Week 4

Tomás

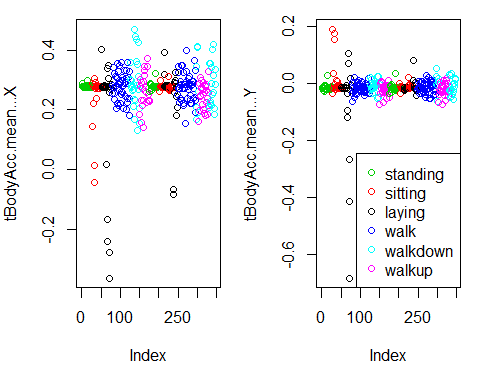
10 de diciembre de 2019

# Clustering Case Study

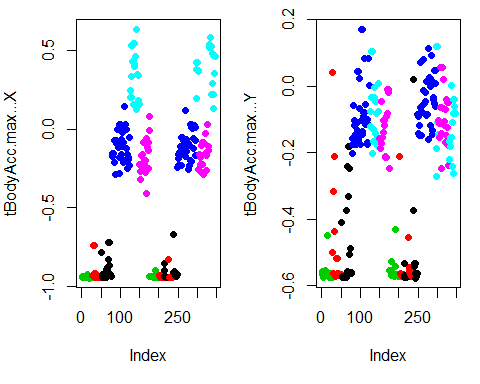
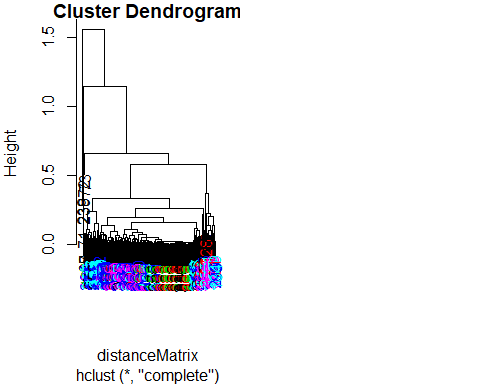
data <- read.csv("SamsungData.csv")  
table(data$activity)

##   
## laying sitting standing walk walkdown walkup   
## 1407 1286 1374 1226 986 1073

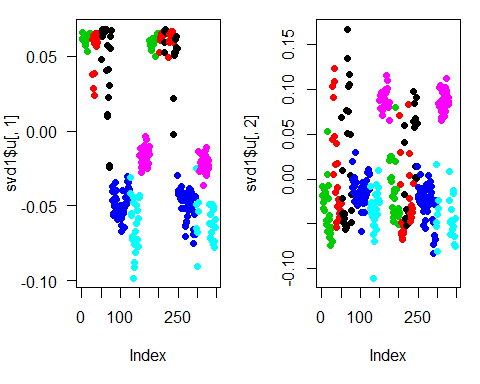
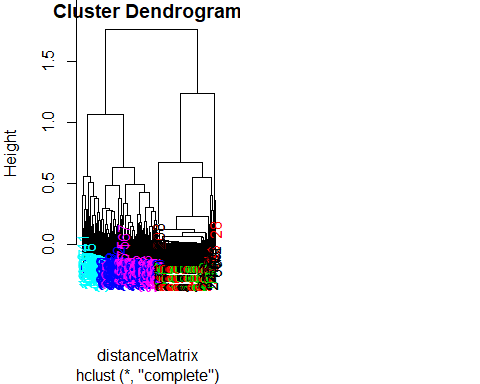
par(mfrow = c(1,2), mar = c(5,4,1,1))  
data <- transform(data, activity = factor(activity))  
sub1 <- subset(data, subject == 1)  
plot(sub1[, 2], col = sub1$activity, ylab = names(sub1)[2])#Ploteo las primeras dos mediciones para el primer sujeto  
plot(sub1[, 3], col = sub1$activity, ylab = names(sub1)[3])  
legend("bottomright", legend = unique(sub1$activity), col = unique(sub1$activity), pch = 1)



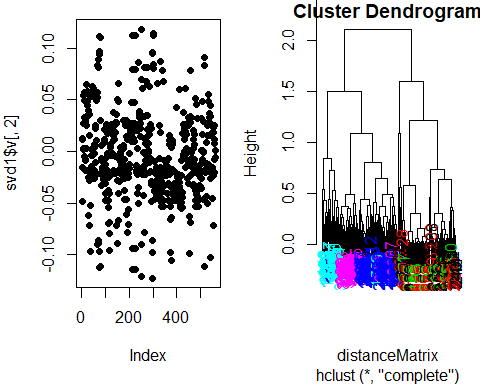
source("myplclust.R")  
distanceMatrix <- dist(sub1[, 2:4])  
hclustering <- hclust(distanceMatrix)  
myplclust(hclustering, lab.col = unclass(sub1$activity))# Creamos un dendograma que no nos entrega mucha informacion  
  
{par(mfrow = c(1,2), xpd = TRUE)  
plot(sub1[, 11], pch = 19, col = sub1$activity, ylab = names(sub1)[11])  
plot(sub1[, 12], pch = 19, col = sub1$activity, ylab = names(sub1)[12])}#Ahora puedo plotear los maximos y encuentro una diferencia mas notoria sentarse y no hacer nada no entrega informacion. Pero moverse subir escaleras existe mas variabilidad



distanceMatrix <- dist(sub1[, 11:13])  
hclustering <- hclust(distanceMatrix)  
myplclust(hclustering, lab.col = unclass(sub1$activity))# Creamos un dendrograma que ahora si nos entrega la informacion de que existe dos clusters mas marcados  
  
svd1 <- svd(scale(sub1[, -c(563,564)]))  
{par(mfrow = c(1,2))  
plot(svd1$u[, 1], col = sub1$activity, pch = 19, las = 1)  
plot(svd1$u[, 2], col = sub1$activity, pch = 19)}



plot(svd1$v[, 2], pch = 19)#Veo cual es la columna que genera mas varianza  
maxContrib <- which.max(svd1$v[, 2]) #Entrega cual es la columna que entrega una mayor variacion a los datos  
distanceMatrix <- dist(sub1[, c(11:13, maxContrib)]) #calculamos la distancia del maximo mas este extra de la maxima contribucion  
hclustering <- hclust(distanceMatrix)  
myplclust(hclustering, lab.col = unclass(sub1$activity))# Las 3 actividades de mas movimiento estan mejor separadas por clusters, las sin movimiento no generan mucha diferencia



names(data)[maxContrib]

## [1] "tBodyGyroMag.arCoeff..2"

kClust <- kmeans(sub1[, -c(563,564)], centers = 6, nstart = 100) #TK cluster puede encontrar distintos dependiendo de donde parta  
table(kClust$cluster, sub1$activity)

##   
## laying sitting standing walk walkdown walkup  
## 1 0 0 0 5 24 27  
## 2 11 1 0 43 0 0  
## 3 10 24 27 0 0 0  
## 4 12 22 26 0 0 0  
## 5 17 0 0 41 0 0  
## 6 0 0 0 6 25 26