods

2.1 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 [5]. Therefore, to overcome this problem, Johnson and Rossell (2012)[5] proposed using non-local prior densities, then use Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution.

Let θ be a vector of regression coefficient, ϕ a dispersion parameter, M_k a submodel, V a p by p positive and definite matrix with default to be the identical matrix, c_E the moment generating function of an inverse chi-square RV evaluated at -1, and $\tau > 0$ is a hyperparameter. From the paper by Johnson and Rossell (2012)[5], the three NLPs are in the following form:

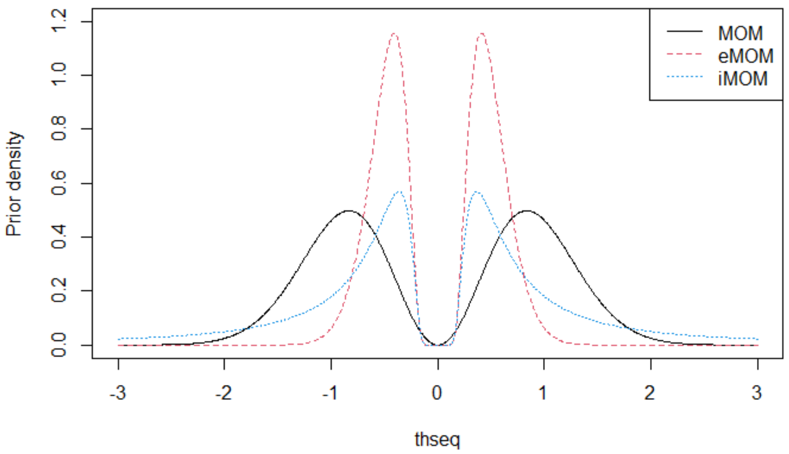
$$p_M(\theta|\phi, M_k) = \frac{\theta^{,}V\theta}{p\tau\phi} N(\theta; 0, \tau\phi V^{-1})$$

$$p_E(\theta|\phi, M_k) = c_E e^{-\frac{\tau\phi}{\theta' V\theta}} N(\theta; 0, \tau\phi V^{-1})$$

$$p_I(\theta|\phi, M_k) = \frac{\gamma(p/2)}{|V|^{1/2}(\tau\phi)^{p/2}\gamma(v/2)\pi^{p/2}} (\theta' V\theta)^{-\frac{v+p}{2}} e^{-\frac{\tau\phi}{\theta' V\theta}}$$

which are called moment (MOM), exponential moment (eMOM), and inverse moment (iMOM) priors respectively. Figure 2 provides a visualization on the densities of the 3 different NLPs.

Figure 2: NLP densities



For this project, we decided to proceed with the MOM prior as from Johnson and Rossell (2012)[5], it induces quadratic penalty as the regression coefficients getting close to 0, and it has a computational advantage of the penalty can often has a closed-form solution. More details can also be found at the vignettes for the mombf package by David Rossell [6].

The computation for Bayesian variable selection is done in R using the *modelSelection()* function in the mombf library. For the priors on the regression coefficients, we specified a MOM prior by using the *momprior()* function with the default value $tausd = 1$ which is the prior dispersion parameter for all the variables. For the prior on model space, we used a Beta-Binomial prior with parameters $\alpha = 1, \beta = 1$ which is a flat prior. We also specified 'family = binomial' to accommodate the binary outcome. We also specified the number of iterations to be 5,000, an arbitrarily large number.

In order to decrease computation time, instead of enumerate all models, we set the argument 'enumerate = F' so that we could use the Gibbs sampling method to search in the model space. Then, the function would return different sets of visited variables with highest posterior probability and the marginal posterior inclusion probabilities for each covariate which are estimated using the Rao-Blackwellization for accuracy improvement [6].

2.2 Bayesian GLM - logistic regression

Since our diagnostic outcome is binary, it is natural to consider using Bayesian logistic regression. We would proceed with the variables selected and fit a 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:

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

and the joint posterior distribution for regression coefficients is proportional to the product of the priors and N likelihood. 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.

For fitting the Bayesian logistic regression, the computation was done in R using the *stan_glm()* function from the rstanarm package. This package is convenient as it allows users to perform Bayesian estimations with relatively simple syntax, and it has many tools to visualize the posterior results and make and evaluate predictions.

For the prior density for the coefficients, we used a student t-prior by using the function *student_t(df = 7, location = 0, scale = 2.5)*. This is the default choice for the prior density for the coefficients, and this is a reasonable choice as we did not have prior confidence that coefficients would be close to 0 and at the same time allowing them to be able to get large. For the prior on the intercept, we used a $N(0, 1)$ prior as we believe it is likely to be positive or negative but less likely to be far from 0 [7].

2.3 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 [8].

For this analysis, we plan to first make predictions on the whole training data set and obtain a training classification accuracy, calculate the area under the curve (AUC), and produce a receiver operating characteristic (ROC) curve to assess the predictive performance. Then, we would use leave-one-out cross validation (loocv) to calculate a more robust classification accuracy.

To make predictions, we used the *posterior_epred()* function from the loo package, which computes posterior draws of the mean of the posterior predictive distribution. To calculate loocv misclassification error rate, we first used the function *E_loo* to calculate the LOO weighted mean of the probabilities, then we use a cutoff of 0.5 to assign predicted class to be either malignant (> 0.5) or benign (< 0.5) and test using the one observation that has been hold out.

In addition, we used the LOOIC, which is analogous to AIC, and ELPD to compare the performance of the fitted model versus an intercept model [9].

To compare with traditional frequentist approach, we also implemented a penalized logistic regression with LASSO using the glmnet package. We also calculated training accuracy, LOOCV accuracy, and AUC.

3 Results

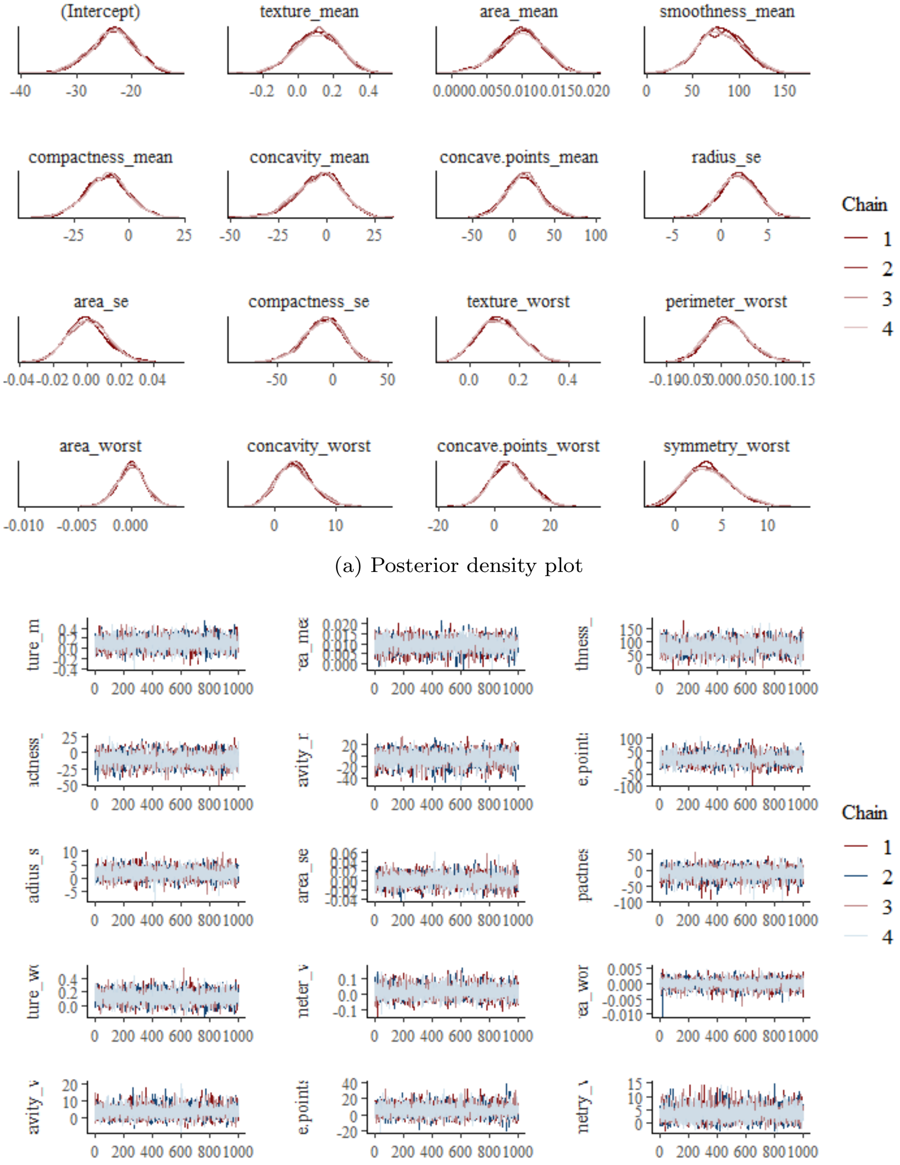
3.1 Variable selection result

After running Bayesian variable selection method as described above from the Method section, we obtained the following variables: texture mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, radius se, area se, compactness se, texture worst, perimeter worst, area worst, concavity worst, concave points worst, and symmetry worst. There are 15 variables in total after selection whereas the original data set contains 30 features.

3.2 Bayesian logistic regression result

We fit the Bayesian logistic regression using the method described above. First, it is a good idea to check model fitting and posterior convergence. Figure 3 contains visualizations for the posterior results. Figure 3 (a) is the plot of the posterior densities for each of the coefficients of the variables selected. We could see that, for each coefficient, the posterior density is approximately bell-curved with 4 chains. Figure 3 (b) is the trace plot for each of the regression coefficients. From the trace plot, we did not observe any sudden jumps or any obvious pattern. All those two plots from Figure 3 indicates the posterior densities for each coefficients converged well.

Table 1 from below shows the posterior median for each of the selected variables. We also computed 90% credible interval for each variables. Note that the variables area se and area worst have a posterior median very close to 0, which indicates that are not very predictive of the diagnosis status outcome. The variables area mean, smoothness mean, and compactness se all have a 90% that does not contain 0. This provide us evidence that those variables have stronger association with the diagnosis outcome.



(a) Posterior density plot

(b) Trace plot

Figure 3: Posterior plots

Variables	Median (90% CI)
Intercept	-23.22 (-30.21, -16.95)
Texture mean	0.11 (-0.10, 0.30)
Area mean	0.0099 (0.004, 0.015)*
Smoothness mean	79.95 (40.43, 121.27)*
Compactness mean	-10.12 (-26.86, 5.78)
Concavity mean	-3.12 (-24.22, 14.99)
Concave points mean	12.25 (-21.52, 50.12)
Radius se	1.91 (-1.25, 4.97)
Area se	0.00003 (-0.019, 0.021)
Compactness se	-7.86 (40.47, 18.37)*
Texture worst	0.12 (-0.02, 0.27)
Perimeter worst	0.01 (-0.048, 0.081)
Area worst	-0.00007 (-0.0024, 0.0021)
Concavity worst	3.14 (-1.46, 8.99)
Concave points worst	5.33 (-5.20, 17.49)
symmetry worst	3.43 (-0.45, 8.38)

Table 1: Posterior median for each coefficients with 90% credible intervals. * indicates the 90% CI does not include 0

3.3 Prediction result

For this section, we first compare the fitted logistic regression with an intercept-only model. We compare those two models using both LOOCI and ELPD [9]. Table 2 shows the LOOIC and the difference in ELPD compering the fitted model and the intercept-only mode. from Table 2, we could see that the fitted model has a lower LOOIC value thus is more preferable. The ELPD for the intercept-only model is estimated to be 307.3 lower (worse) than the fitted model. Thus, we have evidence that our fitted model is better than an intercept-only model [9].

model	LOOIC	ELPD_diff
Fitted model	137.9	
Intercept-only model	752.5	-307.3

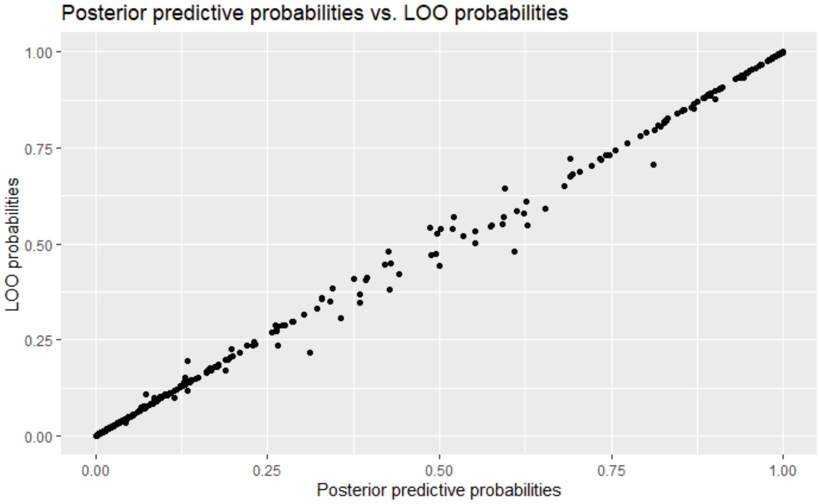
Table 2: Comparing fitted model with intercept-only model using LOOIC and ELPD

In addition, we also produced Figure 4 which is a plot of posterior predictive probabilities and against the LOO probabilities. From figure 4, we could see most of the points fall onto the diagonal straight line meaning that the posterior predictive probabilities agree with the LOO probabilities well.

Next, we fit the model using the whole data set as training data and and make prediction using the whole data set again to obtain a training accuracy (or classification accuracy) and AUC. Since the training accuracy tend to be over-optimistic as the training set and the test set are the same, we also calculated the accuracy using LOOCV. In addition, as a comparison to a frequentist method, we also implemented a popular approach, the penalized logistic regression with LASSO which could also perform variable selection and coefficient estimation. We also calculated training accuracy, LOOCV accuracy, and AUC for the penalized regression with LASSO.

The predictive performance results for the Bayesian logistic regression and the penalized logistic

Figure 4



regression with LASSO have been summarized in Table 3. From the Table 3, we can observe that the Bayesian logistic regression has a training accuracy of 0.967, LOOCV accuracy 0.961, and AUC 0.960. For the penalized logistic regression with LASSO, it has a training accuracy 0.981, LOOCV accuracy of 0.972, and AUC 0.978. The ROC curves of the Bayesian logistic regression and the LASSO can also be found in Figure 5. We could observe that the predictive performance for penalized regression using LASSO is better than the Bayesian logistic regression in terms of training accuracy, LOOCV accuracy, and AUC. However, the difference is only about 0.01 for all the measurements. Since the there was only a small difference in the performance, we would need to consider the advantage of using Bayesian approach such as easier interpretation of the result on the posterior inference. In addition, we could include clinical knowledge by specifying a prior which might be more preferred in a clinical setting. Moreover, with the loo package in R, computing LOOCV for Bayesian methods is very fast. In the contrast, if we would compute LOOCV accuracy for LASSO, it took around 5 hours to finish. Thus, we could reasonably conclude that this Bayesian logistic regression approach is comparable with the penalized logistic regression with LASSO.

model	train accuracy	LOOCV accuracy	AUC
Bayesian logistic	0.967	0.961	0.960
Penalized logistic with LASSO	0.981	0.972	0.978

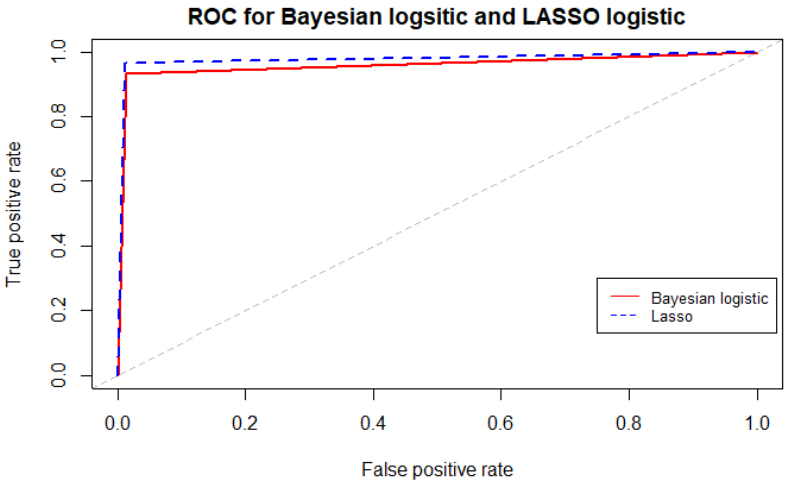
Table 3: Model comparison between Bayesian methods and LASSO

4 Discussion

4.1 Limitation and future direction

This project has several limitations, and could potentially improve in the future. First, for the variable selection, we could potentially use other NLPs and perform a sensitivity analysis and compare

Figure 5



the selected variables. Second, when fitting Bayesian logistic regression, we could potentially consider using a normal prior on the regression coefficients and compare the posterior results to the one presented here. Third, the computation for this project presented here mainly used R and various packages to perform Bayesian variable selection and fitting Bayesian logistic regression. In the future, we could consider fit the same hierarchical Bayesian logistic regression using JAGs/WinBugs just to compare to results across different computation platforms. Moreover, this project only selected LASSO due to its popularity and ability for dimension reduction. In the future, we could consider implementing other popular methods such as Ridge, Elastic net, classification tree, and even Neural Network (NN) and compare the Bayesian approach to those methods, which could provide more insights on the predictive performance of the Bayesian method.

4.2 Things learned

Through this project, I have gained more experience with Bayesian methods and fitting hierarchical Bayesian models especially when the outcome is binary. I have also learned how to perform Bayesian variable selection with NLPs and using the accompanying R package mombf. In addition, I have learned how to fit Bayesian logistic regression using R package rstanarm, produce different types of visualizations, and use different measures to assess predictive performance across different models.

5 Reference

[1] K.P. Bennett, P.S. Bradley, A. Demiriz.(2000). Constrained K-Means Clustering . Microsoft Research Dept. of Mathematical Sciences One Microsoft Way Dept. of Decision Sciences and Eng. Sys. `https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2000-65.pdf`

[2] Gavin Brown.(2004). Diversity in Neural Network Ensembles. The University of Birmingham. `http://www.cs.man.ac.uk/~gbrown/publications/gbrownThesis.pdf`

[3] K.P. Bennett, A. Demiriz, R. Maclin.(2002). Exploiting unlabeled data in ensemble methods. KDD.

[4] Nikunj C. Oza, and Stuart J. Russell. (2001). Experimental comparisons of online and batch versions of bagging and boosting. KDD.

[5] Johnson, V.E., and Rossell, D.(2012). Bayesian Model Selection in High-Dimensional Settings, Journal of the American Statistical Association, 107:498, 649-660.

[6] Rossell, D. Bayesian model selection and averaging with mombf. https://cran.r-project.org/web/packages/mombf/vi

[7] Vehtari, A., Gabry, J., and Goodrich, B., (2022). Bayesian Logistic Regression with rstanarm. https://avehtari.github.io/modelselection/diabetes.html#4_A_Bayesian_logistic_regression_model

[8] Gelman, A., Carlin, J.B., Stern, H.S., et al. (2021). Bayesian Data Analysis 3rd ed. http://www.stat.columbia.edu/~gelman/book/BDA3.pdf

[9] Alicia A. Johnson, Miles Q. Ott, Mine Dogucu. (2021) https://www.bayesrulesbook.com/chapter-11.html

[10] Breast Cancer Wisconsin (Diagnosis) Data set. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data