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)

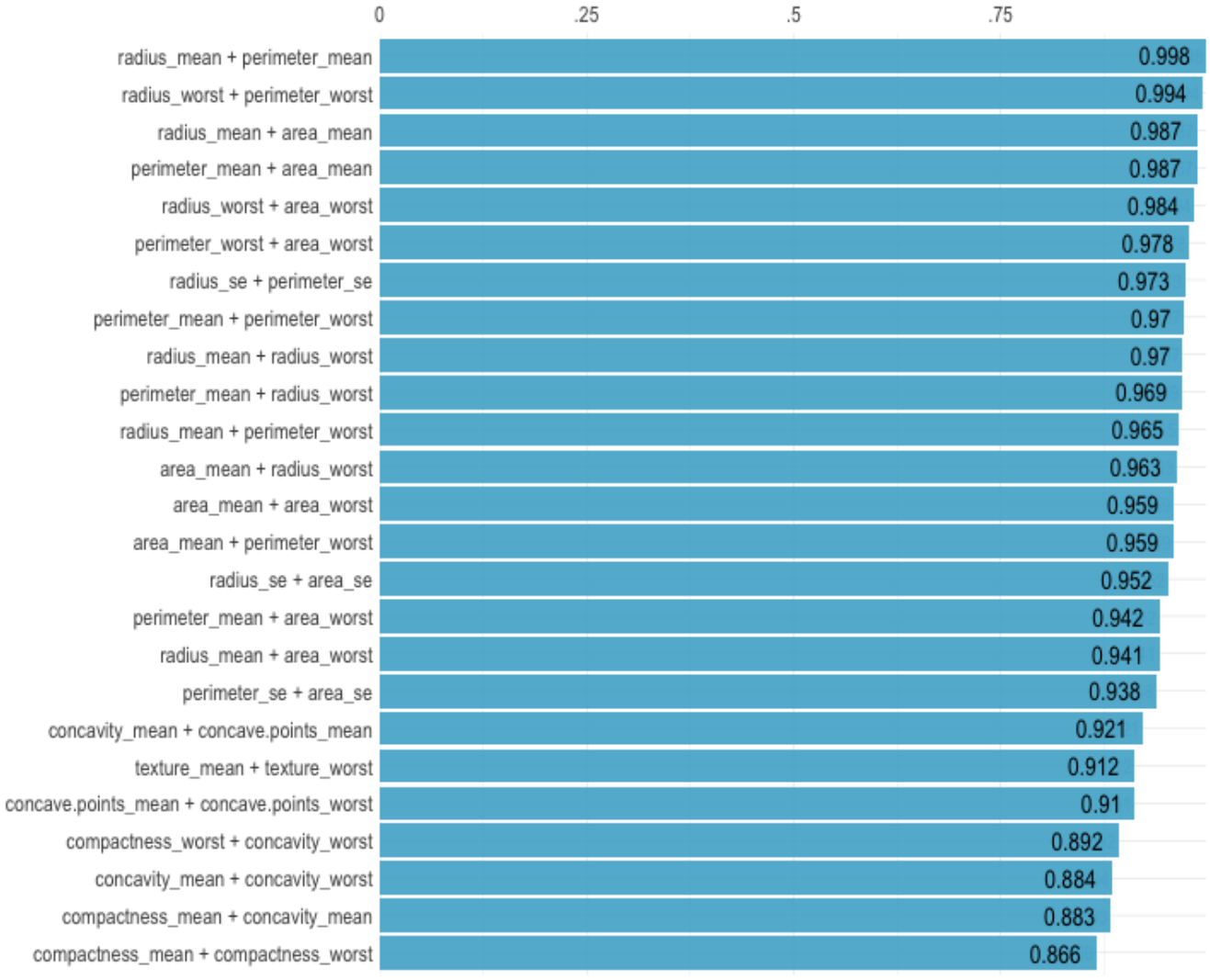
Data

- 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.

Correlations Between Variables

Figure 2: Ranked Cross-Correlations

25 most relevant



Correlation

The correlation pairs can be grouped into equivalence classes. To identify the best proxy for this grouping we look at the highest mean correlation

- First: {`area_mean`, `area_worst`, `perimeter_mean`, `perimeter_worst`, `radius_mean`, `radius_worst`}
  - Winner: `radius_worst`.
- Second: {`radius_sd`, `perimeter_se`, `area_se`}.
  - Winner: `radius_se`.
- Third: {`concavity_mean`,`concavity_worst`, `concave.point_worst`, `concave.point_mean`}
  - Winner: `concave.point_worst`.
- Fourth: {`texture_mean`, `texture_worst`}
  - Winner: `texture_mean`

Therefore 10 variables are removed.

Remaining Variables

Variable	Diagnosis Received		
	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)

To impliment the Newton-Raphson Method we need the likelihood, gradiant, and hessian matrix:

$$\pi_i = P(Y_i = 1 | x_{i,1}, \dots x_{i,20}) = \frac{e^{\beta_0 + \sum_{j=1}^{20} \beta_j x_{i,j}}}{1 + e^{\beta_0 + \sum_{j=1}^{20} \beta_j x_{i,j}}}$$

likelihood function:

$$L(\mathbf{X}|\beta) = \prod_{i=1}^{n} \left[\pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \right]$$

log-likelihood:

$$l(\mathbf{X}|\vec{\beta}) = \sum_{i=1}^{n} \left[y_i \left(\beta_0 + \sum_{j=1}^{20} \beta_j x_{i,j} \right) - \log \left(1 + \exp \left(\beta_0 + \sum_{j=1}^{20} \beta_j x_{i,j} \right) \right) \right]$$

Full Model (Newton-Raphson)

The gradient:

$$\nabla l(\mathbf{X}|\vec{\beta}) = \begin{bmatrix} \sum^n y_i - \pi_i & \sum^n x_{i,1}(y_i - \pi_i) & \dots & \sum^n x_{i,20}(y_i - \pi_i) \end{bmatrix}^T_{(1\times 21)}$$

The hessian matrix (21×21)

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

Optimal Model (Logistic LASSO)

also going to be some math

Optimal Model (Logistic LASSO)

more math

5-fold Cross Validation

Cross Validation Results

Best λ

Coefficient Comparison

AUC

Discussion

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.