P8160 - Breast Cancer Diagnosis

Hongjie Liu, Xicheng Xie, Jiajun Tao, Shaohan Chen, Yujia Li

3/30/2023

Contents

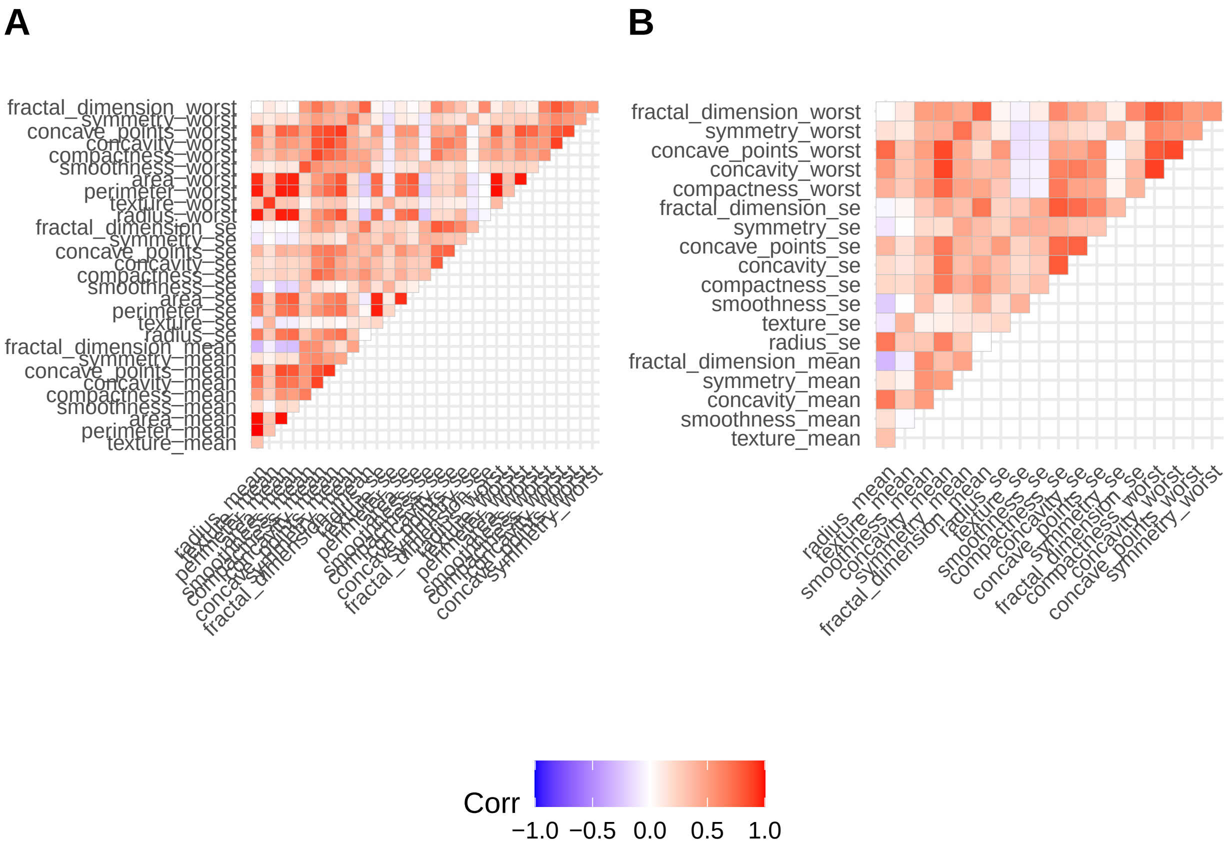
1. Objectives

Mammography is recognized as the most effective screening method for early breast cancer detection, but its accuracy remains limited. And as the number of variables that help predict breast cancer increases, doctors are forced to rely more on their subjective experiences to make decisions. The use of computer models can detect abnormalities in mammograms to aid radiologists in breast cancer diagnosis (Freer et al. 2001). The purpose of this study was to predict breast cancer benign/malignant status by quantitative modeling based on logistic regression, which may help radiologists manage the large amount of available information, make effective decisions to detect breast cancers and reduce unnecessary biopsy.

2. Background

The data given has 569 observations. The column 'Diagnosis' which identifies if the image is coming from cancer tissue or benign cases (M=malignant, B = benign) would be used as outcome for modeling. We denote malignant as 1 and benign cases as 0 for prediction. The other 30 columns correspond to mean, standard deviation and the largest values (points on the tails) of the distributions of the following 10 features computed for the cell nuclei:

- radius: mean of distances from center to points on the perimeter
- texture: standard deviation of gray-scale values
- perimeter: mean size of the core tumor
- area: mean area of the core tumor
- 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: symmetry of the tumor
- fractal dimension: ”coastline approximation” - 1



There are multicollinearity features among our predictors as shown above in the left graph that could cause unstable parameter estimation as well as perplexing the interpretation of logistic model. While we could eliminate some features for modeling as shown in the right graph, regularized logistic regression also help to tackle with it, which would be further explore in the following parts.

2. Methods

2.1. Logistic Model

Logistic model measure the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables, and commonly used in classifying binary response variables.

Write $p(X) = Pr(Y = 1|X)$, logistic regression uses the form $p(X) = \frac{e^{\beta_0+\beta_1X_1+...+\beta_pX_p}}{1+e^{\beta_0+\beta_1X+...+\beta_pX_p}}$. By applying the logit transformation, $log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1X_1 + ... + \beta_pX_p$, we use maximum likelihood to estimate the parameters.

2.2. Newton-Raphson Algorithm

While derivative of the likelihood function with respect to the parameters is nonlinear and difficult to solve analytically for maximum likelihood, it requires Newton-Raphson iterations.

2.3. Logistic-LASSO Model

LASSO is a frequently used model in predictor selection during regression. If we add penalization to the log-likelihood function of logistic regression, then we get: $logl(\beta) - \lambda\{(1-\alpha)\frac{1}{2}\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|\}$. λ represents the total amount of penalization, and by setting $\alpha = 1$, we get the logistic-lasso regression model.

2.4. Five-fold Cross Validation for LASSO

3. Results

5-fold CV Results for Logistic-LASSO

Model Comparison

4. Discussion

4.1. Summary

4.2. Limitations

4.3. Group Contributions

References

Freer, Timothy W., and Michael J. Ulissey. “Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center.” Radiology 220.3 (2001): 781-786.

Appendices