Q5.R

mirrien

2022-04-10

```
################################################
# Question 5

euclid <- function(x,K){
  distance = matrix(NA, nrow= nrow(x), ncol = nrow(K))
  for(j in 1:nrow(K)) {
    for(i in 1:nrow(x)) {
      distance[i,j]<-dist(rbind(x[i,],K[j,]), method = "euclidean")
    }
  }
  return(distance)
}

mykmeans <- function(x,K,iters) {
  # convert df to matrix
  x = as.matrix(x)

  # randomly sample some centers, set a seed 100
  set.seed(100)
  K <- x[sample(nrow(x), K),]

  # empty lists to store outputs
  assignments <- vector(iters, mode = "list")
  locations <- vector(iters, mode = "list")

  for(i in 1:iters) {
    # call euclidean distance helper function
    dists = euclid(x,K)
    # find minimum distance
    clusters <- apply(dists,1,which.min)
    # tapply mean()
    centers <- apply(x,2,tapply,clusters,mean)

    # store outputs
    assignments[[i]] <- clusters
    locations[[i]] <- centers
  }

  # return outputs in list
  return(list(locations=locations[[1]], assignments = assignments[[1]]))
}
```

```r
suppressWarnings(suppressMessages(library(ggpubr)))
suppressWarnings(suppressMessages(library(factoextra)))

# Create a test data frame
set.seed(100)
sample_df = data.frame(V1 = rnorm(50,0,10), V2 = rnorm(50,0,10))
# head(sample_df)

# Use mykmeans()

result01 = mykmeans(scale(sample_df),3,1000)
result01$assignments
```

```
##  [1] 1 1 2 2 1 2 1 2 2 1 1 2 1 2 2 1 2 2 1 3 1 1 1 2 1 1 1 2 1 1 2 3 1 1 2 1 2 2
## [39] 2 3 1 3 1 1 1 3 1 3 1 1
```

```r
# Plot

plot1 = fviz_cluster(list(data = sample_df, cluster = result01$assignments),
                     data = sample_df,
                     palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
                     geom = "point",
                     ellipse.type = "convex",
                     ggtheme = theme_bw()
)


# Use kmeans()

set.seed(222)
result02 = kmeans(scale(sample_df),3,1000)
result02$cluster
```

```
##  [1] 3 3 1 2 3 1 3 2 1 3 1 1 1 2 1 3 1 1 1 2 1 3 3 2 3 1 3 1 1 3 1 2 3 3 1 1 1 1
## [39] 2 2 3 2 3 3 3 2 3 2 3 3
```

```r
# Plot

plot2 = fviz_cluster(result02, data = sample_df,
                     palette = c("#2E9FDF", "#00AFBB", "#E7B800"),
                     geom = "point",
                     ellipse.type = "convex",
                     ggtheme = theme_bw()
)

# Combine two plots to visualize comparison
figure <- ggarrange(plot1, plot2, labels = c("mykmeans", "kmeans"),
                    label.x = 0.35, label.y = 1,
                    ncol=2,nrow=1)
```

```
# Output the combined plot
figure
```