HW 06

TUAN BUI

10/25/2021

# Question 01:

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4  
## ✓ tibble 3.1.4 ✓ dplyr 1.0.7  
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0  
## ✓ readr 2.0.1 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(dummies)

## dummies-1.5.6 provided by Decision Patterns

library(leaps)  
library(bestglm)  
theme\_set(theme\_bw())  
  
banknote\_data <- read.csv('~/OneDrive - Stony Brook University/SBU/MAT + AMS/Fall 2021/AMS 380/hw/06/banknote.csv', header = T)  
  
banknote\_data <- na.omit(banknote\_data)  
  
banknote\_data$class <- as.factor(banknote\_data$class)

## (a): Split the data into 80% training and 20% testing using seed =123

set.seed(123)  
training.samples <- banknote\_data$class %>%   
 createDataPartition(p = 0.8, list = FALSE)  
train.data <- banknote\_data[training.samples, ]  
test.data <- banknote\_data[-training.samples, ]

## (b): Fit a logistic regression model with all 4 predictors using the training data

model <- glm( class ~., data = train.data, family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(model)$coef

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -8.5894550 2.1862852 -3.928790 8.537441e-05  
## variance 9.3610771 2.4703379 3.789391 1.510169e-04  
## skewness 4.6967769 1.2672540 3.706263 2.103398e-04  
## curtosis 6.1372023 1.6413565 3.739104 1.846775e-04  
## entropy 0.5192738 0.4207628 1.234125 2.171565e-01

# logistic equation: p = exp(-7.1001295 + 7.4068618 \* variance + 3.9759205 \* skewness + 4.9812792 \* curtosis + 0.5236681 \* entropy) / [1 + exp(-7.1001295 + 7.4068618 \* variance + 3.9759205 \* skewness + 4.9812792 \* curtosis + 0.5236681 \* entropy)]

## (c): Predict the response variable ‘class’, generate confusion matrix, and report accuracy, sensitivity, specificity for the testing data

probabilities <- model %>% predict(test.data, type = "response")  
predicted.classes <- ifelse(probabilities > 0.5, 1, 0)  
  
mean(test.data$class == predicted.classes)

## [1] 0.9817518

# accuracy of prediction in the test data is 0.9817518  
  
sum((test.data$class == 1)\*(predicted.classes == 1))/sum(test.data$class == 1)

## [1] 0.9868421

# sensitivity in the test data is 0.9868421  
  
sum((test.data$class == 0)\*(predicted.classes == 0))/sum(test.data$class == 0)

## [1] 0.9754098

# specificity in the test data is 0.9754098  
  
# confusion matrix  
table(predicted.classes, test.data$class)

##   
## predicted.classes 0 1  
## 0 119 2  
## 1 3 150

# accuracy of prediction in the test data is 0.9817518  
# sensitivity in the test data is 0.9868421  
# specificity in the test data is 0.9754098

# Question 01 (other):

fit <- glm(class ~ . , data = banknote\_data, family = 'binomial')

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(fit)$coef

## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -7.321805 1.5588603 -4.696896 2.641448e-06  
## variance 7.859330 1.7383123 4.521242 6.147788e-06  
## skewness 4.190963 0.9041488 4.635258 3.564919e-06  
## curtosis 5.287431 1.1611830 4.553486 5.276415e-06  
## entropy 0.605319 0.3307210 1.830301 6.720497e-02

# logistic equation: p = exp(-7.321805 + 7.859330 \* variance + 4.190963 \* skewness + 5.287431 \* curtosis + 0.605319 \* entropy) / [1 + exp(-7.321805 + 7.859330 \* variance + 4.190963 \* skewness + 5.287431 \* curtosis + 0.605319 \* entropy)]  
  
step1 <- stepAIC(fit, trace = T, k = log(nrow(banknote\_data)))

## Start: AIC=86.01  
## class ~ variance + skewness + curtosis + entropy

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Df Deviance AIC  
## - entropy 1 53.30 82.19  
## <none> 49.89 86.01  
## - skewness 1 636.52 665.42  
## - curtosis 1 719.24 748.14  
## - variance 1 1145.48 1174.38

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##   
## Step: AIC=82.19  
## class ~ variance + skewness + curtosis  
##   
## Df Deviance AIC  
## <none> 53.30 82.19  
## - curtosis 1 722.03 743.70  
## - skewness 1 850.17 871.84  
## - variance 1 1399.79 1421.46

step1$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## class ~ variance + skewness + curtosis + entropy  
##   
## Final Model:  
## class ~ variance + skewness + curtosis  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 1367 49.89066 86.01078  
## 2 - entropy 1 3.40798 1368 53.29864 82.19474

BIC(step1)

## [1] 82.19474

# The best predict model using the stepwise variable section method and the BIC is class ~ variance + skewness + curtosis with the associated BIC value is 82.19474

# Question 02:

step2 <- bestglm(banknote\_data , IC = "BIC", family = binomial)

## Morgan-Tatar search since family is non-gaussian.

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred  
  
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

step2$BestModel

##   
## Call: glm(formula = y ~ ., family = family, data = Xi, weights = weights)  
##   
## Coefficients:  
## (Intercept) variance skewness curtosis   
## -6.885 6.783 3.507 4.464   
##   
## Degrees of Freedom: 1371 Total (i.e. Null); 1368 Residual  
## Null Deviance: 1885   
## Residual Deviance: 53.3 AIC: 61.3

BIC(step2$BestModel)

## [1] 82.19474

# The best predict model using the best subset variable selection method and the BIC is class ~ variance + skewness + curtosis with the associated BIC value is 82.19474