Application of Logistic Regression with Breast Cancer Diagnosis

Benjamin Ocampo

Abstract

The diagnosis for breast cancer is important for many women as accurate prediction can lead to a reduced mortality of this illness. Due to the large amount of data available for it, modern statistical methods such as machine learning can be used to improve the accuracy of breast cancer diagnosis. In this paper, we explore the use of the logistic regression method with the classification of breast cancer diagnosis. Two models are used in this paper: first one uses all ten attributes from the dataset collected from the University of Wisconsin while the second model uses only five attributes. Noting that the five chosen attributes do not show any dependencies between each other, both models exhibit a classification success rate of 0.95. However, the second model is more preferred as the deviance information criterion is unable to predict the penalty for the first model. This is most likely due to dependencies being shown between attributes.

1 Introduction

Breast cancer is a significant health concern among women all over the world as it has a tendency towards a high mortality rate (Greenlee, Murray, Bolden, & Wingo, 2000). However, decline in its mortality rate are a result of both treatment improvement and early detection (DeSantis, Siegel, Bandi, & Jemal, 2011). In terms of early detection, self and clinical examinations have been promoted among the public (Kosters & Gotzsche, 2003). However, as a result of the large number of data collected associated with breast cancer diagnosis, statistical methods such as machine learning techniques can be used to improve early detection (e.g., Wolberg, Street, & Mangasarian, 1994).

In this report, we first describe the data used in the model. Then, the logistic regression model is described along with its chosen prior and posterior distributions. Finally, we provide a discussion to model results and how accurately it can predict the outcome of the diagnosis depending on the observations.

2 Data

The University of Wisconsin has collected a vast amount of data associated with breast cancer diagnosis, which is publicly available at `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`. Along with the response variable being the breast cancer diagnosis, the other attributes highlighted in the dataset are mean radius, texture (standard deviation of the grey scale), perimeter, area, smoothness (variations in radii), compactness (perimeter^2 / area - 1.0), concavity (severity of contour concave portions), concave points (number of contour concave portions), symmetry, and fractal dimension ("coastline approximation" - 1) (Kaggle, 2018). The dataset also shows the mean, standard deviation, and "worst" of these attributes, where the "worst" is associated with the mean of the three largest values of the attribute. Note that in this report, we are only interested in the mean of these attributes.

3 Model

Considering that the response variable for breast cancer diagnosis is binary and given that the dataset is large, the logistic regression model is appropriate to use. The benefit for using this type of model is that it is similar to the linear regression models in two aspects: it is not computationally costly and it is easy to implement (Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997; Press, Teukolsky, Vetterling, & Flannery, 2007). However, the tradeoff for this type of model is that it requires a large number of data points and a fairly good estimate of the prior and posterior means to get a relatively decent fit (Hosmer et al., 1997; Press et al., 2007).

Defining the binary diagnosis response as y_i for observations $i = 1, ..., N$, ϕ_i the probability of success of the ith observation, and all of the attributes as x_i, we can assign the Bernoulli likelihood to these observations:

$$y_i|\phi_i \sim \text{Bern}(\phi_i). \tag{1}$$

Modelling the expected value of y_i as

$$E(y_i) = \phi_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_{ji} y_i)]}, \tag{2}$$

we can define the $\text{logit}(\phi_i)$ as

$$\text{logit}(\phi_i) = \log\left(\frac{\phi_i}{1 - \phi_i}\right) = \beta_0 + \sum_{j=1}^{10} \beta_{ji} y_i. \tag{3}$$

Note that $j = 1, ..., 10$, where 10 represents the number of attributes y_i is dependant on. Assuming that the attributes are all independent and identically distributed, the double exponential or Laplacian distribution are assigned as the posterior distributions for the β_0 and β_{ji} coefficients as it is appropriate for a logistic regression model. This can be seen by the following equation:

$$\beta_0 \sim \text{Laplace}(0.0, 1.0) \tag{4}$$

$$\beta_{ji} \sim \text{Laplace}(0.0, 1.0) \tag{5}$$

Note that for each β coefficient, the mean is set to 0.0 and the variance is set to 1.0. This is because the attributes in the data are rescaled by

$$\tilde{x}_i = \frac{x_i - E(x_i)}{SD(x_i)}, \tag{6}$$

where $E(x_i)$ and $SD(x_i)$ represents the expectation value and the mean of x_i, respectively.

Two different logistic regression models are presented in this paper. The first model uses all of the attributes from the dataset while the second model only uses the mean radius, texture, smoothness, compactness, and concavity. The second model only uses these attributes because there are some dependencies exhibited between attributes, such as the radius and the area.

	Intercept	$\beta_{\text{mean radius}}$	β_{texture}	$\beta_{\text{perimeter}}$	β_{area}	$\beta_{\text{smoothness}}$
Mean	-0.5787	0.8801	1.4549	0.8071	1.6886	0.8954
Standard Deviation	0.2410	1.0815	0.2432	1.0852	1.3660	0.3919

	$\beta_{\text{compactness}}$	$\beta_{\text{concavity}}$	$\beta_{\text{concave points}}$	β_{symmetry}	$\beta_{\text{fractal dimension}}$
Mean	-0.2752	0.8538	2.0025	0.3815	-0.2953
Standard Deviation	0.4743	0.5577	0.9096	0.2691	0.4050

Table 1: The mean and standard deviation of the β coefficient results from the first model.

	Intercept	$\beta_{\text{mean radius}}$	β_{texture}	$\beta_{\text{smoothness}}$	$\beta_{\text{compactness}}$	$\beta_{\text{concavity}}$
Mean	-0.7691	4.0875	1.4368	1.4612	-0.2270	1.6602
Standard Deviation	0.2016	0.5031	0.2377	0.3070	0.3737	0.3928

Table 2: The mean and standard deviation of the β coefficient results from the second model.

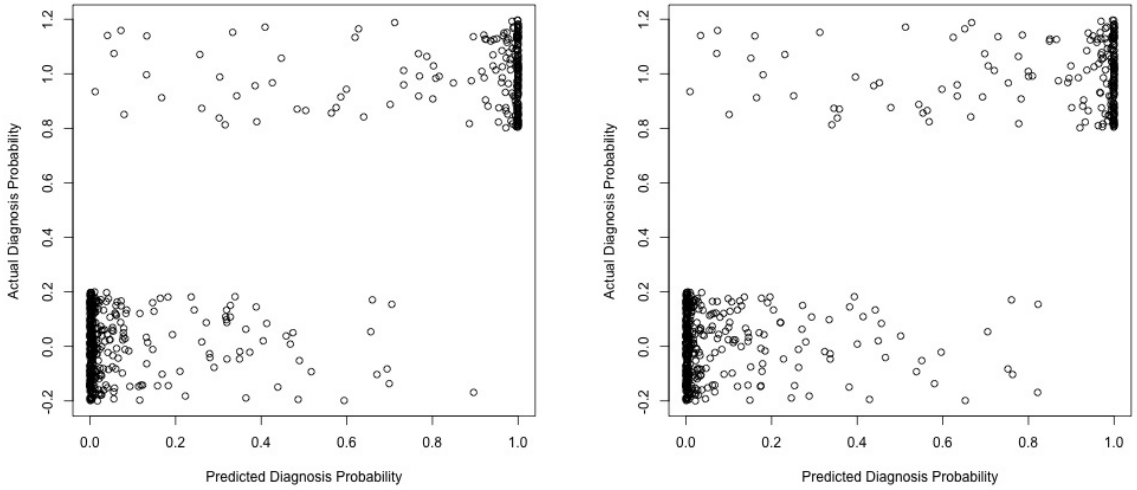
4 Results

The first model provides an accurate prediction for the breast cancer diagnosis based on the dataset. Table 1 presents the mean and standard deviation of the predicted β coefficients. Figure 1a shows a comparison between its predicted probability and the actual probability of the breast cancer diagnosis, showing roughly a 0.95 success rate in classification. While both the Gelman diagnostic and the autocorrelation diagnostic show that there are no patterns and no strong positive or negative correlation for the predicted β coefficients implying convergence, the deviance information criterion (DIC) does show some problems with the model. The DIC mean deviance for the model is 159.9. However, due to computational error, it is unable to predict a penalty or penalized deviance. This shows some problems with the model, most likely due to dependencies between attributes.

The second model, on the other hand, does indeed reach convergence and provides a prediction for the breast cancer diagnosis similar to the first model. Table 1 presents the mean and standard deviation of the predicted β coefficients. Figure 1b shows a comparison between its predicted probability and the actual probability of the breast cancer diagnosis, showing roughly a 0.95 success rate in classification. However, unlike the first model, it does indeed reach convergence based on the Gelman diagnostic and autocorrelation diagnostic. The DIC also shows that the model has a mean deviance of 167.5 with a penalty of roughly 5.6 and penalized deviance of 173.1. Despite the second model having a larger mean deviance than the first model, the DIC penalty justifies the five chosen attributes. Classification accuracy is also maintained despite not using all of the attributes in the data set. Thus, the second model is preferred over the first model.

5 Conclusion

The logistic regression model is appropriate to use on the breast cancer dataset as it provides a good fit and prediction of the cancer diagnosis. Two models are used in this paper. The first model uses all ten attributes from the dataset, showing a classification success rate of 0.95. However, the



(a) Model 1

(b) Model 2

Figure 1: A comparison between the predicted and actual probabilities of the breast cancer diagnosis for the first and second models.

deviance information criterion is unable to predict its penalty, most likely due to computational error. This implies that some attributes show some dependency between each other. The second model is more successful as its deviance information criterion does not display such problems. In fact, the five chosen attributes (mean radius, texture, smoothness, compactness, and concavity) are justified by the penalty of 5.6 outputted by the deviance information criterion. The second model's classification success rate of 0.95 shows that choosing five out of the ten attributes do not reduce accuracy, indicating that the second model is preferable over the first one.

References

DeSantis, C., Siegel, R., Bandi, P., & Jemal, A. (2011). Breast cancer statistics, 2011. *CA: A Cancer Journal for Clinicians*, *61*(6), 408–418. Retrieved from `http://dx.doi.org/10.3322/caac.20134` doi: 10.3322/caac.20134

Greenlee, R. T., Murray, T., Bolden, S., & Wingo, P. A. (2000). Cancer statistics, 2000. *CA: A Cancer Journal for Clinicians*, *50*(1), 7–33. Retrieved from `http://dx.doi.org/10.3322/canjclin.50.1.7` doi: 10.3322/canjclin.50.1.7

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16*(9), 965–980. Retrieved from `http://dx.doi.org/10.1002/(SICI)1097-0258(19970515)16:9<965::AID-SIM509>3.0.CO;2-O` doi: 10.1002/(SICI)1097-0258(19970515)16:9⟨965::AID-SIM509⟩3.0.CO;2-O

Kaggle. (2018). *Breast cancer wisconsin (diagnostic) data set.* Retrieved from `https://www.kaggle`

.com/uciml/breast-cancer-wisconsin-data

Kosters, J., & Gotzsche, P. C. (2003). Regular self-examination or clinical examination for early detection of breast cancer. *Cochrane Library*(Issue 2).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: The art of scientific computing* (3rd ed.). New York, NY: Cambridge University Press.

Wolberg, W. H., Street, W., & Mangasarian, O. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, *77*(2), 163 - 171. Retrieved from http://www.sciencedirect.com/science/article/pii/030438359490099X (Computer applications for early detection and staging of cancer) doi: https://doi.org/10.1016/0304-3835(94)90099-X