ISYE 6420 Final Project: Breast Cancer Classification

Enson Soo

Spring 2025

1 Introduction

In the field of medical diagnostics, accurate classification models play a critical role in identifying diseases early and effectively. For example, breast cancer classification presents a unique problem due to its imbalanced nature. Malignant cases are significantly outnumbered by benign cases - the benign to malignant ratio among breast biopsies is believed to be around 4:1 or 5:1. Traditional machine learning approaches tend to struggle with imbalanced data, as they produce biased predictions that favor the majority class (benign tumors). This leads to reduced sensitivity in detecting the minority class (malignant tumors), which can have serious medical consequences.

Frequentist methods often handle imbalanced data through techniques such as oversampling minority classes, undersampling majority classes, or adjusting class weights. Bayesian modeling approaches offer another promising solution by incorporating prior knowledge into the model parameters, which can help prevent bias and overfitting from the majority class data.

This report explores the application of Bayesian classification methods, specifically Bayesian Logistic Regression, in the context of handling an imbalanced breast cancer diagnosis dataset. By analyzing this approach, this may determine whether Bayesian models can enhance diagnostic accuracy and mitigate the adverse effects of class imbalance in classification tasks.

2 Dataset

This report used data from the Diagnostic Wisconsin Breast Cancer Database. This dataset contained 30 predictors and a binary target representing the cancer diagnosis (either malignant or benign). The predictors represented characteristics and measurements of the cell nuclei taken from a digitized image of a fine needle aspirate (FNA) of the breast mass. The features represented the mean, standard error, and ”worst” or largest (mean of the 3 largest values) of the following 10 characteristics: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. There are a total of 569 records, consisting of 357 benign and 212 malignant records, with no missing values in the predictors or target variables. In the context of this report, malignant diagnoses were considered to be the positive class and benign diagnoses were considered to be the negative class.

3 Approach

3.1 Preprocessing

To prevent data leakage during model training, the dataset was split into a training and testing dataset. This was accomplished by randomly sampling 200 records (100 positive and 100 negative) as the testing dataset, and treating the remaining 369 records as the training dataset.

To simulate the imbalanced classification problem, a subset of the training dataset (of size 150) was sampled such that 20% of the records corresponded to positive classes (and 80% corresponded to negative classes), pertaining to the 4:1 benign to malignant ratio. To reflect the same real world distribution during evaluation and testing, an imbalanced testing dataset (of size 125) was constructed by sampling in a similar manner: 20% of the records correspond to positive classes (and 80% correspond to negative classes). To evaluate the classification performance between the majority and minority classes, a balanced testing dataset (of size 150) was also constructed such that there was an equal number of positive and negative classes in the testing dataset. This hybrid evaluation approach allowed for a fair evaluation of the model performance on real world scenarios, while reducing bias in the evaluation metrics (such as precision, recall, and F1-score).

The predictors in the training dataset were standardized (subtracting the mean and dividing by standard deviation). The predictors in both testing datasets were also standardized, using the mean and standard deviation of the training dataset. This standardization process ensured that all of the features were on the same scale, preventing certain variables from dominating the training process. Because the same prior distribution is applied on each predictor coefficient, it is important that the effects are applied fairly on each predictor. Additionally, this improves numerical stability during Bayesian computations, which allows sampling methods to explore the parameter space efficiently.

3.2 Bayesian Logistic Regression

The data was modeled using a Bayesian logistic regression model. The outcome of a malignant tumor was modeled as a single Bernoulli event:

$$y_i|\beta, x_i \sim \text{Bernoulli}(p_i)$$

$$p_i = f(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})$$

$$\beta_0 \sim N(0, \sigma_0^2)$$

$$\beta_j \sim N(0, \sigma^2),\ j = 1, \ldots, k$$

where $k = 30$ is the number of predictors and $f(\cdot)$ is the logistic function $f(x) = \frac{1}{1+e^{-x}}$
In the experiments, σ_0 was set to 10, which is relatively noninformative, to allow for flexibility in the decision boundary. This is ideal because a noninformative prior on the intercept term allows the data to determine the baseline without being skewed by prior assumptions. This also allows the sampling process to explore a broader posterior distribution without too many constraints.

To observe the effects of different informative Normal priors on the model coefficients, the data was modeled using $\sigma \in [0.1, 1, 5]$. Because the σ represents the prior's standard deviation of the β coefficients, smaller σ values correspond to more informative prior distributions. By setting smaller σ

values, this places a strong prior that the β coefficients should be very close to 0. This effectively shrinks and regularizes the β coefficients, which prevents the model from overfitting to the training data.

Each σ value corresponded to its own Bayesian model, which was used to generate posterior predictions on the training dataset and both testing datasets. These predictions were used to compute the following classification evaluation metrics: accuracy, precision, recall, and F1-score.

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + FP + FN + TN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{F1-score} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned}$$

where:
TP = ”true positives” = number of correctly predicted positive classes
FP = ”false positives” = number of incorrectly predicted positive classes
FN = ”false negatives” = number of incorrectly predicted negative classes
TN = ”true negatives” = number of correctly predicted negative classes

The precision, recall, and F1-score metrics are especially important in evaluating imbalanced data because they provide deeper insights, compared to accuracy, into how well the model is able to classify the minority class. This is because accuracy also accounts for correct majority class predictions, which is biased in imbalanced datasets.

4 Results

The results of the Bayesian logistic regression experiments are displayed in the figures below:

	Testing (imbalanced)				Testing (balanced)			
σ	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
0.1	0.944	1	0.72	0.837	0.847	1	0.693	0.819
1	0.976	1	0.88	0.936	0.967	1	0.933	0.966
5	0.976	1	0.88	0.936	0.96	1	0.92	0.958

Table 1: Evaluation metrics of Bayesian logistic regression on testing datasets (imbalanced and balanced)

Table 1 displays the classification performance metrics (accuracy, precision, recall, and F1-score) of the three Bayesian logistic regression models with varying standard deviation parameters σ on both the imbalanced and balanced testing datasets. This table displays the impact of the standard deviation parameter σ in the model coefficients' prior on the testing dataset performances. The best performance was observed when $\sigma = 1$. On both the balanced and imbalanced testing datasets, this model achieved the highest accuracy, recall, precision, and F1-score. Furthermore, it can be observed that, in the balanced testing dataset, the performance slightly improved as σ decreased from 5 to 1, but

	Training (imbalanced)			
σ	Accuracy	Precision	Recall	F1-score
0.1	0.92	1	0.6	0.75
1	0.9933	1	0.967	0.983
5	1	1	1	1

Table 2: Evaluation metrics of Bayesian logistic regression on training dataset (imbalanced)

degraded when σ decreased to 0.1. In fact, the performance of the model with $\sigma = 0.1$ had the worst performance on both testing datasets, achieving significantly lower recall and F1-scores compared to the other two models.

Table 2 displays the classification performance metrics (accuracy, precision, recall, and F1-score) of the three Bayesian logistic regression models with varying standard deviation parameters σ on the training dataset. This table displays the impact of the standard deviation parameter σ in the model coefficients' prior on the training dataset performances. It can be observed that the model with $\sigma = 5$ achieved the best performance on the training dataset, as it displayed a perfect accuracy, precision, recall, and F1-score. However, it is important to also consider the model's performance on unseen testing data. It can also be observed that the model's training performance decreased as σ decreased.

5 Analysis

The results of the experiments indicate that placing priors with varying levels of informativeness on the model coefficients can influence the predictive ability of the model. Machine learning models often perform poorly on imbalanced datasets because they are trained to minimize overall error. As a result, the majority class dominates the optimization process, leading the model to favor the majority class over the minority class. This causes the model to overfit to the majority class, where the model only learns patterns that are specific to the majority class, but does not generalize to the minority class. This leads to low recall scores for the minority cases because the model struggles to classify minority classes correctly.

In traditional machine learning, regularization plays a critical role in preventing overfitting. By constraining the model's complexity, regularization prevents it from only learning majority class patterns and instead, promotes more balanced learning across classes. In Bayesian logistic regression, the informativeness of the coefficient priors act as a form of regularization. As σ decreases, this corresponds to placing increasingly informative priors on the model coefficients, where the coefficients are encouraged to shrink towards 0. This is analogous to L2 regularization in traditional machine learning models, where a penalty term is added to the loss objective function to limit coefficient growth.

The experiments demonstrated that moderately informative priors improve model performance. As σ decreased from 5 to 1, the recall increased from 0.92 to 0.933 in the balanced testing dataset. This suggests that increasing the informativeness in the prior distribution contributed to alleviating the overfitting to the majority class and helped the model to learn patterns in the minority class. This allowed the model to generalize better and correctly classify more minority cases, which directly improved its recall score without impacting its precision score.

However, overly informative priors degrade the model's performance. When σ decreased to 0.1, which represented a very informative prior, the recall dropped significantly across all datasets: 0.6 in the training dataset, 0.693 in the balanced testing dataset, and 0.72 in the imbalanced testing dataset. This occurred because highly informative priors overly constrain the model's coefficients, forcing them to 0. This leaves the intercept term to dictate the model's predictions. However, because this model becomes overly simple, it is dominated by the majority class, recreating the original overfitting issue.

6 Conclusion

The results and analysis highlight the important role of prior informativeness in Bayesian logistic regression, particularly in the context of imbalanced datasets. By acting as a form of regularization, priors influence the model's ability to generalize across majority and minority classes. The results from the experiments demonstrated that moderately informative priors enhanced the model's performance by addressing the problem of overfitting to the majority class, resulting in improved recall and F1-scores, while preserving precision scores. However, as observed in the results, it is extremely important to strike a balance between moderately informative priors and overly informative priors, as overly informative priors can degrade the model's performance and cause it to underfit the data. These findings emphasize that proper prior selection can heavily influence the predictive capabilities of Bayesian models and can serve as a promising solution for addressing class imbalance problems in critical applications, such as healthcare and diagnostics.

7 References

1. Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. Int. J. Advance Soft Compu. Appl, 5(3), 176-204.

2. Spivey, G. H., Perry, B. W., Clark, V. A., Coulson, A. H., & Coulson, W. F. (1982). Predicting the risk of cancer at the time of breast biopsy. Variation in the benign to malignant ratio. The American surgeon, 48(7), 326–332.

3. Wolberg, W., Mangasarian, O., Street, N., & Street, W. (1993). Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.