P8160 - Breast Cancer Data: To lasso or to not lasso

Amy Pitts, Hun Lee, Jimmy Kelliher,
Tucker Morgan, and Waveley Qiu

2022-03-28

Motivation

Diagnosing breast cancer is extremely important.

According to NIH there has been an estimated:

- 281,550 new cases of breast cancer in women in 2021,
- 43,600 breast cancer in women related deaths in 2021.

American Cancer Society Guideline for Breast Cancer Screening:

- Women between ages 25-40 should have an annual clinical breast examination.
- Women between ages 40-44 should begin annual screening via mammogram
- Women between ages 45-54 should screened annually via mammogram

Goal

With using all the collected imagine data we want to develop an algorithm to predict diagnosis. Since diagnosis is a binary outcome a logistic regression will be utilized.

Methods:

- Newton-Raphson Algorithm (Full Model)
- Logistic LASSO Algorithm (Optimal Model)

Data

- 569 rows and 31 columns all related to breast tissue images
- Outcome of interest: Diagnosis (B or M)
  - 357 benign (B) cases and 212 malignant (M) cases
- The Covariates include information such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

Figure 1: Ranked Cross-Correlations

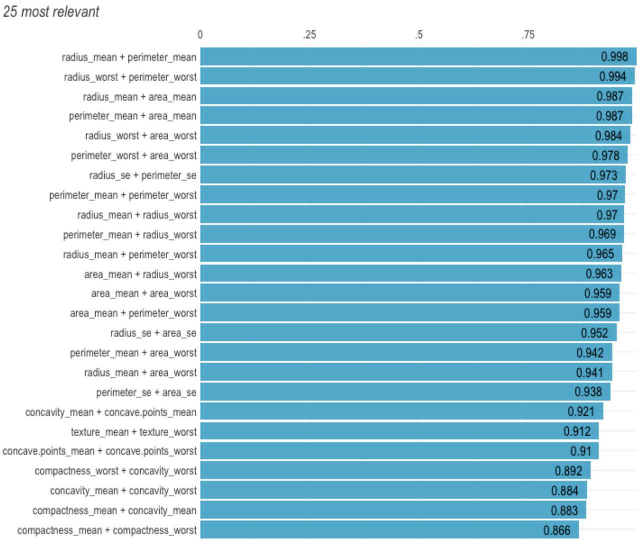
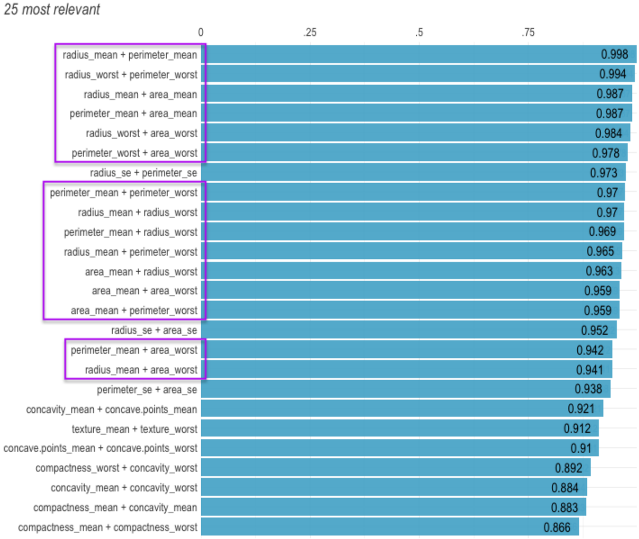
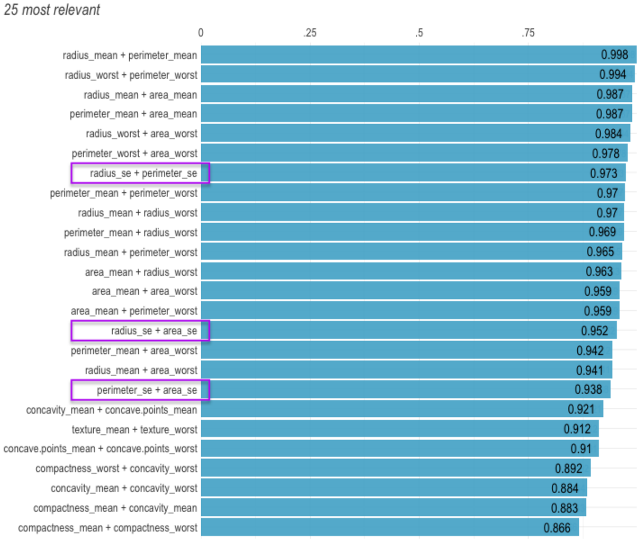


Figure 1: Ranked Cross-Correlations



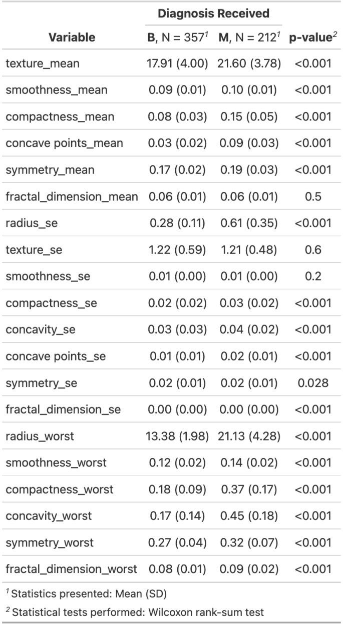
Best Representative `radius_worst`

Figure 1: Ranked Cross-Correlations



Best Representative `radius_se`

Table 1: Remaining Variables



	Diagnosis Received		
Variable	**B**, N = 357[1]	**M**, N = 212[1]	**p-value**[2]
texture_mean	17.91 (4.00)	21.60 (3.78)	<0.001
smoothness_mean	0.09 (0.01)	0.10 (0.01)	<0.001
compactness_mean	0.08 (0.03)	0.15 (0.05)	<0.001
concave points_mean	0.03 (0.02)	0.09 (0.03)	<0.001
symmetry_mean	0.17 (0.02)	0.19 (0.03)	<0.001
fractal_dimension_mean	0.06 (0.01)	0.06 (0.01)	0.5
radius_se	0.28 (0.11)	0.61 (0.35)	<0.001
texture_se	1.22 (0.59)	1.21 (0.48)	0.6
smoothness_se	0.01 (0.00)	0.01 (0.00)	0.2
compactness_se	0.02 (0.02)	0.03 (0.02)	<0.001
concavity_se	0.03 (0.03)	0.04 (0.02)	<0.001
concave points_se	0.01 (0.01)	0.02 (0.01)	<0.001
symmetry_se	0.02 (0.01)	0.02 (0.01)	0.028
fractal_dimension_se	0.00 (0.00)	0.00 (0.00)	<0.001
radius_worst	13.38 (1.98)	21.13 (4.28)	<0.001
smoothness_worst	0.12 (0.02)	0.14 (0.02)	<0.001
compactness_worst	0.18 (0.09)	0.37 (0.17)	<0.001
concavity_worst	0.17 (0.14)	0.45 (0.18)	<0.001
symmetry_worst	0.27 (0.04)	0.32 (0.07)	<0.001
fractal_dimension_worst	0.08 (0.01)	0.09 (0.02)	<0.001

[1] Statistics presented: Mean (SD)

[2] Statistical tests performed: Wilcoxon rank-sum test

Full Model (Newton-Raphson)

Consider the following log-likelihood, gradient, and hessian matrix. First Let

$$\pi_i = P(Y_i = 1 | x_{i,1}, \dots x_{i,p}) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}}}.$$

The log-likelihood:

$$l(\mathbf{X}|\vec{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) - \log \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \right) \right) \right]$$

The gradient:

$$\nabla l(\mathbf{X}|\vec{\beta}) = \left[\sum^{n} y_i - \pi_i \quad \sum^{n} x_{i,1}(y_i - \pi_i) \quad \dots \quad \sum^{n} x_{i,p}(y_i - \pi_i) \right]^{T}_{1 \times (p+1)}$$

The hessian: produces a matrix $(p + 1 \times p + 1)$

$$\nabla^2 l(\mathbf{X}|\vec{\beta}) = -\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X \end{pmatrix} \begin{pmatrix} 1 & X \end{pmatrix} \pi_i(1 - \pi_i)$$

Optimal Model (Logistic LASSO)

Lemma 1. Consider the optimization problem

$$\min_{x \in \mathbb{R}} \left\{ \frac{1}{2}(x-b)^2 + c|x| \right\}$$

for $b \in \mathbb{R}$ and $c \in \mathbb{R}_{++}$. It follows that the minimizer is given by

$$\hat{x} = S(b, c),$$

where S is the soft-thresholding operator.

Lemma 2. Consider the optimization problem

$$\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\}$$

for some $k \in \{1, \dots, p\}$. It follows that the minimizer is given by

$$\hat{\beta}_k = \left(\sum_{i=1}^{n} w_i x_{ik}^2 \right)^{-1} \sum_{i=1}^{n} w_i x_{ik} \left(z_i - \sum_{j \neq k} \beta_j x_{ij} \right).$$

Optimal Model (Logistic LASSO)

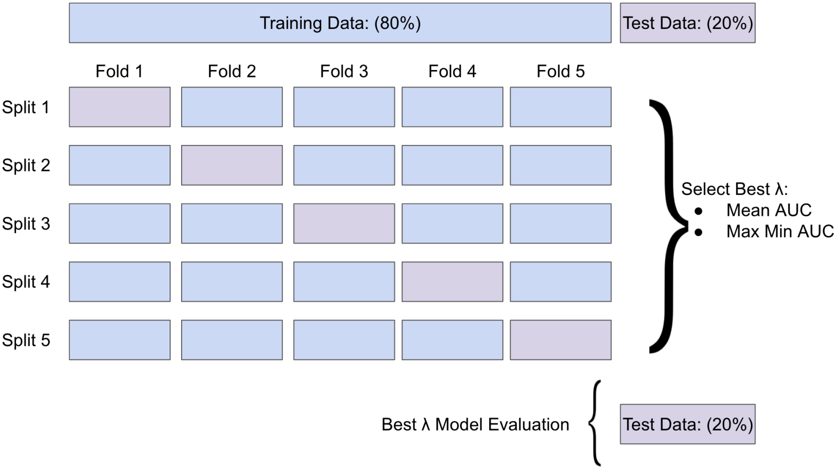
Lemma 3. With $\hat{\beta}_k$ defined as above,

$$\min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$
$$= \min_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2} (\beta_k - \hat{\beta}_k)^2 + \left(\frac{1}{n} \sum_{i=1}^{n} w_i x_{ik}^2 \right)^{-1} \lambda |\beta_k| \right\}.$$

By Lemma 1 and Lemma 3,

$$\operatorname*{arg\,min}_{\beta_k \in \mathbb{R}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} w_i \left(z_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$
$$= S \left(\hat{\beta}_k, \left(\frac{1}{n} \sum_{i=1}^{n} w_i x_{ik}^2 \right)^{-1} \lambda \right)$$

Figure 2: 5-fold Cross Validation



Training Data: (80%)

Test Data: (20%)

Fold 1

Fold 2

Fold 3

Fold 4

Fold 5

Split 1

Split 2

Split 3

Split 4

Split 5

Select Best λ:

- Mean AUC
- Max Min AUC

Best λ Model Evaluation

Test Data: (20%)

Cross Validation Results

Best λ using AUC

LASSO Coefficients

Best λ using beta plot

Coefficients Comparison

AUC

Discussion

- Goal is accurately classify every patient
- Balancing Sensitivity vs. Specificity.
  - In first screening cases want to catch every case. Maximize Sensitivity.

Resources

Cancer Stat Facts: Female Breast Cancer. *National Cancer Institute - NIH* https://seer.cancer.gov/statfacts/html/breast.html

American Cancer Society. (2019). Breast cancer facts & figures 2019–2020. Am Cancer Soc, 1-44.