**R project**

**Data set**: Use the data from UC Irvine Machine Learning Repository.

**Brief description:**

It is an Iris Plants Database, which include 150 data sets.

There are four features:

1. sepal length in cm

2. sepal width in cm

3. petal length in cm

4. petal width in cm

Three classes:

-- Iris Setosa

-- Iris Versicolour

-- Iris Virginica

**Code:**

> Iris=read.csv("Desktop/Iris.csv")

> Iris.feature = Iris

> Iris.feature$class <- NULL

> results <- kmeans(Iris.feature, 3)

> results

**Results** (I explain the results in green words, which are not from R)**:**

K-means clustering with 3 clusters of sizes 38, 62, 50 //Three classes are identified: first cluster: 38; second cluster: 62; third cluster: 50

Cluster means:

sepal.length sepal.width petal.length petal.width

1 6.850000 3.073684 5.742105 2.071053 //means of first cluster for each feature

2 5.901613 2.748387 4.393548 1.433871 //means of second cluster for each feature

3 5.006000 3.418000 1.464000 0.244000 //means of third cluster for each feature

Clustering vector:// which cluster the data belongs to. e.g. “3” means the value place belongs to the third cluster.

[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 3 3 3

[38] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

[75] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1

[112] 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1

[149] 1 2

Within cluster sum of squares by cluster:// C = Σ(X(i)-mj)2

[1] 23.87947 39.82097 15.24040

(between\_SS / total\_SS = 88.4 %)

Available components:// The available components that can be obtained from this function

[1] "cluster" "centers" "totss" "withinss" "tot.withinss"

[6] "betweenss" "size" "iter" "ifault"

**More code:**

>results$size //want to know the size of each cluster

38 62 50 // size results

> plot(Iris[c("petal.length", "petal.width")], col = results$cluster) //plot petal length vs. petal width in each cluster



>table(Iris$class, results$cluster) //learning results (clusters) vs. data collected

1 2 3

Iris-setosa 0 0 50

Iris-versicolor 2 48 0

Iris-virginica 36 14 0