Recap on Partitioning Clustering and Exploratory Analysis Example

Lecture 20

Dr. Emmanuel Papadakis

University of HUDDERSFIELD
Inspiring global professionals

HEA Global Teaching Excellence Award

Race Equality Charter Working Towards

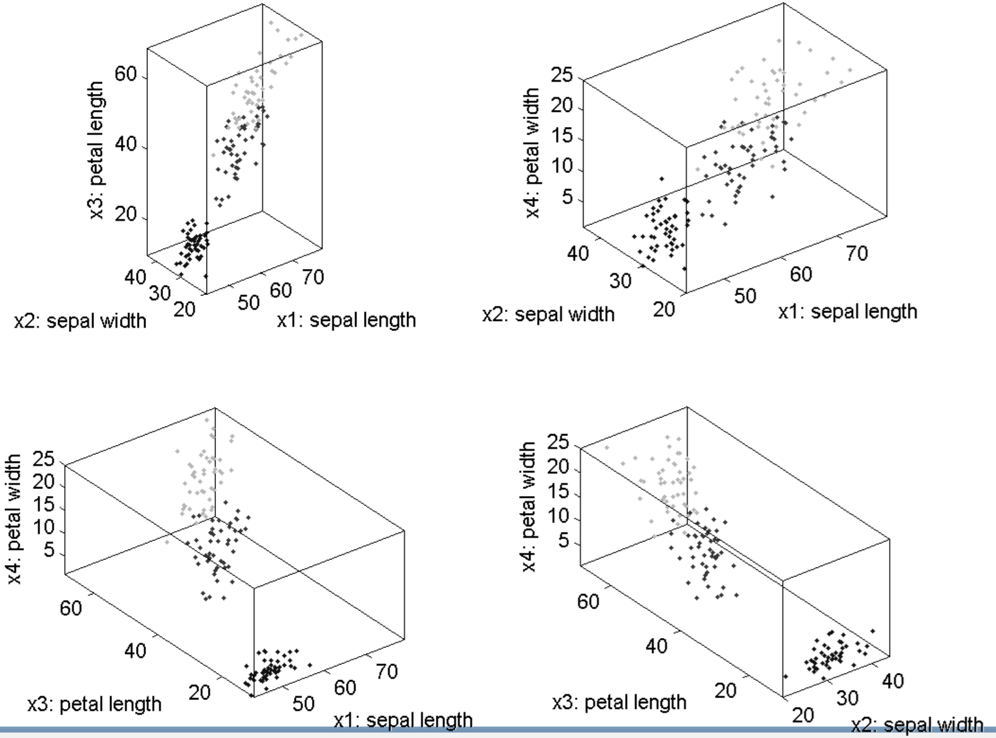
Example: Iris dataset

Example: Iris dataset

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa

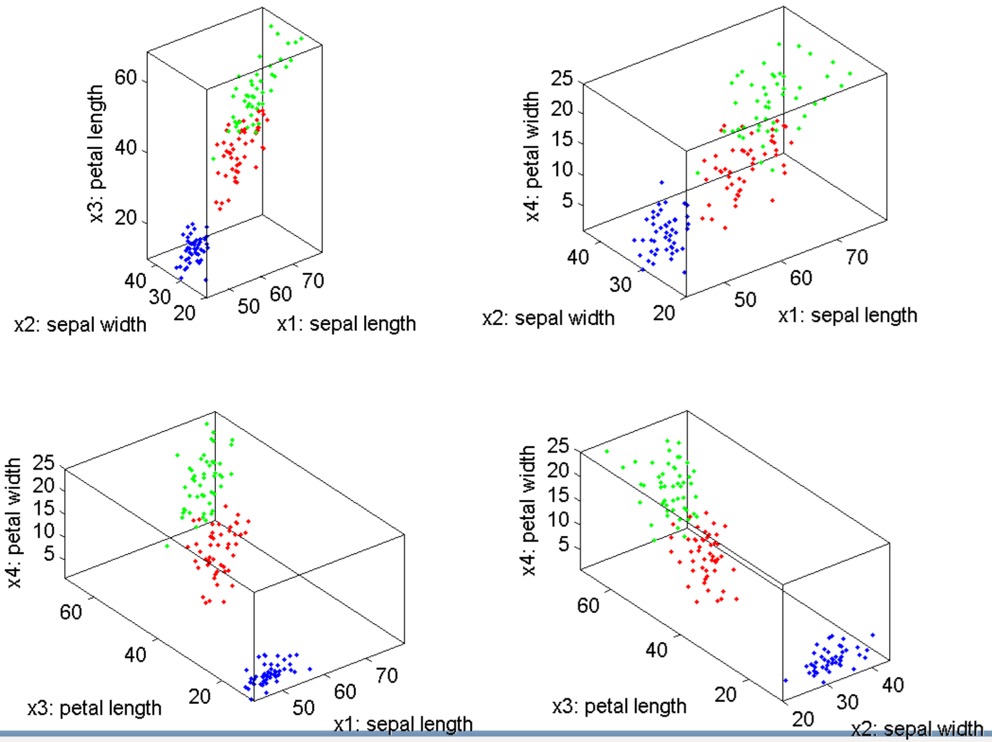
Example: Iris dataset

University of HUDDERSFIELD
Inspiring global professionals



HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

Example: Iris dataset



What is clustering?

- **Clustering**: the process of grouping a set of objects into classes of similar objects

- Most common form of ***unsupervised learning***
  - Unsupervised learning = learning from raw data
  - …as opposed to supervised data where a classification of examples is given

Clustering considerations

- **Clustering**: the process of grouping a set of objects into classes of similar objects

- What does it mean for objects to be similar? How do we measure this?

- What algorithm and approach do we take?
  - Partitional
  - Hierarchical

k–means algorithm(s)

University of HUDDERSFIELD
Inspiring global professionals

- Terminology: **centroid** = a point that is considered to be the center of a cluster
- Start by picking *k*, the number of clusters (centroids)
- Initialise clusters by picking one point per cluster (seeds)
  - E.g., pick data points at random
  - Could also generate these randomly

HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

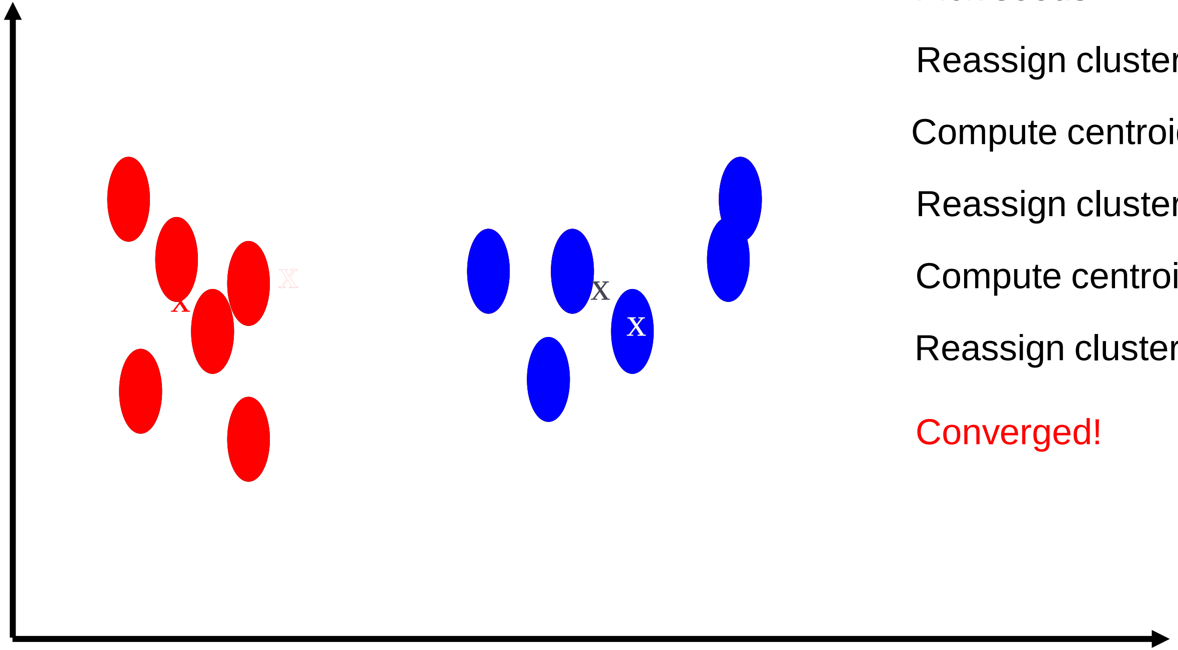
Populating clusters

Iterate until converged

1. Compute **distance** from all data points to all *k* centroids
2. For each **data point**, assign it to the cluster whose current centroid it is nearest
3. For each **centroid**, compute the average (mean) of all points assigned to it
4. Replace the *k* centroids with the new averages

By converged, we mean that a new iteration will not change the arrangement of points into clusters.

k-means example ($k = 2$)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

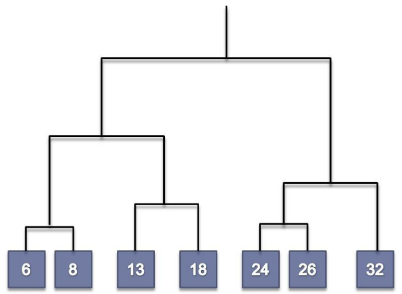
Compute centroids

Reassign clusters

Converged!

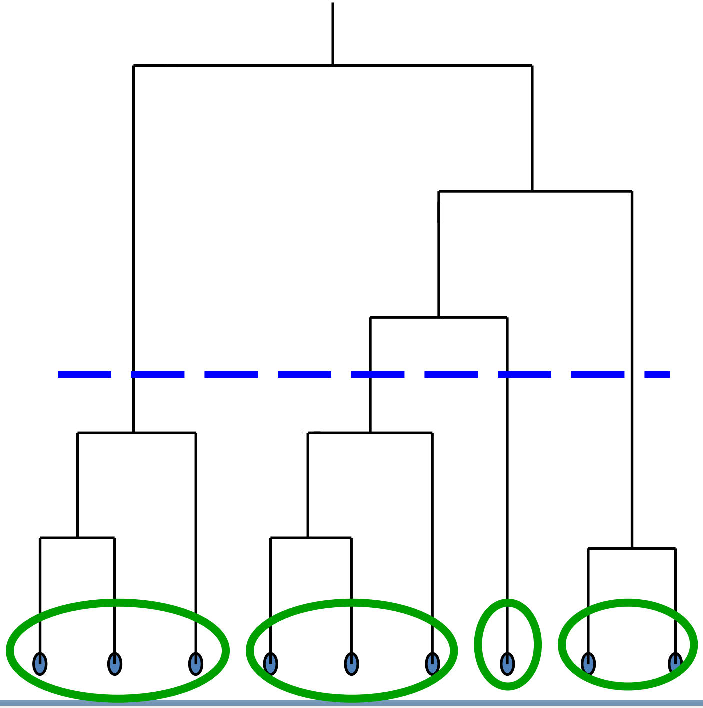
Hierarchical Agglomerative Clustering

- Assumes a similarity function for determining the similarity of two data points
  - = distance function in a n-dimensional space
- Starts with all points in separate clusters
  - Then repeatedly joins the clusters that are most similar until there is only one cluster
- The history of merging forms a binary tree or hierarchy



Hierarchical Agglomerative Clustering

- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster



HEA Global Teaching Excellence Award
Race Equality Charter Working Towards
RACE CHARTER

Hierarchical Agglomerative Clustering

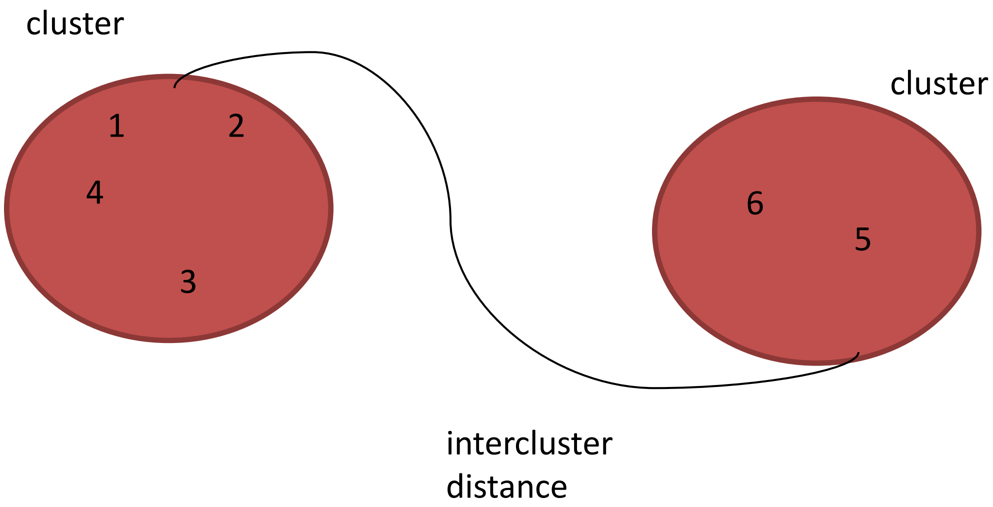
- Basic algorithm is straightforward
  - Compute the distance matrix (= distance between any 2 data points)
  - Let each data point be a cluster
  - Repeat
    - Merge the two (or more) closest clusters
    - Update the distance matrix
  - Until only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

Hierarchical clustering

- Two important questions:
  - How do you determine the “nearness” of clusters?
  - How do you represent a cluster of more than one point?

Example



University of HUDDERSFIELD
Inspiring global professionals
HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

Closest pair of clusters

Many variants to defining closest pair of clusters

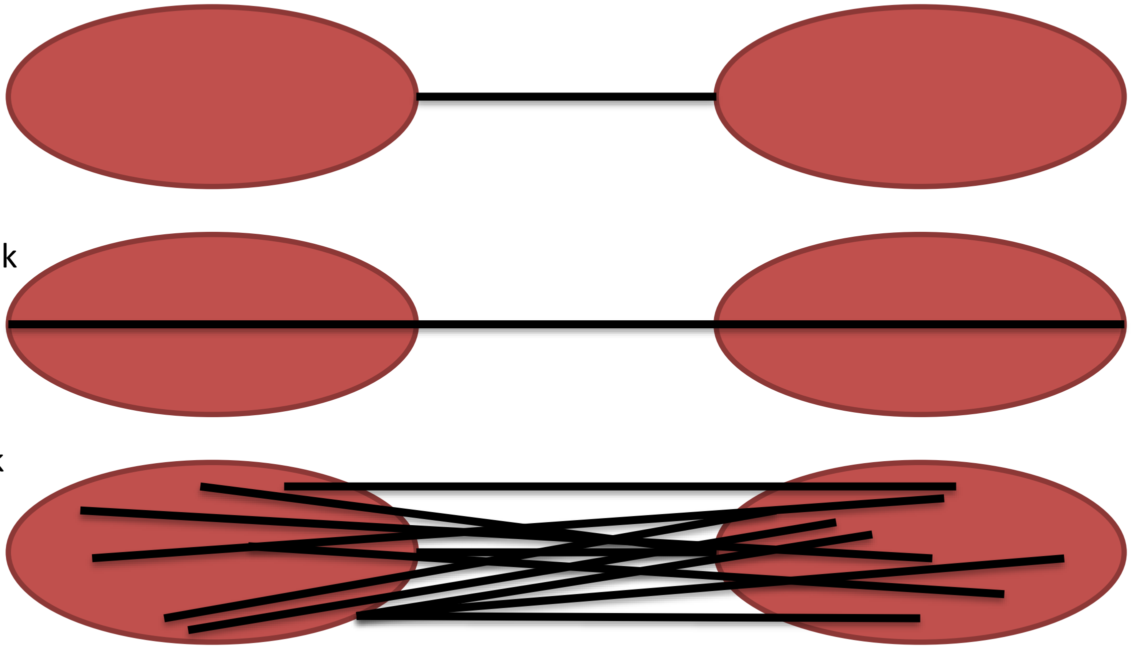
- **Single-link**
  - Distance of the *“closest” points*
- **Complete-link**
  - Distance of the “*furthest*” points
- **Centroid**
  - Distance of the centroids (centers of gravity)
- **Average-link**
  - Average distance between pairs of elements

Examples

Single-link

Complete-link

Average-link



HEA Global Teaching Excellence Award
Race Equality Charter Working Towards
RACE CHARTER

Exercise

- Given the following 1D data: {6, 8, 18, 26, 13, 32, 24}, perform *complete-link HAC*

  - Compute the distance matrix (= distance between any 2 points)
  - Let each data point be a cluster
  - Repeat
    - Merge the two (or more) closest clusters
    - Update the distance matrix
  - Until only a single cluster remains

Distance matrix

	6	8	18	26	13	32	24
6	0						
8	2	0					
18	12	10	0				
26	20	16	8	0			
13	7	5	5	13	0		
32	26	24	14	5	19	0	
24	18	16	6	2	11	8	0

Let each data point be a cluster
Repeat
    Merge the two (or more) closest clusters
    Update the distance matrix
Until only a single cluster remains

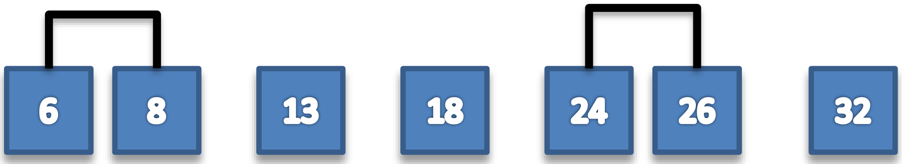
Distance matrix

University of HUDDERSFIELD
Inspiring global professionals

	6	8	18	26	13	32	24
6	0						
8	2	0					
18	12	10	0				
26	20	16	8	0			
13	7	5	5	13	0		
32	26	24	14	5	19	0	
24	18	16	6	2	11	8	0

HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

Dendogram



Distance matrix: complete-link

	6,8	18	24,26	13	32
6,8	0				
18	12	0			
24,26	20	8	0		
13	7	5	13	0	
32	26	14	8	19	0

Distance matrix: complete-link

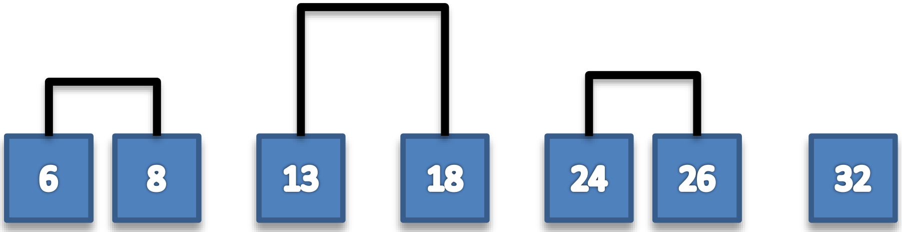
University of HUDDERSFIELD
Inspiring global professionals

	6,8	18	24,26	13	32
6,8	0				
18	12	0			
24,26	20	8	0		
13	7	5	13	0	
32	26	14	8	19	0

HEA Global Teaching Excellence Award

Race Equality Charter Working Towards

Dendogram



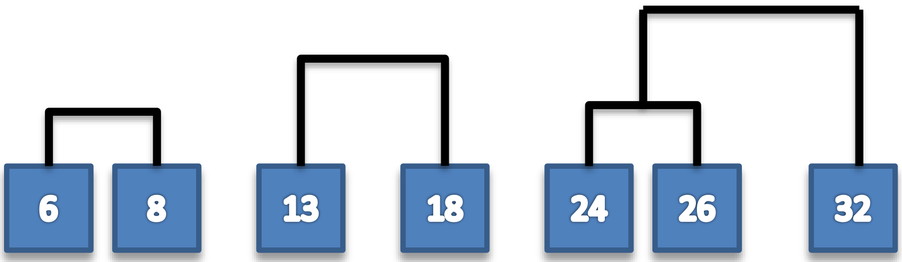
Distance matrix: complete-link

	6,8	13,18	24,26	32
6,8	0			
13,18	12	0		
24,26	20	13	0	
32	26	19	8	0

Distance matrix: complete-link

	6,8	13,18	24,26	32
6,8	0			
13,18	12	0		
24,26	20	13	0	
32	26	19	8	0

Dendogram

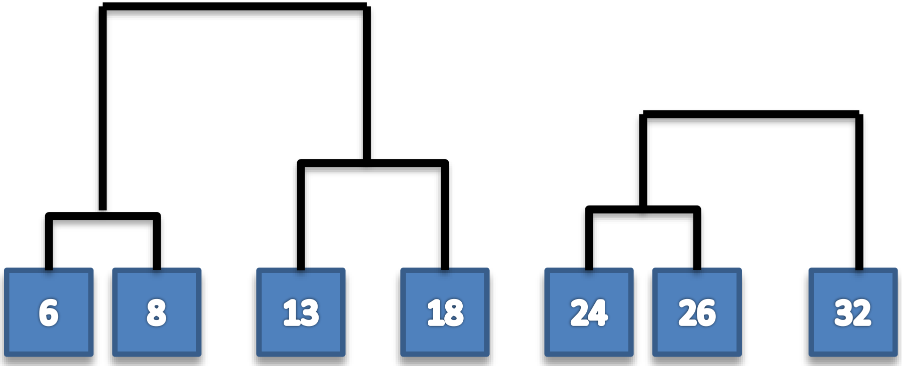


HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

Distance matrix: complete-link

	6,8	13,18	24,26,32
6,8	0		
13,18	12	0	
24,26,32	26	19	0

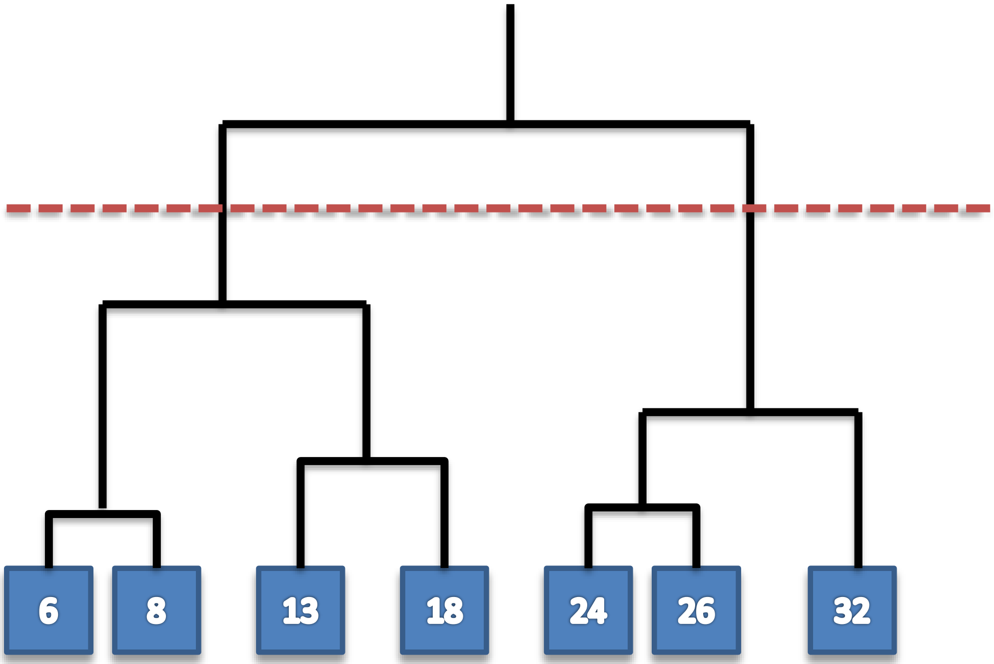
Dendogram



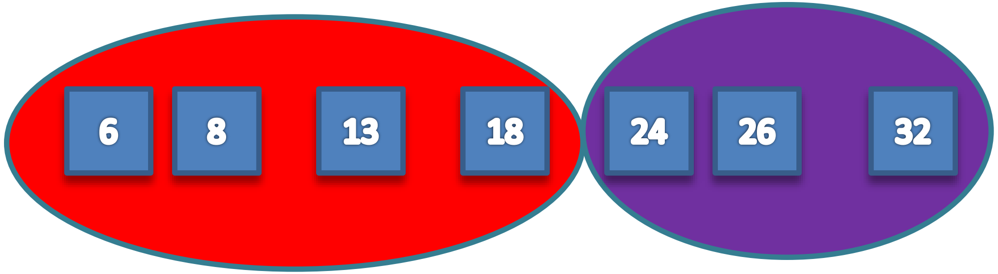
University of HUDDERSFIELD
Inspiring global professionals
HEA Global Teaching Excellence Award
Race Equality Charter Working Towards

Final dendogram

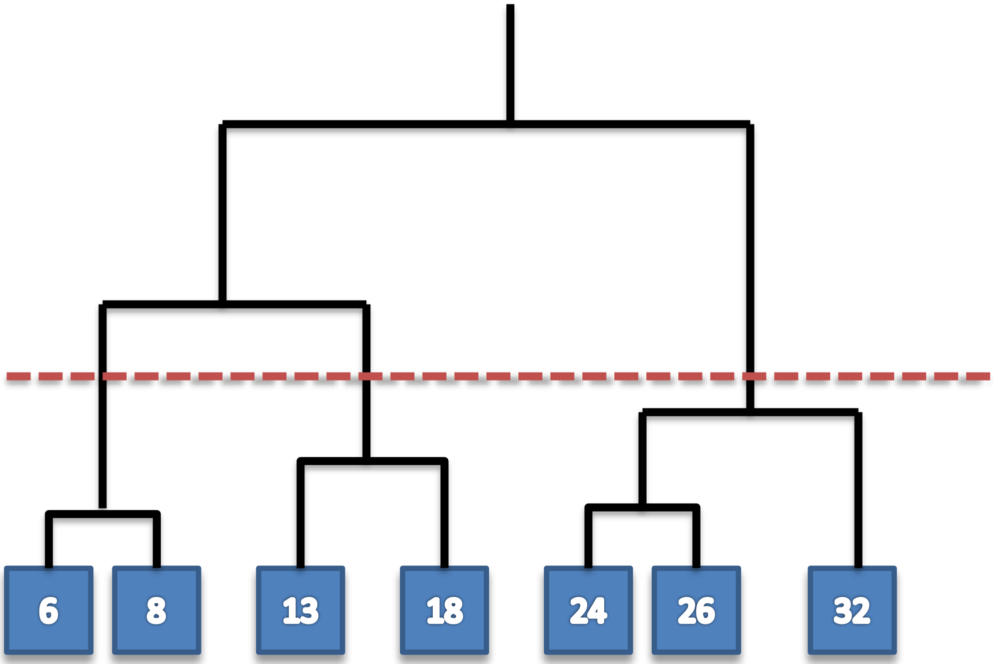
University of HUDDERSFIELD
Inspiring global professionals



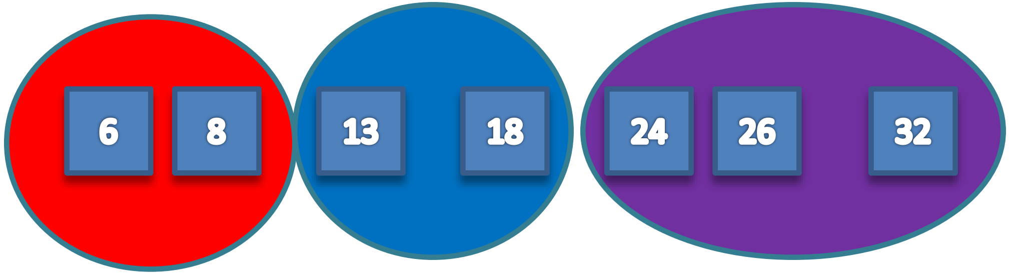
Final clustering: HAC



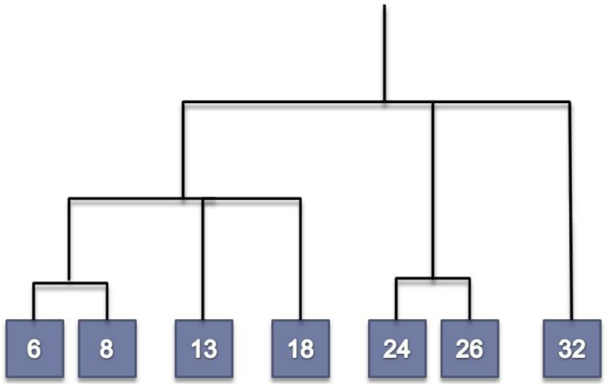
Final dendogram



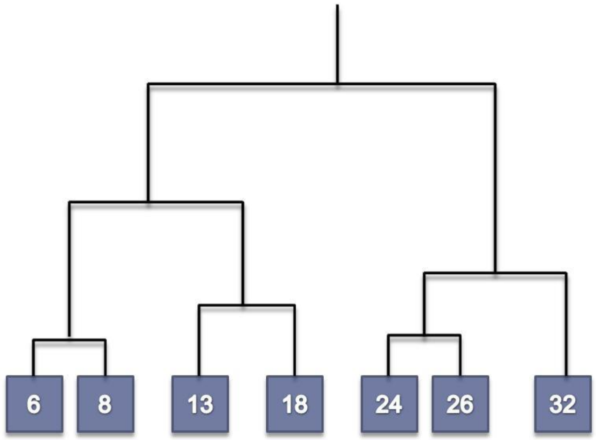
Final clustering: HAC



Compare dendograms



single-link

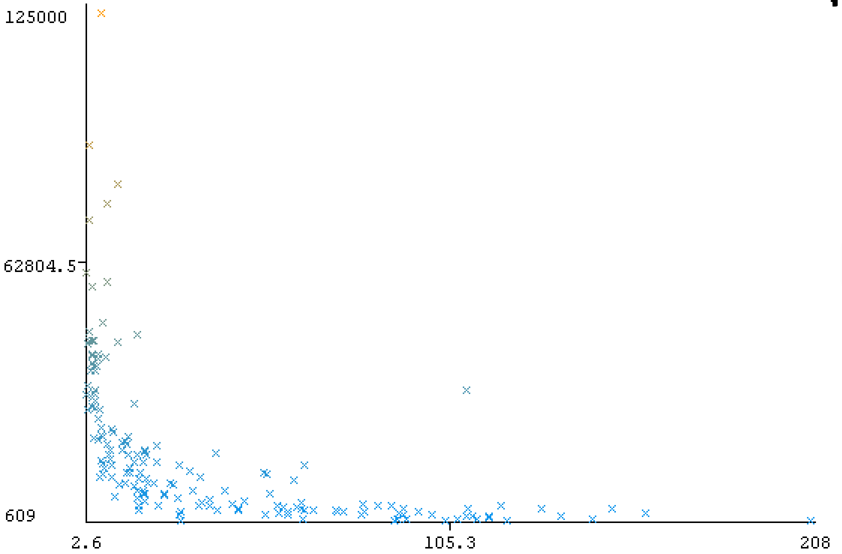


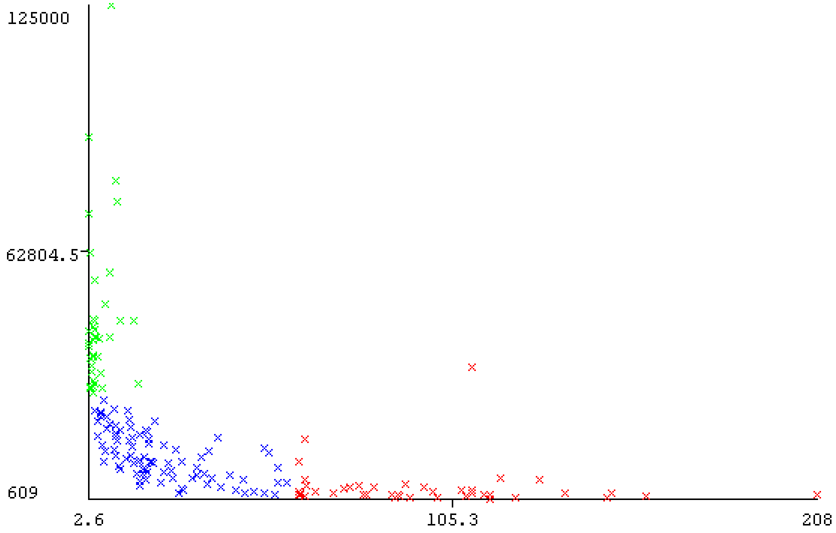
complete-link

Exploratory analysis - Weka

University of HUDDERSFIELD
Inspiring global professionals

- Real data about child mortality compared to income and health per Country





HEA Global Teaching Excellence Award
Race Equality Charter