# P8160 Group Project 2: Breast Cancer Diagnosis and Optimizations

Wenhan Bao | Tianchuan Gao | Jialiang Hua | Qihang Wu | Paula Wu

3/25/2022

## Objective

The main objective of our project is to build an accurate predictive model based on logistic regression that classifies malignant and benign images of breast tissue. Using the Breast Cancer Diagnosis dataset, we will implement logistic model and logistic-LASSO model 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. 5-fold cross-validation will be applied to find the optimal value for the tuning parameter. Our aim is to find the model with the best performance in predicting the diagnosis of breast tissue images.

## Background & Methods

### Background

Breast cancer, which affects 1 in 7 women worldwide, is the most common invasive cancer in women around the world. It can start from different parts of the breast and is marked by the uncontrolled growth of breast cells. Benign breast tumors do not metastasize and are usually not life-threatening, while malignant tumors are aggressive and deadly. Nowadays, substantial support for breast cancer awareness and research funding has created great advances in the diagnosis and treatment of breast cancer. The prognosis of the disease has been greatly improved once it is detected and treated early. Therefore, it's important to have breast lumps accurately diagnosed so that timely clinical intervention can be conducted.

The dataset we uses contains 569 observations and 33 columns, including 357 benign and 212 malignant cases. 30 out of the 33 columns are predictors that contain the mean, standard deviation and the largest value of the distributions of 10 distinct features listed below:

- 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$\hat{2}$ / 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)

Our goal is to build a predictive model based on logistic regression to facilitate cancer diagnosis.

# Methods

## Exploratory Data Analysis (EDA)

Firstly, we imported and cleaned the data and conducted exploratory data analysis. We plotted the correlation plot of all the predictors by R. From the plot, we can find that there is a high correlation between many of the predictors. This is because our predictors include mean, standard deviation and the largest values of the distributions of 10 features, which means that some predictors can be calculated by other predictors. Then, we made the feature plots to explore the relationship between binary diagnosis outcome and all our predictors to determine the potential effective predictors.

## Modified Newton-Raphson Algorithm

***Model Parameters***

The logistic model can be defined as:

$$log(\frac{\pi_i}{1-\pi_i}) = \mathbf{X}\boldsymbol{\beta}$$

$\pi_i$ stands for the probability that the tissue is malignant in the $i_{th}$ observation and it can be written as:

$$\pi_i = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1+e^{\mathbf{X}\boldsymbol{\beta}}}$$

The likelihood function for vector $\beta$ is:

$$L(\beta) = \prod_{i=0}^{n}(\frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}})^{y_i}(\frac{1}{1+e^{\mathbf{X}\beta}})^{1-y_i}$$

and thus we can calculate the log-likelihood function, gradient and Hessian matrix respectively.:

$$l(\beta) = \sum_{i=0}^{n}(y_i * X\beta - log(1+e^{X\beta}))$$

$$\nabla l(\beta) = \begin{pmatrix} \sum_{i=1}^{n}(y_i - \pi_i) \\ \sum_{i=1}^{n} x_{i1} \times (y_i - \pi_i) \\ . \\ . \\ . \\ \sum_{i=1}^{n} x_{in} \times (y_i - \pi_i) \end{pmatrix}$$

$$\begin{aligned} \nabla^2 f(\beta) &= -\sum_{i=1}^{n} \begin{pmatrix} 1 \\ x_i \end{pmatrix} \begin{pmatrix} 1 & x_i \end{pmatrix} \pi_i (1-\pi_i) \\ &= -\begin{pmatrix} \sum \pi_i (1-\pi_i) & \sum x_i \pi_i (1-\pi_i) \\ \sum x_i \pi_i (1-\pi_i) & \sum x_i^2 \pi_i (1-\pi_i) \end{pmatrix} \end{aligned}$$

***Model Explanation***

With those parameters above, we can apply the Newton-Raphson algorithm to calculate $\beta$ with the following equation in each iteration:

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

We also apply two modifications: step-halving and re-direction. Both modification are meant to ensure the likelihood is increasing the the ascent direction at each iteration.

***Results***

We apply the modified Newton-Raphson algorithm on the first ten features(mean) and get the following result:

## Logistic-LASSO

***Least Absolute Shrinkage and Selection Operator (LASSO)*** To estimate coefficients through Newton-Raphson method, it is necessary to compute the corresponding inverse of Hessian Matrix $[\nabla^2 f(\theta_{i-1})]^{-1}$. However, the computational burden of calculation will increase as the dimension of predictors increases and the collinearity will also be a problem. Therefore, we use a regularization method, LASSO, to shrink coefficients and perform variable selections. For regression lasso, the objective function is:

$$f(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda\sum_{i=1}^{p}|\beta_j|,$$

where the first term is residual sum of squares (RSS) and the second term is the lasso l1 penalty. Noted that the $x_{i,j}$ needs to be standardized before LASSO so that the penalty will be equally applied to all predictors. For each single predictor, the LASSO solution is like:

$$\hat{\beta}^{lasso}(\lambda) = S(\hat{\beta}, \lambda) = \begin{cases} \hat{\beta} - \lambda, \text{ if } \hat{\beta} > 0 \text{ and } \lambda < |\hat{\beta}| \\ \hat{\beta} + \lambda, \text{ if } \hat{\beta} < 0 \text{ and } \lambda < |\hat{\beta}| \\ 0, \text{ if } \lambda > |\hat{\beta}|, \end{cases}$$

where $S(\hat{\beta}, \lambda)$ is called soft threshold. The basic idea of this function is to shrink all $\beta$ coefficients between $-\lambda$ and $\lambda$.

***Coordinate-wise Descent Algorithm*** Another approach to solve the complex computation of inverse Hessian Matrix is coordinate descent approach, which starts with an initial guess of parameters $\theta$, optimize one parameter at one time based on the best knowledge of other parameters, and use the results as the start values for the next iteration. We then repeat the above steps until convergence. Finally, this approach tries to minimize the following objective function when considering a lasso penalty:

$$f(\beta_j) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k\neq j} x_{i,j}\tilde{\beta}_k - x_{i,j}\beta_j)^2 + \lambda\sum_{k\neq j}|\tilde{\beta}_k| + \lambda|\beta_j|,$$

where $\tilde{\beta}$ represents the current estimates of $\beta$ and therefore constants. If we also consider a weight $\omega_i$ is associated with each observation, then the updated $\beta_j$ in this case is:

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

where $\tilde{y}_i^{(-j)} = \sum_{k\neq j} x_i\tilde{\beta}_k$. From here, we use the Taylor expansion. So the log-likelihood around "current estimate" is:

$$l(\beta) = -\frac{1}{2n}\sum_{i=1}^{n}\omega_i(z_i - X_i\beta)^2,$$

where working weights $\omega_i = \tilde{\pi}_i(1-\tilde{\pi}_i)$, working response $z_i = X_i\tilde{\beta} + \frac{y_i - \tilde{\pi}_i}{\tilde{\pi}_i(1-\tilde{\pi}_i)}$, and $\tilde{\pi}_i = \frac{exp(X_i\tilde{\beta})}{1+exp(X_i\tilde{\beta})}$.
Finally, similar to regression lasso, the logistic lasso can be written as a penalized weighted least-squares problem like this:

$$min_\beta L(\beta) = -l(\beta) + \lambda\sum_{j=0}^{p}|\beta_j|$$

.

***Pathwise Coordinate-wise Algorithm*** The difference between pathwise coordinate-wise method and coordinate-wise method is that a sequence value of lambda is required to input. There are some steps taken as followed:

- Select a $\lambda_{max}$ for all the estimate $\beta = 0$ which is the inner product ($max_l \langle X_l, y\rangle$).

- Compute the solution with a sequence of descending $\lambda$ from maximum to zero ($\lambda_{max} > \lambda_k > \lambda_{k-1} ... > 0$).
- Initialize coordinate descent algorithms for $\lambda_k$ by the calculated estimate $\beta$ from previous $\lambda_{k+1}$ as a warm start
- By repeating the two steps above, a sequence of optimal coefficients $\beta$ for each descending $\lambda$. When objective function is taken the minimum value ($min_\beta L(\beta)$), the best $\lambda$ could be identified and optimal coefficients $\beta$ could be selected under this best $\lambda$

**5-fold Cross-validation**

In a nutshell, a 5-fold cross-validation first splits the shuffled training data (80% of the whole dataset) into 5 parts and takes one group as the hold-out set (validation set) while fitting the model using the remaining 4 training sets. After a model is fitted, we evaluate and retain the model performance, namely AUC and RSS, and move on to the next iteration. After 5 iterations, we calculated the mean RSS and the mean AUC with standard deviations. Our goal is to find an optimal $\lambda$ that maximizes the mean AUC.

In our implementation, we initially choose 30 equally spaced $\lambda$ from $e^{(-4,3)}$. The reason why the $\lambda$'s are widespread is that we would like to determine the rough range, instead of the precise value, of the optimum of this tuning parameter first. For each $\lambda$, we ran the 5-fold cross-validation. Within each fold, we use Coordinate-wise optimization to update the coefficients vectors $\boldsymbol{\beta}$ and evaluate the performance by calculating AUC and RSS. Among 30 $\lambda$, the optimal $\lambda$ that maximizes the AUC equals 2.3681. We then narrow down the range to (3, 0) and choose 150 equally spaced $\lambda$ in between. The best $\lambda$ selected by cross-validation equals 1.5705.

# Results

## Newton-Raphson Methods

## Pathwise Logistic-LASSO Model

In the part of Pathwise logistic-lasso model, we firstly select the maximum $\lambda$ which is the inner product of response and all variables and $\lambda_{max} = 218.13$. By using a sequence of $\lambda$ to minimize the objective function and have a relatively higher AUC (Figure.1), the minimum value would be taken under when $\lambda = 1.57$ which means the optimal solution will be computed for this optimal $\lambda$ value.

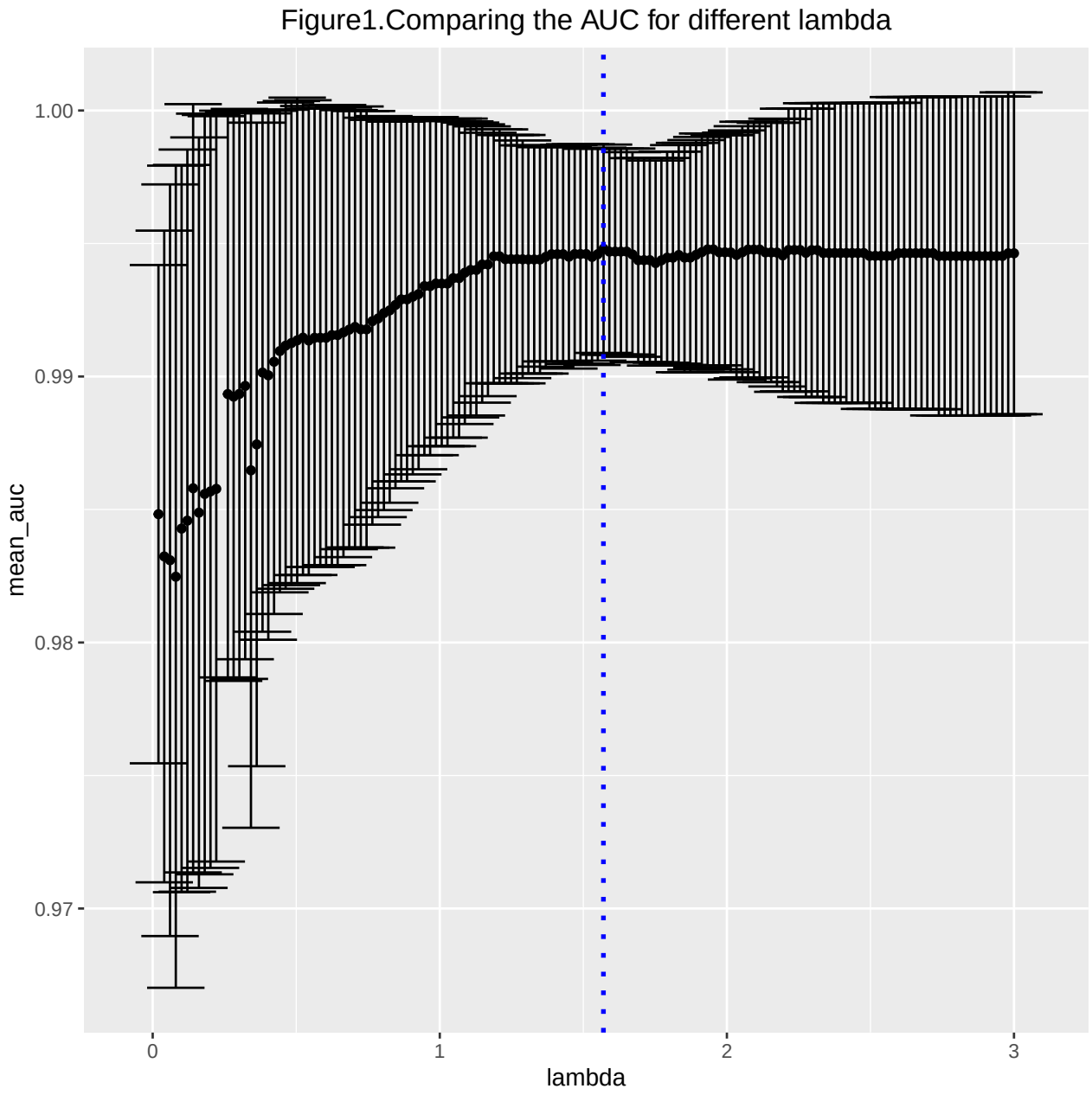


Table 1: Comparing the coefficients for optimal and full model

| beta | full_coefficients | optimal_coefficients |
|---|---|---|
| V1 | -6.7120020 | -0.2307176 |
| V2 | 15.3615092 | 0.0000000 |
| V3 | 1.2214100 | 0.0000000 |
| V4 | 6.6009820 | 0.0000000 |
| V5 | 2.0001412 | 0.0000000 |
| V6 | 3.8129595 | 0.0000000 |
| V7 | 3.6996305 | 0.0000000 |
| V8 | 6.3315880 | 0.1514531 |
| V9 | -4.5445495 | 0.9240514 |
| V10 | 6.0177074 | 0.1887078 |
| V11 | -3.7150361 | 0.0000000 |

| beta | full_coefficients | optimal_coefficients |
|---|---|---|
| V12 | 0.8185762 | 0.0000000 |
| V13 | -1.9716802 | 0.0000000 |
| V14 | 8.6455676 | 0.0000000 |
| V15 | 0.7532164 | 0.0000000 |
| V16 | -7.2858931 | 0.0000000 |
| V17 | -2.9384468 | 0.0000000 |
| V18 | 1.4394241 | 0.0000000 |
| V19 | -13.9518630 | 0.0000000 |
| V20 | -3.8392690 | 0.0000000 |
| V21 | 3.4956263 | 0.0000000 |
| V22 | 18.2512163 | 0.0000000 |
| V23 | 6.7310378 | 0.1302448 |
| V24 | 19.3803433 | 3.7063624 |
| V25 | 7.7215211 | 0.0000000 |
| V26 | 17.3699841 | 0.3410588 |
| V27 | 2.7156803 | 0.0000000 |
| V28 | 7.1459593 | 0.2560776 |
| V29 | -3.1117336 | 0.0196025 |
| V30 | -4.6306316 | 0.0203774 |
| V31 | 6.3252661 | 0.0000000 |

In the above table, we directly compare the coefficients estimates $\beta$ for all 30 variables and $\beta_0$ for intercept (V1-V31 represent the $\beta$). In the optimal lasso model, there are only ten variables included while the rest predictors are all shrinked towards zero.

### Comparsion: "Optimal" model vs. "Full" model

As the name suggested, the "optimal" model has $\lambda = 1.5705$ as its tuning parameter while the "full" model has $\lambda = 0$. We evaluate both models and obtain their AUC scores: optimal model has an AUC of 0.9994 and the full model has an AUC of 1. Thus, we concluded that the full model has a better prediction, even though the prediction performance of these two models only differ by 0.0006.

## Conclusion and Discussion

It's quite interesting that the full model has a perfect classification result. One plausible reason could be that the data is well separated for benign and malignant breast cancer.

## Contributions

We contributed to this project evenly.

## Appendix

## Reference