UNIVERSIDADE FEDERAL DE MINAS GERAIS
GRADUAÇÃO EM ENGENHARIA DE SISTEMAS
Introdução ao Reconhecimento de Padrões - Aula 6
Nikolas Dias Magalhães Fantoni - 2018019400

Classificador de Bayes aplicado a um problema multivariado (Base Spam)

No exercício proposto, a base de dados *spambase.data*, fornecida pelo professor, foi utilizada para demonstrar um classificador bayesiano. Primeiro, as amostras foram separadas em classes do tipo 1 (consideradas spam) ou do tipo 2 (não spams), de acordo com a última coluna de cada amostra observada. Depois, os dados foram separados em 10 *folds* e foi utilizada a técnica de validação cruzada, em que a cada iteração, um dos folds (correspondente a aproximadamente 10% dos dados) é utilizado como teste e o restante serve de base para o treinamento do classificador. Para uma implementação correta da separação dos *folds* de maneira aleatória, foi utilizada a biblioteca do *R caret*.

Após realizados os 10 testes, a acurácia e o desvio padrão médios dos testes da classe 1 e da classe 2, separadamente, são mostrados na figura 1.

```
media1:  95.64018 %  media2:  68.36982 %
desvio1:  2.170583 %  desvio2:  23.99426 %
```

Figura 1: Resultados obtidos para as classes separadas.

A acurácia total, juntamente com o desvio padrão total, estão mostrados na figura 2. Logo abaixo, tem-se o código utilizado neste exercício.

```
Acuracia total:  82.005 % e Desvio Padrao Total:  21.69439 %.
```

Figura 2: Resultados obtidos para o treinamento das classes em conjunto.

```
#Universidade Federal de Minas Gerais
#Introducao ao Reconhecimento de Padroes
#Nikolas Dias Magalhaes Fantoni
#AULA 6 - Base Spam
#2019/2

#Limpando o ambiente
rm(list=ls())

#Adicionando biblioteca
library(caret)

#Funcao estimativa densidade para n vari veis
pdfnvar <- function(x, m, K, n) {
  if (det(K) == 0) 10^(50) else (1/(sqrt((2*pi)^(n) *
  (det(K)))))*exp(-0.5*(t(x-m)%*%(solve(K))%*%(x-m)))
}

#Lendo os dados
spam <- read.csv( "spambase.data", sep=",",header = FALSE)
ic1 <- which(spam$V58 == 1)
ic2 <- which(spam$V58 == 0)
spamc1 <- spam[ic1,]
spamc2 <- spam[ic2,]
```

```
#Criando os folds que separao os dados
fl1 <- createFolds(spamc1[,1], k = 10, list = TRUE, returnTrain = FALSE)
fl2 <- createFolds(spamc2[,1], k = 10, list = TRUE, returnTrain = FALSE)

#Looping para os 10 testes
acuracia1 <- NULL
acuracia2 <- NULL
for (j in 1:10){
  test1 <- spamc1[fl1[[j]],]
  test2 <- spamc2[fl2[[j]],]
  trainc1 <- spamc1[-fl1[[j]],]
  trainc2 <- spamc2[-fl2[[j]],]

  #Calculando as medias
  u1<- NULL
  u2<- NULL
  for (m in 1:57){
    u1 <- rbind(u1, mean(trainc1[,m]))
    u2 <- rbind(u2, mean(trainc2[,m]))
  }

  #Covariancias dos dados
  cov1 <- cov(trainc1[,1:57])
  cov2 <- cov(trainc2[,1:57])

  #Testando o algoritmo para classe 1
  erro <- 0
  for (i in 1:length(test1[,1])){
    l <- t(test1[i,1:57])
    f1 <- pdfnvar(l, u1 , cov1, 57)
    f2 <- pdfnvar(l, u2 , cov2, 57)
    if (f2 == 0){
      c <- 1
    } else {
      c <- if ((f1/f2 > 1) == TRUE) 1 else 0
      if ((c-test1[i,58]) != 0) erro <- erro +1
    }
  }
  #Obtendo a acuracia da classe 1
  acerto <- 1 - erro/length(test1[,1])
  acuracia1 <- c(acuracia1, acerto)

  #Testando o algoritmo para classe 2
  erro <- 0
  for (i in 1:length(test2[,1])){
    l <- t(test2[i,1:57])
    f1 <- pdfnvar(l, u1 , cov1, 57)
    f2 <- pdfnvar(l, u2 , cov2, 57)
    if (f2 == 0){
      c <- 1
    } else {
      c <- if ((f1/f2 >= 1) == TRUE) 1 else 0
      if ((c-test2[i,58]) != 0) erro <- erro +1
    }
  }

  #Obtendo a acuracia da classe 2
```

```
  acerto <- 1 - erro/length(test2[,1])
  acuracia2 <- c(acuracia2, acerto)
}

#Imprimindo a saida
cat("\n acuracia classe 1: [", acuracia1, "]\n",
    "acuracia classe 2: [", acuracia2, "]\n\n", "media1: ",
    mean(acuracia1)*100,"%  media2: ", mean(acuracia2)*100, "%",
    "\n", "desvio1: ", sd(acuracia1)*100,"%  desvio2: ",
    sd(acuracia2)*100,"%", "\n\n")
cat("\n\nAcuracia total: ", mean(c(acuracia1,acuracia2))*100,
    "% e Desvio Padrao Total: ", sd(c(acuracia1, acuracia2))*100, "%.\n\n")
```