methods

Yuqi Miao ym2771

3/23/2020

Methods

Data

The mean statistics for every feature are selected as predictors in the model. The sample size for the data is 569. The response variable is "diagnosis", which is transformed to a binary variable (Benign = 0, Malignant = 1),and the 11 predictors are standardised.

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   diagnosis = col_character(),
##   X33 = col_character()
## )

## See spec(...) for full column specifications.
```

description table?

Model define

For i_{th} observation, the response variable Y_i follows binary distribution:

$$Y_i \sim Bin(\pi_i)$$

the log-liklihood function:

$$l(\mathbf{Y}, \boldsymbol{\pi}) = \sum_{i=1}^{n} l(y_i, \pi_i) = \sum_{i=1}^{n} (y_i log \frac{\pi_i}{1-\pi_i} + log(1-\pi_i)) = \sum_{i=1}^{n} (y_i log \frac{\pi_i}{1-\pi_i} + log(1-\pi_i))$$

Where π_i denotes the probablity of the i_{th} observation to be maglinant.

To build relationship between the response and predictors, the GLM is defined as:

$$\log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}_i \boldsymbol{\beta} = \theta_i$$

Full model

Firstly, Newton-Rapson method is used to fit the full model. To find the maximum likelihood estimation of coefficients, iteration process is set as follows:

$$\theta_{i+1} = \theta_i - \delta(\nabla^2 l(\theta_i|\boldsymbol{X}) - \gamma I)^{-1} \nabla l(\theta_i|\boldsymbol{X})$$

where δ is the step coefficient to ensure the increasing of likelihood funciton, and γ is the modification coefficient to ensure the aescent direction of the iteration vector.

$$\nabla l(\theta|\boldsymbol{X}) = \mathbf{X}^T(\mathbf{Y} - \boldsymbol{\pi})$$

$$\nabla^2 l(\theta|\boldsymbol{X}) = -\mathbf{X}^T diag(\pi_i(1-\pi_i))\mathbf{X}$$

logit-lasso pathwise coordinate-wise update algorithm

To select variables and increase the prediction efficiency, lasso was integrated into the coordinate-wise logit regression.

The target function:

$$min\{-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

$$l(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n} \omega_i(z_i - \boldsymbol{X_i}\boldsymbol{\beta})$$

$$\pi_i = \frac{exp(\boldsymbol{X_i}\boldsymbol{\beta})}{1 + exp(\boldsymbol{X_i}\boldsymbol{\beta})}$$

$$\omega_i = \pi_i(1-\pi_i)$$

$$z_i = \boldsymbol{X_i}\boldsymbol{\beta} + \frac{y_i - \pi_i}{\pi_i(1-\pi_i)}$$

Pre-define the tuning parameter sequence $\{\lambda_1, ..., \lambda_s\}$, starting point $\boldsymbol{\beta_{start=}}\{\beta_0^{(0)}, ..., \beta_p^{(0)}\}$. Here we elaborately explain the optimal process for λ_u: Using the optimal $\boldsymbol{\beta_{u-1}}$ from last iteration as the warm start. within every iteration, find the optimal β coordinate-wise. For β_j in t_th iteration

$$\beta_j^{(t)} = \begin{cases} \sum_{i=1}^{n} \omega_i(z_i - \sum_{j=1}^{p} \boldsymbol{X_i}\beta_j), & j = 0 \\ s(\beta_j^{(t*)}, \lambda_u), & j = 1, 2, ..., p \end{cases}$$

$$\beta_j^{(t*)} = \frac{\sum_{i=1}^{n} \omega_i x_{ij} z_{ij}^*}{\sum_{i=1}^{n} \omega_i x_{ij}^2}$$

$$z_{ij}^* = z_t - \sum_{\substack{k=0 \\ \beta_k \neq 0}}^{j-1} \beta_k^{(i)} x_{ik} - \sum_{\substack{k=j+1 \\ \beta_k \neq 0}}^{p} \beta_k^{(i-1)} x_{ik}$$

Cross Validation

In order to check the model performance, we use 5-fold cross validation. Using training dataset to choose the best tuning parameter and fit model, and using test dataset to evaluate the final prediction. The statistics we use to compare the validation is MSE and person chi-square statistics, which are defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\pi}_i)$$

$$\widehat{\pi}_i = log\frac{exp(\boldsymbol{X}_i\beta)}{1 + exp(\boldsymbol{X}_i\beta)}$$

$$G = \sum_{i=1}^{n}\frac{y_i - \widehat{\pi}_i}{\widehat{\pi}_i(1 - \widehat{\pi}_i)}$$

By taking average of above 2 statistics of 5 fold, we get the final index to evaluate the model fitting.

contingency table