Cluster Analysis

With an Application to NBA data

Patrick Simpson

Xavier University

February 2, 2020

Outline

- What cluster analysis is
- A short example
- An application to NBA data

Background

- Clustering is ubiquitous in everyday life
- Cluster analysis methods attempt to group objects based on quantitative data
- Early uses

Types of Cluster Analysis

- Centroid Clustering
- Density Clustering
- Distribution Clustering
- Connectivity Clustering

Hierarchical Clustering

- Two different approaches
    - Agglomerative
    - Divisive

Agglomerative Hierarchical Clustering Algorithm

- Begin with an NxN proximity matrix
- Merge the most similar clusters N-1 times until there is only one cluster remaining
- After each interation, the proximity matrix is updated with N-1 rows and columns

Creating the Proximity Matrix

There are different measures of distance between two objects; in this model Euclidean Distance is used.

Definition

Let x and y be two points. The Euclidean Distance function between x and y can be expressed as: d(x,y) $= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

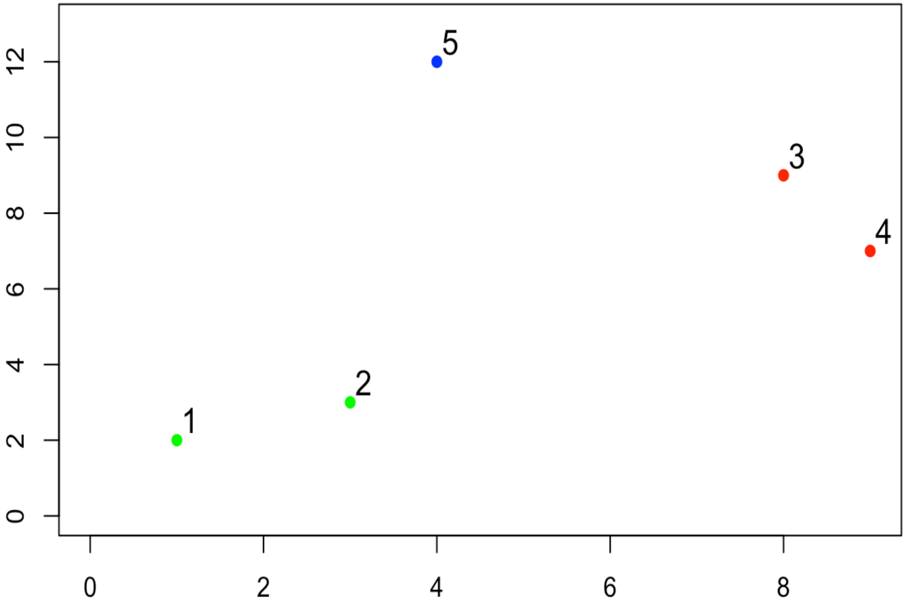
Creating the Proximity Matrix (cont.)

- Single Linkage
- Average Linkage
- Complete Linkage

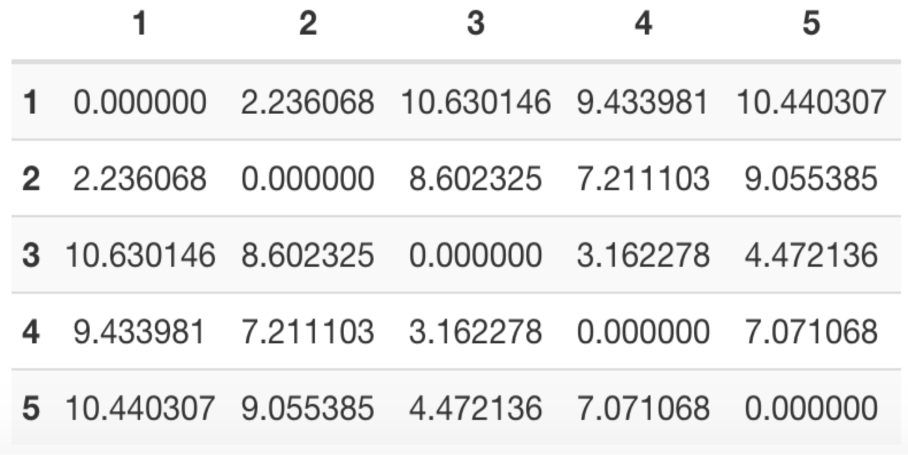
Definition

Let G and H represent two clusters. The dissimilarity d(G,H) between G and H is computed from the set of pairwise observation dissimilarities ij where one member of the pair i is in G and the other j is in H. The dissimilarity of G and H with complete linkage is computed as follow:

$$d_{CL}(G,H) = \max_{i \in G, j \in H} d_{ij}$$

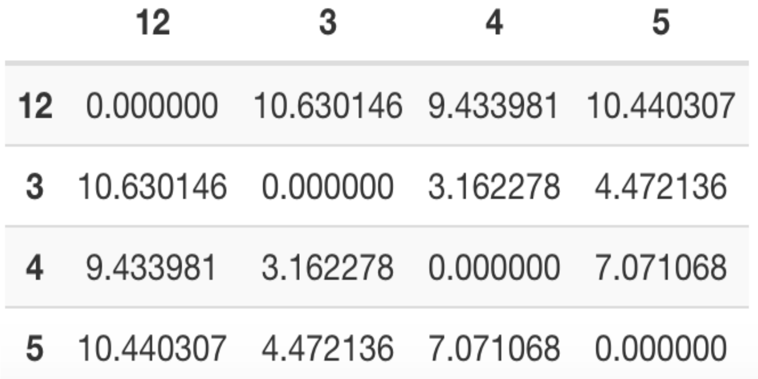

5
3
4
2
1
0
2
4
6
8
10
12
0
2
4
6
8

Proximity Matrix



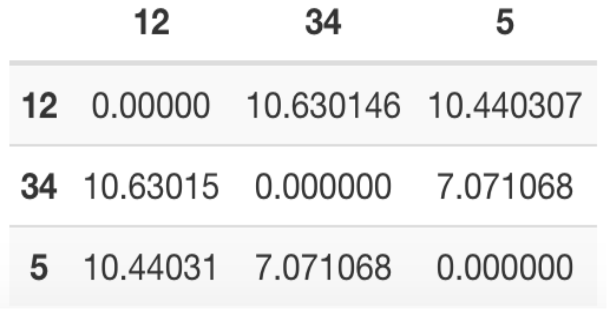
	1	2	3	4	5
1	0.000000	2.236068	10.630146	9.433981	10.440307
2	2.236068	0.000000	8.602325	7.211103	9.055385
3	10.630146	8.602325	0.000000	3.162278	4.472136
4	9.433981	7.211103	3.162278	0.000000	7.071068
5	10.440307	9.055385	4.472136	7.071068	0.000000

Proximity Matrix 2



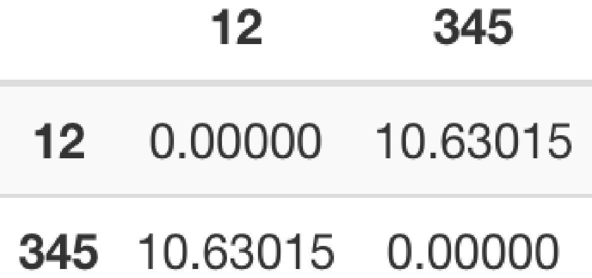
	12	3	4	5
12	0.000000	10.630146	9.433981	10.440307
3	10.630146	0.000000	3.162278	4.472136
4	9.433981	3.162278	0.000000	7.071068
5	10.440307	4.472136	7.071068	0.000000

Proximity Matrix 3



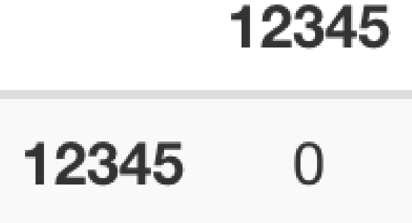
	12	34	5
12	0.00000	10.630146	10.440307
34	10.63015	0.000000	7.071068
5	10.44031	7.071068	0.000000

Proximity Matrix 4



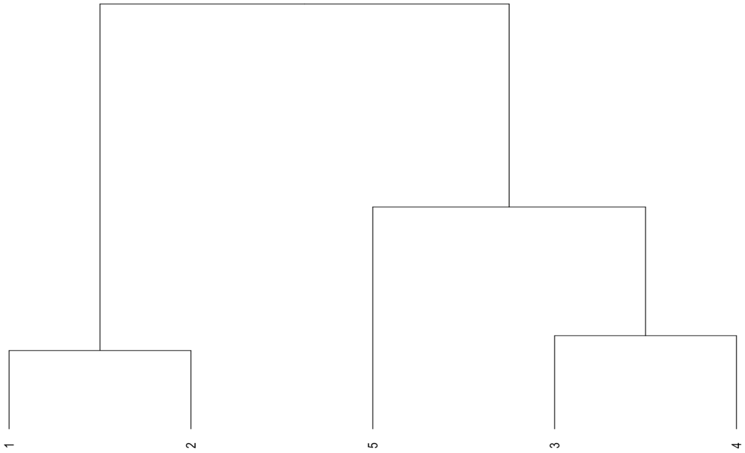
	12	345
12	0.00000	10.63015
345	10.63015	0.00000

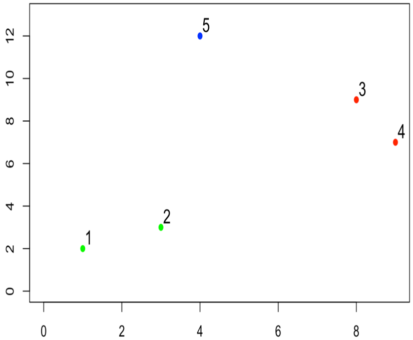
Proximity Matrix 5



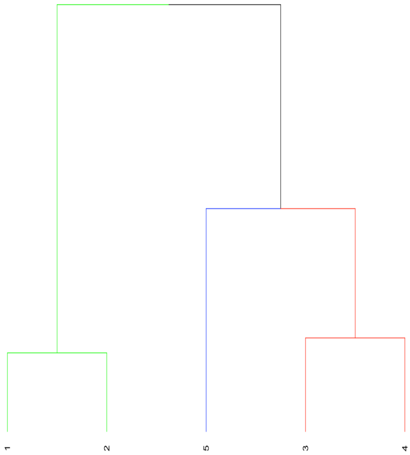
	12345
12345	0

Dendogram





Graph



Dendogram