Kmeans

Subbu Kandhaswamy

10/30/2019

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

datazoo <- read.csv('C:/Users/Administrator/Documents/MSDatascience/IST707/Week4/zoo.csv')  
str(datazoo)

## 'data.frame': 101 obs. of 18 variables:  
## $ animal.name: Factor w/ 100 levels "aardvark","antelope",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ hair : int 1 1 0 1 1 1 1 0 0 1 ...  
## $ feathers : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ egg : int 0 0 1 0 0 0 0 1 1 0 ...  
## $ milk : int 1 1 0 1 1 1 1 0 0 1 ...  
## $ airborne : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ aquatic : int 0 0 1 0 0 0 0 1 1 0 ...  
## $ predator : int 1 0 1 1 1 0 0 0 1 0 ...  
## $ toothed : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ backbone : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ breathes : int 1 1 0 1 1 1 1 0 0 1 ...  
## $ venomous : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ fins : int 0 0 1 0 0 0 0 1 1 0 ...  
## $ legs : int 4 4 0 4 4 4 4 0 0 4 ...  
## $ tail : int 0 1 1 0 1 1 1 1 1 0 ...  
## $ domestic : int 0 0 0 0 0 0 1 1 0 1 ...  
## $ catsize : int 1 1 0 1 1 1 1 0 0 0 ...  
## $ type : int 1 1 4 1 1 1 1 4 4 1 ...

datazoocleaned <- datazoo[, 2:17]  
library(RWeka)  
model\_rweka <- SimpleKMeans(datazoocleaned, control=Weka\_control(N=7, S=100))  
model\_rweka

##   
## kMeans  
## ======  
##   
## Number of iterations: 5  
## Within cluster sum of squared errors: 94.70379505535756  
##   
## Initial starting points (random):  
##   
## Cluster 0: 0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,1  
## Cluster 1: 1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1  
## Cluster 2: 0,0,1,0,0,1,1,0,0,0,0,0,6,0,0,0  
## Cluster 3: 0,0,1,0,0,1,1,1,1,1,1,0,4,0,0,0  
## Cluster 4: 0,1,1,0,0,0,1,0,1,1,0,0,2,1,0,0  
## Cluster 5: 0,0,1,0,0,0,0,0,1,1,0,0,4,1,0,1  
## Cluster 6: 0,0,0,1,0,1,1,1,1,1,0,1,0,1,0,1  
##   
## Missing values globally replaced with mean/mode  
##   
## Final cluster centroids:  
## Cluster#  
## Attribute Full Data 0 1 2 3 4 5 6  
## (101.0) (13.0) (37.0) (18.0) (8.0) (14.0) (7.0) (4.0)  
## ===================================================================================================  
## hair 0.4257 0 1 0.2222 0 0 0 0.5  
## feathers 0.198 0 0 0 0 1 0.8571 0  
## egg 0.5842 1 0.027 0.9444 0.875 1 1 0  
## milk 0.4059 0 1 0 0 0 0 1  
## airborne 0.2376 0 0.0541 0.3333 0 0.9286 0.4286 0  
## aquatic 0.3564 1 0.0541 0.3333 0.625 0.2857 0.2857 1  
## predator 0.5545 0.6923 0.4865 0.5 0.875 0.4286 0.4286 1  
## toothed 0.604 1 0.973 0 1 0 0 1  
## backbone 0.8218 1 1 0 1 1 1 1  
## breathes 0.7921 0 1 0.6111 0.875 1 1 1  
## venomous 0.0792 0.0769 0 0.2222 0.375 0 0 0  
## fins 0.1683 1 0 0 0 0 0 1  
## legs 2.8416 0 3.6757 4.7222 2.5 2 2.2857 0.5  
## tail 0.7426 1 0.8649 0.0556 0.625 1 1 0.75  
## domestic 0.1287 0.0769 0.2162 0.0556 0 0.2143 0 0  
## catsize 0.4356 0.3077 0.7568 0.0556 0 0 1 1

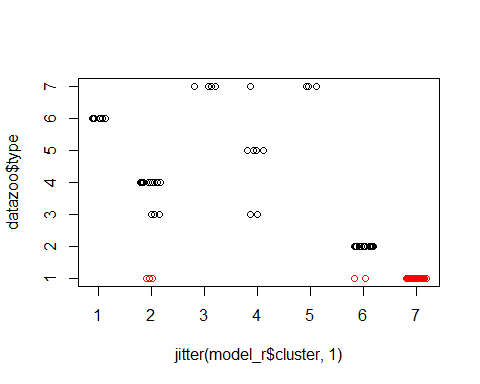
model\_r <- kmeans(datazoocleaned,7)  
model\_r

## K-means clustering with 7 clusters of sizes 8, 19, 4, 8, 4, 22, 36  
##   
## Cluster means:  
## hair feathers egg milk airborne aquatic  
## 1 0.50000000 0.0000000 1.00000000 0.00000000 0.7500000 0.00000000  
## 2 0.05263158 0.0000000 0.78947368 0.15789474 0.0000000 0.89473684  
## 3 0.00000000 0.0000000 1.00000000 0.00000000 0.0000000 0.25000000  
## 4 0.00000000 0.0000000 1.00000000 0.00000000 0.0000000 0.75000000  
## 5 0.00000000 0.0000000 0.75000000 0.00000000 0.0000000 0.75000000  
## 6 0.09090909 0.9090909 0.90909091 0.09090909 0.8181818 0.27272727  
## 7 1.00000000 0.0000000 0.02777778 1.00000000 0.0000000 0.08333333  
## predator toothed backbone breathes venomous fins legs  
## 1 0.1250000 0.00000000 0.00 1.0000000 0.2500000 0.00000000 6.000000  
## 2 0.7894737 1.00000000 1.00 0.2631579 0.1578947 0.84210526 0.000000  
## 3 0.5000000 0.00000000 0.00 0.5000000 0.2500000 0.00000000 0.000000  
## 4 0.7500000 0.62500000 0.75 0.7500000 0.1250000 0.00000000 4.125000  
## 5 1.0000000 0.00000000 0.00 0.2500000 0.2500000 0.00000000 7.000000  
## 6 0.4090909 0.09090909 1.00 1.0000000 0.0000000 0.00000000 2.000000  
## 7 0.5277778 0.97222222 1.00 1.0000000 0.0000000 0.02777778 3.722222  
## tail domestic catsize  
## 1 0.0000000 0.12500000 0.0000000  
## 2 0.9473684 0.05263158 0.3684211  
## 3 0.0000000 0.00000000 0.0000000  
## 4 0.3750000 0.00000000 0.1250000  
## 5 0.2500000 0.00000000 0.2500000  
## 6 1.0000000 0.13636364 0.2727273  
## 7 0.8611111 0.22222222 0.8055556  
##   
## Clustering vector:  
## [1] 7 7 2 7 7 7 7 2 2 7 7 6 2 3 4 5 6 7 2 2 6 6 7 6 1 4 4 6 7 7 1 7 7 6 2  
## [36] 7 7 6 2 1 1 6 1 6 7 7 5 7 7 7 7 1 4 5 7 7 6 6 6 6 2 2 2 7 7 7 2 7 7 7  
## [71] 7 6 5 2 2 7 2 3 6 6 2 3 2 6 7 4 2 6 1 4 4 4 2 6 7 6 7 1 7 3 6  
##   
## Within cluster sum of squares by cluster:  
## [1] 6.75000 26.63158 3.50000 12.37500 8.50000 29.00000 48.02778  
## (between\_SS / total\_SS = 80.8 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

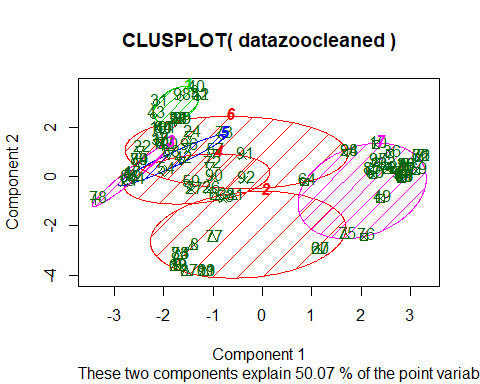
model\_r$centers

## hair feathers egg milk airborne aquatic  
## 1 0.50000000 0.0000000 1.00000000 0.00000000 0.7500000 0.00000000  
## 2 0.05263158 0.0000000 0.78947368 0.15789474 0.0000000 0.89473684  
## 3 0.00000000 0.0000000 1.00000000 0.00000000 0.0000000 0.25000000  
## 4 0.00000000 0.0000000 1.00000000 0.00000000 0.0000000 0.75000000  
## 5 0.00000000 0.0000000 0.75000000 0.00000000 0.0000000 0.75000000  
## 6 0.09090909 0.9090909 0.90909091 0.09090909 0.8181818 0.27272727  
## 7 1.00000000 0.0000000 0.02777778 1.00000000 0.0000000 0.08333333  
## predator toothed backbone breathes venomous fins legs  
## 1 0.1250000 0.00000000 0.00 1.0000000 0.2500000 0.00000000 6.000000  
## 2 0.7894737 1.00000000 1.00 0.2631579 0.1578947 0.84210526 0.000000  
## 3 0.5000000 0.00000000 0.00 0.5000000 0.2500000 0.00000000 0.000000  
## 4 0.7500000 0.62500000 0.75 0.7500000 0.1250000 0.00000000 4.125000  
## 5 1.0000000 0.00000000 0.00 0.2500000 0.2500000 0.00000000 7.000000  
## 6 0.4090909 0.09090909 1.00 1.0000000 0.0000000 0.00000000 2.000000  
## 7 0.5277778 0.97222222 1.00 1.0000000 0.0000000 0.02777778 3.722222  
## tail domestic catsize  
## 1 0.0000000 0.12500000 0.0000000  
## 2 0.9473684 0.05263158 0.3684211  
## 3 0.0000000 0.00000000 0.0000000  
## 4 0.3750000 0.00000000 0.1250000  
## 5 0.2500000 0.00000000 0.2500000  
## 6 1.0000000 0.13636364 0.2727273  
## 7 0.8611111 0.22222222 0.8055556

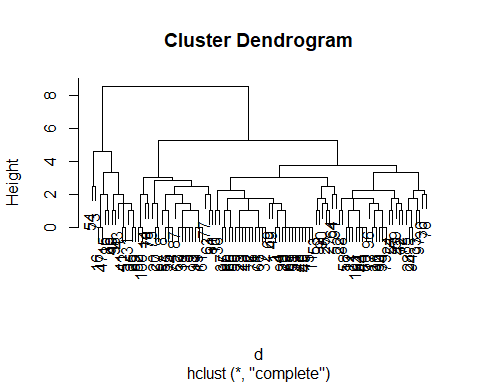
cluster\_assignment <- data.frame(datazoo,model\_r$cluster)  
View(cluster\_assignment)



library(cluster)  
clusplot(datazoocleaned,model\_r$cluster, color=TRUE, shade = T, labels = 2, lines = 0) # plot clusters



d=dist(as.matrix(datazoocleaned))  
hc=hclust(d)  
plot(hc)



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.