ams580\_quiz4

Mustafa Yigit Isik

2023-03-01

#**Problem 1**

#import libraries  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0   
## ✔ readr 2.1.4 ✔ forcats 1.0.0   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(glmnet)

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loaded glmnet 4.1-6

library(caTools)  
#I did text-to-column in excel based on semicolons to make our data usable in R  
#Maybe there are methods to do the same in R as well but I didn't know for now  
#read data  
data <- read.csv('/Users/mustafayigitisik/Desktop/stuff/semesters/spring 2023/ams 580/quizzes/quiz4/wine.csv')  
str(data)

## 'data.frame': 1599 obs. of 12 variables:  
## $ fixed\_acidity : num 7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...  
## $ volatile\_acidity : num 0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...  
## $ citric\_acid : num 0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...  
## $ residual\_sugar : num 1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...  
## $ chlorides : num 0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...  
## $ free\_sulfur\_dioxide : num 11 25 15 17 11 13 15 15 9 17 ...  
## $ total\_sulfur\_dioxide: num 34 67 54 60 34 40 59 21 18 102 ...  
## $ density : num 0.998 0.997 0.997 0.998 0.998 ...  
## $ pH : num 3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...  
## $ sulphates : num 0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...  
## $ alcohol : num 9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...  
## $ quality : int 5 5 5 6 5 5 5 7 7 5 ...

#set seed  
set.seed(123)  
#divide to 75%train and 25%test  
split <- sample.split(1:dim(data)[1], SplitRatio = 0.75)  
train.data <- data[split,]  
test.data <- data[!split,]  
str(train.data) # 1199 obs

## 'data.frame': 1199 obs. of 12 variables:  
## $ fixed\_acidity : num 7.4 7.8 7.4 7.9 7.8 7.5 7.5 5.6 7.8 8.9 ...  
## $ volatile\_acidity : num 0.7 0.76 0.66 0.6 0.58 0.5 0.5 0.615 0.61 0.62 ...  
## $ citric\_acid : num 0 0.04 0 0.06 0.02 0.36 0.36 0 0.29 0.18 ...  
## $ residual\_sugar : num 1.9 2.3 1.8 1.6 2 6.1 6.1 1.6 1.6 3.8 ...  
## $ chlorides : num 0.076 0.092 0.075 0.069 0.073 0.071 0.071 0.089 0.114 0.176 ...  
## $ free\_sulfur\_dioxide : num 11 15 13 15 9 17 17 16 9 52 ...  
## $ total\_sulfur\_dioxide: num 34 54 40 59 18 102 102 59 29 145 ...  
## $ density : num 0.998 0.997 0.998 0.996 0.997 ...  
## $ pH : num 3.51 3.26 3.51 3.3 3.36 3.35 3.35 3.58 3.26 3.16 ...  
## $ sulphates : num 0.56 0.65 0.56 0.46 0.57 0.8 0.8 0.52 1.56 0.88 ...  
## $ alcohol : num 9.4 9.8 9.4 9.4 9.5 10.5 10.5 9.9 9.1 9.2 ...  
## $ quality : int 5 5 5 5 7 5 5 5 5 5 ...

str(test.data) # 400 obs

## 'data.frame': 400 obs. of 12 variables:  
## $ fixed\_acidity : num 7.8 11.2 7.4 7.3 6.7 8.9 7.9 8.9 8.5 6.7 ...  
## $ volatile\_acidity : num 0.88 0.28 0.7 0.65 0.58 0.62 0.32 0.22 0.49 0.675 ...  
## $ citric\_acid : num 0 0.56 0 0 0.08 0.19 0.51 0.48 0.11 0.07 ...  
## $ residual\_sugar : num 2.6 1.9 1.9 1.2 1.8 3.9 1.8 1.8 2.3 2.4 ...  
## $ chlorides : num 0.098 0.075 0.076 0.065 0.097 0.17 0.341 0.077 0.084 0.089 ...  
## $ free\_sulfur\_dioxide : num 25 17 11 15 15 51 17 29 9 17 ...  
## $ total\_sulfur\_dioxide: num 67 60 34 21 65 148 56 60 67 82 ...  
## $ density : num 0.997 0.998 0.998 0.995 0.996 ...  
## $ pH : num 3.2 3.16 3.51 3.39 3.28 3.17 3.04 3.39 3.17 3.35 ...  
## $ sulphates : num 0.68 0.58 0.56 0.47 0.54 0.93 1.08 0.53 0.53 0.54 ...  
## $ alcohol : num 9.8 9.8 9.4 10 9.2 9.2 9.2 9.4 9.4 10.1 ...  
## $ quality : int 5 6 5 7 5 5 6 6 5 5 ...

#**Problem 2**

x <- model.matrix(quality~., train.data)[,-1]  
y <- train.data$quality  
cv <- cv.glmnet(x, y, alpha = 0)  
cv$lambda.min

## [1] 0.05455629

Best lambda for ridge regression is 0.0545563

model <- glmnet(x, y, alpha = 0, lambda = cv$lambda.min)  
coef(model)

## 12 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 27.409163429  
## fixed\_acidity 0.017112529  
## volatile\_acidity -1.063242689  
## citric\_acid -0.051669202  
## residual\_sugar 0.013641201  
## chlorides -1.820771557  
## free\_sulfur\_dioxide 0.002185583  
## total\_sulfur\_dioxide -0.003035716  
## density -22.970469483  
## pH -0.391806527  
## sulphates 0.832378522  
## alcohol 0.243665707

x.test <- model.matrix(quality ~., test.data)[,-1]  
predictions <- model %>% predict(x.test) %>% as.vector()  
data.frame(  
 RMSE = RMSE(predictions, test.data$quality),  
 Rsquare = R2(predictions, test.data$quality)  
)

## RMSE Rsquare  
## 1 0.6518861 0.3638303

RMSE is 0.6518861 and coefficient of determination is 0.3638303

#**Problem 3**

cv <- cv.glmnet(x, y, alpha = 1)  
cv$lambda.min

## [1] 0.01320332

Best lambda for LASSO is 0.0132033

model <- glmnet(x, y, alpha = 1, lambda = cv$lambda.min)  
coef(model)

## 12 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 4.274090224  
## fixed\_acidity .   
## volatile\_acidity -1.091529254  
## citric\_acid .   
## residual\_sugar .   
## chlorides -1.527390489  
## free\_sulfur\_dioxide .   
## total\_sulfur\_dioxide -0.002259329  
## density .   
## pH -0.342044940  
## sulphates 0.738734818  
## alcohol 0.271267910

x.test <- model.matrix(quality ~., test.data)[,-1]  
predictions <- model %>% predict(x.test) %>% as.vector()  
data.frame(  
 RMSE = RMSE(predictions, test.data$quality),  
 Rsquare = R2(predictions, test.data$quality)  
)

## RMSE Rsquare  
## 1 0.6543806 0.3592586

RMSE is 0.6543806 and coefficient of determination is 0.3592586

#**Problem 4**

model <- train(  
 quality ~., data = train.data, method = "glmnet",  
 trControl = trainControl("cv", number = 10),  
 tuneLength = 10  
)  
model$bestTune

## alpha lambda  
## 86 1 0.01143081

coef(model$finalModel, model$bestTune$lambda)

## 12 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 4.3166018824  
## fixed\_acidity .   
## volatile\_acidity -1.0917799993  
## citric\_acid .   
## residual\_sugar .   
## chlorides -1.5895709770  
## free\_sulfur\_dioxide 0.0000649628  
## total\_sulfur\_dioxide -0.0023237929  
## density .   
## pH -0.3583086427  
## sulphates 0.7529681452  
## alcohol 0.2721771151

x.test <- model.matrix(quality ~., test.data)[,-1]  
predictions <- model %>% predict(x.test)  
data.frame(  
 RMSE = RMSE(predictions, test.data$quality),  
 Rsquare = R2(predictions, test.data$quality)  
)

## RMSE Rsquare  
## 1 0.6541453 0.3593835

RMSE is 0.6541453 and coefficient of determination is 0.3593835