Breast Cancer Prediction Using Bayesian Logistic Regression

Michael Chang¹, Rohan J Dalpatadu¹, Dieudonne Phanord¹ and Ashok K Singh²*

¹Department of Mathematical Sciences, University of Nevada, USA
²William F Harrah College of Hotel Administration, University of Nevada, USA

*Corresponding author: Ashok K Singh, William F Harrah College of Hotel Administration, University of Nevada, Las Vegas, USA.

Submission: July 26, 2018; Published: September 25, 2018

Abstract

Prediction of breast cancer based upon several features computed for each subject is a binary classification problem. Several discriminant methods exist for this problem, some of the commonly used methods are: Decision Trees, Random Forest, Neural Network, Support Vector Machine (SVM), and Logistic Regression (LR). Except for Logistic Regression, the other listed methods are predictive in nature; LR yields an explanatory model that can also be used for prediction, and for this reason it is commonly used in many disciplines including clinical research. In this article, we demonstrate the method of Bayesian LR to predict breast cancer using the Wisconsin Diagnosis Breast Cancer (WDBC) data set available at the UCI Machine Learning Repository.

Introduction

Breast cancer is a group of diseases characterized by the uncontrolled growth and spread of abnormal cells [1]. Globally, breast cancer is the most frequently diagnosed cancer and the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths [2]. In US as well, breast cancer is the most frequent type of cancer (Figure 1). Bozorgi et al. [3] used logistic regression for the prediction of breast cancer survivability using the SEER (Surveillance, Epidemiology, and End Results) database NCI (2016) of 338,596 breast cancer patients. Salama et al. [4] compared different classifiers (decision tree, Multi-Layer Perception, Naive Bayes, Sequential Minimal Optimization, and K-Nearest neighbor) on three different data sets of breast cancer and found a hybrid of the four methods to be the best classifier. Delen et al. [5] used artificial neural networks (ANN), decision trees (DT) and logistic regression (LR) to predict breast cancer survivability using a dataset of over 200,000 cases, using 10-fold cross-validation for performance comparison. The overall accuracies of the three methods turned out to be 93.6%(ANN), 91.2%(DT), and 89.2%(LR). Peretti & Amenta [6] used logistic regression to predict breast cancer tumor on a data set with 569 cases and obtained overall accuracy of 85%. Barco et al. [7] used LR on a data set of 1254 breast cancer patients to predict high tumour burden (HTB), as defined by the presence of three or more involved nodes with macro-metastasis. Three predictors (tumour size, lymphovascular invasion and histological grade) were found to be statistically significant. LR and ANN are commonly used in many medical data classification tasks. Dreiseitl & Ohno-Machado [8] summarize the differences and similarities of these models and compare them with a few other machine learning algorithms. Van Domelen et al. [9] estimated the LR model from a Bayesian approach when the predictors are known to have errors.

Figure 1: Estimated number of new cases in US for selected cancers-2018.
ical cystectomy [10], multivariate logistic regression was used to show that the main causes of complications were anemia before surgery, weight loss, intraoperative blood loss, intra-abdominal infection. In the present article, we use the Wisconsin Diagnostic Breast Cancer Data Set of 569 observations on 32 variables [11] to predict breast cancer using the method of Bayesian LR. We provide a description of the Bayesian LR in the next section.

**Bayesian Estimation of Logistic Regression Model**

The Logistic Regression (LR) model is a special type of regression model fitted to a binary (0-1) response variable $Y$, which relates the probability that $Y$ equals 1 to a set of predictor variables:

$$ P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}} \quad (1) $$

where $X_1, \ldots, X_p$ are $p$ predictors, which can be continuous or discrete. The above model can be expressed in terms of log-odds as follows [12]:

$$ \log \left( \frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \quad (2) $$

In the frequentist approach, given the random sample, $(Y_j, X_{1j}, X_{2j}, \ldots, X_{pj}), j = 1, 2, \ldots, n$

are $n$ independent realizations of a Bernoulli experiment with probability of success $P(Y=1)$ given by (1); the model coefficients $\beta_j$ are unknown constants to be estimated from data. The likelihood function of the sample is

$$ L(\beta; Y) = \prod_{i=1}^{n} P_i^{y_i} (1 - P_i)^{1-y_i} \quad (3) $$

The LR model parameters are determined by the method of maximum likelihood estimation (MLE), which finds the $\beta$-coefficients that maximize the logarithm of the likelihood function

$$ \sum_{i=1}^{n} [y_i \log(P_i) + (1-y_i) \log(1-P_i)] \quad (4) $$

In the Bayesian approach, the model coefficients $(\beta_0, \beta_1, \ldots, \beta_p)$ are realizations of a $P$-variate random vector generated from the joint prior distribution; any prior knowledge about the $\beta$-coefficients can be incorporated in this joint prior distribution. All inferences drawn using Bayesian approach are conditional on data, and large sample theory of estimates is not needed. The conditional sample likelihood given by expression (3) is combined with the joint prior distribution of the parameters via the Bayes theorem [13] to obtain the joint posterior distribution of the model parameters, as shown below.

$$ g^*(\beta|Y) = (\prod_{i=1}^{n} P_i^{y_i} (1-P_i)^{1-y_i}) g(\beta) $$

“where $g^* (\beta|Y)$ is the joint posterior distribution” and “ $g(\beta)$ the joint posterior distribution”

“of the parameters $\beta_j$ if very little prior knowledge exists about the model parameters, we can use a vague prior. The marginal posterior distributions are numerically computed from the joint posterior distribution, and the means of these distributions are the parameter estimates. We can also obtain 95% confidence intervals of the parameters from these marginal posterior distributions. In Bayesian framework, these confidence intervals are called credible sets. In computing a credible set, it is desirable to obtain a credible set with shortest interval. The 95% highest posterior density (HPD) credible set contains only those points with largest posterior probability distribution [14]. A comparison of Bayesian and Frequentist approaches for estimation of predictive models is provided in [15-18].

**Performance Measures for Prediction of a Binary Response**

A large number of performance measures for multi-level classifiers exist in machine learning literature [19]. Commonly used performance measures of classifiers are accuracy, precision, recall and the geometric mean $F_1$ of precision and recall [20,21]. To compute these measures, we first need to calculate the 2x2 confusion matrix in Table 1 The performance measures accuracy, precision, recall and $F_1$ are calculated for each category 0 and 1 from the following formulas:

$$ \text{Accuracy} = \frac{\sum_{j=0}^{1} \sum_{k=0}^{1} C_{j,k}}{\sum_{j=0}^{1} \sum_{k=0}^{1} C_{j,k}} $$

$$ \text{Precision}_j = \frac{C_{j,j}}{\sum_{k=0}^{1} C_{j,k}} $$

$$ \text{Recall}_j = \frac{C_{j,j}}{\sum_{i=0}^{1} C_{i,j}} $$

$$ F_1 = \frac{2 \times \text{Precision}_j \times \text{Recall}_j}{(\text{Precision}_j + \text{Recall}_j)} $$

<table>
<thead>
<tr>
<th>$j$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed $Y$</td>
<td>Predicted $Y$</td>
<td>Predicted $Y$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>C_{0,0}</td>
</tr>
<tr>
<td>1</td>
<td>C_{1,1}</td>
<td>C_{1,1}</td>
</tr>
</tbody>
</table>

**Bayesian Prediction of Breast Cancer**

The data set used here is the Wisconsin Diagnostic Breast Cancer (WDBC) Data Set, which is well-known in Machine Learning literature [9]. This data set has 569 observations on 32 variables including the binary response variable “Diagnosis” which takes values $M$ (malignant) and $B$ (benign). There are 10 features are computed for each cell nucleus:

1. Radius (average distance from center to points on the perimeter).
2. Texture (standard deviation of gray-scale values).
3. Perimeter.
4. Area.
5. Smoothness (local variation in radius lengths).
6. Compactness (perimeter^2 / area - 1.0).
7. Concavity (severity of concave portions of the contour).
8. Concave points (number of concave portions of the contour).
10. Fractal dimension ("coastline approximation" - 1).

The mean, standard error, and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in a total of 30 features for each of the 569 patients. Detailed descriptions of how these features are computed can be found in [22,23]. Since 20 of the 30 predictors were computed from data, high multicollinearity is expected in this data set. This can be seen in Figure 2, which is a plot of the correlations among the predictors in the WDBC data set.

![Figure 2: Correlation plot of 30 predictors in WDBC data set.](image)

There are three common approaches for fitting a LR model when high multicollinearity exists in the data. Aguilera et al. [24] used Principal Components Analysis (PCA) to obtain independent predictors (Principal Components) and then used LR; simulated data was used in this study. Asar [25] proposed shrinkage type estimators for fitting LR models and used Monte Carlo simulation experiments to show that the shrinkage estimators perform better than the standard MLE estimator. Another simpler and more common approach is to drop predictors with high variance inflation factor (VIF) values and obtain a model in which largest VIF is 5 [26]. This is the approach taken in this article.

**Result for WDBC Data Set**

All of the analyses presented here are performed using the statistical software environment R [27]. The WDBC data set of 569 cases was first split into a 75% training set of 427 observations and 25% test set of 142 observations. The LR Model for the training set, with all 30 predictors in the model had VIF falling in the range 78 to 123630, with none of the predictors significant (Table 2); this is due to extremely high multicollinearities among the 30 predictors. After eliminating predictors with VIF>5 one by one, the final LR model was obtained (Table 3) with Texture, Area, Concavity, and Symmetry in the model. A comparison of Table 2 & 3 shows how multicollinearities affect the estimation of LR model coefficients:

1. In the LR model with all predictors, all P-values are 1 i.e., none of the predictors are significant.
2. The estimated coefficients of the final predictors in the LR model with all predictors are all negative, when these coefficients should all be positive.
3. The standard errors (SE) of the final predictors in the LR model with all predictors are orders of magnitude higher than the corresponding estimates.
4. The final LR model, which has Texture, Area, Concavity, and Symmetry as the significant predictors, does not suffer from any of the above three issues; each coefficient is positive as it should be, and each predictor is highly significant.

**Table 2: Bayesian LR model with all 30 predictors in the model fitted to the training set.**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>SE</th>
<th>z Value</th>
<th>P-Value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2968.33</td>
<td>118926.4</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Radius</td>
<td>-110.8</td>
<td>20409.25</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Texture</td>
<td>-0.43</td>
<td>1609.57</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Perimeter</td>
<td>30.78</td>
<td>4840.38</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Area</td>
<td>-1.07</td>
<td>2357.23</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Smoothness</td>
<td>262.66</td>
<td>482463.59</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Compactness</td>
<td>-4846.98</td>
<td>127885.25</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Concavity</td>
<td>-938.94</td>
<td>76622.712</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>N Concave</td>
<td>8703.04</td>
<td>1884638.69</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Symmetry</td>
<td>-619.86</td>
<td>58801.99</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Fractal Dim</td>
<td>4286.86</td>
<td>336578.33</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Radius SE | 1307.2 | 836904.03 | 0 | 1 | 6244.44
Texture SE | -36.76 | 138213.51 | 0 | 1 | 3272.97
Perimeter SE | -46.95 | 49083.59 | 0 | 1 | 1334.69
Area SE | -1.97 | 10112.03 | 0 | 1 | 649.77
Smoothness SE | 9958.43 | 606290.39 | 0 | 1 | 182.61
Compactness SE | 2104.2 | 328412.37 | 0 | 1 | 2212.24
Concavity SE | 3543.98 | 250799.37 | 0 | 1 | 1488.06
N Concave SE | 1017.04 | 131351.45 | 0 | 1 | 2677.67
Symmetry SE | -1398.05 | 316909.88 | 0 | 1 | 189.51
Fractal Dim SE | -87436.83 | 255544.67 | 0 | 1 | 1169.2
Radius worst | -17.55 | 221557.85 | 0 | 1 | 58635.27
Texture worst | 11.33 | 20078.63 | 0 | 1 | 8625.44
Perimeter worst | 8.8 | 5050.34 | 0 | 1 | 1760.05
Area worst | -0.02 | 2742.31 | 0 | 1 | 8248.72
Smoothness worst | 269.41 | 174399.91 | 0 | 1 | 408.94
Compactness worst | -582.97 | 490340.38 | 0 | 1 | 2872.22
Concavity worst | 352.13 | 668403.99 | 0 | 1 | 5241.94
N Concave worst | -1317.63 | 1509411.14 | 0 | 1 | 1163.37
Symmetry worst | 937.3 | 490396.22 | 0 | 1 | 357.43
Fractal Dim worst | 11727.58 | 1821720.52 | 0.01 | 0.99 | 402.7

Note: VIF values for LR model with all predictors in are very high: minimum(VIF)=78, max(VIF)=123630.

The values of precision, recall and F1 measures for both training and test data are all quite high, as shown in Table 6.

Table 3: Final Bayesian LR model fitted to the training set.

| Estimate | SE | z Value | Pr(>|z|) | VIF |
|----------|----|---------|----------|-----|
| (Intercept) | -20.38 | 3.1 | -6.57 | 0 |
| Texture | 0.28 | 0.06 | 4.94 | 0 | 1.31 |
| Area | 0.01 | 0.06 | 6.9 | 0 | 1.45 |
| Concavity | 28.32 | 5.64 | 5.02 | 0 | 1.49 |
| Symmetry | 24.14 | 10.42 | 2.32 | 0.02 | 1.68 |

Note: Each of the four VIF values is <5

Table 4: Training set. Overall accuracy for the training set=93.2%.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>249</td>
</tr>
<tr>
<td>M</td>
<td>18</td>
</tr>
</tbody>
</table>

Table 5: Test set. Overall accuracy for the test set=93.0%.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>91</td>
</tr>
<tr>
<td>M</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 6:

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category 1</td>
<td>0.93</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>Category 0</td>
<td>0.93</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Category 1</td>
<td>0.87</td>
<td>0.91</td>
<td>0.89</td>
</tr>
<tr>
<td>Category 0</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
</tr>
</tbody>
</table>

References