Assignment 10.2 on Logistic Regression Model (Thoracic Surgery and Binary Classifier Data)

Basak Atanu

05-17-2022

library(foreign)  
  
setwd("C:\\Users\\atanu\\Documents\\BellevueUniversity\_MSDS\\DSC520\\Repository\\dsc520\_")  
thoraric <- read.arff("data\\ThoraricSurgery.arff")  
#head(thoraric)

### lets split the data into 80-20 ratio for testing.

library(caTools)  
split <- sample.split(thoraric, SplitRatio = 0.8)  
split

## [1] TRUE TRUE FALSE TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE TRUE  
## [13] TRUE TRUE TRUE TRUE TRUE

train <- subset(thoraric, split=='TRUE')  
test <- subset(thoraric, split=='FALSE')

### lets factorize the categorical data.

model <- glm(Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + + PRE8 + PRE9 + PRE10 + PRE11 +   
 PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE, data=train, family='binomial')  
summary(model)

##   
## Call:  
## glm(formula = Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + +PRE8 +   
## PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +   
## PRE32 + AGE, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5346 -0.5563 -0.4311 -0.2901 2.5471   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -15.95838 2399.54540 -0.007 0.99469   
## DGNDGN2 14.35079 2399.54481 0.006 0.99523   
## DGNDGN3 13.92418 2399.54478 0.006 0.99537   
## DGNDGN4 14.54057 2399.54482 0.006 0.99517   
## DGNDGN5 16.06382 2399.54486 0.007 0.99466   
## DGNDGN6 0.14372 2676.32090 0.000 0.99996   
## DGNDGN8 34.41982 3393.46876 0.010 0.99191   
## PRE4 -0.24341 0.21101 -1.154 0.24869   
## PRE5 -0.02306 0.01849 -1.247 0.21247   
## PRE6PRZ1 -0.04854 0.61247 -0.079 0.93684   
## PRE6PRZ2 -0.02341 0.92970 -0.025 0.97991   
## PRE7T 0.99244 0.62979 1.576 0.11506   
## PRE8T -0.22754 0.45669 -0.498 0.61832   
## PRE9T 1.46366 0.52215 2.803 0.00506 \*\*  
## PRE10T 0.55947 0.56044 0.998 0.31814   
## PRE11T 0.42908 0.43567 0.985 0.32468   
## PRE14OC12 0.47369 0.38227 1.239 0.21530   
## PRE14OC13 0.91113 0.67291 1.354 0.17573   
## PRE14OC14 1.66628 0.66065 2.522 0.01166 \*   
## PRE17T 0.89076 0.51567 1.727 0.08410 .   
## PRE19T -14.08992 2399.54475 -0.006 0.99531   
## PRE25T 1.13392 1.27934 0.886 0.37544   
## PRE30T 0.73131 0.53158 1.376 0.16890   
## PRE32T -13.98362 1683.34321 -0.008 0.99337   
## AGE -0.01328 0.02070 -0.641 0.52130   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 307.13 on 357 degrees of freedom  
## Residual deviance: 264.92 on 333 degrees of freedom  
## AIC: 314.92  
##   
## Number of Fisher Scoring iterations: 15

### according to the summary, PRE9 (Dyspnoea before surgery), PRE14: T in clinical TNM - size of the original tumour with OC13 and OC14 which means large size of tumour have greater affect on servical rate.

res <- predict(model, train, type='response')  
confmatrix <- table(actual\_value=train$Risk1Yr, predicted\_value = res>0.5)

### Accuracy

(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.8379888

### Reading binary classifier data.

setwd("C:\\Users\\atanu\\Documents\\BellevueUniversity\_MSDS\\DSC520\\Repository\\dsc520\_")  
classifier\_data <- read.csv("data\\binary-classifier-data.csv")

### Fitting logicstic regression model.

model <- glm(label ~ x + y, data=classifier\_data, family='binomial')  
summary(model)

##   
## Call:  
## glm(formula = label ~ x + y, family = "binomial", data = classifier\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3728 -1.1697 -0.9575 1.1646 1.3989   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.424809 0.117224 3.624 0.00029 \*\*\*  
## x -0.002571 0.001823 -1.411 0.15836   
## y -0.007956 0.001869 -4.257 2.07e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2075.8 on 1497 degrees of freedom  
## Residual deviance: 2052.1 on 1495 degrees of freedom  
## AIC: 2058.1  
##   
## Number of Fisher Scoring iterations: 4

### Accuracy of logistic regression classifier.

res <- predict(model, classifier\_data, type='response')  
confmatrix <- table(actual\_value=classifier\_data$label, predicted\_value = res>0.5)  
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.5834446