Using Logistic Regression to Improve Quality Assurance in Detecting Cancerous Breast Tissue

Philip Terrien and Alex Scharp
Department of Biomedical Engineering, Department of Computer Science, University of Wisconsin Madison, Madison, WI

Abstract - This study develops an algorithm using logistic regression and alternative classification boundaries to provide a more accurate and reliable diagnosis of breast cancer from fine needle aspiration data. The data used for training and testing the predictive model come from the Wisconsin Diagnostic Breast Cancer dataset [4]. The study found that by basing decision boundaries on a lower probabilistic threshold of 0.2920 and upper probabilistic threshold of 0.6820, classification accuracy can be improved to 99.044%. However, this improvement comes at a cost; not all input receives classification, and some classifications are rejected because of insufficient confidence. This predictive model could still provide a useful supplement to the tools used by physicians to assure the quality of a breast cancer diagnosis.

1 INTRODUCTION

In the United States, breast cancer is diagnosed in approximately 230,000 women and 2,000 men annually and results in a total of 41,000 deaths per year [1]. Noninvasive techniques such as mammograms are effective for detecting abnormalities, and several successful non-operative methods exist for determining whether the region is benign or malignant. One common method, fine needle aspiration (FNA), involves the extraction of tissue from a suspicious area through a hollow needle syringe. Despite its popularity, FNA has been criticized for suboptimal accuracy and false positive rates [2]. Given the demanding treatment for breast cancer, a high degree of quality assurance against any false positives is required.

Machine learning could be used to improve this technique. In particular, a classification model can be learned to predict whether a cell is benign or malignant based on the cell's features. Classification occurs in two steps. First, a model is trained by providing the machine a training dataset with known classifications. This paper evaluates two specific models: Ridge Regression with a classification function, and Logistic Regression. In both cases, Stochastic Gradient Descent is used to learn the models. Second, the model's accuracy is evaluated on a testing dataset. The holdout approach, which estimates the model's predictive accuracy by measuring its accuracy on a test dataset which is not involved in the learning process, is used for both models [3].

While binary classification is a powerful tool for predicting labels, it may not always be appropriate. One weakness of this approach is its failure to convey levels of confidence for its predictions. In practice, overreliance on the model's predictions may cause misdiagnosis in fringe cases (predictions that occur near the decision boundary). One approach to curb this risk is to increase the number of categories so as to reflect degrees of certainty. This paper explores the application of this idea to the logistic regression model by learning additional decision boundaries (thresholds) in an attempt to improve FNA's quality assurance.

2 THE DATASET

The University of Wisconsin Madison hosts a dataset of features extracted from digitized FNA images from 569 instances of potentially cancerous cells. Each instance is labeled as benign or malignant and contains 30 features describing the nucleus of the cell including radius, area, texture (standard deviation of gray-scale values), etc [4, 5].

3 METHODS

This study developed the following models in MATLAB®, which was licensed through the University of Wisconsin Madison. The scripts can be found in the following repository:

https://github.com/terrienphilip/cancerdetector.git.

3.1 Data Preprocessing

Prior to learning each model, the data was standardized using Eq. 1, and a feature vector of ones was added to each sample to provide an offset for all linear fits [6].

$$z = \frac{X - \mu}{\sigma} \quad (1)$$

The input X takes the same form, $X \in \Re^{nxp}$, for both linear and logistic regression where n = number of samples and p = 31 for the 30 image features plus an additional offset feature. The output changes form to accommodate the different classification rules for linear and logistic regression. For ridge regression, it exists as $y \in \{-1, +1\}^n$; for logistic regression, it exists as $y \in \{0, 1\}^n$.

3.2 Stochastic Gradient Descent Parameters

The models for ridge regression and logistic regression were learned using Stochastic Gradient Descent (SGD). In both cases, the tuning parameter τ, convergence parameter ε, and number of iterations (epochs) were determined observationally. Their specific values, as well as the equations for the gradients, are provided in the models' respective sections.

3.3 Ridge Regression

Ridge regression was performed to provide a comparison to the results of logistic regression. The weight vector for the hypothesis model from Eq. 2 was learned using mean-squared error as the loss function with Tikhonov regularization (see Eq. 3). The optimal λ ($\lambda = 0.1$), τ ($\tau = 10^{-4}$), ε ($\varepsilon = 10^{-6}$), and epochs (epochs=10^4) were selected based on the best classification accuracy from a range of different values.

$$\hat{y} = \sum_{i=0}^{N} x_i w_i \quad (2)$$

$$\hat{w} = argmin\left(\frac{1}{N}\sum_{i=0}^{N}(x_i w_i - y)^2 + \lambda||w||_2^2\right) \quad (3)$$

The classification rule in Eq. 4 was used to label $\hat{y}$ as benign (-1) or malignant (+1).

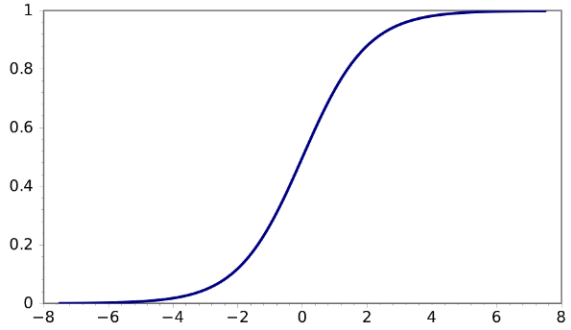
$$f(\hat{y}) = \begin{cases} +1 & \hat{y} \geq 0 \\ -1 & \hat{y} < 0 \end{cases} \quad (4)$$

3.4 Logistic Regression

Logistic regression is a probabilistic linear classifier. It computes the probability that an outcome belongs to a default class. Just as in ridge regression, Eq. 2 is used to make predictions. However, the results are then transformed by the sigmoid function [8]:

$$\sigma(z) = (1 + e^{-z})^{-1} \quad (5)$$

The sigmoid function has several favorable properties. One such property can be observed from the function's graph.



Because most inputs are mapped to values close to either 0 or 1, the sigmoid function is particularly useful for classification. Another important benefit of the sigmoid function is the simplicity of its derivative [7]:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)) \quad (6)$$

This property is advantageous when computing the gradient of the cost function for SGD.

Combining Equations 2 and 5 yields the logistic regression prediction equation given in Eq. 7 (Note that $x^T w$ is equivalent to the right-hand side of Eq. 2).

$$\hat{y} = (1 + e^{-x^T w})^{-1} \quad (7)$$

The output of this equation can be interpreted as the probability that the corresponding data point belongs to the default class. For the FNA model, malignant is chosen to be the default class. According to the model, data points with predicted outcomes close to 1 are likely to be

malignant, whereas those close to 0 are likely to be benign. The classification function for logistic regression is defined to reflect this trend:

$$f(\hat{y}) = \begin{cases} +1 \text{ (malignant)} & \hat{y} \geq 0.5 \\ 0 \text{ (benign)} & \hat{y} < 0.5 \end{cases} \quad (8)$$

For now the decision boundary is assumed to be 0.5, per standard practice (refer to section 3.6 for further exploration of this topic).

In order to train the model with SGD, it is essential to define an appropriate cost function. One might be tempted to use mean-squared error, the cost function that is used in ridge regression. However, the mean-squared error of the sigmoid function is a non-convex function with many local minima. Consequently, gradient descent may not find the optimal global minimum [8]. Though the stochastic nature of SGD does somewhat offset this danger, a vast number of iterations would be needed to guarantee the optimal result.

A better approach is to use the cost function known as Cross-Entropy or Log Loss, which applies a logarithmic transformation to the prediction error. Eq. 9 shows the resulting piecewise cost function.

$$Cost(\hat{y}, y) = \begin{cases} -log(\hat{y}) & y = 1 \\ -log(1-\hat{y}) & y = 0 \end{cases} \quad (9)$$

Transforming the output logarithmically and assigning the corresponding signs has the effect of penalizing confident, wrong predictions more than rewarding confident, right predictions [9]. Therefore, the model will likely have high sensitivity and specificity (refer to section 4).

The gradient (with respect to the weight vector w) of the Cross-Validation cost function is surprisingly simple and elegant. Eq. 10 is a direct result of simplifying the gradient of Eq. 9. Substituting Eq. 7 for $\hat{y}$ yields Eq. 11.

$$Cost'(\hat{y}, y) = (\hat{y} - y) * x \quad (10)$$

$$Cost'(\hat{y}, y) = ((1 + e^{-x^T w})^{-1} - y) * x \quad (11)$$

Applying SGD to this gradient provides a more optimal solution [8]. In this study's analysis, the optimal τ ($\tau = 10^{-4}$), ε ($\varepsilon = 10^{-7}$), and epochs (epochs=10^5) were selected based on the best classification accuracy from a range of different values.

3.5 Classification Metrics

Accuracy of each system will be assessed by considering the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) of each classifier. The performance metrics calculated from these characteristics are accuracy, sensitivity, specificity, positive predictive rate (PPR), and negative predictive rate (NPR) [10].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (12)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \quad (14)$$

$$PPR = \frac{TP}{TP + FP} \times 100\% \quad (15)$$

$$NPR = \frac{TN}{TN + FN} \times 100\% \quad (16)$$

3.6 Learning Logistic Threshold Boundaries

Traditional binary classification rules for logistic regression use a decision boundary of 0.5 (see Eq. 8) [9]. However, alternative thresholds can be used to yield more accurate classifications. The optimal upper and lower thresholds will produce the desired PPR and NPR respectively while providing the maximum number of confident classifications. In this paper, confident classifications are defined as the classifications where $\hat{y}$ is above or below the specified thresholds. All values between the thresholds in the sigmoid graph are ignored and reported as indeterminate classifications. The desired PPR and NPR are specified as 99%. These quantitative constraints are used to learn the best thresholds. This is done by averaging the classification statistics for a range of threshold values over 1000 iterations of model approximations. When determining the upper threshold, a range of thresholds between [0.5, 1.0] are tested while the lower threshold is held at 0. This allows the algorithm to specifically

evaluate the model's ability to correctly classify only positive (malignant) diagnoses. Alternatively, the lower threshold is determined by testing a range of thresholds between [0.0, 0.5] while the upper threshold is held at 1.0.

4 RESULTS

The logistic regression (with and without thresholding) and ridge regression predictive models were approximated 1000 times each. The classification metrics for each set of approximations were averaged, and the results are reported in Table I. The optimal upper and lower thresholds for the logistic regression with thresholding were calculated as 0.6820 and 0.2920 respectively. The last row represents the percentage of the validation set that was confidently classified.

	Ridge Regression	Logistic Regression	Logistic Regression w/ Thresh.
Accuracy	95.072%	95.496%	99.044%
Sensitivity	93.138%	93.886%	98.443%
Specificity	96.242%	96.482%	99.417%
PPR	93.663%	94.051%	99.075%
NPR	95.909%	96.342%	99.033%
Percent Classified	N/A	100%	82.021%

Table I: Averaged classification results for 1000 model approximations

The resultant sigmoid graph for the 1000th model approximation can be seen in Figure I where the y-axis represents the probability that $\hat{y}$ matches the default case (malignant), and the x-axis represents the respective $\hat{y}$.

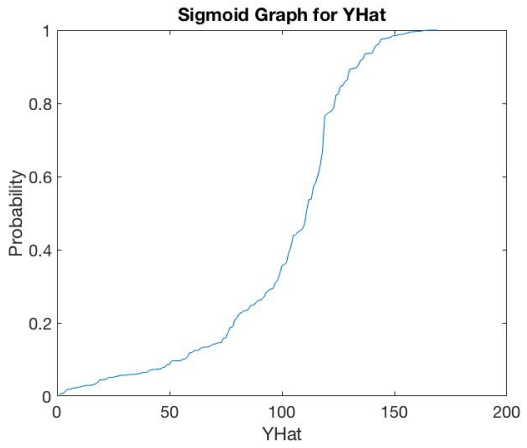


Figure I: Resultant sigmoid graph for the 1000th model approximation

Figure II and Figure III show the trade-off between the percentage of the validation set that is confidently classified and the predictive rates for different threshold values. When the thresholds are close to either 0 or 1, the likelihood of confident classification decreases. At times, there are no confident classifications; these data are reported as 100% in the graph.

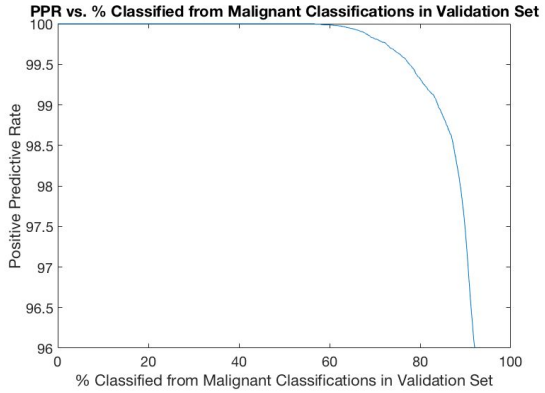


Figure II: Trade-off between percent of confident classifications and PPR for different upper threshold values

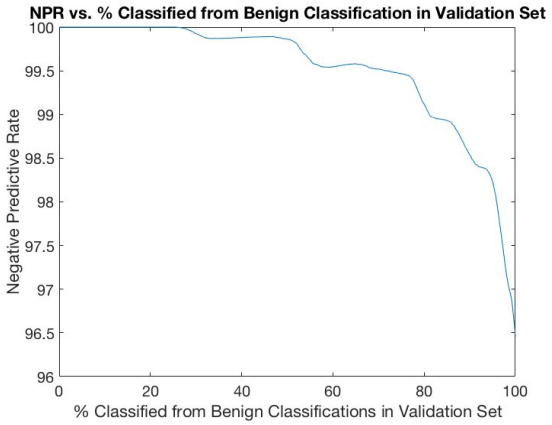


Figure III: Trade-off between percent of confident classifications and NPR for different lower threshold values

5 DISCUSSION

The logistic regression with thresholding demonstrated better performance in most metrics when compared to standard ridge regression and logistic regression without thresholding. The logistic regression without thresholding performed with effectiveness similar to that of standard ridge regression, which is to be expected because the sigmoid function is just a probabilistic transform of the linear regression.

The two trade-off curves each demonstrate that the percentage of confidently classified data increases as the requirement imposed by the threshold becomes less stringent (i.e. upper threshold decreases or lower threshold increases). The reported sensitivity and specificity, along with the sigmoid graph of the 1000^{th} iteration, demonstrate that logistic regression on this particular data set is less effective at reliably identifying TP than reliably identifying TN.

The results for linear regression and logistic regression agree with performance metrics from previous studies that use the Wisconsin Diagnostic Database or other similar datasets [6, 10].

6 CONCLUSION

By using logistic regression with the thresholding techniques outlined in this paper, a machine learning algorithm is produced that offers greater accuracy and reliability in diagnosing breast cancer. The cost of using this technique is that not all input receives classification, and some classifications are rejected because of insufficient confidence. However, this lack of confidence is made known to the physicians and not hidden in a boolean output, which provides transparency for when the output cannot be trusted as a final means of quality assurance. This algorithm could be used as a valuable tool for informing physicians when additional assessment may be needed in diagnosing a patient of breast cancer. For instance, if a physician's diagnosis is in disagreement with a confident classification or the algorithm does not yield a confident classification, another physician should be consulted for a second opinion. As the model becomes more robust with future FNA training data, logistic regression with thresholding could prove to be another valuable tool in improving breast cancer diagnostics to administer early and appropriate treatment.

7 REFERENCES

[1] "Breast Cancer," *Centers for Disease Control and Prevention*, 07-Jun-2017. [Online]. Available: https://www.cdc.gov/cancer/breast/statistics/index.htm. [Accessed: 31-Oct-2017].

[2] E. Manfrin, R. Mariotto, A. Remo, D. Reghellin, D. Dalfior, F. Falsirollo, and F. Bonetti, "Is there still a role for fine-needle aspiration cytology in breast cancer screening?," Cancer Cytopathology, vol. 114, no. 2, pp. 74–82, 2008.

[3] D. Mittal, D. Gaurav, and S. S. Roy, "An effective hybridized classifier for breast cancer diagnosis," *Advanced Intelligent Mechatronics (AIM), 2015 IEEE International Conference*, pp. 1026–1031, IEEE, 2015.

[4] U. C. I. M. Learning, *Breast Cancer Wisconsin (Diagnostic) Data Set* | Kaggle, 25-Sep-2016. [Online]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. [Accessed: 31-Oct-2017].

[5] M. Kumar, *Basic Machine Learning with Cancer* | Kaggle, 1-Jun-2017. [Online].

Available: https://www.kaggle.com/gargmanish/basic-machine-learning-with-cancer/notebook. [Accessed: 31-Oct-2017].

[6] A. F. Agarap, "On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset," *arXiv preprint* arXiv:1711.07831, 2017.

[7] D. Gualtieri, Tikalon Blog by Dev Gualtieri. [Online]. Available: http://tikalon.com/blog/blog.php?article=2011%2Fsigmoid. [Accessed: 19-Dec-2017].

[8] "Logistic Regression," Logistic Regression — ML Cheatsheet documentation. [Online]. Available: http://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html#gradient-descent. [Accessed: 19-Dec-2017].

[9] P. Rao and J. Manikandan, "Design and evaluation of logistic regression model for pattern recognition systems," in *India Conference (INDICON), 2016 IEEE Annual*, pp. 1–6, IEEE, 2016.

[10] A. Azar and S. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition.," *Neural Computing & Applications,* vol. 24, pp. 1163–1177, Apr 2014. Article.