Intro_logistic

Anyu Zhu

3/24/2022

Introduction

Background and Objective

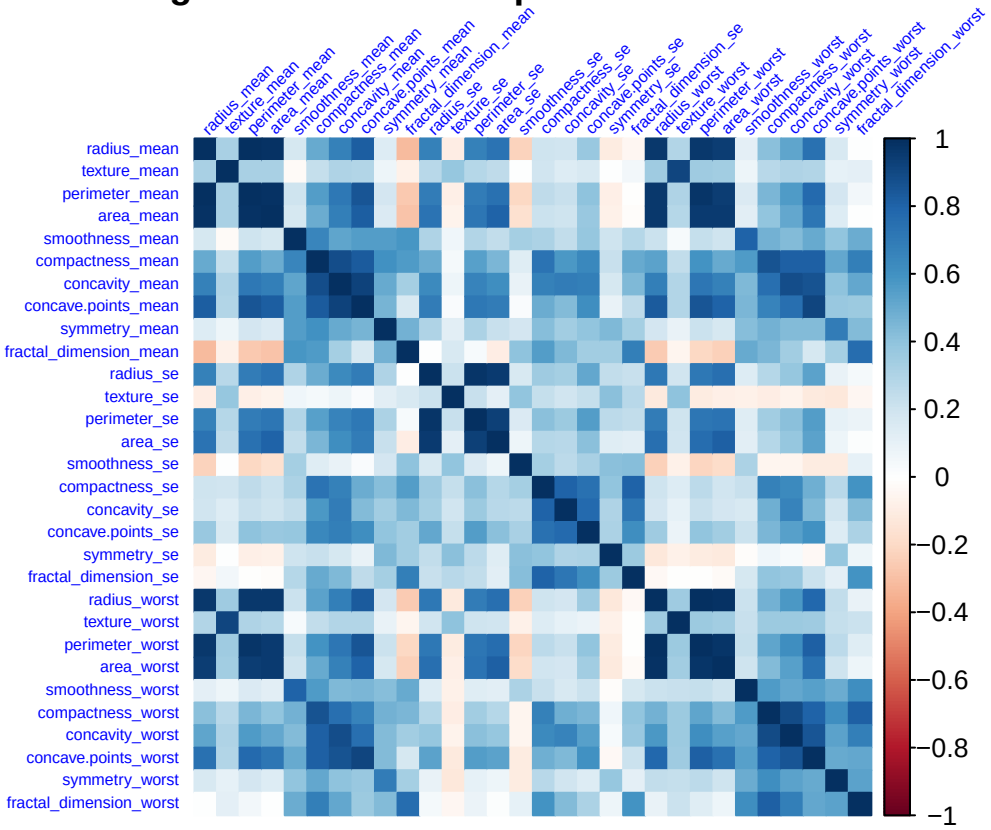
Breast cancer mainly occurs in middle-aged and older women. The median age at the time of breast cancer diagnosis is 62. This means half of the women who developed breast cancer are 62 years of age or younger when they are diagnosed. The goal of the project is to build a predictive model based on logistic regression to facilitate cancer diagnosis. We first build a logistic model to classify the images, then developed a Newton-Raphson algorithm to estimate the parameters of the logistic model. Then, we built a logistic-LASSO model to select features. Finally, we applied 5-fold cross-validation to select the best λ for the logistic-LASSO model.

Data Preprocessing

The dataset 'breast-cancer'we used contains 569 rows and 32 columns. The variable `diagnosis` identifies if the image is coming from cancer tissue or benign. We labeled `malignant` as 1 and `benign` as 0. In total there are 212 malignant cases and 357 benign cases. There are 30 variables corresponding to mean, standard deviation and the largest values (points on the tails) of the distributions of 10 features: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

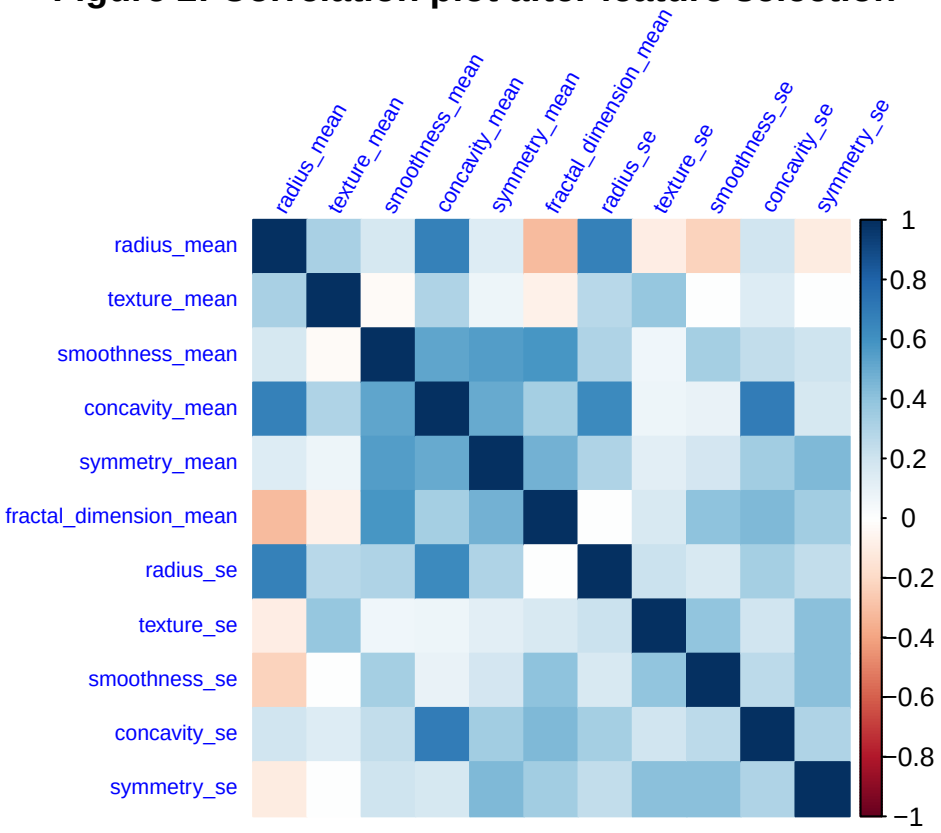
Figure q displays the correlation between variables. We can see the correlation coefficient is large between several pairs of variables, which will potentially cause the problem of not converging in Newton-Raphson algorithm and Logistic Lasso model.

Figure 1: Correlation plot of all variables



To reduce the multicolinearity effect, we conducted feature selection by removing variables with correlation coefficient > 0.7 and keep the rest 11 variables. After the adjustment, the correlation plot between variables change to:

Figure 2: Correlation plot after feature selection



We standardized the data by the `scale()` function in R, take the first 80% of observations as training dataset, and the rest 20% of observations as testing dataset for model comparison.

Method

Logistic Model

Take Y_i as the response of i_{th} observation and follows binary distribution $Bin(\pi_i)$. π_i is the probability of i_{th} observation being malignant. By applying the logit link:

$$g(\mu) = \text{logit}(\mu) = \log \frac{\mu}{1-\mu}$$

we have the logistic regression model:

$$\log \frac{\pi_i}{1-\pi_i} = X_i\beta$$

Thus we have the likelihood function of logistic regression

$$L(\pi) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

$$L(\beta; X, y) = \prod_{i=1}^{n} \left\{ \left(\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \right)^{y_i} \left(\frac{1}{1+\exp(X_i\beta)} \right)^{1-y_i} \right\}$$

Then maximize the log likelihood:

$$l(\beta) = \sum_{i=1}^{n} \{ y_i (X_i\beta) - \log(1+\exp(X_i\beta)) \}$$

By taking derivative with respect to β, the gradient is:

$$\nabla l(\beta) = \sum_{i=1}^{n} (y_i - \pi_i) \boldsymbol{x}_i = X^T (Y - \boldsymbol{\pi})$$

where $\pi_i = \frac{e^{\beta_i}}{1+e^{\beta_i}}$

By taking the second derivative, the Hessian matrix can be represented by:

$$\nabla^2 l(\beta) = -X^T \operatorname{diag}(\pi_i (1 - \pi_i)) X$$

i = 1, ... n. Hessian matrix is negative definite.