Project2 :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 cancer death 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 a predictive model based on binary logistic regression that classifies between malignant and benign cases. Using the Breast Cancer Diagnosis dataset, logistics model and logistic-LASSO model will be implemented to predict the diagnosis. As two important methods for numerical optimization, Newton-Raphson algorithm and Pathwise Coordinate optimization will be developed to estimate the logistic model and the lasso model respectively.

Breast Cancer Diagnosis Dataset

The Breast Cancer Diagnosis dataset contains the diagnosis and a set of 30 features capturing the characteristics of the cell nuclei present in the digitized image of breast mass. Ten features are collected for each cell nucleus:

- 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 (perimeterˆ2 / 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 error (SE) and largest values of these features are computed for each image, resulting in 30 features, which will be analyzed to understand the predictive value for diagnosis of cancer(malignant) or benign cases.

Methods

Model define

For i_{th} observation, the response variable Y_i follows binary distribution:

$$Y_i \sim Bin(\pi_i)$$

the log-liklihood function:

$$l(\mathbf{Y}, \boldsymbol{\pi}) = \sum_{i=1}^{n} l(y_i, \pi_i) = \sum_{i=1}^{n} (y_i log \frac{\pi_i}{1-\pi_i} + log(1-\pi_i)) = \sum_{i=1}^{n} (y_i log \frac{\pi_i}{1-\pi_i} + log(1-\pi_i))$$

Where π_i denotes the probablity of the i_{th} observation to be maglinant.

To build relationship between the response and predictors, the GLM is defined as:

$$\log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}_i \boldsymbol{\beta} = \theta_i$$

Full model

Firstly, Newton-Rapson method is used to fit the full model. To find the maximum likelihood estimation of coefficients, iteration process is set as follows:

$$\theta_{i+1} = \theta_i - \delta(\nabla^2 l(\theta_i|\boldsymbol{X}) - \gamma I)^{-1} \nabla l(\theta_i|\boldsymbol{X})$$

where δ is the step coefficient to ensure the increasing of likelihood funciton, and γ is the modification coefficient to ensure the aescent direction of the iteration vector.

$$\nabla l(\theta|\boldsymbol{X}) = \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\pi})$$

$$\nabla^2 l(\theta|\boldsymbol{X}) = -\mathbf{X}^T diag(\pi_i(1-\pi_i))\mathbf{X}$$

logit-lasso pathwise coordinate-wise update algorithm

To select variables and increase the prediction efficiency, lasso was integrated into the coordinate-wise logit regression.

The target function:

$$min\{-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

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

$$\pi_i = \frac{exp(\boldsymbol{X_i \beta})}{1 + exp(\boldsymbol{X_i \beta})}$$

$$\omega_i = \pi_i(1-\pi_i)$$

$$z_i = \boldsymbol{X_i\beta} + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)}$$

Pre-define the tuning parameter sequence $\{\lambda_1, ..., \lambda_s\}$, starting point $\boldsymbol{\beta_{start=}}\{\beta_0^{(0)}, ..., \beta_p^{(0)}\}$. Here we elaborately explain the optimal process for λ_u: Using the optimal $\boldsymbol{\beta_{u-1}}$ from last iteration as the warm start. within every iteration, find the optimal β coordinate-wise. For β_j in t_th iteration

$$\beta_j^{(t)} = \begin{cases} \frac{\sum_{i=1}^n \omega_i(z_i - \sum_{j=1}^p \boldsymbol{X_i}\beta_j)}{\sum_i^n \omega_i}, & j=0 \\ \frac{s(\beta_j^{(t*)}, \lambda_u n)}{\sum_{i=1}^n \omega_i x_{ij}^2}, & j = 1, 2, ..., p \end{cases}$$

$$\beta_j^{(t*)} = \sum_{i=1}^n \omega_i x_{ij} z_{ij}^*$$

$$z_{ij}^* = z_i - \sum_{\substack{k=0 \\ \beta_k \neq 0}}^{j-1} \beta_k^{(i)} x_{ik} - \sum_{\substack{k=j+1 \\ \beta_k \neq 0}}^{p} \beta_k^{(i-1)} x_{ik}$$

Cross Validation

In order to check the model performance, we use 5-fold cross validation. Using training dataset to choose the best tuning parameter and fit model, and using test dataset to evaluate the final prediction. The statistics we use to compare the validation is SSE and person chi-square statistics, which are defined as:

$$SSE = \sum_{i=1}^n (y_i - \widehat{\pi}_i)^2$$

$$\widehat{\pi}_i = log\frac{exp(\boldsymbol{X}_i\beta)}{1 + exp(\boldsymbol{X}_i\beta)}$$

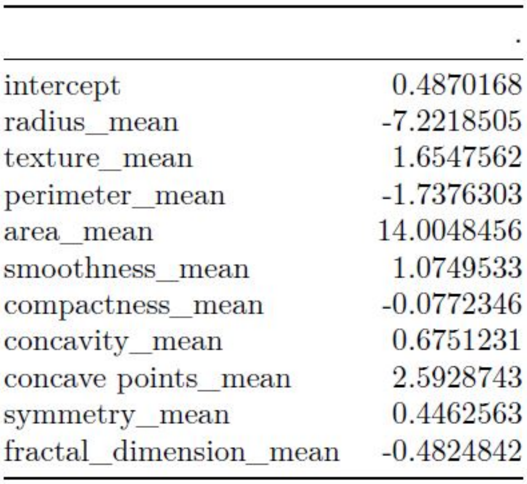
$$G = \sum_{i=1}^n \frac{y_i - \widehat{\pi}_i}{\widehat{\pi}_i(1-\widehat{\pi}_i)}$$

By taking average of above 2 statistics of 5 fold, we get the final index to evaluate the model fitting.

Results

Newton-Raphson

After doing the Newton-Raphson modified with step and direction, the estimation of coefficients are:

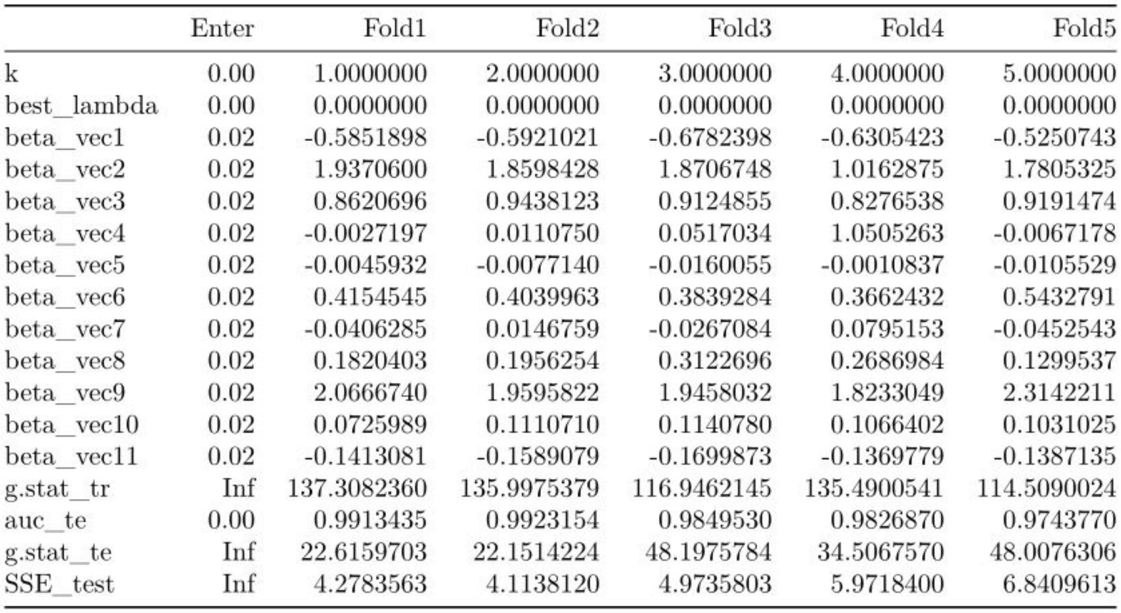


	.
intercept	0.4870168
radius_mean	-7.2218505
texture_mean	1.6547562
perimeter_mean	-1.7376303
area_mean	14.0048456
smoothness_mean	1.0749533
compactness_mean	-0.0772346
concavity_mean	0.6751231
concave points_mean	2.5928743
symmetry_mean	0.4462563
fractal_dimension_mean	-0.4824842

Table 1. Estimated coefficients under Newton-Raphson method

Coordinate-Wise

The range of λ we tried is (3,0) with length 100. The initial guess of all β including the intercept is 0.02. The g.statistics, SSE, AUC are introduced to select best λ in Cross Validation. The optimal λ would be selected based on the minimum SSE, maxmimum AUC and minimum g-statistics in test data. Below is the results in 5-Fold Cross Validation:



	Enter	Fold1	Fold2	Fold3	Fold4	Fold5
k	0.00	1.0000000	2.0000000	3.0000000	4.0000000	5.0000000
best_lambda	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
beta_vec1	0.02	-0.5851898	-0.5921021	-0.6782398	-0.6305423	-0.5250743
beta_vec2	0.02	1.9370600	1.8598428	1.8706748	1.0162875	1.7805325
beta_vec3	0.02	0.8620696	0.9438123	0.9124855	0.8276538	0.9191474
beta_vec4	0.02	-0.0027197	0.0110750	0.0517034	1.0505263	-0.0067178
beta_vec5	0.02	-0.0045932	-0.0077140	-0.0160055	-0.0010837	-0.0105529
beta_vec6	0.02	0.4154545	0.4039963	0.3839284	0.3662432	0.5432791
beta_vec7	0.02	-0.0406285	0.0146759	-0.0267084	0.0795153	-0.0452543
beta_vec8	0.02	0.1820403	0.1956254	0.3122696	0.2686984	0.1299537
beta_vec9	0.02	2.0666740	1.9595822	1.9458032	1.8233049	2.3142211
beta_vec10	0.02	0.0725989	0.1110710	0.1140780	0.1066402	0.1031025
beta_vec11	0.02	-0.1413081	-0.1589079	-0.1699873	-0.1369779	-0.1387135
g.stat_tr	Inf	137.3082360	135.9975379	116.9462145	135.4900541	114.5090024
auc_te	0.00	0.9913435	0.9923154	0.9849530	0.9826870	0.9743770
g.stat_te	Inf	22.6159703	22.1514224	48.1975784	34.5067570	48.0076306
SSE_test	Inf	4.2783563	4.1138120	4.9735803	5.9718400	6.8409613

Table 2. Cross validatin results

In all 5 folds, all three critria indicates the same optimal λ, whcih is 0.In the test data of each fold, the AUC ranges from 0.99 to 0.97, g-statistics ranges from 22.6 to 48.0, while SSE changes from 4.1 to 6.8.

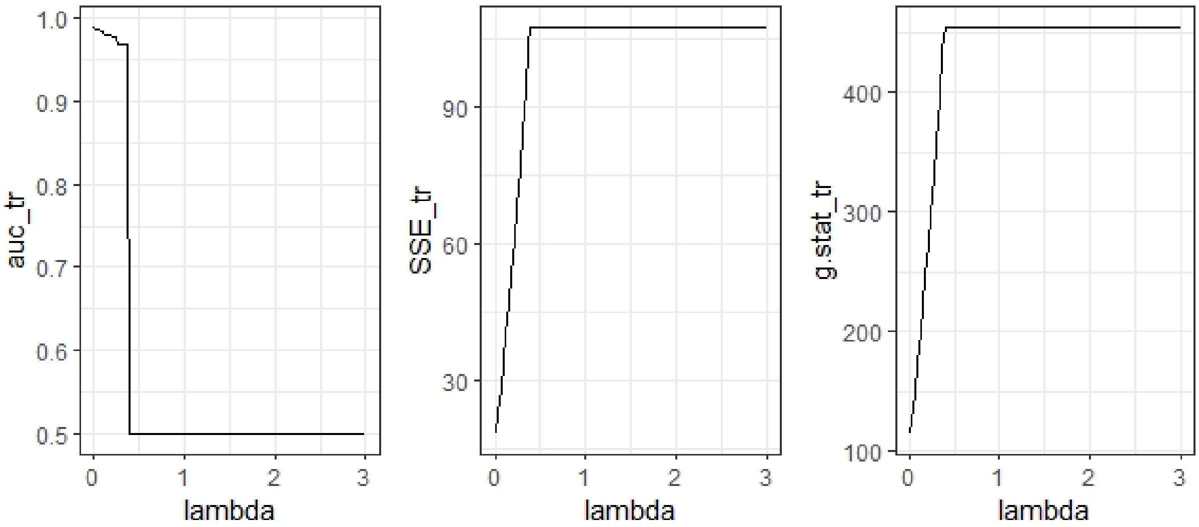


Fig 1. Changes in three criteria vs λ

The Figure 1 shows the trend of AUC SSE and g-statistics in train data. With the increse of λ, both SSE and g-statistics have a dramatic soar. While AUC has a great drop from 1 to 0.5. All of them indicate that the bigger the λ is, the worse the model would perform.

Discussion