Clustering

Lecture(Practical-19)

18/09/2021

Single Linkage Clustering

- Consider the data

2, 5,9, 15,16, 18,25,33,33,45.

Suppose we are interested in using single linkage agglomerative clustering on this dataset

R code

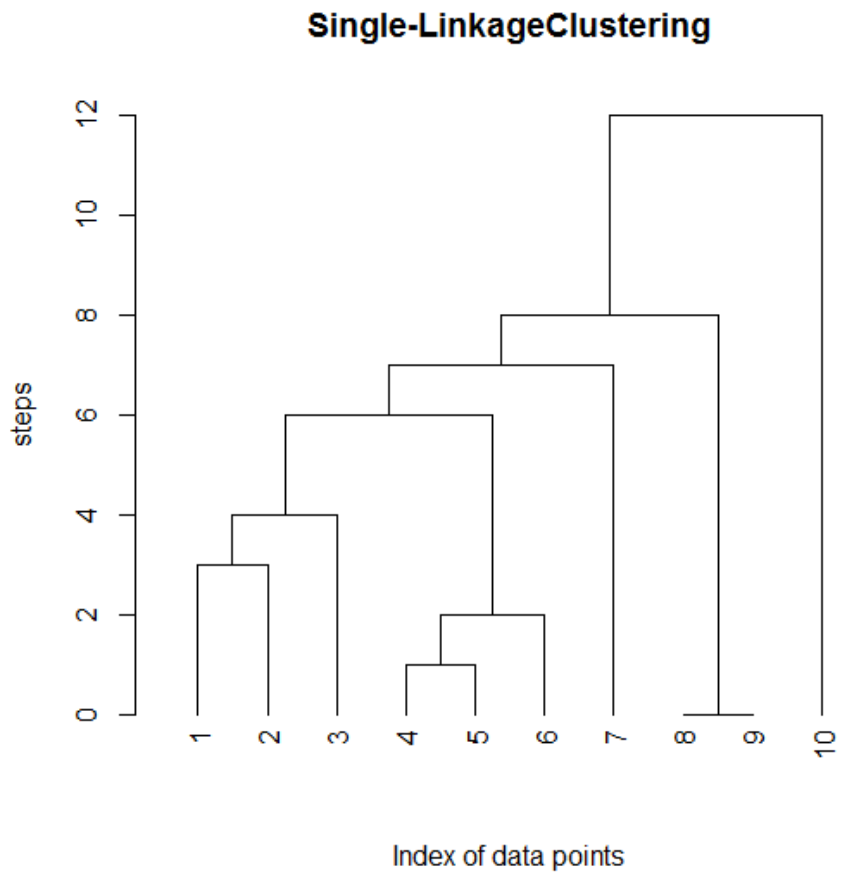
```
library(cluster)
data<-c(2,5,9,15,16,18,25,33,33,45)
#single linkage clustering
```

- #Computes agglomerative hierarchical clustering of the dataset. If diss= FALSE, then x is treated as a matrix of

  observations by variables.

```
agn<-agnes(data, diss=FALSE, stand=FALSE, method="single")
#Make and plot the dendrogram
dend_agn<-as.dendrogram(agn)
plot(dend_agn,xlab="Index of data points",ylab="steps", main="Single-LinkageClustering")
```

Dendrogram

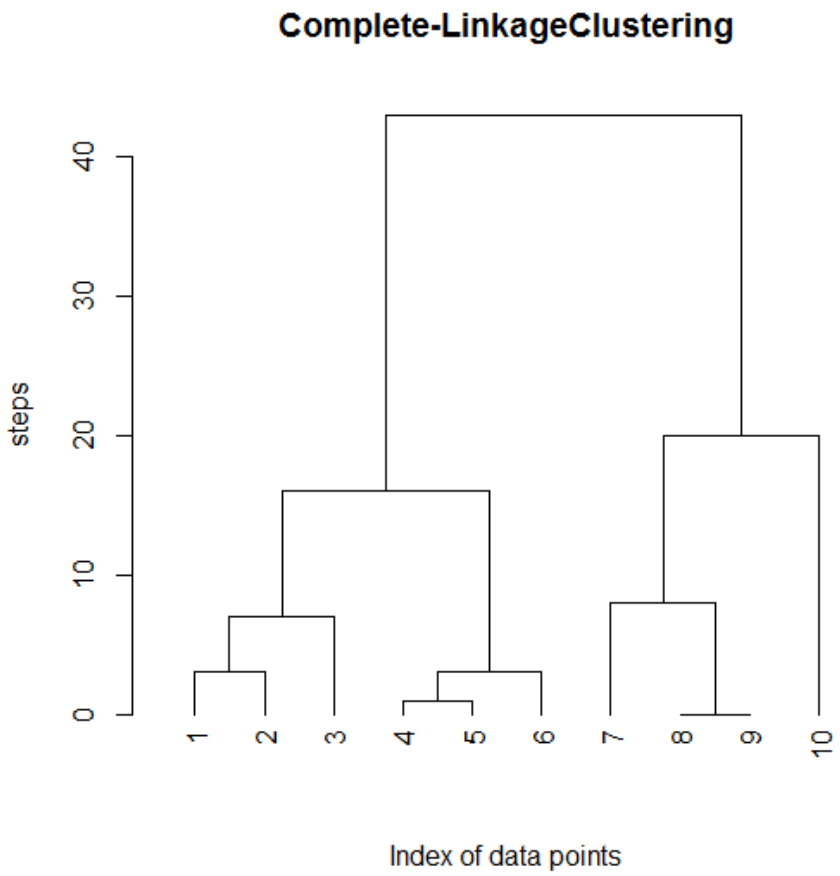

Single-LinkageClustering
steps
0
2
4
6
8
10
12
1
2
3
4
5
6
7
8
9
10
Index of data points

Complete- Linkage Clustering

- Let us examine whether using a complete linkage criterion would result in a different clustering of the same dataset.

```
#Complete Linkage Clustering

agn_complete<-agnes(data,diss=FALSE,stand=
FALSE,method="complete")

#Make and plot the dendrogram

dend_agn_complete<-as.dendrogram(agn_complete)

plot(dend_agn_complete,xlab="Index of data
points",ylab="steps", main="Complete-
LinkageClustering")
```

Dendrogram


Complete-LinkageClustering
steps
0
10
20
30
40
1
2
3
4
5
6
7
8
9
10
Index of data points

Complete- Linkage Clustering

- Now we will see whether using an average linkage criterion would result in a different clustering of the same dataset.

#Average Linkage Clustering

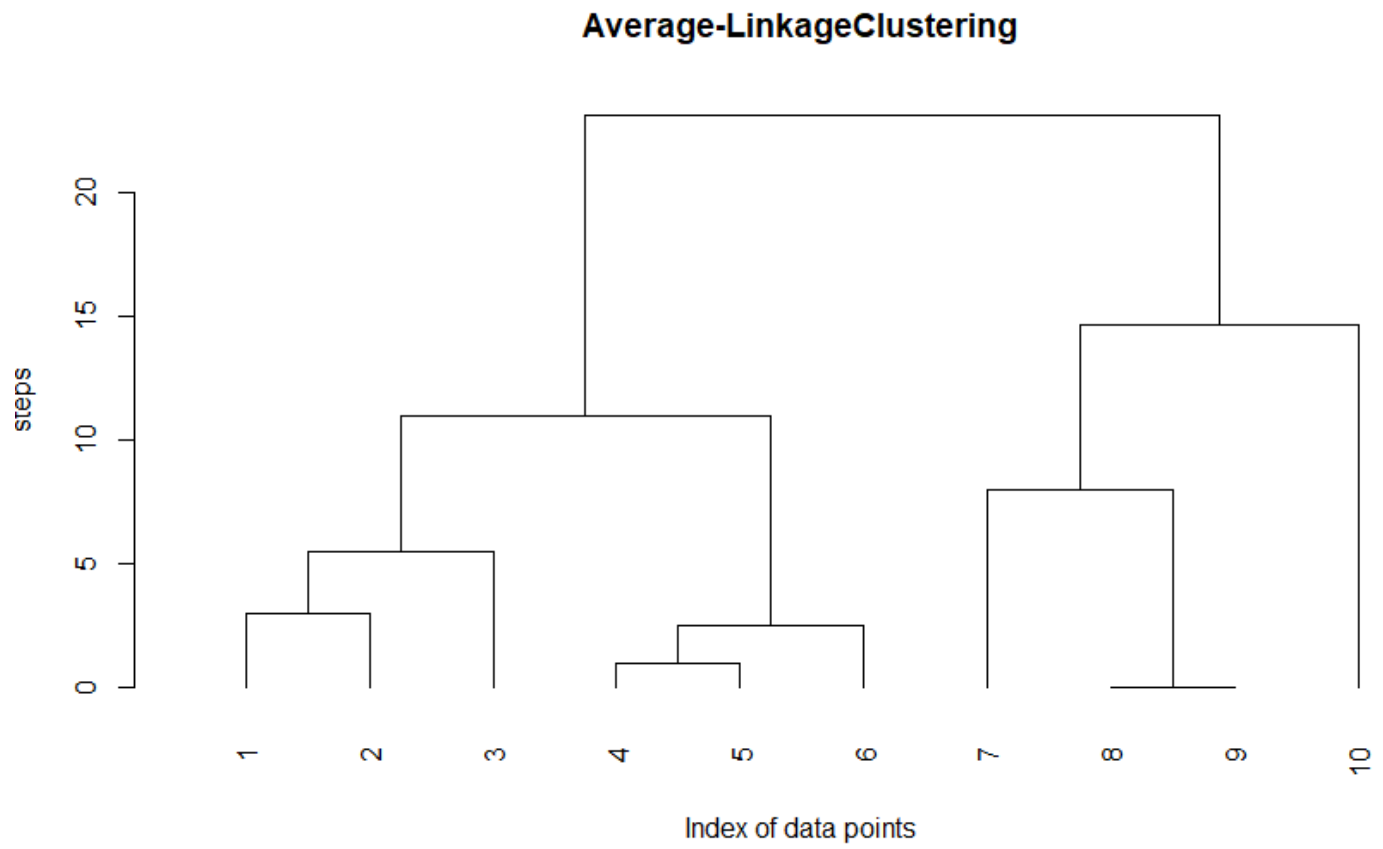
agn_complete<-agnes(data,diss=FALSE,stand= FALSE,method="average")

#Make and plot the dendrogram

dend_agn_average<-as.dendrogram(agn_average)

plot(dend_agn_complete,xlab="Index of data points",ylab="steps", main="Average-LinkageClustering")

Dendrogram


Average-LinkageClustering
steps
0
5
10
15
20
1
2
3
4
5
6
7
8
9
10
Index of data points

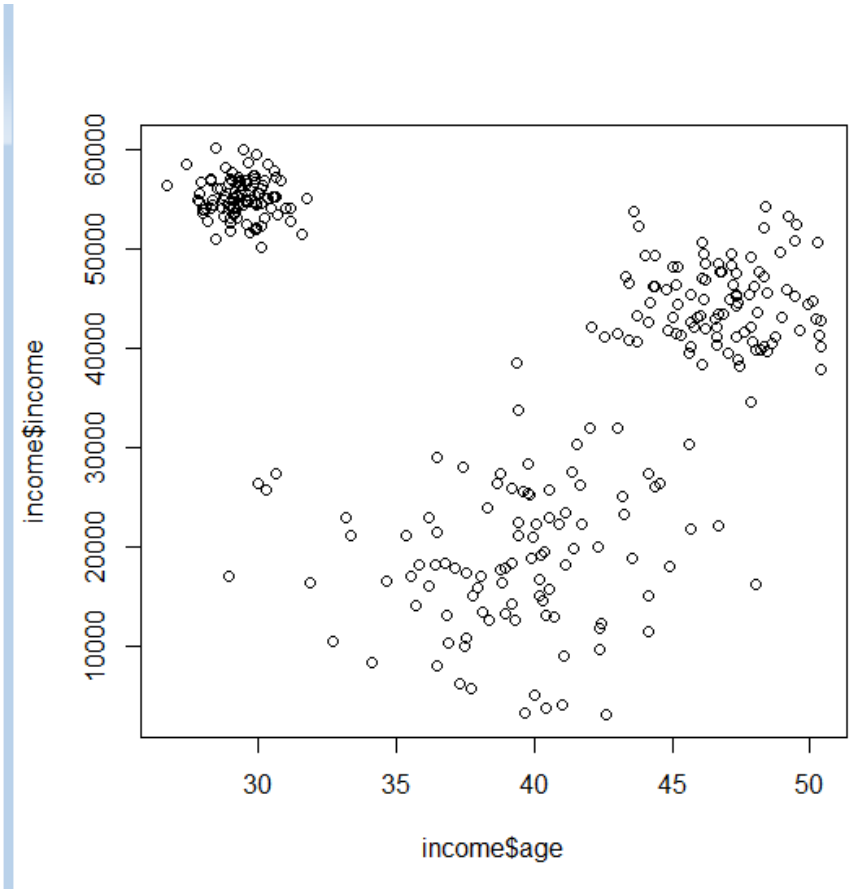
K-Means clustering: Income data

- Consider Income data.csv, which contains customer's age and income information. Analyze the customer segments that might exist and identify the key attributes of each segment using cluster analysis.
- Draw a scatterplot with age and income of customers. Could you visualize any customer segments from the plot?
- Use k-means algorithm and find out the optimum number of clusters?

R-zone Income data

- income<-read.csv(file.choose())
- income
- plot(income$age,income$income)

Scatter plot



K-means:R zone

- km<-kmeans(scale(income),centers=3)
- Km
- write.csv(km$cluster,file="income cl membership.csv")

K-means summary

```
K-means clustering with 3 clusters of sizes 100, 103, 97

Cluster means:
      income        age
1  0.9730910 -1.2061643
2  0.2963336  1.0857230
3 -1.3178500  0.0905872

Clustering vector:
  [1] 2 1 2 3 3 3 1 2 3 2 1 1 1 2 1 3 2 2 3 1 3 2 1 2 3 3 2 3 1 1 3 2 2 1 1 3 1
 [38] 3 2 1 3 1 2 1 1 3 1 3 3 1 3 1 3 3 1 2 2 3 3 2 1 1 2 3 1 3 2 1 2 1 3 2 3 3
 [75] 1 2 1 3 2 2 3 2 1 2 2 2 1 3 2 2 3 1 3 2 1 1 2 1 3 2 1 3 2 1 2 2 3 2 1 1 2
[112] 3 2 2 3 3 2 2 1 1 1 3 1 1 1 3 1 1 1 3 3 3 2 3 3 1 3 2 3 3 1 2 1 2 3 3 2 3
[149] 3 1 2 3 2 1 3 3 1 1 2 1 2 2 1 2 3 2 2 2 2 3 1 2 3 1 1 1 2 1 2 2 1 3 2 2 2
[186] 2 1 3 2 3 2 2 1 1 3 2 1 3 2 3 1 3 2 3 1 3 2 3 2 1 2 2 3 1 1 1 1 2 3 1 2 1
[223] 1 1 2 3 3 2 2 3 2 1 1 2 1 3 3 3 2 2 1 3 3 3 3 1 3 3 2 1 1 2 1 1 3 2 1 3 2
[260] 2 3 2 3 3 2 3 2 1 1 1 1 2 2 2 2 2 3 3 1 2 2 1 1 1 3 1 3 3 1 1 3 3 3 1 2 2
[297] 3 1 2 2

Within cluster sum of squares by cluster:
[1]  2.940366 15.927255 39.239252
 (between_SS / total_SS =  90.3 %)
```

Cluster plot

- clusplot(income, km$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)