hmwrk4.2

student

2023-09-06

Clustering

Summary

R code performs k-means clustering on the Iris dataset to find the best combination of predictors and the optimal number of clusters(k). Best result found that using "Sepal.Width" as a predictor and best k=2 clusters minimized the within-cluster sum of squares (WSS), sum of the squared distances between each data point and the centroid of its cluster of 12.32171, finds Setosa as best flower type.

Begin Code

Load Libraries

```
library(ggplot2)
```

Load Data

```
# load the Iris data
data(iris)
```

Elbow Method: find_elbow function attempts to find a point where the decrease in WSS starts to slow down looks like an "elbow", and this is taken as a good indication of the appropriate number of clusters to use.

```
# create the find_elbow function
find_elbow <- function(wss_values) {
  # elbow finding logic here
  diff_wss <- diff(wss_values)
  diff_ratio <- diff(diff_wss)
  # +1 as diff reduces the length by 1
  return (which.max(diff_ratio) + 1)
}
```

Create Variables

```
# variables to store the best results
best_wss <- Inf
best_k <- NA
best_predictors <- NULL
```

Number of predictors: 4, "Sepal.Length", "Sepal.Width", "Petal.Length", and "Petal.Width".

```r
# number of predictors
n_predictors <- ncol(iris[, 1:4])
```

Loop through predictions, use kmeans, and find the best k, print graph.

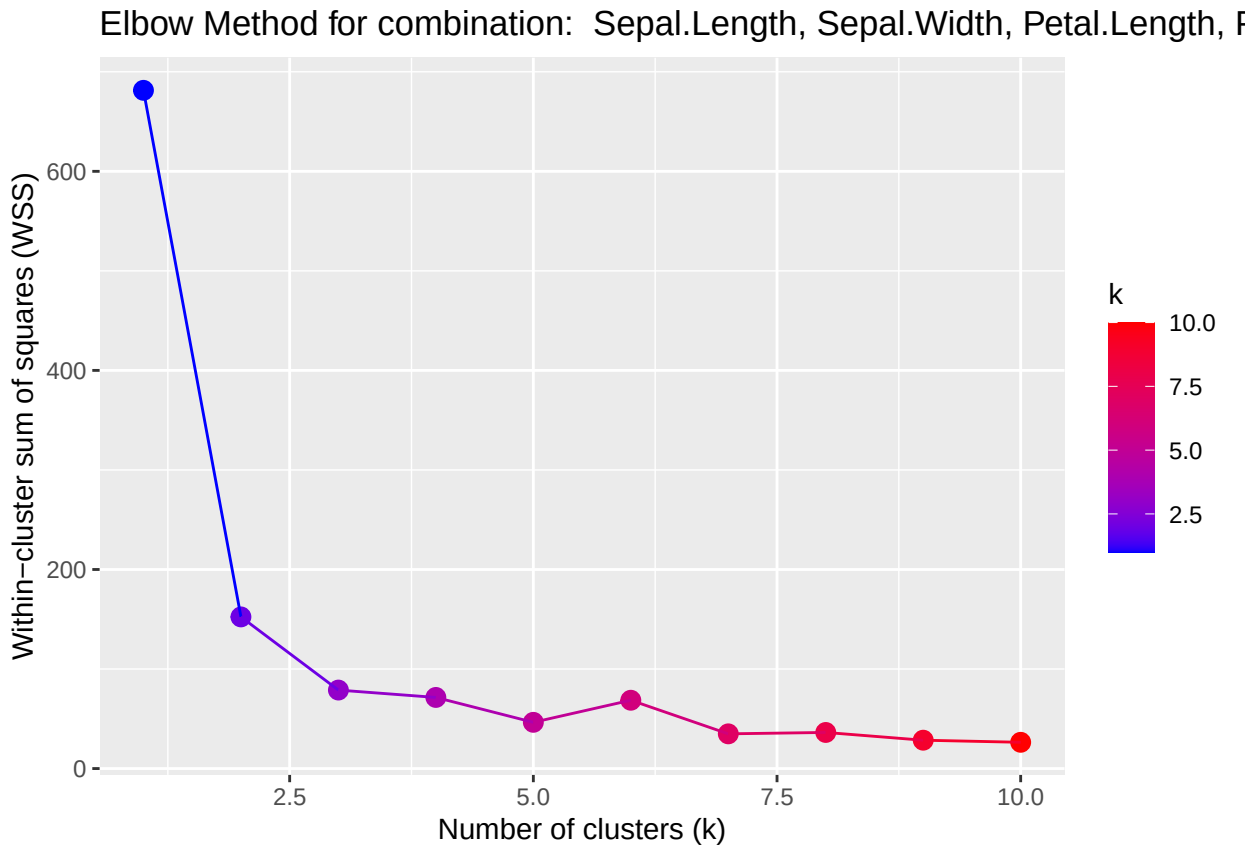
X-axis: the number of clusters you are using for that particular k-means run.

Y-axis:Within-cluster sum of squares (WSS): the total distance of each point to its cluster centroid, lower values may mean that the data points are closer to the centroids, which usually means better clustering.

Points/Lines: Each point represents the WSS value for a specific k.

Color Gradient: The color indicates the value of k, going from blue (smaller k) to red (larger k).

```r
set.seed(123)

# loop through all non-empty combinations of predictors
for (n in 1:n_predictors) {
  combinations <- combn(1:n_predictors, n)
  n_combinations <- ncol(combinations)

  for (i in 1:n_combinations) {
    predictors <- combinations[, i]
    # initialize WSS array
    wss <- numeric(10)

    # find the best 'k' for the current combination
    for (k in 1:10) {
      km_out <- kmeans(iris[, predictors], centers=k)
      wss[k] <- km_out$tot.withinss
    }

    # use the elbow method to find the optimal k for this combination
    optimal_k <- find_elbow(wss)

    # plot elbow graph
    plot = ggplot(data.frame(k = 1:10, WSS = wss), aes(x = k, y = WSS)) +
    geom_point(aes(color = k), size = 3) +
    geom_line(aes(color = k)) +
    scale_color_gradient(low = "#0000FF", high = "#FF0000") +
    ggtitle(paste("Elbow Method for combination: ", paste(colnames(iris)[predictors], collapse = ", "))
    xlab("Number of clusters (k)") +
    ylab("Within-cluster sum of squares (WSS)")

    # check if the current combination is better
    if (wss[optimal_k] < best_wss) {
      best_wss <- wss[optimal_k]
      best_k <- optimal_k
      best_predictors <- predictors
    }
  }
}
print(plot)
```



Print Results

```
# print the best results
cat("Best WSS:", best_wss, "\n")
```

```
## Best WSS: 12.32171
```

```
cat("Best k:", best_k, "\n")
```

```
## Best k: 2
```

```
cat("Best predictors:", colnames(iris)[best_predictors], "\n")
```

```
## Best predictors: Sepal.Width
```

Use kmeans and view Possible Flower Type: Setosa is found to have the best match according my findings.

```
# load the Iris data
data(iris)

# perform k-means clustering using Sepal.Width and k = 2
km_out <- kmeans(iris[, "Sepal.Width", drop = FALSE], centers = 2)
```

```r
# add the cluster assignments back to the original Iris data frame
iris$Cluster <- km_out$cluster

# view the data with cluster assignments
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species Cluster
## 1          5.1         3.5          1.4         0.2  setosa       1
## 2          4.9         3.0          1.4         0.2  setosa       2
## 3          4.7         3.2          1.3         0.2  setosa       2
## 4          4.6         3.1          1.5         0.2  setosa       2
## 5          5.0         3.6          1.4         0.2  setosa       1
## 6          5.4         3.9          1.7         0.4  setosa       1
```

```r
# cluster assignments vs actual flower types
table(iris$Cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1     33          2         8
##   2     17         48        42
```