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.
```

ones	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	conca
1	1.0960995	-2.0715123	1.2688173	0.9835095	1.5670875	3.2806281	
1	1.8282120	-0.3533215	1.6844726	1.9070303	-0.8262354	-0.4866435	
1	1.5784992	0.4557859	1.5651260	1.5575132	0.9413821	1.0519999	
1	-0.7682333	0.2535091	-0.5921661	-0.7637917	3.2806668	3.3999174	
1	1.7487579	-1.1508038	1.7750113	1.8246238	0.2801253	0.5388663	
1	-0.4759559	-0.8346009	-0.3868077	-0.5052059	2.2354545	1.2432416	

Model define

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

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

the log-liklihood function:

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

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:

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

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:

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

where λ 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.

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

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

logit-lasso coordinate-wise update algorithm

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

The target function:

$$min - l(\beta)$$