USING NEWTON-RAPHSON AND LOGISTIC-LASSO FOR PREDICTING BREAST CANCER STATUS

A PREPRINT

Xinyi Lin
xl2836

Junting Ren
jr3755

Christian Pascual
cbp2128

July 17, 2019

1 Introduction

The diagnosis of breast cancer at an early stage is an important goal of screening. While many different screening tests exist today, there is still room for improvement. A promising avenue for breast cancer detection is prediction based off of breast cancer tissue images. For example, a study by Kim et al. demonstrated that regression models could be helpful in diagnosing breast cancer using Logistic LASSO and stepwise logistic regression. [1]

This report seeks to validate their findings and use regression methods to predict breast cancer status from image data. The two techniques of interest will be a logistic model using the Newton-Raphson optimization and another using coordinate descent in a LASSO context.

2 Methods

2.1 The data set

Our data set contains 569 observations and 33 rows. The response is a binary variable, $diagnosis$, that takes either M for malignant tissue or B for benign tissue. We take "malignant" to take on the 1 value. There are 30 potential predictors that are derived from the image data, including the mean, standard deviation and largest values of various features in the images. Important examples of features included in the data set include cell nuclei radius, texture (derived from gray-scale values), and concavity.

The data was centered and scaled before use in any model to ensure comparability between models and to prevent undue influence from differences in magnitude between predictors.

2.2 Full model

Due to the highly correlated nature of the data set, we have chosen a subset of 11 predictors of the original 31 to represent a full model. The full set of chosen predictors is given in Table 1.

2.3 Newton-Raphson model

For logistic regression, the probability of a malignant tissue is given by:

$$P(Y_i = 1|X_i) = p_i = \frac{e^{\beta_0 + X_i\boldsymbol{\beta}}}{1 + e^{\beta_0 + X_i\boldsymbol{\beta}}}$$

With n observations, we can derive the likelihood $L(y; \boldsymbol{\beta})$ and the log-likelihood $l(y; \boldsymbol{\beta})$ for data $(\boldsymbol{x_1}, Y_1)$, ... , $(\boldsymbol{x_n}, Y_n)$:

$$L(\boldsymbol{\beta}) = \prod[(\frac{e^{\boldsymbol{x_i^T}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x_i^T}\boldsymbol{\beta}}})^{Y_i}(\frac{1}{1 + e^{\boldsymbol{x_i^T}\boldsymbol{\beta}}})^{1-Y_i}]$$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i(\beta_0 + X_i\boldsymbol{\beta}) - log(1 + e^{\beta_0 + X_i\boldsymbol{\beta}})$$

From the log-likelihood, we can derive both the gradient $\nabla l(y;\beta)$ and the Hessian $\nabla^2 l(y;\beta)$

$$\nabla l(y;\beta) = \begin{bmatrix} \sum_{i=1}^{n}(Y_i - p_i) \\ \sum_{i=1}^{n} x_{i1}(Y_i - p_i) \\ . \\ . \\ . \\ \sum_{i=1}^{n} x_{in}(Y_i - p_i) \end{bmatrix}$$

$$\nabla^2 l(y;\beta) = \begin{bmatrix} \sum_{i=1}^{n} p_i(1-p_i) & \sum_{i=1}^{n} X_i^T p_i(1-p_i) \\ \sum_{i=1}^{n} X_i p_i(1-p_i) & \sum_{i=1}^{n} X_i X_i^T p_i(1-p_i) \end{bmatrix}$$

With these components, we can use the Newton-Raphson algorithm to calculate a set of β coefficients with the following equation:

$$\boldsymbol{\beta_{i+1}} = \boldsymbol{\beta_i} - [\nabla^2 l(\boldsymbol{\beta_i})]^{-1} \nabla l(\boldsymbol{\beta_i})$$

for which at each iteration, the log-likelihood will be recalculated. The Newton-Raphson algorithm requires an initial guess for the β vector. For our initial guess, we used a vector of size 13, with each element equal to 0.001. We defined convergence as when the difference between the current log-likelihood and the last iteration's log-likelihood reached below 10^{-5}.

2.4 Logistic-LASSO Model

For this model, we sought to minimize the objective function, derived from the quadratic Taylor approximation to the binomial log-likelihood function:

$$\underset{(\beta_0,\boldsymbol{\beta_1})}{min} \left(\frac{1}{2n} \sum_{i=1}^{n} \omega_i (z_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta_1})^2 + \lambda \sum_{j=0}^{p} |\beta_j| \right)$$

ω_i is the working weight, z_i is the working response, β_0 is the intercept and $\boldsymbol{\beta_1}$ is the set of β coefficients.

This objective function was minimized using coordinate descent. The intercept term was not penalized. Each β_k was optimized using the following equation:

$$\tilde{\beta}_j = \frac{S(\sum_i \omega_i x_{i,j}(y_i - \tilde{y}_i^{(-j)}), \gamma)}{\sum \omega_i x_{i,j}^2}$$

where S is the weighted soft threshold function, y^{-j} is the response omitting β_j and γ is the threshold to be used on all the β_k.

Similar to our Newton-Raphson algorithm, our Logistic-LASSO defined convergence to occur when the Frobenius norm between the calculated β and the β from the last iteration to drop below 1^{-5}. The same 0.001 vector will also be used to initialize the algorithm.

2.5 Assessing optimal parameters and predictive ability

5-fold cross validation will be used to find an optimal λ value for the Logistic-LASSO model. The optimal λ will be defined as the λ that minimizes the average test MSE across the tested λ values.

In order to assess predictive ability, we will use 10-fold cross validation to derive an average test MSE for each of the 2 models (Newton-Raphson, and Logistic LASSO). The best model will be defined as the one with the lowest average test MSE among the folds.

Table 1: Comparison of model coefficient estimates

Coefficient	Newton-Raphson	Logistic-LASSO
Iterations	9	119
Intercept	-0.2878910	-0.5106712
Mean Radius	2.4041094	2.2687699
Mean Texture	2.0356397	1.6859265
Mean Smoothness	0.8524514	0.7658457
Mean Concavity	5.9133791	4.4779677
Mean Symmetry	0.7718074	0.6047137
Mean Fractal Dimension	-0.9057417	-0.6121695
SE Radius	1.9680360	1.3665148
SE Texture	-0.8839497	-0.6389937
SE Smoothness	-0.0380407	-0.0830219
SE Concavity	-2.6835101	-1.8917712
SE Symmetry	-0.5401464	-0.4096745

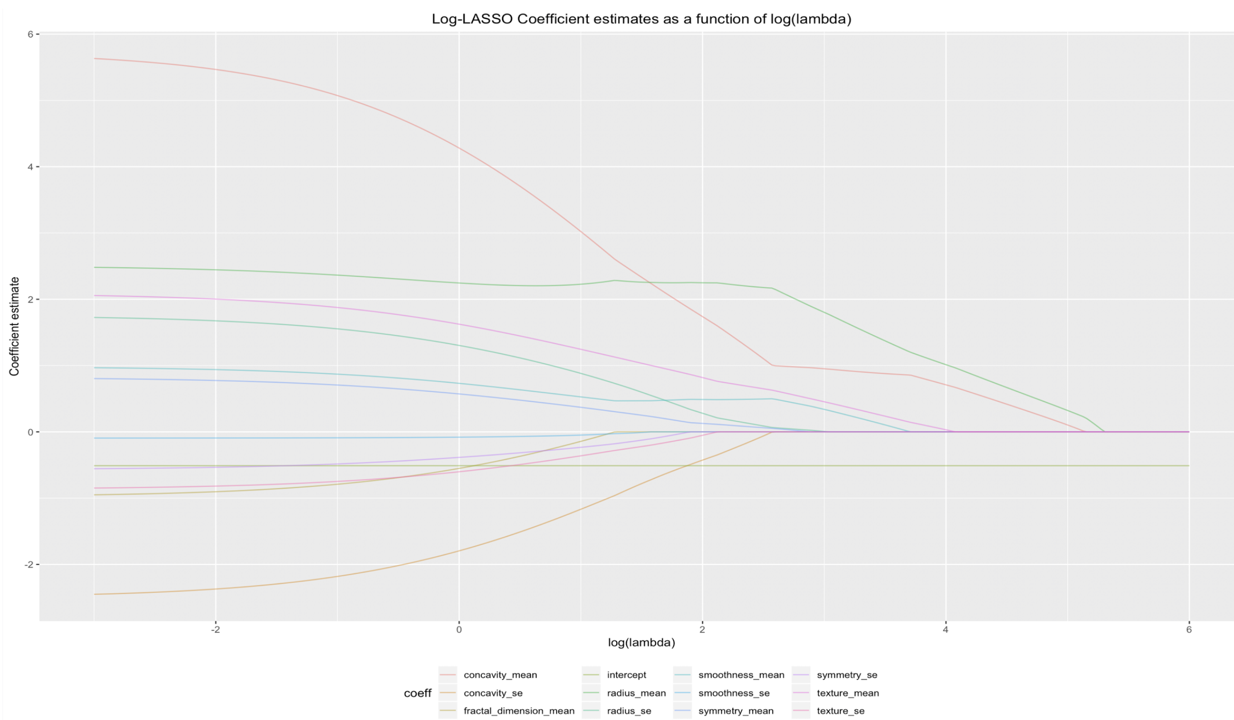

Figure 1: Comparing average test MSE against λ

3 Results

3.1 Model coefficient estimates

The estimated coefficients for each model are listed in Table 1. The estimates for the Newton-Raphson algorithm match what is estimated in the glm implementation of logistic regression. The estimates for the Logistic-LASSO match up closely against the implementation in $glmnet$. Newton-Raphson reaches convergence much faster than the Logistic LASSO.

Figure 1 illustrates a path of solutions along increasing λ's. The constant line represents the intercept since we chose not to penalize it in our implementation.

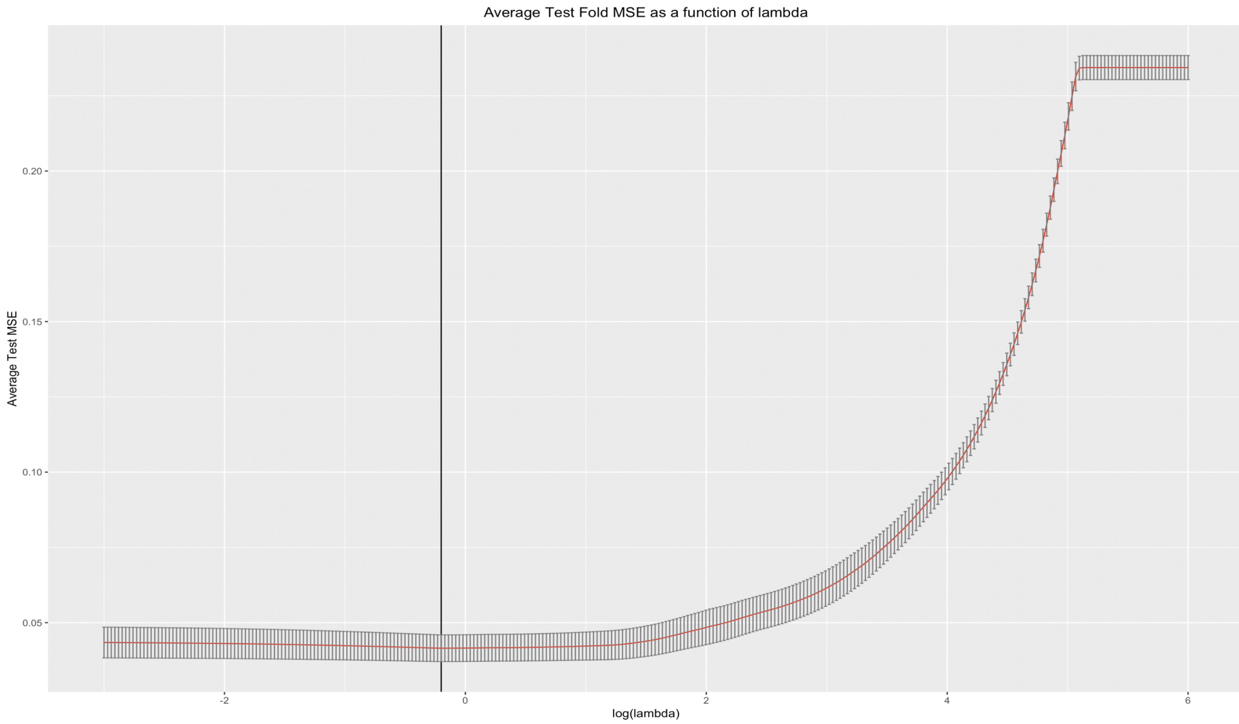

Figure 2: Comparing average test MSE against λ

3.2 Optimal lambda for Logistic-LASSO

After the 5-fold cross validation, we found that a λ value of 0.818 had the lowest MSE between the range of e^{-3} to e^{6} Figure 2 shows how the average test MSE among the 5 folds changed as a function of $log(\lambda)$, the minimum is depicted at the horizontal black line.

3.3 Model with best predictive ability

Figure 3 graphs the distribution of test MSEs created in the 10-fold cross validation. The distribution of test MSEs is similar for both Newton-Raphson and Logistic LASSO, with Newton-Raphon having a slightly lower distribution.

4 Conclusion

This report sought to explore and compare how two different models are able to predict cancer malignancy given various image data. The optimization of these models requires the use of the Newton-Raphson and coordinate descent algorithms in a logistic regression context. We used cross validation to optimize the regularization parameter λ and to judge the predictive ability of both models. In the end, Logistic-LASSO produced a model that was more effective at predicting cancer malignancy in the data. These types of models have promise in improving healthcare through improved diagnostics.

References

[1] Kim, Sun Mi et al. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography *Ultrasonography (Seoul, Korea)* vol. 37,1 (2017): 36-42.

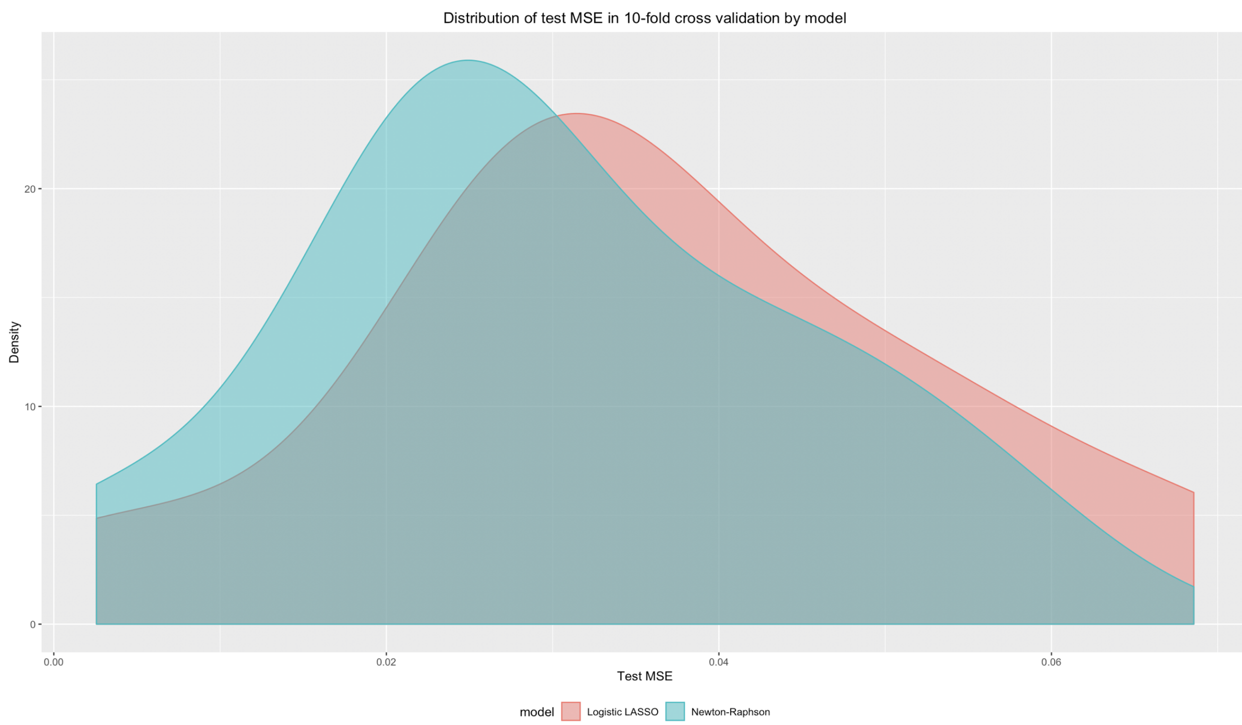

Figure 3: Comparing average test MSE against λ