Project 2: Breast Cancer Prediction Model

Group 7: Melanie Mayer, Yuqi Miao, Sibei Liu, Xue Jin

March 31, 2020

### Introduction

Breast cancer is the most common invasive cancer and the second leading cause of death from cancer in women worldwide, marked by the uncontrolled growth of breast cells. Non-cancerous breast tumors do not metastasize and are usually not life-threatening, while malignant tumors are cancerous, aggressive and deadly. Therefore, it’s important to have breast lumps accurately diagnosed so that decision with regard to medical treatment, rehabilitation and personal matters can be made appropriately.

### Objectives

The main objective of this project is to build an accurate predictive model based on logistic regression that classifies between malignant and benign images of breast tissue. Using the Breast Cancer Diagnosis dataset, logistics model and logistic-LASSO model will be implemented to predict the diagnosis. A Newton-Raphson algorithm and Pathwise Coordinate optimization will be developed to estimate the logistic model and the lasso model respectively. We aim to find the model with the best performance in terms of predicting when breast tissue is malignant.

### Breast Cancer Diagnosis Dataset

The Breast Cancer Diagnosis dataset, with 569 observations, contains the diagnosis and a set of 10 features capturing the characteristics of the cell nuclei present in the digitized image of breast mass. The ten features collected for each cell nucleus are:

* radius (mean of distances from center to points on the perimeter)
* texture (standard deviation of gray-scale values)
* perimeter
* area
* smoothness (local variation in radius lengths)
* compactness (perimeter^2 / area - 1.0)
* concavity (severity of concave portions of the contour)
* concave points (number of concave portions of the contour)
* symmetry
* fractal dimension (“coastline approximation” - 1)

The mean, standard deviation (SD) and largest values of these features are computed for each image, resulting in 30 possible predictive variables. We will analyze the predictive ability for diagnosis of malignant or benign cases of these covariates. The mean of each feature will be included in our models and presented in the result section. We will also create models with all 30 predictors however, these results will be reported in the supplement section.

### Methods

#### Model Parameters

For logistic regression we assume the response variable for the observation follows a binary distribution:

where denotes the probablity that the observation’s tissue is malignant. We can assume all observations are independent from one another, hence the likelihood function for the vector can be written as:

For logistic regression, the logit link function is used:

where and . One can solve for . We aim to find the best estimate of the vector of coefficients, . The log-likelihood function for this vector can be written as:

$$l(\boldsymbol\theta)=logL(\boldsymbol\pi) =\sum\_{i=1}^n(Y\_ilog\frac{\pi\_i}{1-\pi\_i}+log(1-\pi\_i)) = \sum\_{i=1}^n(Y\_i\theta\_i - log(1 + e^{\mathbf \theta\_i}))\\$$

The maximum likelihood is thus achieved when the gradient is equal to zero and the Hessian is negative definite. The gradient can be found to be:

The Hessian matrix is thus:

#### Full model

In order to estimate we need to maximize the loglikelihood function. There is no closed form, hence we turn to numerical methods. The Newton-Raphson method is used to fit the full logistic model. To find the maximum likelihood estimate of each element of , an iterative process is set as follows:

This is the Newton-Raphson method with two modifications. is included in the process to accomplish the step-halving modification, the step coefficient ensures the likelihood is always increasing in order to achieve quicker convergence. Once the likelihood approaches convergence the steps become smaller until convergence is reached. is the modification coefficient to ensure the aescent direction of the iteration vector at .

#### Logit-lasso Pathwise Coordinate-wise Update Algorithm

In the case of large dimensionality or multicollinearity of predictors it can be beneficial to perform a regularization method which shrinks coefficients and can perform variable selection. Here we implement the Least Absolute Shrinkage and Selection Operator (LASSO) method for logistic regression with a path-wise coordinate-wise optimization algorithm to select variables from the full model, increase the prediction efficiency and avoid overfitting. With the ten predictor model, LASSO may not be as beneficial compared to the 30 predictor model where we have 30 predictors describing 10 features hence are likely to experience multicollinearity and expect LASSO to outperform the classical logistic regression.

In LASSO we add an L1 penalization term to the loss function, such that we try to find the coefficients to minimize:

With the pathwise coordinate-wise update algorithm, we find the likelihood to be:

where,

One must pre-define the tuning parameter sequence and a starting vector . When choosing the sequence of lambdas, it is best to define , where refers to the estimated coefficients from the logistic regression, such that all coefficients will shrink to zero. The optimal process for each is then found by using the optimal from the previous iteration as a warm start in order to reach the optimal values quicker. Within every iteration for each lambda, the optimal is searched using coordinate-wise updating, that is it is minimized over one parameter at a time while keeping all others fixed.

For in iteration

#### Cross Validation

In order to find the optimal lambda, we use 5-fold cross validation. The dataset is divided into five subdatasets. The optimal coefficients is then found five times by running the logit-LASSO on a combination of four of the five subsets, leaving a different subset out each time. The subset left out is then used to estimate the model performance. This is done for all lambdas in the pre-defined sequence in order to search for the lambda with the highest average predictive ability. The statistics we use to compare predictive ability is SSE and pearson chi-square statistic. SSE is defined as:

where . Pearson chi-square statistic is defined as:

We want both of these statistics to be minimized. By taking the average of the above statistics over the value found for each of the five folds, we get a value to evaluate the model fit for each lambda.

### Results

#### Logistic Regression Model

The results from the logistic model estimated using the modified Newton-Raphson algorithm with step-halving and gradient ascent can be found in Table 1. Here we are showing the results when only using the means of the ten features as predictors. We found average area had the largest coefficient. For a one unit increase in the average area of the tissue, we expect to see a change in the log odds of having a malignant tutor of 14, holding all other predictors constant. The predictors were all standardized so caution must be used when analyzing these coefficients however, this seems to be very high however. The smallest coefficient, -7.22, was estimated for average radius. The results from logistic regression using the mean, SD and largest values of the 10 features (30 predictors) are shown in Supplementary Table 1.

#### Logistic-LASSO Model

The sequence of we compared ranged from three to zero, with a length of 100. The pre-defined values for all coefficients including the intercept was 0.02. The optimal was selected based on the minimum SSE, maxmimum AUC and minimum g-statistics estimated on the test dataset. The results from each of the five folds can be found in Table 2.

We find the results are consistant across all five folds based on all three statistics. The optimal appears to be zero. This implies the L1 penalization term added to the loss function is not effecting our model. Using the test data of each fold, the AUC ranges from 0.99 to 0.97, the g-statistic ranges from 22.6 to 48.0, while SSE ranges from 4.1 to 6.8.

Figure 1 shows the trend in AUC, SSE and g-statistic over the range of lambdas tested. With the increase of , both SSE and g-statistics have a dramatic increase and become constant around 0.4. AUC decreases drastically from 1 to 0.4. All of them indicate that the bigger the is, the worse the model performs.

Since the optimal = 0, none of the coefficients of the logistic-lasso is shrunk to zero, as seen in Table 3. The coefficient terms are much smaller compared to those estimated by the classic logistic regression however.

The results when all 30 predictors are considered, and their corresponding Cross Validation results in each fold, are shown in Supplementary Table 2. The optimal under that scenario is 0.03, which shrinks about 20 coefficients to 0, as presented in Supplementary Table 3.

### Discussion

When attempting to estimate a binary outcome, there are many other models which could have been used. We could have compared our results to other generalized linear models using a probit or complementary log-log link function.

### Supplement