##########################################################

####### ALGUNS COMANDOS R - Apoio à Aula de An. Agrupamento

####### IBI5086 – IME - 2o.Sem-2023

####### Profa. Júlia Maria Pavan Soler

##########################################################

#Dados do cão pré-histórico

dat <- matrix(c(9.7, 21.0, 19.4, 7.7, 32.0, 36.5,

8.1, 16.7, 18.3, 7.0, 30.3, 32.9,

13.5, 27.3, 26.8, 10.6, 41.9, 48.1,

11.5, 24.3, 24.5, 9.3, 40.0, 44.6,

10.7, 23.5, 21.4, 8.5, 28.8, 37.6,

9.6, 22.6, 21.1, 8.3, 34.4, 43.1,

10.3, 22.1, 19.1, 8.1, 32.2, 35.0),6,7)

caes <- t(dat)

colnames(caes) <- c("X1","X2","X3","X4","X5","X6")

caes

dim(caes) #n=7, p=6

##Estatísticas descritivas multivariadas

mi<-colMeans(caes) #centróide dos dados

mi

s<-cov(caes) #matriz de covariância

s

r<-cor(caes) # matriz de correlação

r

#Padronização das variáveis (subtrai a média e divide pelo desvio padrão)

caesz<-scale(caes,center=TRUE,scale=TRUE)

caesz

colMeans(caesz) #próximo de 0

cov(caesz) #é a matriz de correlação de "caes"

cor(caesz) #cov(caesz) = cor(caesz)

#Distância Euclidiana (ordinária) entre observações

de<-dist(caes)

de

de<-round(de,2)

de

dim(de)

de <- as.matrix(de)

dim(de)

de37<-de[3,7]

de37

#Distância Euclidiana Ponderada (dist. Pearson) entre observações

caesp<-scale(caes, center = FALSE, scale = apply(caes, 2, sd, na.rm = TRUE))

dep<-dist(caesp)

dep

round(dep,2)

dep<-as.matrix(dep)

dep37<-dep[3,7]

dep37

#ESCALONAMENTO MULTIDIMENSIONAL - Solução Métrica

#cmdscale(d, k = 2, eig = FALSE, add = FALSE, x.ret = FALSE)

esc <- cmdscale(de)

round(esc,2)

#x <- esc[,1]

#y <- esc[,2]

#plot(x,y)

plot(esc, main="Escalonamento Multidimensional")

text(esc, labels=c("C1", "C2","C3", "C4","C5", "C6", "C7"), lwd=3)

difd<-as.matrix(dist(caes))-as.matrix(dist(esc))

difd #compara as matrizes de distância "de" com a estimada por EM (k=2)

esc2<-cmdscale(de, k = 2, eig = TRUE, add = FALSE, x.ret = FALSE)

esc2 #saída com os autovalores

(esc2$eig[1]+esc2$eig[2])/sum(esc2$eig)

#proporção da variância total explicada

# ANÁLISE DE AGRUPAMENTO

caes

d<-dist(caes) #distância Euclidiana entre as obs

round(d,2)

hc1 <- hclust(d)

hc1 #default usa o agrupamento hierárquico Ligação Completa

plot(hc1,main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Lig. Completa - Dist. Euclidiana")

rect.hclust(hc1, k = 2, border = "red")

x1<-cutree(hc1,2)

x1

#names(hc1)

#hc1$order

#par(mfrow=c(2,2))

hc2<- hclust(d, method='single')

hc2

plot(hc2,main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Lig. Simples - Dist. Euclidiana")

rect.hclust(hc2, k = 2, border = "red")

x2<-cutree(hc2,2)

x2

hc3<- hclust(d, method='average')

hc3

plot(hc3,main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Lig. Média - Dist. Euclidiana")

rect.hclust(hc3, k = 2, border = "red")

x3<-cutree(hc3,2)

x3

hc4<- hclust(d, method='centroid')

hc4

plot(hc4,main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Lig. Centróide - Dist. Euclidiana")

rect.hclust(hc4, k = 2, border = "red")

x4<-cutree(hc4,2)

x4

hcW.1 <- hclust(d,method="ward.D")

hcW.1

plot(hcW.1, main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Ward.D - Dist. Euclidiana")

rect.hclust(hcW.1, k = 2, border = "red")

xW1<-cutree(hcW.1,2)

hcW.2 <- hclust(d, method="ward.D2")

hcW.2

plot(hcW.2, main= "Dendrograma - Dados dos cães", xlab="Cão", sub="Ward.D2 - Dist. Euclidiana")

rect.hclust(hcW.2, k = 2, border = "red")

xW2<-cutree(hcW.2,2)

#Ward.D e Ward.D2 são algoritmos diferentes

#Ward.D2 tem sido o mais recomendado

#Obter as Coordenadas Principais e comparar as soluções

#par(mfrow=c(2,2))

esc <- cmdscale(de)

plot(esc, main="Escalonamento Multidimensional")

text(esc, labels=c("C1", "C2","C3", "C4","C5", "C6", "CPre"), lwd=3)

#Método da Partição K-Means - Hartigan-Wong

library(graphics)

cl.1 <- kmeans(caes, 2) #default: algorithm="Hartigan-Wong")

names(cl.1)

cl.1$cluster

cl.1$centers

cl.1$totss #Soma de Quadrados total

cl.1$withinss #Soma de Quadrados Dentro (SQD) de grupos

cl.1$tot.withinss #Soma das SQD

cl.1$betweenss #Soma de Quadrados Entre (SQE) grupos

cl.1$size

#caracterizando os grupos formados

pairs(caes, pch=23, bg=c('red', 'blue')[as.factor(cl.1$cluster)],

main="K-means - Hartigan-Wong")#matrix de gráficos de dispersão

plot(caes, col = cl.1$cluster)

points(cl.1$centers, col = 1:3, pch = 8, cex = 2)

plot(caes[,c(1,3)], col = cl.1$cluster)

points(cl.1$centers[,c(1,3)], col = 1:3, pch = 8, cex = 2)

#Método da Partição K-Means - Lloyd

cl.2 <- kmeans(caes, 2, algorithm = "Lloyd")

cl.2

cl$cluster

cl.2$cluster

pairs(caes, pch=23, bg=c('blue', 'red')[as.factor(cl.2$cluster)],

main="K-means - Lloyd")#matrix de gráficos de dispersão

#Dados: pardais

pard<-read.table('C:\\Users\\jpsol\\OneDrive\\Documents\\Julia\\IBI5085-2023\\pardais.csv', header = TRUE, sep=';', dec=',')

pard

head(pard)

attach(pard)

names(pard)

str(pard)

ct <- table(grup); ct

#Visualizando os dados

X2<- cmdscale(dist(pard))

plot(X2,col= factor(grup), main="Escalonamento Multidimensional", pch=16)

legend("bottomleft",levels(factor(grup)),fill=seq\_along(levels(factor(grup))))

#Classificação dos pardais - Ward

hcW.3 <- hclust(dist(pard), method="ward.D2")

hcW.3

plot(hcW.3, main= "Dendrograma - Dados dos pardais", xlab="pardais", sub="Ward.D2 - Dist. Euclidiana")

rect.hclust(hcW.3, k = 2, border = "red")

xW3<-cutree(hcW.3,2)

table(grup,xW3)

(sum(diag(table(grup,xW3))))/49

#library(rafalib)

#myplclust(hcW.3, labels=grup, lab.col=as.numeric(grup), cex=0.5)

#Classificando os pardais via K-means

set.seed(2913)

cl.3 <- kmeans(pard[,3:7], 2, algorithm = "Lloyd")

cl.3

table(grup,cl.3$cluster)

(sum(diag(table(grup,cl.3$cluster))))/49 #taxa de classificação correta

plot(X2, col=cl.3$cluster, main="K-Means-LLoyd",pch=16)

set.seed(2913)

cl.4 <- kmeans(pard[,3:7], 2, algorithm = "Hartigan-Wong")

cl.4

table(grup,cl.4$cluster)

(sum(diag(table(grup,cl.4$cluster))))/49

plot(X2, col=cl.4$cluster, main="K-Means-Hartigan-Wong",pch=16)

#HeatMap

#agrupamento hierárquico Ligação Completa das unidades amostrais

hc12 <- hclust(dist(pard[,3:7]))

plot(hc12,main= "Dendrograma - Dados dos pardais", xlab="pardais", sub="Lig. Completa - Dist. Euclidiana")

rect.hclust(hc12, k = 2, border = "red")

library(gplots)

hv <- heatmap.2(as.matrix(X), col=bluered, scale="column", tracecol="#303030")

hv

names(hv)

heatmap.2(as.matrix(X))

#heatmap.2(as.matrix(X), reorderfun=function(d, w) reorder(d, w, agglo.FUN = mean) )