SciMedCentral

Annals of Community Medicine and Practice

Research Article

Breast Cancer Prediction Using Bayesian Logistic Regression

Michael Chang, Rohan J. Dalpatadu, Dieudonne Phanord, and Ashok K. Singh*

Department of Mathematical Sciences, University of Nevada Las Vegas, USA

*Corresponding author

Ashok K. Singh, Department of Mathematical Sciences, University of Nevada Las Vegas, USA, Tel: 702-845-039; Email: ashok.singh@unlv.nevada.edu

Submitted: 01 October 2018

Accepted: 26 October 2018

Published: 29 October 2018

ISSN: 2475-9465

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

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, Taghva, and Singh [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, Abdelhalim and Zeid [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, Walker and Kadam [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 and 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, and 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 in situations when the predictors are random variables with measurement errors. In a study to determine the main causes of complications after radical cystectomy (urinary bladder removal) [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:

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

Where $X1, \ldots, X_p$ are K predictors, which can be continuous or discrete. The above model can be expressed in terms of log-odds as follows [12]

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

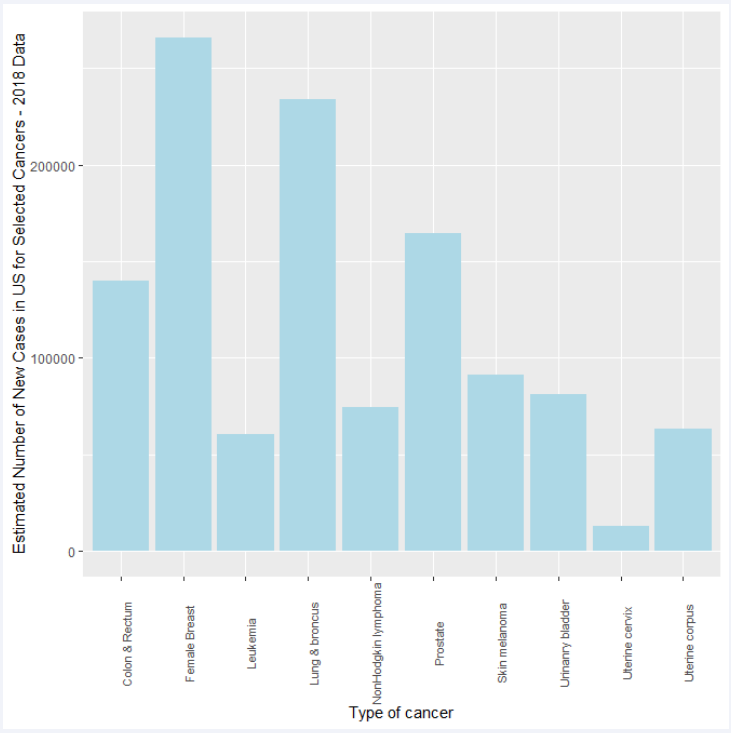

Figure 1 Estimated number of new cases in US for selected cancers – 2018.

In the frequentist approach, given the random sample

$(Y_j, X_{1j}, X_{2j}, ..., X_{Kj}),\ j = 1, 2, ..., n,$

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

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

where $\beta = (\beta_0, \beta_1, ..., \beta_K)$ in which

β_0 is the intercept term, and β_j is the coefficient of the

j-th predictor X_j, $j = 1, 2, ..., K$..

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

$$\sum_{j=1}^{n} [Y_j \log(P_j) + (1 - Y_j) \log(1 - P_j)] \quad (4)$$

In the Bayesian approach, the model coefficients $(\beta_1, \beta_2, ..., \beta_K)$ are realizations of a K-variate random vector generated from the joint prior distribution; any prior knowledge about the β-coefficients can be incorporated in this joint prior distribution. All inferences drawn using the Bayesian approaches 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.

$$g^*(\underline{\beta} \mid \underline{Y}) = \left(\prod_{j=1}^{n} P_i^{Y_i} (1 - P_i)^{1-Y_i} \right) \times g(\underline{\beta}) \quad (5)$$

where $g^*(\underline{\beta} \mid \underline{Y})$ is the joint posterior distribution,

and $g(\underline{\beta})$ the joint posterior distribution

of the parameters $\underline{\beta}$.

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 F1 of precision and recall [20,21]. To compute these measures, we first need to calculate the 2x2 confusion matrix shown in Table (1).

Here $C_{i,j}$ = number of times true response of j get predicted as i (i, j = 0, 1).

Table 1: Confusion matrix for a binary classifier.

	Observed Y	
Predicted Y	0	1
0	$C_{0,0}$	$C_{0,1}$
1	$C_{1,0}$	$C_{1,1}$

The performance measures accuracy, precision, recall and F1 are calculated for each category 0 and 1 from the following formulas:

$$\text{Accuracy} = \frac{\sum_{j=0}^{1} C_{j,j}}{\sum_{i=0}^{1}\sum_{j=0}^{1} C_{j,j}} \quad (6)$$

$$\text{Precision}_j = \frac{C_{j,j}}{\sum_{k=0}^{1} C_{j,k}} \quad (7)$$

$$\text{Recall}_j = \frac{C_{j,j}}{\sum_{k=0}^{1} C_{k,j}} \quad (8)$$

$$F1_j = \frac{2 \times \text{Precision}_j \times \text{Recall}_j}{(\text{Precision}_j + \text{Recall}_j)}, j = 0, 1$$

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 computed for each cell nucleus:

a) Radius (average distance from center to points on the perimeter)

b) Texture (standard deviation of gray-scale values)

c) Perimeter

d) Area

e) Smoothness (local variation in radius lengths)

f) Compactness (perimeter^2 / area - 1.0)

g) Concavity (severity of concave portions of the contour)

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

	Estimate	SE	z value	P-value	VIF
(Intercept)	-2968.33	1189296.4	0	1	
Radius	-110.8	204090.25	0	1	44754.48
Texture	-0.43	16095.7	0	1	2307.93
Perimeter	30.78	48403.8	0	1	123629.76
Area	-1.07	2357.23	0	1	41688.84
Smoothness	2626.6	4824631.59	0	1	995.55
Compactness	-4846.98	1278852.25	0	1	1477.60
Concavity	-938.94	766227.12	0	1	543.40
N.Concave	8703.04	1884638.69	0	1	476.13
Symmetry	-619.86	588019.99	0	1	78.01
Fractal.Dim	4286.86	3366578.33	0	1	102.07
Radius.SE	1307.2	836904.03	0	1	6244.44
Texture.SE	-36.76	138213.51	0	1	3327.97
Perimeter.SE	-46.95	49083.59	0	1	1334.69
Area.SE	-1.97	10112.03	0	1	6439.77
Smoothness.SE	9958.43	6060290.39	0	1	182.61
Compactness.SE	2104.2	3284120.37	0	1	2212.24
Concavity.SE	3543.98	2507993.37	0	1	1488.06
N.Concave.SE	1017.04	13135157.45	0	1	2677.67
Symmetry.SE	-1398.05	3169097.88	0	1	189.51
Fractal.Dim.SE	-87436.83	25555442.67	0	1	1169.20
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	82482.72
Smoothness.worst	269.41	1743939.91	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	402.70

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

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	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: Confusion Matrix for the Training set.

	Predicted	
Observed	B	M
B	249	11
M	18	149
Overall accuracy for the training set = 93.2%		

Table 5: Confusion Matrix for the Test set.

	Predicted	
Observed	B	M
B	91	6
M	4	41
Overall accuracy for the test set = 93.0%		

Table 6: Precision, recall and F1 measures for both training and test data sets.

Data set		Precision	Recall	F1
Training	Category 1	0.93	0.89	0.91
	Category 0	0.93	0.96	0.94
Test	Category 1	0.87	0.91	0.89
	Category 0	0.96	0.94	0.95

h) Concave points (number of concave portions of the contour)

i) Symmetry

j) 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.

There are three common approaches for fitting a LR model when high multicollinearity exists in the data. Aguilera, Escabias, Valderrama [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.

RESULTS 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 predictor's significant (Table 1); 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 2) with Texture, Area, Concavity, and Symmetry in the model. A comparison of Tables 2 and 3 shows how multicollinearities affect the estimation of LR model coefficients:

I. In the LR model with all predictors, all P-values are 1 i.e., none of the predictors are significant,

II. The estimated coefficients of the final predictors in the LR model with all predictors are all negative, when these coefficients should all be positive,

III. The standard errors (SE) of the final predictors in the LR model with all predictors are orders of magnitude higher than the corresponding estimates, and

IV. 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.

The Figure 3 shows the posterior distributions and the 95% HPD credible sets for the coefficients of the predictors in the final LR model; the 95% HPD credible sets are:

$\beta_{Texture}$: (0.16, 0.37), β_{Area}: (0.008, 0.016), $\beta_{Concavity}$: (16.65, 36.30), β_{Area}: (3.22, 40.28).

Observe that all four 95% HPD credible sets fall to the right of 0.

Elimination of predictors with large VIF values leads to the final Bayesian LR model, given in Table 2.

The final LR model was next used to predict response "Diagnosis" for both the training and test data sets. The confusion matrices and overall accuracies for the training and test sets are shown in Tables 4 and 5.

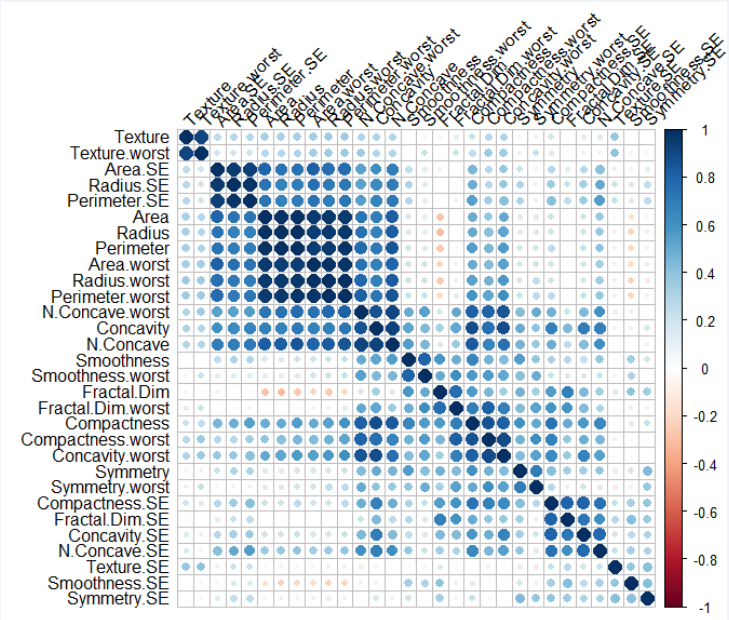

Figure 2 Correlation plot of 30 predictors in WDBC data set.

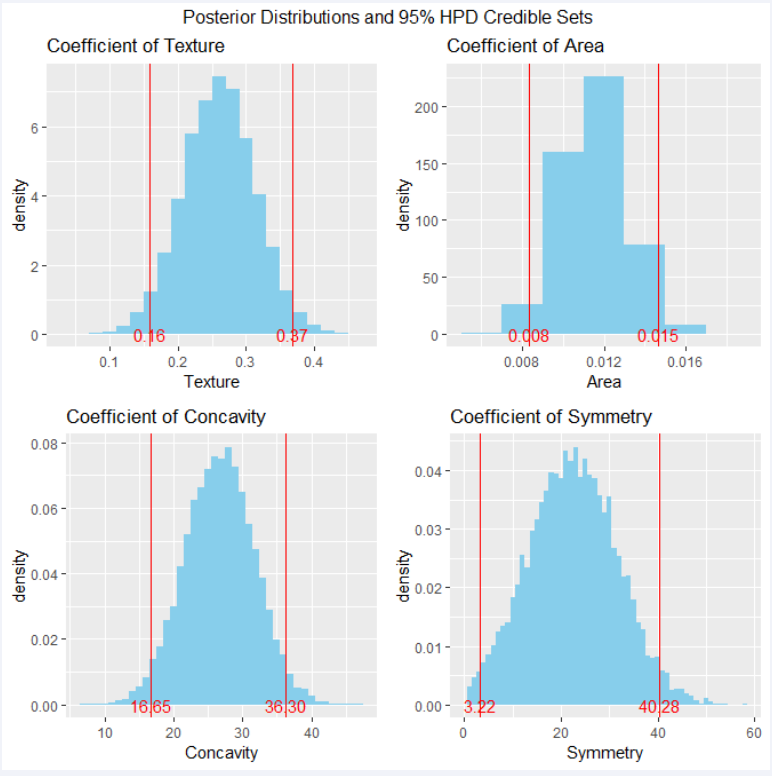

Figure 3 Posterior Distributions of Bayes Estimates of Logistic Regression Model Coefficients and their 95% HPD Credible Sets.

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

DISCUSSION

The fitted Bayesian LR model has a total of four significant predictors: texture, area, concavity, and symmetry, with each predictor coefficient positive, as to be expected; the 95% HPD credible sets for these coefficients are shown in Figure 3; in each case, the entire 95% credible set falls to the right of 0, showing statistical significance of these predictors. Note that the Bayesian credible sets have a simple explanation – for example, we can say with 95% confidence that the random parameter $\beta_{Texture}$ falls

inside the interval (0.16, 0.37) with the most likely value of 0.28.

CONCLUSION

We have used the Bayesian method for estimating the LR model for prediction of breast cancer; the Bayesian method comes with a much higher computational cost but has certain advantages over the classical method. The classical or frequentist approach to fitting an LR model is more common but has two major disadvantages: (i) it does not allow the user to formally incorporate any prior knowledge into parameter estimation [28], and (ii) it yields confidence intervals that are harder to interpret [29], with confidence going with the method or formula of computing the confidence interval, and not with the calculated confidence interval itself. Bayesian LR allows for formally using expert opinion and prior knowledge in the estimation of parameters, and typically yields better results than the classical method (Gordóvil-Merino et al., and Ogunsakin and Siaka).

REFERENCES

1. American Cancer Society. Cancer Facts and Figures. 2018.
2. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global Cancer Statistics. Ca Cancer J Clin. 2011; 61: 69-90.
3. Bozorgi M, Taghva K, Singh AK. Cancer Survivability with Logistic Regression. Computing Conference, 2017-18- July 20. UK. London: IEEE. 2018.
4. Salama GI, Abdelhalim MB, Zeid MA. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. Int J Comp Infor Technol. 2012; 1: 2277-2764.
5. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005; 34: 113-127.
6. Peretti A, Amenta F. Breast Cancer Prediction by Logistic Regression with CUDA Parallel Programming Support. Breast Can Curr Res: 2016; 1: 111.
7. Barco I, Garcia Font M, Garcia Fernandez A, Giménez N, Fraile M, Lain JM, et al. A logistic regression model predicting high axillary tumour burden in early breast cancer patients. Clin Transl Oncol. 2007; 19: 1393-1399.
8. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002; 35: 352-359.
9. Van Domelen DR, Mitchell EM, Perkins NJ, Schisterman EF, Manatunga AK. Logistic regression with a continuous exposure measured inpools and subject to errors. Stat Med. 2018; 1-15.
10. Atduev V, Gasrataliev V, Ledyaev D, Belsky V, Lyubarskaya Y, Mamedov H. Predictors of 30-Day Complications after Radical Cystectomy. Exp Tech Urol Nephrol. 2018; 1.
11. Dua D, Karra Taniskidou E. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
12. Kleinbaum DG, Klein M. Logistic Regression - A Self-Learning Text. 3rd Edn. New York: Springer. 2007.
13. Kruschke J. Doing Bayesian Data Analysis. 2nd Edn. Netherlands: Elsevier. 2014.
14. Rohan D, Gewali L, Singh AK. Computing the Bayesian highest posterior density credible sets for the lognormal mean. Environmetrics. 2002; 13: 465-472.
15. Newcombe PJ, Reck BH, Sun J, Platek GT, Verzilli C, Kader AK, et al. A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. Genet Epidemiol. 2012; 36: 71-83.
16. Ambrose PG, Hammel JP, Bhavnani SM, Rubino CM, Ellis-Grosse EJ, Drusano GL. Frequentist and Bayesian Pharmacometric-Based Approaches To Facilitate Critically Needed New Antibiotic Development: Overcoming Lies, Damn Lies, and Statistics. Antimicrob Agents Chemother. 2012; 3: 1466-1470.
17. Austin PC, Naylor CD, Tu JV. A comparison of a Bayesian vs. a frequentist method for profiling hospital performance. J Eval Clin Pract. 2001; 7: 35-45.
18. Grzenda W. The advantages of Bayesian methods over Classical methods in the context of credible intervals. Infor Sys Manag. 2015; 4: 53-63.
19. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Infor Process Manag. 2009; 45: 427-437.
20. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer. 2013.
21. Guillet F, Hamilton HJ. Quality measures in data mining. New York: Springer. 2007.
22. Street WN, Wolberg WH, Mangasarian OL. Nuclear Feature Extraction for Breast Tumor Diagnosis. Int Sym on Electron Imaging: Sci Technol. 1905: 1993; 861-870.
23. Wolberg WH, Street WN, Mangasarian OL. Machine learning techniques to diagnose breast cancer from: image-processed nuclear features of fine needle aspirates. Cancer Lett. 1994; 77: 163-171.
24. Ana M. Aguilera, Manuel Escabias, Mariano J. Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput. Stat Data Anal. 2006; 50: 1905-1924.
25. Yasin Asar. Some new methods to solve multicollinearity in logistic regression. Commun Stat Simul Comput. 2014; 46: 2576-2586.
26. Montgomery DC, Peck EA, Vining GG. Introduction to Linear regression Analysis. 3rd Edn. New York: Wiley. 2012.
27. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.
28. Amalia Gordóvil-Merino, Joan Guàrdia-Olmos, Maribel Peró-Cebollero, Emilia I. de la Fuente-Solanas. Classical and Bayesian estimation in the logistic regression model applied to diagnosis of child attention deficit hyperactivity disorder. Psychol Rep. 2010; 106: 519-533.
29. Grzenda W. The advantages of Bayesian methods over Classical methods in the context of credible intervals. Infor Sys Manag. 2015; 4: 53-63.

Cite this article

Chang M, Dalpatadu RJ, Phanord D, Singh AK (2018) Breast Cancer Prediction Using Bayesian Logistic Regression. Ann Community Med Pract 4(3): 1039.