Grip Task #2

Insha Shah

6/17/2021

Topic: Prediction Using Unsupervised Machine Learning

Problem: From the 'Iris' dataset, predict the optimum number of clusters and represent it visually.

Load the required packages

```
library(cluster)
```

Load the Iris Dataset

```
data<-iris[,-5]
head(iris)
```

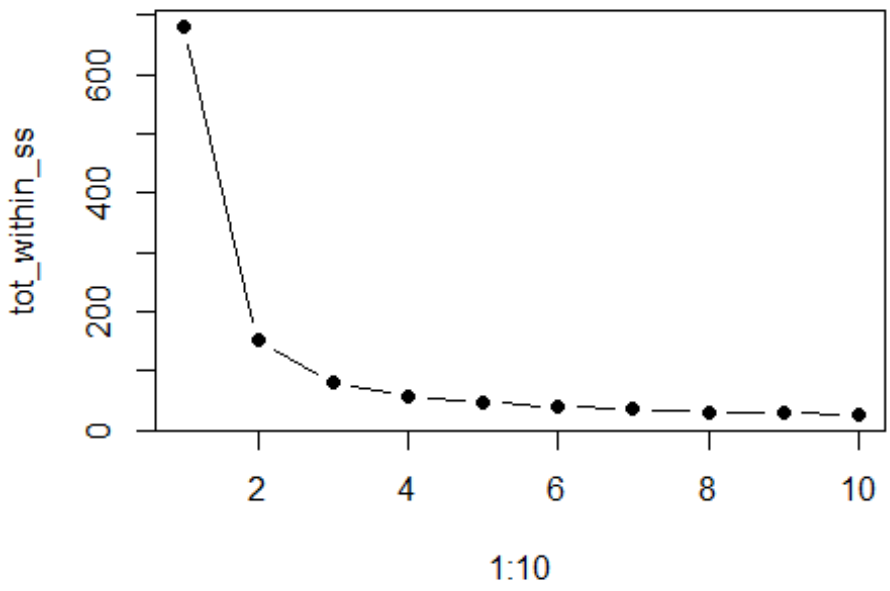
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5  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

Finding the optimum number of clusters for k-means classification

```
set.seed(123)
tot_within_ss<- vector(mode = "character", length = 10)
for (i in 1:10) {
  iriscluster<- kmeans(data[,1:4], centers = i, nstart = 20)
  tot_within_ss[i] <- iriscluster$tot.withinss
}
```

Plotting the results onto a line graph

```
plot(1:10, tot_within_ss, type = "b", pch = 19)
```



The optimum clusters is where the elbow occurs. Thus, the number of clusters is '3'.

Applying Kmeans to the dataset

```
set.seed(123)
kmean <- kmeans(data[,1:4], centers = 3, iter.max = 300)
kmean

## K-means clustering with 3 clusters of sizes 50, 62, 38
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.006000    3.428000     1.462000    0.246000
## 2     5.901613    2.748387     4.393548    1.433871
## 3     6.850000    3.073684     5.742105    2.071053
##
## Clustering vector:
##   [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
## [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## [75] 2 2 2 3 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 2 3 3 3
## [112] 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 3 3 2 3 3 3 2
## [149] 3 2
##
```

```
## Within cluster sum of squares by cluster:
## [1] 15.15100 39.82097 23.87947
##  (between_SS / total_SS =  88.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Confusion Matrix-Comparing the predicted clusters with the original data

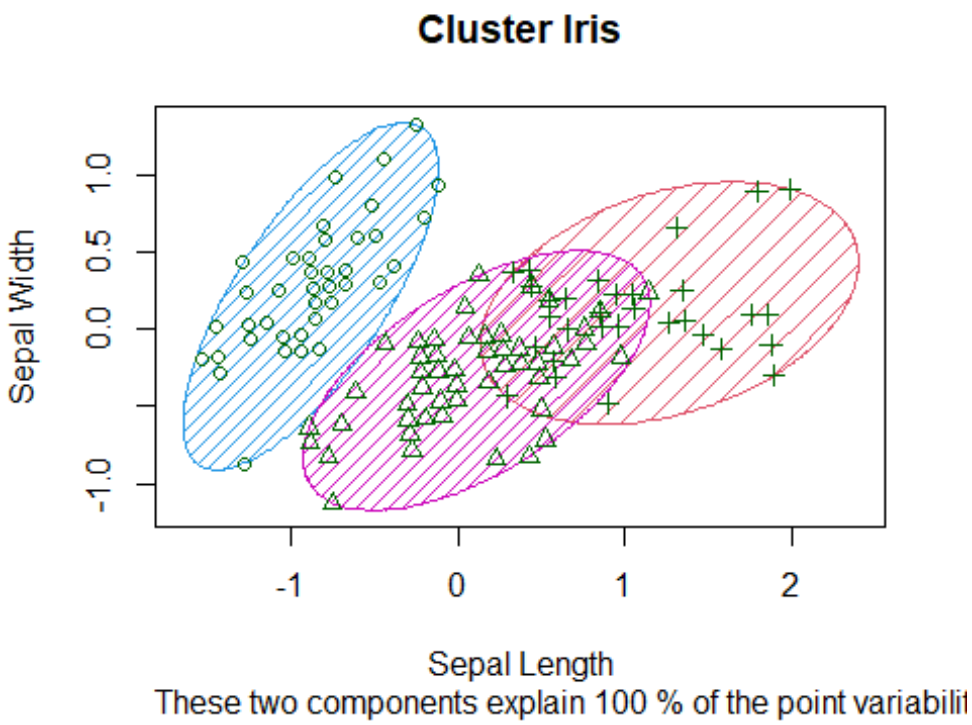
```
table(kmean$cluster, iris$Species)

##
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         48        14
##   3      0          2        36
```

All 50 Setosa are correctly classified as Setosa. Out of 62 Versicolor,48 are correctly classified as Versicolor and 14 are classified as Virginica Out of 38 Virginica, 36 are correctly classified as Virginica and 2 are classified as Versicolor

Visualizing the cluster - First two columns

```
clusplot(data[,c(1,2)],
         kmean$cluster,
         shade = T,
         color = T,
         lines = 0,
         span = T,
         main = "Cluster Iris",
         xlab = "Sepal Length",
         ylab = "Sepal Width",
         plotchar = T)
```


Cluster Iris
Sepal Width
1.0
0.5
0.0
-1.0
-1
0
1
2
Sepal Length
These two components explain 100 % of the point variabilit