Final Report

Ngoc Duong, Cui Sitong (sc4636), Xinru Wang, Jin Ge

4/15/2020

Objective

Breast cancer is one of the most common cancers in women. However, early diagnoses of breast cancer can aid in reducing the mortatlity rate. Additionally, advances in imaging technologies and statistical methodologies have allowed for higher-quality data and novel models that could improve the precision of breast cancer diagnoses. The purpose of our project is to build and compare different models in classifying breast cancer tumor as benignant or malignant based on image-based predictors. Specifically, we are look to build a logistic regression model using Newton Raphson method, as a regularized logistic-LASSO usisng coordinate-wise optimization algorithm.

Dataset

The data was obtained from ...

There were 569 images collected independently from different patients, 212 of whom had malignant tumor and 357 were benign cases. The images were broken down into 30 predictors, corresponding to the mean, standard deviation, and largest values (points on the tails) of the following 10 features:

- 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 (perimeter$\hat{2}$ / 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)

Data cleaning

As shown in the pairwise correlation plot (Figure 1), we can observe the presence of some multicollinearity among the predictors. For instance, the radius_mean variable has almost perfect correlation of 1 and 0.99 with perimeter_mean and area_mean variables, respectively. We then left out variables that are correlated by more than 85% with other predictors. The final dataset contained 13 predictors.

Next, considering the LASSO is not scale-invariant, we standardized the design matrix. This is to ensure comparability of estimates by the logistic-LASSO model and Newton-Raphson/logistic regression model. The standardization formula is as follows:

$standardized(x_{ij}) = \frac{x_{ij} - \bar{x_j}}{std(x_j)}$ for $i = 1, 2, ...30$ and $j = 1, 2, ..., 569$

Finally, we recoded response variable such that "malignancy" = 1, and "benign" = 0.

Newton-Raphson model

We used logistic regression to classify the malignancy of tissue. Malignancy corresponds to response variable being 1 ($y^{(i)} = 1$).

Log likelihood is

$$l(y; \beta) = \sum_{i=1}^{n} \{y_{(i)} log \mu_{(i)} + (1 - y_{(i)}) log(1 - \mu_{(i)})\}$$

Its gradient is given by

$$g : \bigtriangledown l(y; \beta) = \sum_{i=1}^{n} (y_{(i)} - \mu_{(i)}) x_{(i)} = X^T (y - \mu)$$

Its Hessian matrix is given by

$$H : \bigtriangledown^2 l(y; \beta) = -\sum_{i=1}^{n} \mu_{(i)}(1 - \mu_{(i)}) x_{(i)} (x_{(i)})^T = -X^T S X$$

where $S = diag(\mu_{(i)}(1 - \mu_{(i)}))$, and

$$\mu_{(i)} = p_\theta(y = 1|x) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$$

Since we have several predictors, we want to optimize several likehood functions simultaneously. This is equivalent to solving a system of log-likelihood equations $\bigtriangledown l(y; \beta_j) = 0$ where $j = 1, 2, ...13$. To achieve this, we use the Newton Raphson algorithm.

Newton-Raphson Algorithm

Starting at a current point β_i, we can expand the log-likelihood function around this point using Taylor's expansion, which gives a neighborhood of β_i containing β_{i+1} which increases the likelihood. The equation below can be used to iteratively update β_i until the sequence converges and $\bigtriangledown l(y; \beta_j) = 0$ is satisfied:

$\beta_{i+1} = \beta_i - [\bigtriangledown^2 l(\beta_i)]^{-1} \bigtriangledown l(\beta_i)$.

However, when implementing the algorithm, we need to check at every step, that the updating direction (for β_{i+1}) is heading to a maximum, and that the point is moving sufficient distances towards the maximum so we do not miss it. Therefore, we also implemented some modifications, including gradient descent and step-halving.

- For step-halving, we modified the updating function for β_{i+1} as follows:

$\beta_{i+1} = \beta_i - \lambda[\bigtriangledown^2 l(y; \beta_i)]^{-1} \bigtriangledown l(y; \beta_i)$, where $\lambda = 1$ until $l(\beta_{i+1}) \leq l(\beta_i)$, which means the new point would have gone too far. Then, we can search for a value λ such that $l(\beta_{i+1}, \lambda) \geq l(\beta_i)$. At this step, we can cut the step, or λ in half for each sub-iteration.

- For gradient descent, at every iteration, we checked whether $\bigtriangledown^2 l(y; \beta)$ is negative definite (signifying the point is moving in the right direction). If $\bigtriangledown^2 l(y; \beta)$ is not, we replace it with a similar negative definite matrix, such as $\bigtriangledown^2 l(y; \beta) - \gamma I$ where γ is chosen such that the resulting matrix is negative definite. Naturally, this γ must be greater than any of the elements of the diagonal matrix D obtained by eigendecomposing $\bigtriangledown^2 l(y; \beta) = P^T D P$.

Logistic-LASSO model

The LASSO is a high-dimensional method to handle high-dimensional data, which adds a penalty term to the loss function of a regular linear regression model. In linear regression, the LASSO minimizes:

$f(\beta) = \frac{1}{2n}\Sigma_{i=1}^{n} w_i(y_i - \Sigma_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|$ for some $\lambda \geq 0$

where w_i is the working weights, defined as $p(1-p)$ (p is the probability of event for each observation).

Coordinate-wise Descent algorithm is used to minimize this objective function. Each β_i is optimized using this equation:

$\tilde{\beta}_i = \frac{S(\Sigma_i w_i x_{i,j}(y_i - \tilde{y}_i^{(-j)}), \lambda)}{\Sigma_i w_i x_{i,j}^2}$

where $S(\hat{\beta}, \gamma)$ is called soft-threshold and is defined as

...

Results

The coefficient estimates can be found in table 1. Newton-Raphson algorithm gives quite similar estimates to those in the logistic regression model produced by GLM package. For the logistic-LASSO model, we can see the coefficient estimates are approximately close to the ones produced by GLMnet with 5-fold cross-validation. They do not exactly match, however, due to the potential dissimilarities in set-up conditions in our implementation and theirs.

The path of solution could be found in Figure 2, and the distribution of cross-validated MSEs produced by the hand-built logistic-LASSO model could be found in Figure 3. Five-fold cross-validation suggested the best λ is 0.00454, which corresponds to the lowest cross-validated MSE. A similar distribution of cross-validated MSEs produced by the GLMNet package in R can be found in Figure 4. Here, 5-fold cross-validation suggested the best λ is 0.0037.

Lastly, we wanted to compare the prediction performance using MSE as a criteria. We examined this undertwo scenarios, one is with 5-fold cross-validation, and the other is 5-fold repeated cross-validation (number of repeats is 5). These distributions can be found in Figure 3 and 4, respectively. We noticed that the cross-validated MSE of both Newton Raphson and logistic-LASSO are similary distributed, although the latter seems to perform slightly less well. On the other hand, GLMnet gives the most consistently well performance under both scenarios. Nonetheless, the errors were all in close and acceptable proximity with one another so we are confident they all very good discriminatory power.

Conclusions

The report aimed to explore how different models perform at the same task of classifying breast cancer tumors into benign and malignant types using various predictors derived from the tumor images. Since we have eliminated most multicollinearity at the beginning, it is reasonable to expect Newton-Raphson and logistic-LASSO to have quite similar discriminatory performance. On the other hand, we would expect to see logistic-LASSO to perform better than Newton Raphson in terms of predictive ability in the presence of higher-dimensional data, and with more correlated predictors. We will aim to explore this further as more time and bandwidth allow. All in all, these models we looked at in this report nevertheless perform decently in classifying correctly each type of breast cancer.