ass1.R

augjo318

2019-11-24

library("kknn")  
library(readxl)  
data <- read\_excel("/home/augjo318/Desktop/TDDE01\_Lab1/spambase.xlsx")  
  
data$Spam <- as.factor(data$Spam)  
n=dim(data)  
set.seed(12345)  
id=sample(1:n, floor(n\*0.5))

## Warning in 1:n: numerical expression has 2 elements: only the first used

train=data[id,]  
test=data[-id,]  
  
#data[data$. > 0.5,]$.<- 1  
  
logistic <- glm(Spam ~ ., data = train, family = "binomial")

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

#1.2  
res\_train <- predict(logistic, train, type = "response")  
res\_test <- predict(logistic, test, type="response")  
  
res1 <- ifelse(res\_train > 0.5, 1, 0)  
tbl <- table(train$Spam, res1)  
colnames(tbl) = c("Predicted Non-spam", "Predicted spam")  
rownames(tbl) = c("Actual Non-spam", "Actual Spam")  
tbl

## res1  
## Predicted Non-spam Predicted spam  
## Actual Non-spam 803 142  
## Actual Spam 81 344

MisClassification\_train <- (tbl[1, "Predicted spam"] + tbl[2, "Predicted Non-spam"])/(sum(tbl))  
print(MisClassification\_train)

## [1] 0.1627737

res2 <- ifelse(res\_test > 0.5, 1, 0)  
tbl1 <- table(test$Spam, res2)  
  
colnames(tbl1) = c("Predicted Non-spam", "Predicted spam")  
rownames(tbl1) = c("Actual Non-spam", "Actual Spam")  
tbl1

## res2  
## Predicted Non-spam Predicted spam  
## Actual Non-spam 791 146  
## Actual Spam 97 336

MisClassification\_test <- (tbl1[1, "Predicted spam"] + tbl1[2, "Predicted Non-spam"])/(sum(tbl1))  
print(MisClassification\_test)

## [1] 0.1773723

#1.3  
res3 <- ifelse(res\_train > 0.8, 1, 0)  
tbl2 <- table(train$Spam, res3)  
colnames(tbl2) = c("Predicted Non-spam", "Predicted spam")  
rownames(tbl2) = c("Actual Non-spam", "Actual Spam")  
tbl2

## res3  
## Predicted Non-spam Predicted spam  
## Actual Non-spam 941 4  
## Actual Spam 335 90

MisClassification\_train <- (tbl2[1, "Predicted spam"] + tbl2[2, "Predicted Non-spam"])/(sum(tbl2))  
print(MisClassification\_train)

## [1] 0.2474453

res4 <- ifelse(res\_test > 0.8, 1, 0)  
tbl3 <- table(test$Spam, res4)  
colnames(tbl3) = c("Predicted Non-spam", "Predicted spam")  
rownames(tbl3) = c("Actual Non-spam", "Actual Spam")  
tbl3

## res4  
## Predicted Non-spam Predicted spam  
## Actual Non-spam 926 11  
## Actual Spam 367 66

MisClassification\_test <- (tbl3[1, "Predicted spam"] + tbl3[2, "Predicted Non-spam"])/(sum(tbl3))  
print(MisClassification\_test)

## [1] 0.2759124

kknn\_model <- kknn(Spam ~., train, test, k = 30)   
kknn\_model0 <- kknn(Spam ~., train, train, k = 30)  
fit0 <- fitted(kknn\_model0)  
fit <- fitted(kknn\_model)  
  
tbl4 <- table(test$Spam, fit)  
tbl5 <- table(train$Spam, fit0)  
  
  
MisClassification <- (tbl4[1, "1"] + tbl4[2, "0"])/(sum(tbl4))  
MisClassification2 <- (tbl5[1, "1"] + tbl5[2, "0"])/(sum(tbl5))  
  
print(MisClassification)

## [1] 0.329927

print(MisClassification2)

## [1] 0.1722628

kknn\_model2 <- kknn(Spam ~., train, test, k =1)  
kknn\_model3 <- kknn(Spam ~., train, train, k =1)  
  
fit2 <- fitted(kknn\_model2)  
fit3 <- fitted(kknn\_model3)  
  
tbl6 <- table(test$Spam, fit2)  
tbl7 <- table(train$Spam, fit3)  
  
MisClassification3 <- (tbl6[1, "1"] + tbl6[2, "0"])/(sum(tbl6))  
MisClassification4 <- (tbl7[1, "1"] + tbl7[2, "0"])/(sum(tbl7))  
  
print(MisClassification3)

## [1] 0.3459854

print(MisClassification4)

## [1] 0