Week 10 & 11 clustering

Pooja Chopra

7 June 2019

library(“cluster”) library(“NbClust”) library(“clustertend”) library(“factoextra”) library(“NbClust”) library(“purrr”) library(“tibble”) library(“dplyr”)

Preparation of data

CLus_data <- na.omit(CLus_data) #Remove any missing values

CLus_data <- scale(CLus_data) # Scaling Dataframe for standardize variable

Hierarchical Clustering , Commonly used functions are hclust, agnes , diana

d <- dist(CLus_data, method = “euclidean”) # Dissimilarity matrix

HCLUST

hc1 <- hclust(d, method = “ward”) # Ward method Linkage plot(hc1) # display dendogram groups <- cutree(hc1, k=7) # cut tree into 7 clusters rect.hclust(hc1, k=7, border=“blue”) # dendogram with blue borders around the 7 clusters

agnes

hc2 <- agnes(d, method = “complete / single”) # agglomerative clustering using agnes() with a method for complete linkage clustering. hc2$ac # Agglomerative coefficient

divisive

hc4 <- diana(d) hc4$dc # Agglomerative coefficient pltree(hc4, cex = 0.6, hang = -1, main = “Dendrogram of diana”) # dendogram

methods to assess

m <- c(“average”, “single”, “complete”, “ward”) names(m) <- c(“average”, “single”, “complete”, “ward”)

ac <- function(x) {agnes(d, method = x)$ac} # function to compute coefficient map_dbl(m, ac) ## Package reqired purrr

To determine optimal clusture we execute Elbow Method , Average Silhouette Method , Gap Statistic Method

Elbow Method

fviz_nbclust(CLus_data, FUN = hcut, method = “wss”)

Average Silhouette Method

fviz_nbclust(CLus_data, FUN = hcut, method = “silhouette”)

Gap Statistic Method

gap_stat <- clusGap(CLus_data, FUN = hcut, nstart = 25, K.max = 10, B = 50) fviz_gap_stat(gap_stat) # Package required stats

Kmeans Clustering

kmeans two parameters are required x : matrix or DF and 2: center : initial cluster centroids , n start =22 which means it create 22 intial configuration

k2 <- kmeans(CLus_data, centers = 3 , nstart=22) fviz_cluster(k2, data = CLus_data) # to get better view at clusture we use fviz_cluster

CLus_data %>%as_tibble() %>%mutate(cluster = k2$cluster,state = row.names(CLus_data)) # second option to use standard pairwise scatter plots ggplot(aes(UrbanPop, Murder, color = factor(cluster), %>% label = state)) +
geom_text()

To determine optimal clusture we execute Elbow Method , Average Silhouette Method , Gap Statistic Method

elbow method

fviz_nbclust(CLus_data, kmeans, method = “wss”)

Average Silhouette Method

fviz_nbclust(CLus_data, kmeans, method = “silhouette”)

Gap Statistic Method NbClust Package

gap_stat <- clusGap(CLus_data, FUN = kmeans, nstart = 25,K.max = 10, B = 50) # clus gap function provide gap statistics fviz_gap_stat(gap_stat)