MSDS_650_KMeans_Practice.R

SeanOMalley1

Tue Nov 29 18:39:42 2016

```
#MSDS_650
#Week7_Exercise_1
#Sean O'Malley

require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
# load and verify
data(iris)

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
```

```
# set seed to ensure reproducible results
set.seed(25)

# apply kmeans function, choose 3 clusters (3 iris species)
km <- kmeans(iris[,1:4], 3, nstart = 25)

km
```

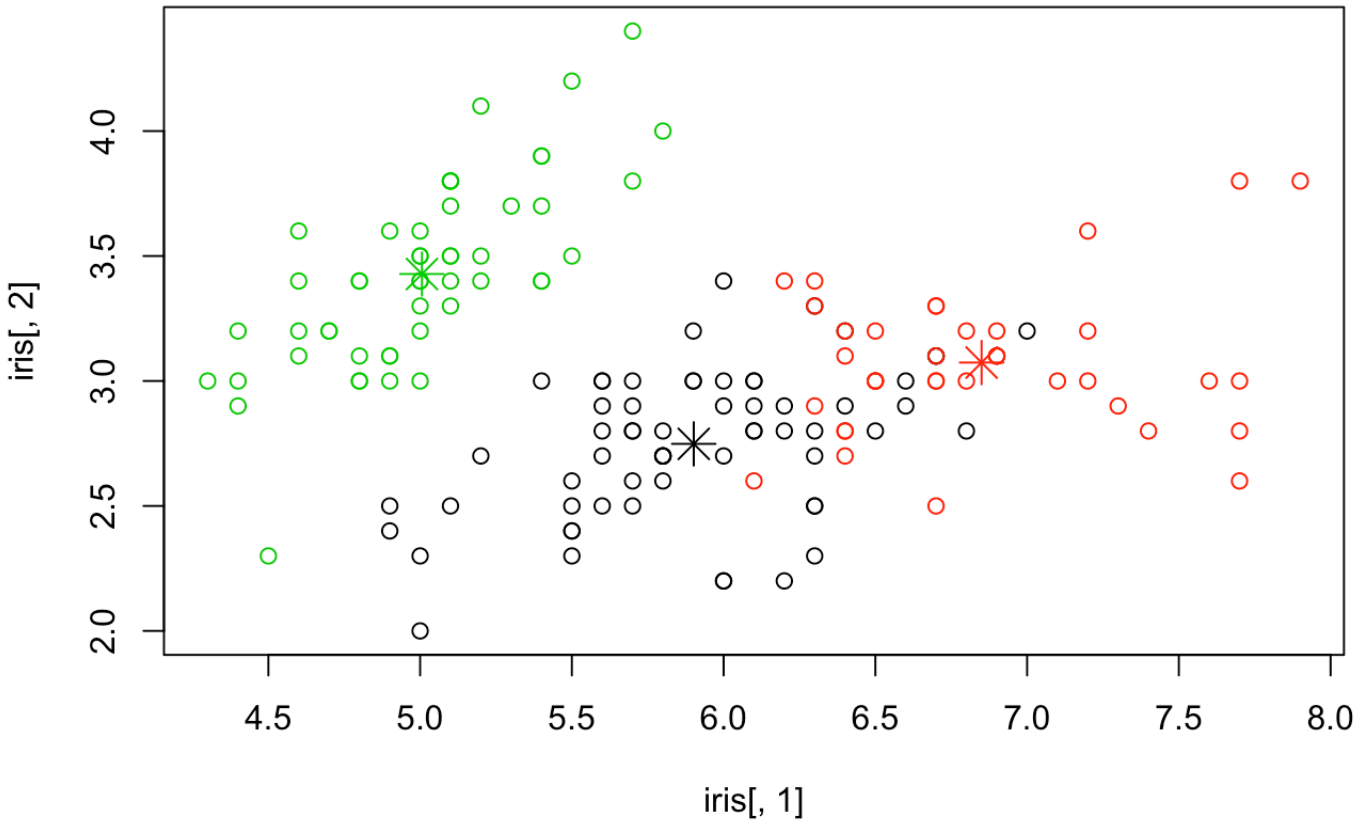
```
## K-means clustering with 3 clusters of sizes 62, 38, 50
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.901613    2.748387     4.393548    1.433871
## 2     6.850000    3.073684     5.742105    2.071053
## 3     5.006000    3.428000     1.462000    0.246000
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [36] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [71] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2
## [106] 2 1 2 2 2 2 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2
## [141] 2 2 1 2 2 2 1 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 39.82097 23.87947 15.15100
##  (between_SS / total_SS =  88.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
# compare clusters with species and plot results
table(km$cluster, iris$Species)
```

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

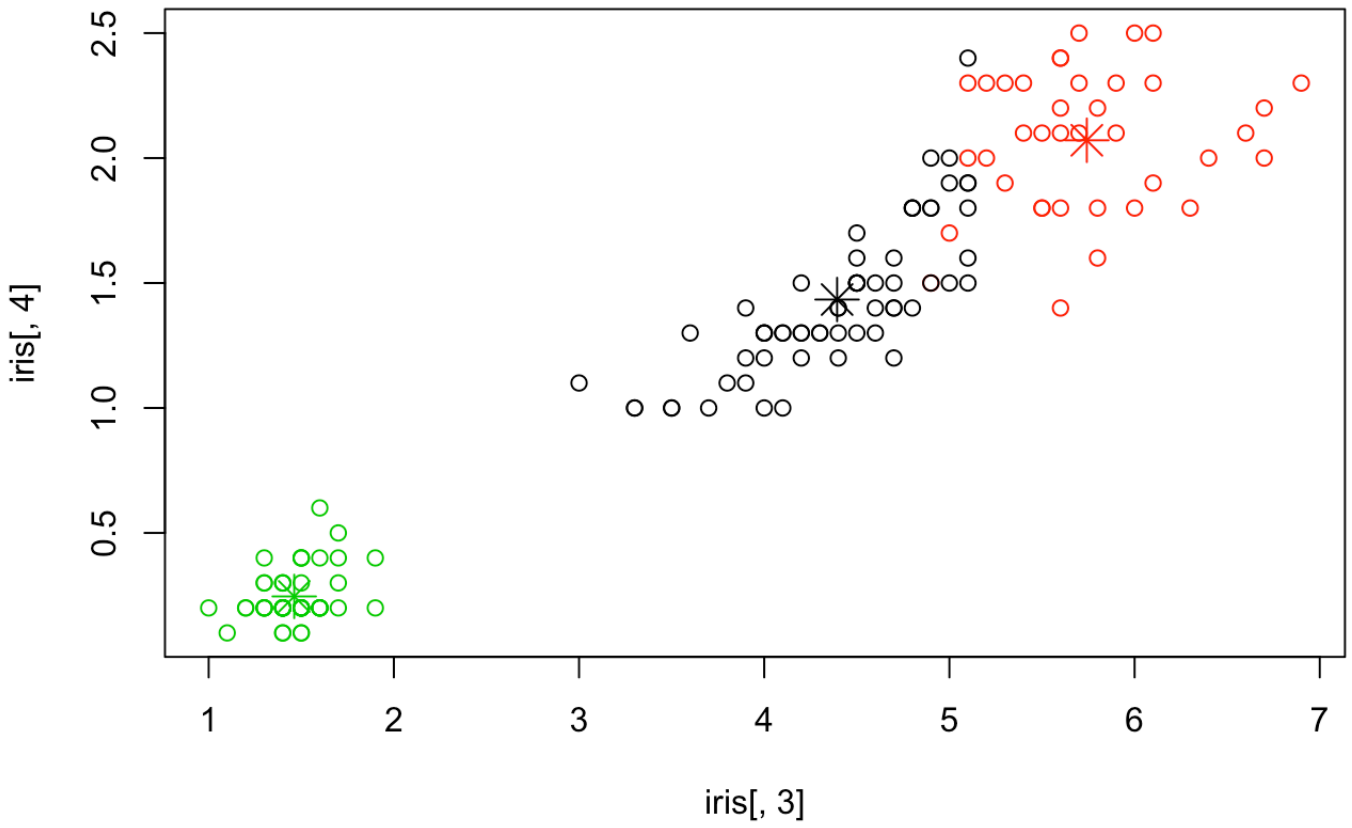
```
# plot sepal length by sepal width
plot(iris[,1], iris[,2], col = km$cluster)

# add points to the cluster centers
points(km$centers[,c(1,2)], col = 1:3, pch = 8, cex = 2)
```




```
# plot petal length and width
plot(iris[,3],iris[,4], col = km$cluster)

# add points to the cluster centers
points(km$centers[,c(3,4)], col = 1:3, pch = 8, cex = 2)
```


iris[, 4]
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
iris[, 3]