Lab11.R

rstudio-user

2021-05-05

```
#pseudorandom number generator
set.seed(11)

# Attach Packages
library(tidyverse)    # data manipulation and visualization
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(kernlab)      # SVM methodology
```

```
## 
## Attaching package: 'kernlab'
```

```
## The following object is masked from 'package:purrr':
## 
##     cross
```

```
## The following object is masked from 'package:ggplot2':
## 
##     alpha
```

```
library(e1071)        # SVM methodology
library(ISLR)         # contains example data set "Khan"
library(RColorBrewer) # customized coloring of plots
library(caret)
```

```
## Loading required package: lattice
```

```
## 
## Attaching package: 'caret'
```

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

```
#Read file
fl <- read.csv("heart.csv")
head(fl)
```

```
##   age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal
## 1  63   1  3      145  233   1       0     150     0     2.3     0  0    1
## 2  37   1  2      130  250   0       1     187     0     3.5     0  0    2
## 3  41   0  1      130  204   0       0     172     0     1.4     2  0    2
## 4  56   1  1      120  236   0       1     178     0     0.8     2  0    2
## 5  57   0  0      120  354   0       1     163     1     0.6     2  0    2
## 6  57   1  0      140  192   0       1     148     0     0.4     1  0    1
##   target
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
## 6      1
```

```
#Data structure
str(fl)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trestbps: int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalach : int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exang   : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slope   : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ ca      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thal    : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ target  : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
#Summary
summary(fl)
```

```
##       age             sex               cp           trestbps
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##       chol            fbs            restecg          thalach
##  Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000   Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485   Mean   :0.5281   Mean   :149.6
##  3rd Qu.:274.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
##      exang           oldpeak         slope             ca
##  Min.   :0.0000   Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267   Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :6.20   Max.   :2.000   Max.   :4.0000
##       thal           target
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:2.000   1st Qu.:0.0000
##  Median :2.000   Median :1.0000
##  Mean   :2.314   Mean   :0.5446
##  3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :3.000   Max.   :1.0000
```

```
#Empty values
colSums(fl==" ")
```

```
##      age      sex       cp trestbps     chol      fbs  restecg  thalach
##        0        0        0        0        0        0        0        0
##    exang  oldpeak    slope       ca     thal   target
##        0        0        0        0        0        0
```

```
#So there are no empty values

#Null values
colSums(is.na(fl))
```

```
##      age      sex       cp trestbps     chol      fbs  restecg  thalach
##        0        0        0        0        0        0        0        0
##    exang  oldpeak    slope       ca     thal   target
##        0        0        0        0        0        0
```

```
#So there are no null values

#MODEL BUILDING
#Split dataset into train and test
index <- sample(1:nrow(fl), 0.75*nrow(fl))
train <- fl[index,]
test <- fl[-index,]

#Convert target into factor
train$target <- as.factor(train$target)
test$target <- as.factor(test$target)

training <- trainControl(method="repeatedcv", number=10, repeats=3)

grid <- expand.grid(C=c(0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 5))
svmgrid <- train(target~., data=train, method="svmLinear", trControl=training, preProcess=c("center", "

svmgrid
```

```
## Support Vector Machines with Linear Kernel
##
## 227 samples
##  13 predictor
##   2 classes: '0', '1'
##
## Pre-processing: centered (13), scaled (13)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 205, 203, 205, 203, 205, 204, ...
## Resampling results across tuning parameters:
```

```
##
##   C     Accuracy   Kappa
##   0.01  0.8247969  0.6364641
##   0.05  0.8263834  0.6419405
##   0.10  0.8249945  0.6398748
##   0.25  0.8233476  0.6367264
##   0.50  0.8278272  0.6466485
##   0.75  0.8293423  0.6497273
##   1.00  0.8293423  0.6497273
##   1.25  0.8293423  0.6497273
##   1.50  0.8263779  0.6437773
##   1.75  0.8263779  0.6437773
##   2.00  0.8263779  0.6437773
##   5.00  0.8263779  0.6437773
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 0.75.
```

```
#Training with gamma=0.2 and cost=0.25
m1 <- svm(target~., data=train, kernel="linear", gamma=0.2, cost=0.25)
m1
```

```
##
## Call:
## svm(formula = target ~ ., data = train, kernel = "linear", gamma = 0.2,
##     cost = 0.25)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.25
##
## Number of Support Vectors:  95
```

```
#Training with gamma=0.5 and cost=0.05
m2 <- svm(target~., data=train, kernel="linear", gamma=0.5, cost=0.05)
m2
```

```
##
## Call:
## svm(formula = target ~ ., data = train, kernel = "linear", gamma = 0.5,
##     cost = 0.05)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.05
##
## Number of Support Vectors:  111
```

```
#Test with gamma=0.2 and cost=0.25
pred1 <- predict(m1, newdata=test)
confusionMatrix(pred1, test$target)
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction  0  1
##          0 27  2
##          1  9 38
##
##                Accuracy : 0.8553
##                  95% CI : (0.7558, 0.9255)
##     No Information Rate : 0.5263
##     P-Value [Acc > NIR] : 1.432e-09
##
##                   Kappa : 0.7069
##
##  Mcnemar's Test P-Value : 0.07044
##
##             Sensitivity : 0.7500
##             Specificity : 0.9500
##          Pos Pred Value : 0.9310
##          Neg Pred Value : 0.8085
##              Prevalence : 0.4737
##          Detection Rate : 0.3553
##    Detection Prevalence : 0.3816
##       Balanced Accuracy : 0.8500
##
##        'Positive' Class : 0
##
```

```r
#Test with gamma=0.5 and cost=0.05
pred2 <- predict(m2, newdata=test)
confusionMatrix(pred2, test$target)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 27  2
##          1  9 38
##
##                Accuracy : 0.8553
##                  95% CI : (0.7558, 0.9255)
##     No Information Rate : 0.5263
##     P-Value [Acc > NIR] : 1.432e-09
##
##                   Kappa : 0.7069
##
##  Mcnemar's Test P-Value : 0.07044
##
##             Sensitivity : 0.7500
##             Specificity : 0.9500
##          Pos Pred Value : 0.9310
##          Neg Pred Value : 0.8085
##              Prevalence : 0.4737
##          Detection Rate : 0.3553
##    Detection Prevalence : 0.3816
##       Balanced Accuracy : 0.8500
```

```
##
##        'Positive' Class : 0
##
```

```
#We trained two different models with two sets of gamma and cost values.
#Upon evaluating the model, we see that model 2 yeilds better accuracy of 84%,
#than model 1, which is 82%.
```