Hierarchical Clustering

Mengta Chung, PhD

Department of Management Sciences
Tamkang University

Example 1 (One Attribute)

Salaries of 5 people are 7, 10, 20, 28, 35. How would you cluster them into 2 groups?

Objective of Clustering

The objective of cluster analysis is to assign observations to clusters:

- Observations within each group are similar to one another (homogeneous).
- Clusters stand apart from one another (heterogeneous).

Hierarchical Clustering

Hierarchical cluster analysis forms clusters iteratively by successively joining (agglomerative) or splitting (divisive) groups.

Agglomerative vs Divisive

- Agglomerative

  In agglomerative, each observation starts in its own group, and groups are successively paired until at the end every observation is in the same large group.

- Divisive

  This method starts with the entire data set in one large group and then successively splits it into smaller groups until each observation is its own group.

Agglomerative vs Divisive

- Agglomerative methods have been implemented in many standard software packages.
- Divisive methods are computationally intensive and have had limited applications in the social sciences.

Agglomerative

Single linkage and complete linkage are two algorithms of agglomerative hierarchical clustering:

- Single linkage: shortest distance between a point and a cluster
- Complete linkage: longest distance between a point and a cluster

Agglomerative (Single Linkage)

Sort observations first: 7, 10, 20, 28, 35

7 10 20 28 35

3 10 8 7

(7 10) 20 28 35

Single Linkage (iteration 2)

$$d_{20\rightarrow(7,10)} = min\,(20-7, 20-10) = 10$$

(7 10) 20 28 35

10 8 7

(7 10) 20 (28 35)

Single Linkage (iteration 3)

$$d_{20 \to (28,35)} = min\,(28 - 20, 35 - 20) = 8$$

(7 10) 20 (28 35)

10 **8**

(7 10) (20 28 35)

Agglomerative (Complete Linkage)

7		10		20		28		35
	3		10		8		7	
(7		10)		20		28		35

Complete Linkage (iteration 2)

$$d_{20 \rightarrow (7,10)} = max\,(20 - 7, 20 - 10) = 13$$

(7 10) 20 28 35

13 8 7

(7 10) 20 (28 35)

Complete Linkage (iteration 3)

$$d_{20 \to (28,35)} = max\,(28 - 20, 35 - 20) = 15$$

(7 10) 20 (28 35)

13 **15**

(7 10 20) 28 35)

Results

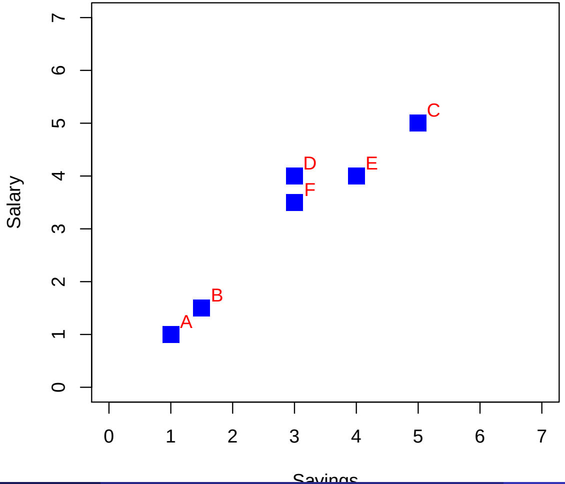
- Single linkage: (7, 10) (20, 28, 35)
- Complete linkage: (7, 10, 20) (28, 35)

Example 2 (Two Attributes)

Cluster observations into two groups using savings and salary:

#	Balance	Income
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Plot



Single Linkage (iteration 1)

Euclidean distance:

- $d_{AB} = \sqrt{(1-1.5)^2 + (1-1.5)^2} = 0.71$
- $d_{DF} = \sqrt{(3-3)^2 + (4-3.5)^2} = 0.50$

\ dist	A	B	C	D	E	F
A	0	0.71	5.66	3.61	4.24	3.20
B	0.71	0	4.95	2.92	3.54	2.50
C	5.66	4.95	0	2.24	1.41	2.50
D	3.61	2.92	2.24	0	1.00	**0.50**
E	4.24	3.54	1.41	1.00	0	1.12
F	3.20	2.50	2.50	**0.50**	1.12	0

Single Linkage (iteration 2)

$$d_{A \to (D,F)} = min\left(d_{A \to D}, d_{A \to F}\right) = min\left(3.61, 3.20\right) = 3.20$$

\ dist	A	B	C	(D, F)	E
A	0	**0.71**	5.66	3.20	4.24
B	**0.71**	0	4.95	2.50	3.54
C	5.66	4.95	0	2.24	1.41
(D, F)	3.20	2.50	2.24	0	1
E	4.24	3.54	1.41	1.00	0

Single Linkage (iteration 3)

$$d_{(A,B)\to(D,F)} = min\left(d_{A\to D}, d_{A\to F}, d_{B\to D}, d_{B\to F}\right) = 2.50$$

\ dist	(A, B)	C	(D, F)	E
(A, B)	0	4.95	2.50	3.54
C	4.95	0	2.24	1.41
(D, F)	2.50	2.24	0	**1.00**
E	3.54	1.41	**1.00**	0

Single Linkage (iteration 4)

\ dist	(A, B)	C	[(D, F), E]
(A, B)	0	4.95	2.50
C	4.95	0	**1.41**
[(D, F), E]	2.50	**1.41**	0

Results from Single Linkage

\ dist	(A, B)	{[(D, F), E], C}
(A, B)	0	2.50
{[(D, F), E], C}	2.50	0

Two clusters are: (A, B) and (D, F, E, C)

Lab

Analyze HC.csv using hierarchical clustering in Python