Lab 12: Clustering

Donggyun Kim
27008257

4/16/2018

K-means clustering

```
my_kmeans <- function(X, k) {
  X <- as.matrix(X)
  N <- nrow(X)
  P <- ncol(X)
  K <- k
  cluster_sizes <- numeric(K)
  cluster_means <- matrix(0, nrow = K, ncol = P)
  clustering_vector <- numeric(N)
  wss_cluster <- numeric(K)
  bss_over_tss <- numeric(1)

  index <- sample(N, K)
  centroids <- X[index, ]
  centroids_new <- X[index, ]
  distant2 <- matrix(0, nrow = N, ncol = K)

  l <- 1
  iteration <- 100
  while (l <= iteration) {
    for (i in 1:N) {
      for(j in 1:K) {
        distant2[i, j] <- as.numeric(t(X[i, ] - centroids[j, ]) %*% (X[i, ] - centroids[j, ]))
      }
      min_k <- which.min(distant2[i, ])
      clustering_vector[i] <- min_k
    }
    for (k in 1:K) {
      index <- clustering_vector == k
      centroids_new[k, ] <- apply(X[index, ], 2, mean)
    }
    if (sum(!centroids_new == centroids) == 0) {
      break
    }
    centroids <- centroids_new
    l <- l + 1
  }

  cluster_means <- centroids_new

  for (k in 1:K) {
    cluster_sizes[k] <- sum(clustering_vector == k)
  }
```

```
  for (k in 1:K) {
    index <- clustering_vector == k
    wss_cluster[k] <- sum(apply(sweep(X[index, ], 2, centroids[k, ], "-"), 1, function(x) sum(x^2)))
  }

  TSS <- sum(apply(sweep(X, 2, apply(X, 2, mean), "-"), 1, function(x) sum(x^2)))
  BSS <- TSS - sum(wss_cluster)
  bss_over_tss <- BSS / TSS

  list(cluster_sizes = cluster_sizes,
       cluster_means = cluster_means,
       clustering_vector = clustering_vector,
       wss_cluster = wss_cluster,
       bss_over_tss = bss_over_tss)
}
set.seed(1991)
my_kmeans(iris[, 1:4], k = 3)
```

```
## $cluster_sizes
## [1] 32 96 22
##
## $cluster_means
##      Sepal.Length Sepal.Width Petal.Length Petal.Width
## [1,]     5.193750    3.631250     1.475000    0.271875
## [2,]     6.314583    2.895833     4.973958    1.703125
## [3,]     4.731818    2.927273     1.772727    0.350000
##
## $clustering_vector
##   [1] 1 3 3 3 1 1 3 1 3 3 1 1 3 3 1 1 1 1 1 1 1 1 1 3 3 1 1 1 3 3 1 1 1 3
##  [36] 1 1 1 3 1 1 3 3 1 1 3 1 3 1 1 2 2 2 2 2 2 2 3 2 2 3 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 3 2 2 2 2 2 2
## [106] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [141] 2 2 2 2 2 2 2 2 2 2
##
## $wss_cluster
## [1]   6.032188 118.651875  18.070000
##
## $bss_over_tss
## [1] 0.7904898
```

```
set.seed(1991)
kmeans(iris[, 1:4], centers = 3)
```

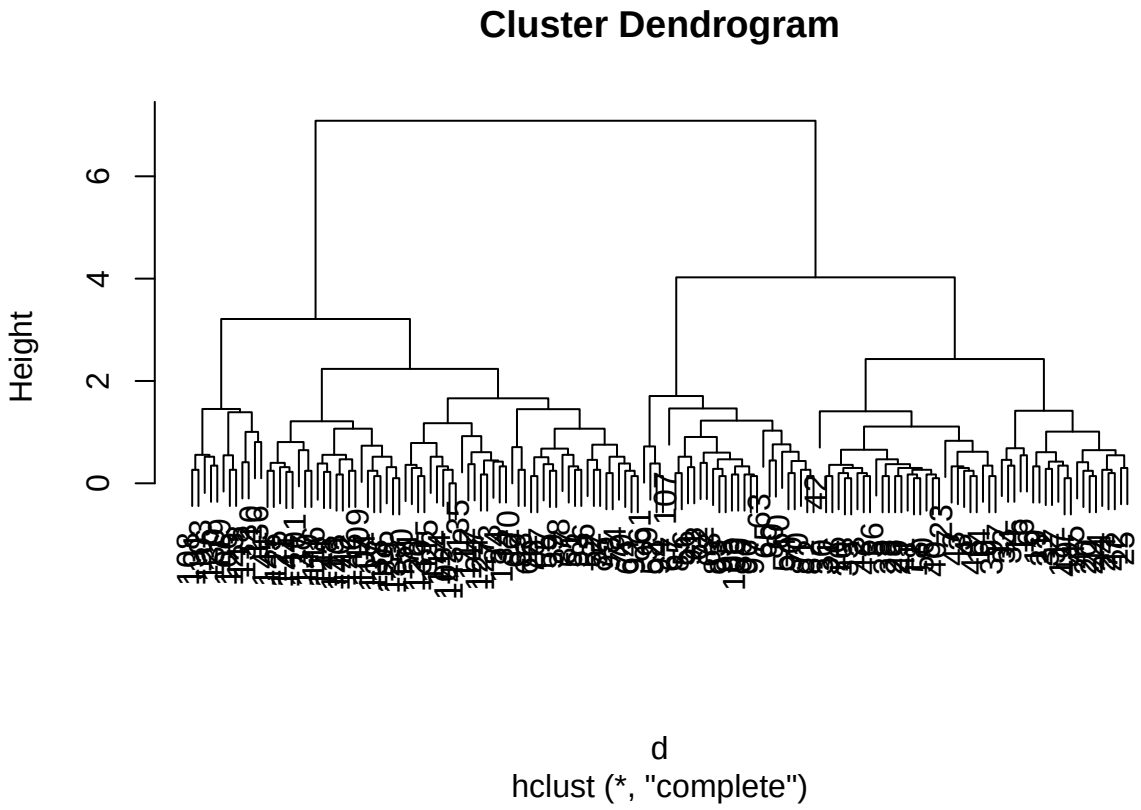
```
## K-means clustering with 3 clusters of sizes 33, 96, 21
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.175758    3.624242     1.472727   0.2727273
## 2     6.314583    2.895833     4.973958   1.7031250
## 3     4.738095    2.904762     1.790476   0.3523810
##
## Clustering vector:
##   [1] 1 3 3 3 1 1 1 1 3 3 1 1 3 3 1 1 1 1 1 1 1 1 1 3 3 1 1 1 3 3 1 1 1 3
##  [36] 1 1 1 3 1 1 3 3 1 1 3 1 3 1 1 2 2 2 2 2 2 2 3 2 2 3 2 2 2 2 2 2 2 2
```

```
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 3 2 2 2 2 2 2
## [106] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [141] 2 2 2 2 2 2 2 2 2 2
##
## Within cluster sum of squares by cluster:
## [1]   6.432121 118.651875  17.669524
##  (between_SS / total_SS =  79.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Hierarchical clustering

```
d <- dist(iris[, 1:4], method = "euclidean")
hc.complete <- hclust(d, method = "complete")
hc.average <- hclust(d, method = "average")
hc.single <- hclust(d, method = "single")

plot(hc.complete)
```

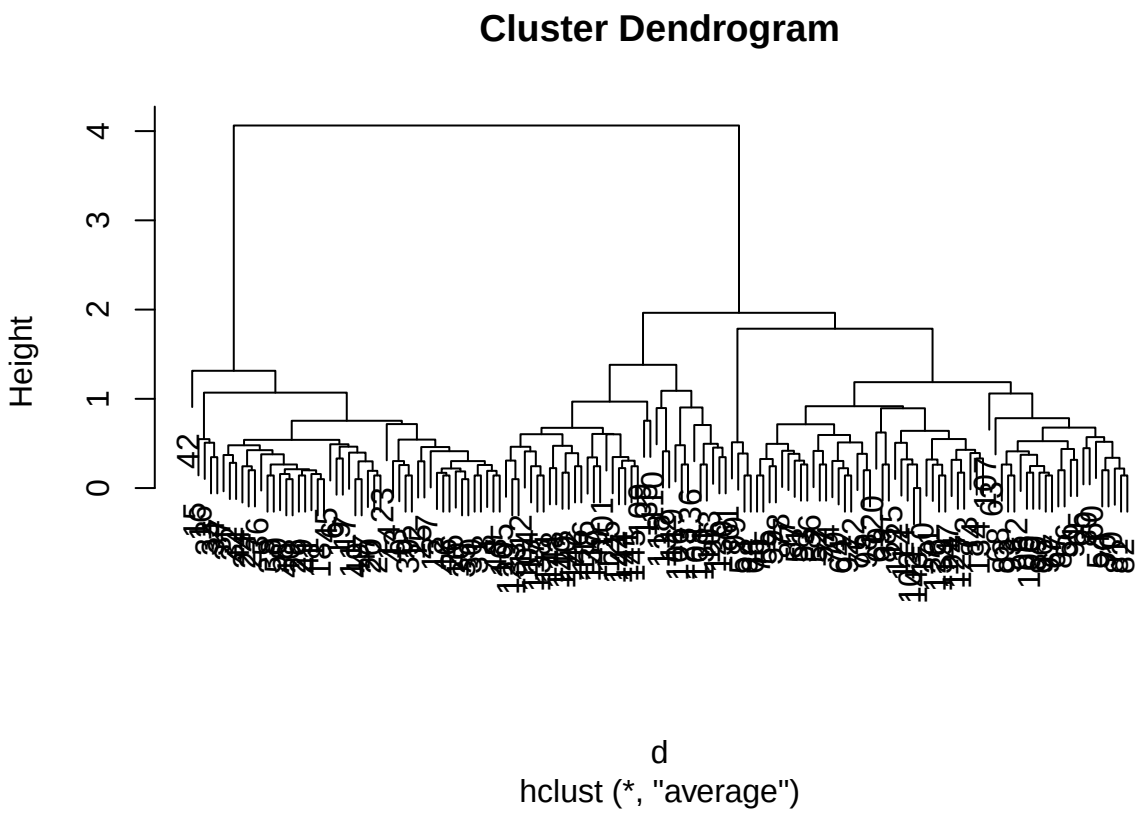



```
cutree(hc.complete, k = 3)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 2 3 2 3 2 3 3 3 3 2 3 2 3 3 2 3
```

```
##  [71] 2 3 2 2 2 2 2 2 2 3 3 3 3 2 3 2 2 2 3 3 3 2 3 3 3 3 3 2 3 3 2 2 2 2 2
## [106] 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [141] 2 2 2 2 2 2 2 2 2 2
```

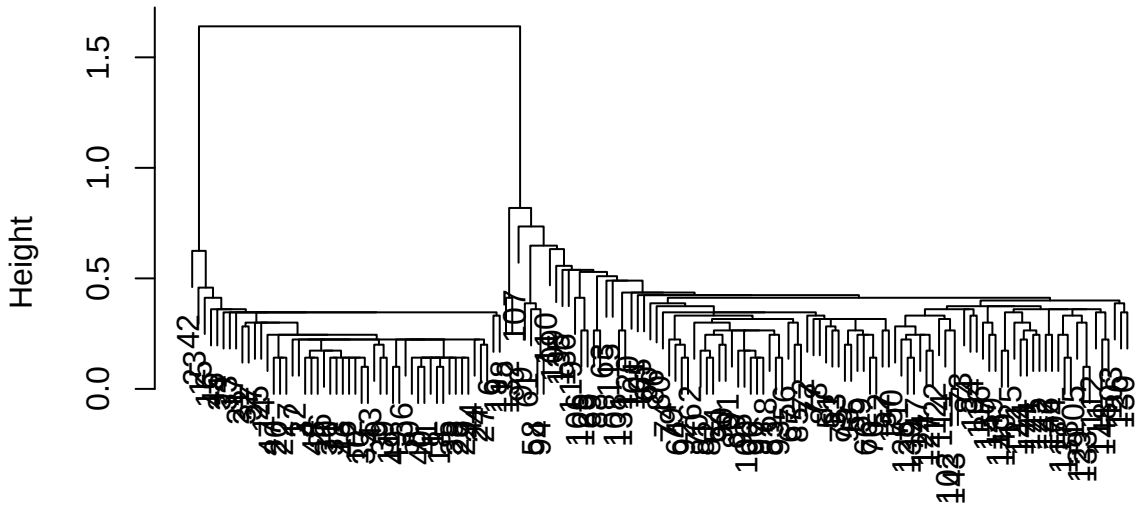
```
plot(hc.average)
```




```
cutree(hc.average, k = 3)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3
## [106] 3 2 3 3 3 3 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3
## [141] 3 3 2 3 3 3 2 3 3 2
```

```
plot(hc.single)
```




```
cutree(hc.single, k = 3)
```

```
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [106] 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2
## [141] 2 2 2 2 2 2 2 2 2 2 2
```