Workshop 4.12 Solution

Veerasak Kritsanapraphan

2/1/2020

library(stats)  
library(caret)  
library(e1071)  
library(fastDummies)  
library(dplyr)  
library(BBmisc)  
telco <- read.csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv", header=TRUE)

### Remove NA

telco <- na.omit(telco)

### Dummy Column

telcodum <- select(telco, Churn, gender,SeniorCitizen,Partner,tenure,PhoneService, MultipleLines,OnlineBackup,Contract,PaperlessBilling,PaymentMethod, MonthlyCharges ,TotalCharges)  
telcodum <- dummy\_cols(telcodum)

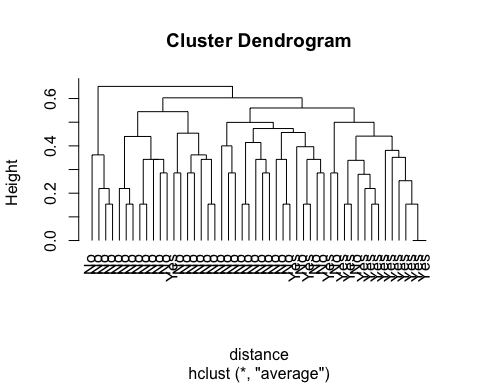
### Hirarchical Clustering

set.seed(123)  
sampletelco <- telcodum[sample(1:nrow(telcodum), 50),]  
  
output\_telco\_train <- sampletelco[,'Churn']  
input\_telco\_train <- sampletelco[, c(5,12:36)]  
input\_telco\_train$tenure <- normalize(input\_telco\_train$tenure, method="standardize")  
input\_telco\_train$MonthlyCharges <- normalize(input\_telco\_train$MonthlyCharges, method="standardize")  
input\_telco\_train$TotalCharges <- normalize(input\_telco\_train$TotalCharges, method="standardize")  
  
str(input\_telco\_train)

## 'data.frame': 50 obs. of 26 variables:  
## $ tenure : num -0.189 -0.355 -0.687 -0.936 -0.106 ...  
## $ MonthlyCharges : num -1.429 -1.447 0.448 1.023 0.659 ...  
## $ TotalCharges : num -0.7289 -0.7736 -0.4691 -0.683 0.0471 ...  
## $ Churn\_No : int 1 1 1 0 1 0 0 1 1 1 ...  
## $ Churn\_Yes : int 0 0 0 1 0 1 1 0 0 0 ...  
## $ gender\_Female : int 1 0 0 1 0 1 1 1 0 1 ...  
## $ gender\_Male : int 0 1 1 0 1 0 0 0 1 0 ...  
## $ Partner\_Yes : int 1 0 0 0 0 0 1 1 1 0 ...  
## $ Partner\_No : int 0 1 1 1 1 1 0 0 0 1 ...  
## $ PhoneService\_No : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PhoneService\_Yes : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ MultipleLines\_No phone service : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MultipleLines\_No : int 1 1 1 0 0 0 0 0 0 1 ...  
## $ MultipleLines\_Yes : int 0 0 0 1 1 1 1 1 1 0 ...  
## $ OnlineBackup\_Yes : int 0 0 0 1 1 1 1 0 0 0 ...  
## $ OnlineBackup\_No : int 0 0 1 0 0 0 0 1 1 0 ...  
## $ OnlineBackup\_No internet service : int 1 1 0 0 0 0 0 0 0 1 ...  
## $ Contract\_Month-to-month : int 0 1 1 1 1 0 1 0 0 0 ...  
## $ Contract\_One year : int 1 0 0 0 0 1 0 0 1 1 ...  
## $ Contract\_Two year : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ PaperlessBilling\_Yes : int 0 1 1 1 1 0 1 1 1 0 ...  
## $ PaperlessBilling\_No : int 1 0 0 0 0 1 0 0 0 1 ...  
## $ PaymentMethod\_Electronic check : int 0 0 0 1 0 0 1 0 0 0 ...  
## $ PaymentMethod\_Mailed check : int 1 0 0 0 0 0 0 0 0 0 ...  
## $ PaymentMethod\_Bank transfer (automatic): int 0 0 1 0 0 1 0 1 0 0 ...  
## $ PaymentMethod\_Credit card (automatic) : int 0 1 0 0 1 0 0 0 1 1 ...

### Distance Function

distance <- dist(input\_telco\_train, method="binary")  
cluster <- hclust(distance, method="average")  
plot(cluster, hang=-1, sampletelco$Churn)



### See how it cluster

group.2 <- cutree(cluster, k = 2) # prune the tree by 3 clusters  
table(group.2, sampletelco$Churn) # compare with known classes

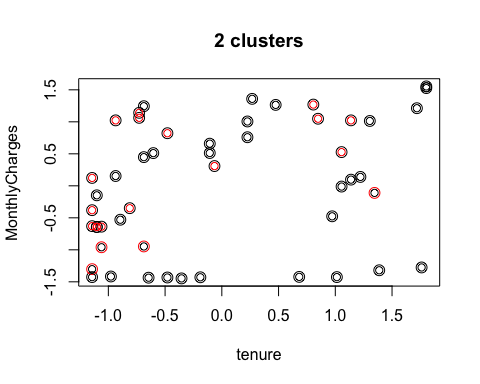
##   
## group.2 No Yes  
## 1 30 16  
## 2 4 0

group.2

## 2023 5543 2876 6207 6610 321 3711 6270 3874 3207 6719 3183 4757 4020 723   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 6315 1727 296 2301 6694 6238 4858 4490 6969 4596 4965 3812 4162 2026 1031   
## 1 1 1 1 1 1 1 1 2 2 1 1 2 1 1   
## 6744 6317 4835 5568 173 3344 5307 1514 2226 1620 999 2899 2892 2578 1066   
## 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1   
## 970 1628 3255 1858 5991   
## 1 1 1 1 1

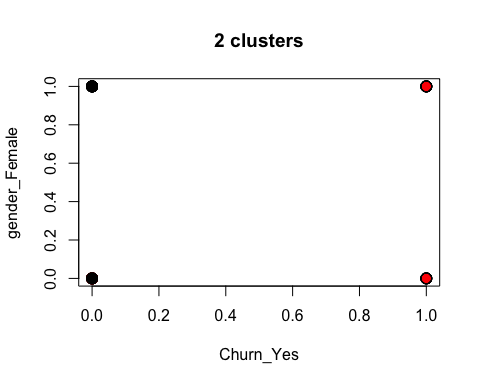
## Plot Graph

plot(input\_telco\_train[,c(1,2)], col=group.2, pch=21, cex=1.5, main="2 clusters")  
points(input\_telco\_train[,c(1,2)], col=sampletelco$Churn, pch=21, cex=1)



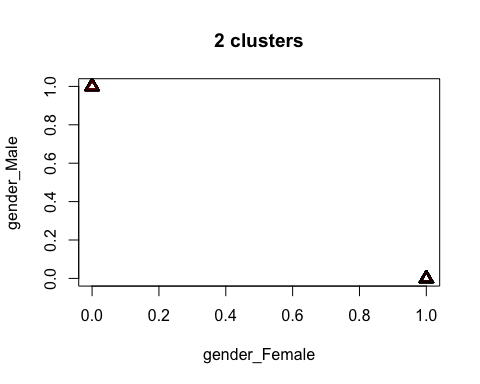
## Plot Graph

plot(input\_telco\_train[,c(5,6)], col=group.2, pch=19, cex=1.5, main="2 clusters")  
points(input\_telco\_train[,c(5,6)], col=sampletelco$Churn, pch=19, cex=1)



## Plot Graph

plot(input\_telco\_train[,c(6,7)], col=group.2, pch=24, cex=1.5, main="2 clusters")  
points(input\_telco\_train[,c(6,7)], col=sampletelco$Churn, pch=24, cex=1)



## Plot Graph

plot(input\_telco\_train[,c(8,9)], col=group.2, pch=25, cex=1.5, main="2 clusters")  
points(input\_telco\_train[,c(8,9)], col=sampletelco$Churn, pch=25, cex=1)

