Problem 6.3

Chen Bo Calvin Zhang

09/11/2020

Import and preprocess the data

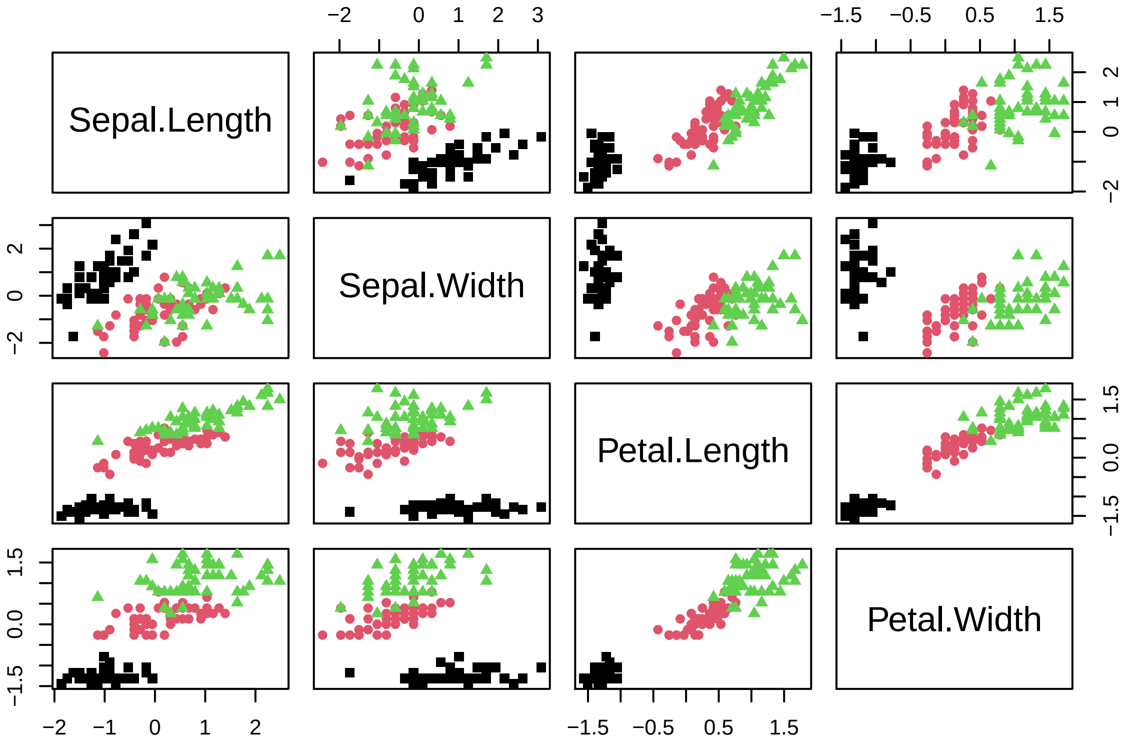
```r
data(iris)

# preprocess
X.iris = scale((iris[, 1:4]), scale=TRUE) # center and standardise
L.iris = iris[, 5]

table(L.iris)
```

```
## L.iris
##     setosa versicolor  virginica
##         50         50         50
```

```r
pairs(X.iris, col=as.integer(L.iris), pch=as.integer(L.iris)+14)
```



Now, perform K-means with K = 3.

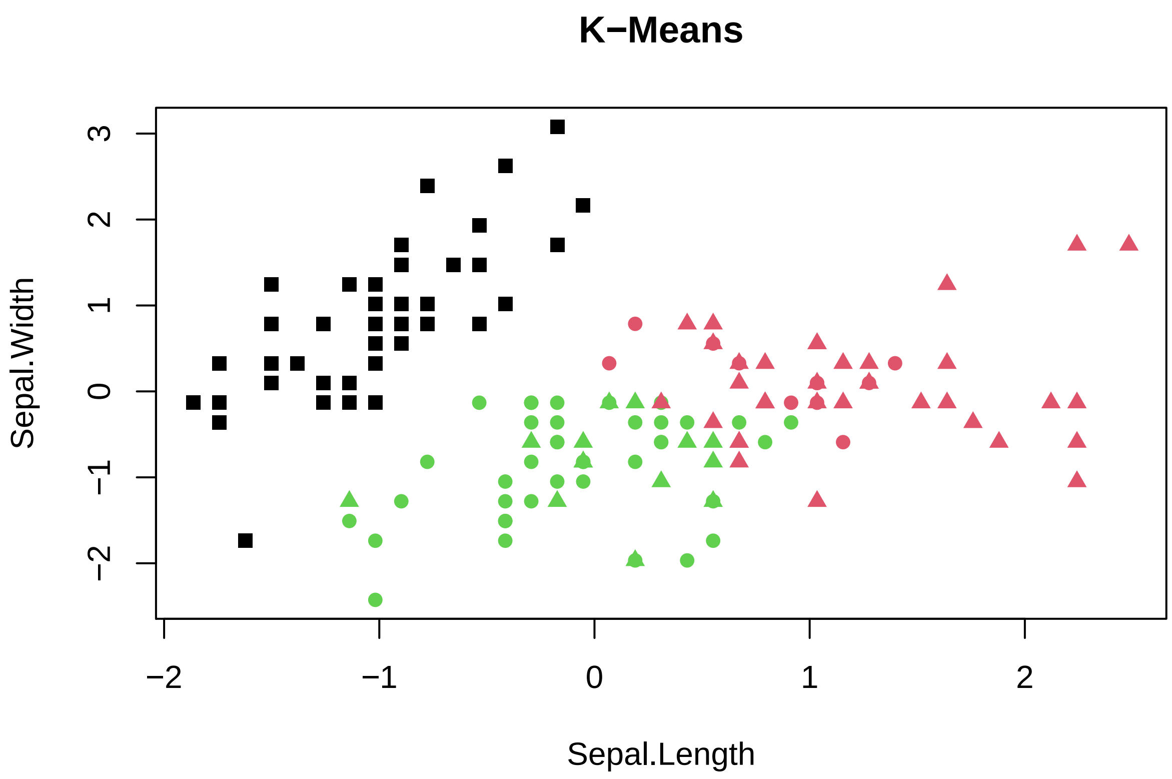
```
kmeans.out3 = kmeans(X.iris, centers = 3)
kmeans.out3
```

```
## K-means clustering with 3 clusters of sizes 50, 47, 53
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1  -1.01119138  0.85041372   -1.3006301  -1.2507035
## 2   1.13217737  0.08812645    0.9928284   1.0141287
## 3  -0.05005221 -0.88042696    0.3465767   0.2805873
##
## 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 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 3 3 3 2 3 3 3 3 3 3 3 3 2 3 3 3 3 2 3 3 3
##  [75] 3 2 2 2 3 3 3 3 3 3 3 2 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 2 2 2 2 3 2 2 2 2 2
## [112] 2 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 2 2 2 2 2 2 3 3 2 2 2 3 2 2 2 3 2 2 2 3 2
## [149] 2 3
##
## Within cluster sum of squares by cluster:
## [1] 47.35062 47.45019 44.08754
##  (between_SS / total_SS =  76.7 %)
##
## Available components:
##
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

And plot the clusters obtained with K-means.

```
plot(X.iris, col = kmeans.out3$cluster, pch = as.integer(L.iris) + 14, main = "K-Means")
```



Compare the predicted clusters with the original ones.

```
table(L.iris, kmeans.out3$cluster)
```

```
##
## L.iris        1  2  3
##   setosa     50  0  0
##   versicolor  0 11 39
##   virginica   0 36 14
```

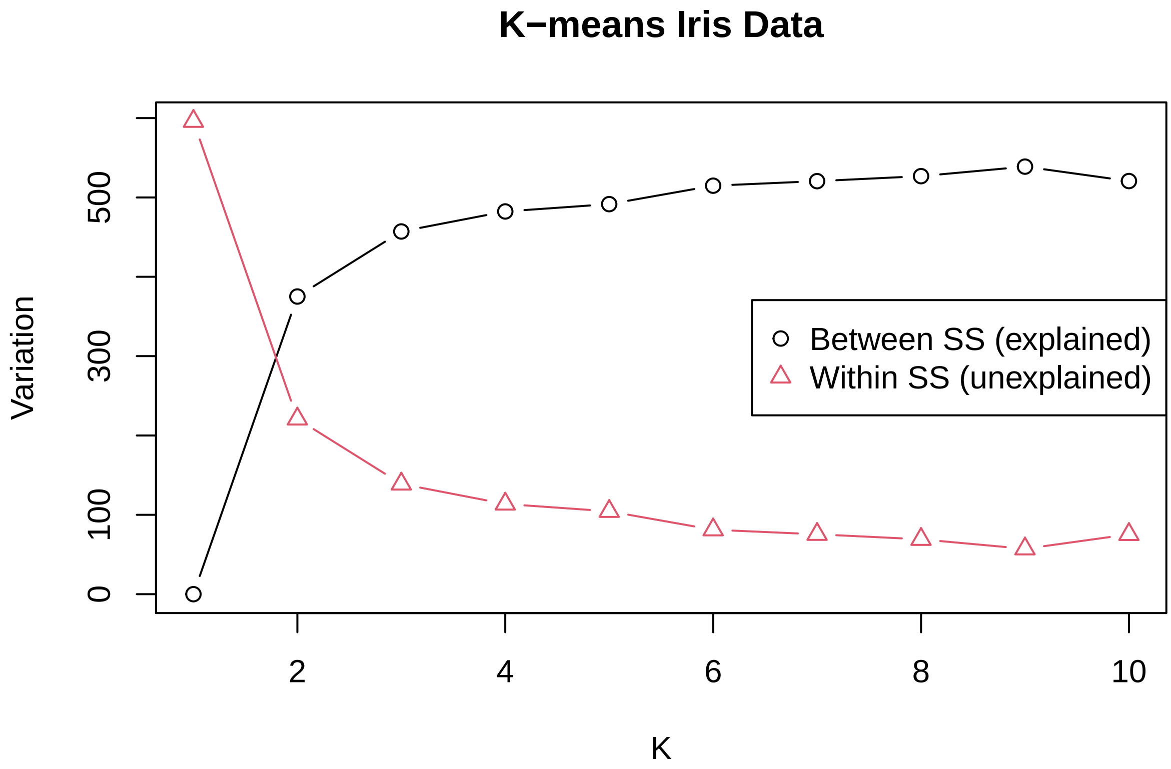
Let us apply K-means with varying values of K and check the between group and within group variation.

```
between_var = numeric(10)
within_var = numeric(10)

for (k in 1:10)
{
  kmeans.out = kmeans(X.iris, centers = k)
  between_var[k] = kmeans.out$betweenss
  within_var[k] = kmeans.out$tot.withinss
}
```

Lastly, let us plot the variations.

```
plot(1:10, between_var, ylim = c(0, max(c(max(between_var, within_var)))),
     type = "b", xlab = "K", ylab = "Variation", main = "K-means Iris Data")
points(1:10, within_var, type = "b", col = 2, pch = 2)
legend("right", c("Between SS (explained)", "Within SS (unexplained)"),
       col=c(1,2), pch=c(1,2))
```


K−means Iris Data
Variation
K
○ Between SS (explained)
△ Within SS (unexplained)