

Big Data Analytics Using R

Session 9 Clustering In-Class Exercise Solution

November 25, 2014

Question: Suppose that we have 5 observations, for which we compute a similarity (distance) matrix as follows:

$$\mathbf{D}_0 = \begin{matrix} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix} \end{matrix}$$

On the basis of the similarity matrix, sketch the dendrogram that results from hierarchically clustering these 5 observations using complete linkage.

Solution:

- Step 1: Initially, there are 5 clusters: $\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$. Find the smallest distance in the distance matrix \mathbf{D}_0 : $d(C, E) = 2$. So we fuse C and E . There are 4 clusters left: $\{A\}, \{B\}, \{C, E\}, \{D\}$.
- Step 2(a): Now we construct the new distance matrix:

$$\mathbf{D}_1 = \begin{matrix} & \begin{matrix} A & B & CE & D \end{matrix} \\ \begin{matrix} A \\ B \\ CE \\ D \end{matrix} & \begin{pmatrix} 0 & & & \\ 9 & 0 & & \\ * & \dagger & 0 & \\ 6 & 5 & \# & 0 \end{pmatrix} \end{matrix}$$

The values shown above are derived from the previous matrix \mathbb{D}_0 . We are to calculate the three unknown values.

$$* : d(\{A\}, \{C, E\}) = \max\{d(A, C), d(A, E)\} = \max\{3, 11\} = 11$$

$$\dagger : d(\{B\}, \{CE\}) = \max\{d(B, C), d(B, E)\} = \max\{7, 10\} = 10$$

$$\# : d(\{D\}, \{C, E\}) = \max\{d(D, C), d(D, E)\} = \max\{9, 8\} = 9$$

Thus, the new distance matrix is

$$\mathbf{D}_1 = \begin{matrix} & \begin{matrix} A & B & CE & D \end{matrix} \\ \begin{matrix} A \\ B \\ CE \\ D \end{matrix} & \begin{pmatrix} 0 & & & \\ 9 & 0 & & \\ 11 & 10 & 0 & \\ 6 & 5 & 9 & 0 \end{pmatrix} \end{matrix}$$

- Step 2(b): Find the smallest distance in the distance matrix \mathbf{D}_1 : $d(B, D) = 5$. So we fuse B and D . There are 3 clusters left: $\{A\}, \{B, D\}, \{C, E\}$.
- Step 3(a): Now we construct the new distance matrix:

$$\mathbf{D}_2 = \begin{matrix} & \begin{matrix} A & BD & CE \end{matrix} \\ \begin{matrix} A \\ BD \\ CE \end{matrix} & \begin{pmatrix} 0 & & \\ ** & 0 & \\ 11 & \dagger\dagger & 0 \end{pmatrix} \end{matrix}$$

The values shown above are derived from the previous matrix \mathbb{D}_1 . We are to calculate the two unknown values.

$$** : d(\{A\}, \{BD\}) = \max\{d(A, B), d(A, D)\} = \max\{9, 6\} = 9$$

$$\dagger\dagger : d(\{B, D\}, \{C, E\}) = \max\{d(B, \{C, E\}), d(D, \{C, E\})\} = \max\{10, 9\} = 10$$

Thus, the new distance matrix is

$$\mathbf{D}_2 = \begin{matrix} & \begin{matrix} A & BD & CE \end{matrix} \\ \begin{matrix} A \\ BD \\ CE \end{matrix} & \begin{pmatrix} 0 & & \\ 9 & 0 & \\ 11 & 10 & 0 \end{pmatrix} \end{matrix}$$

- Step 3(b): Find the smallest distance in the distance matrix \mathbf{D}_2 : $d(A, \{B, D\}) = 9$. So we fuse A and $\{B, D\}$. There are 2 clusters left: $\{A, B, D\}, \{C, E\}$.
- Step 4: There is only one way to fuse the two clusters. It remains to calculate the distance $d(\{A, B, D\}, \{C, E\})$.

$$d(\{A, B, D\}, \{C, E\}) = \max\{d(A, \{C, E\}), d(\{B, D\}, \{C, E\})\} = \max\{11, 10\} = 11$$

As a result, we can obtain the following dendrogram:

