



## Classificador de Bayes aplicado a um problema multivariado (Base Heart)

No exercício proposto, foram amostrados inicialmente da base de dados *heart.dat*, fornecida pelo professor, 90% dos dados para o treinamento de um classificador bayesiano e outros 10% para testar o classificador. A amostra total continha 270 dados distribuídos em 13 variáveis descritivas e uma variável classificatória, que definia como 1 o paciente que não possuía a doença cardíaca e 2 o paciente que a possuía. A função abaixo foi utilizada pra estimar as densidades de probabilidades para cada classe:

```
pdfnvar <- function(x, m, K, n) {  
  (1/(sqrt((2*pi)^(n) * (det(K))))) *  
  exp(-0.5 * (t(x-m) %*% (solve(K)) %*% (x-m)))  
}
```

No primeiro teste, foi obtida uma acurácia de 88,88%. Para um segundo teste, a mesma base foi dividida em uma proporção de 70%-30% para as amostras de treinamento e teste, respectivamente. A nova acurácia passou a ser de 82,71%. Por fim, os dados foram divididos numa proporção em que o número de amostras de teste eram maiores que as amostras de treinamento, sendo 80% e 20% suas respectivas partes do total. Neste último teste, a acurácia obtida foi de 64%.

Estes pequenos testes mostram que para garantir uma eficácia maior de um classificador, é necessário que a amostra de treinamento seja suficientemente grande em relação à amostra a ser testada. O teste foi repetido diversas vezes e, em geral, a quantidade de amostras separadas para treinamento era diretamente proporcional à acurácia obtida do teste.

Ainda que em dados fictícios, cuja variância seja baixa, a relação entre a quantidade de amostras usadas para treinar e a acurácia possa não existir (ou seja, pode-se obter bons resultados mesmo com um número baixo de amostras), para dados reais, cuja variabilidade é alta devido à diversos fatores, é imprescindível que se obtenha o máximo de valores possíveis para o treinamento correto do algoritmo. O código do algoritmo é mostrado abaixo.

```
#Universidade Federal de Minas Gerais  
#Introducao ao Reconhecimento de Padroes  
#Nikolas Dias Magalhaes Fantoni  
#AULA 5 - Base Heart  
#2019/2  
  
#Limpando o ambiente  
rm(list=ls())  
  
#Funcao estimativa densidade para n variaveis  
pdfnvar <- function(x, m, K, n) {(1/(sqrt((2*pi)^(n)*(det(K))))) *  
  exp(-0.5 * (t(x-m) %*% (solve(K)) %*% (x-m)))}  
  
#Percentual amostra para treinamento  
per <- 0.9  
  
#Lendo os dados  
heart <- read.csv("heart.dat", sep="_", header = FALSE)  
ic1 <- which(heart$V14 == 1)  
ic2 <- which(heart$V14 == 2)
```

```

heartc1 <- heart[ic1,]
heartc2 <- heart[ic2,]

#Separando em treinamento e teste
iseq1 <- sample(length(heartc1[,1]))
iseq2 <- sample(length(heartc2[,1]))
trainc1 <- heartc1[iseq1[1:(per*length(iseq1))],]
trainc2 <- heartc2[iseq2[1:(per*length(iseq2))],]
test <- rbind(heartc1[iseq1[(per*length(iseq1)+1):length(iseq1)],],
              heartc2[iseq2[(per*length(iseq2)+1):length(iseq2)],])

#Calculo das medias
for (m in 1:13){
  u1 <- rbind(u1, mean(trainc1[,m]))
  u2 <- rbind(u2, mean(trainc2[,m]))
}

#Calculo das covariancias
cov1 <- cov(trainc1[,1:13])
cov2 <- cov(trainc2[,1:13])

#Teste do classificador
erro <- 0
for (i in 1:length(test[,1])){
  l <- t(test[i,1:13])
  f1 <- pdfnvar(l, u1, cov1, 13)
  f2 <- pdfnvar(l, u2, cov2, 13)
  c <- if ((f1/f2 >= 1) == TRUE) 1 else 2
  if ((c-test[i,14]) != 0) erro <- erro +1
}

#Imprimindo a saida
acertoporcento <- 100 - erro/length(test[,1])*100
cat("\n_Percentual_de_acertos:_",acertoporcento,"\n")

```