BUSINESS INTELLIGENCE

G. MUDARE

Cluster Analysis

- In Cluster analysis ,unlike in classification, the class label of each object is not known.(unsupervised learning)
- *Clustering* is the process of grouping the data into classes or *clusters*, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.
- The objective is to divide the objects into homogeneous and distinct groups.
- Dissimilarities are assessed based on the attribute values describing the objects.
- Observations within each group are similar to one another with respect to variables or attributes of interest, and the groups themselves Stand apart from one another.

CLUSTER ANALYSIS EXPLAINED

BELGIUM CAMPUS iTversity
It's the way we're wired

- In clustering you first partition the set of data into groups based on data similarity and then assign labels to the relatively small number of groups.
- In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups.
- Classification tries to **predict** the label of (unlabeled) data.
- Clustering is grouping things into "natural" categories when **no class label** is available
- Clustering is **unsupervised** learning.
- Need to automatically decide on the grouping structure

Why clustering?

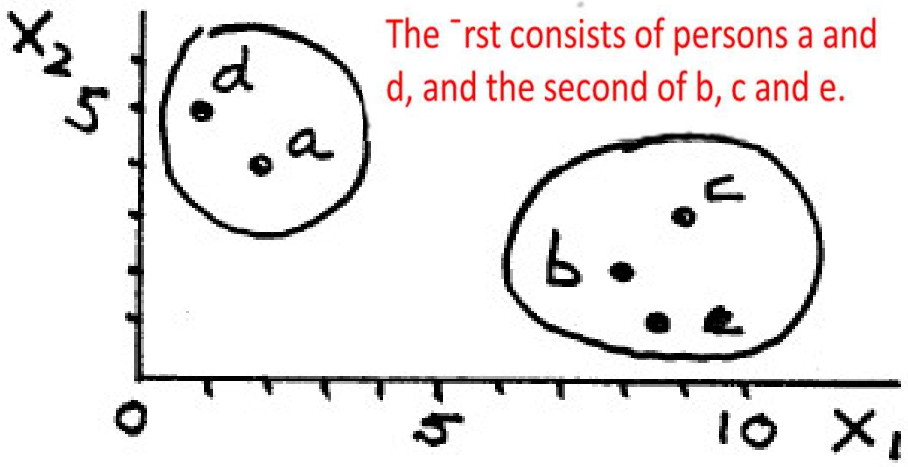
- Labeling a large set of sample patterns can be costly.
- The contents of the database may not be known.
- Clustering can be used for finding features that will later be useful for categorization.
- It may help to gain insight into the nature of the data.
- It may lead to discovery of distinct subclasses or similarities among patterns.

CLUSTER ANALYSIS EXAMPLE

BELGIUM CAMPUS iTversity
It's the way we're wired

- Cluster analysis embraces a variety of techniques, the main objective of which is to group observations or variables into homogeneous and distinct clusters.
- A simple numerical example will help explain these objectives.
- The daily expenditures on food (X1) and clothing (X2) of five persons

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1



The first consists of persons a and d, and the second of b, c and e.

Clustering
Challenges

BELGIUM CAMPUS iTversity
It's the way we're wired

- What cost function to use?
- What underlying structure to assume?
- How to measure similarity?
- How to decide on the number of clusters?
- Different answers to these questions may lead to different clustering algorithms and different clustering outcomes.
- Common objective: **generalize well**.

CLUSTERING GENERAL APPROACH FOR LEARNING

BELGIUM CAMPUS iTversity
It's the way we're wired

- For a given set of points, x_i, $i \in \{1, \ldots, N\}$ learn a class **Assignment** $y_i \in \{1, \ldots, K\}$ for each data point.
- Describe each cluster c using a set of **parameters** Θ_c
- Use an **objective (cost) function** to measure the quality of clustering.
- A function of model parameters and assignment variables.

$$f(O, Y)$$

- Clustering is the process of optimizing the objective function.

$$\underset{\theta, Y}{\operatorname{argmin}} \; f(\theta, Y)$$

Applications of Clustering

BELGIUM CAMPUS iTversity
It's the way we're wired

- Clustering has wide applications in
  - Market research, pattern recognition, data analysis, and image processing
  - Help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns
  - Identification of areas of similar land use
  - In an earth observation database
  - The identification of groups of houses in a city according to house type, value, and geographic location,
  - The identification of groups of automobile insurance policy holders with a high average claim cost.
  - Help classify documents on the web for information discovery
  - used for outlier detection (detection of credit card fraud and the monitoring of criminal activities in electronic commerce.

Requirements of clustering in data mining:

BELGIUM CAMPUS iTversity
It's the way we're wired

- **Scalability:** Clustering on a *sample* of a given large data set may lead to biased results. Highly scalable clustering algorithms are needed.
- **Ability to deal with different types of attributes**: applications may require clustering other types of data different from numeric,, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.
- **Discovery of clusters with arbitrary shape:**Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. It is important to develop algorithms that can detect clusters of arbitrary shape.
- **Minimal requirements for domain knowledge to determine input parameters:** Many clustering algorithms require users to input certain parameters in cluster analysis. Parameters are often difficult to determine, especially for data sets containing high-dimensional objects, makes the quality of clustering difficult to control

Requirements of clustering in data mining

BELGIUM CAMPUS iTversity
It's the way we're wired

- **Ability to deal with noisy data:**Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
- **Incremental clustering and insensitivity to the order of input records:** develop incremental clustering algorithms and algorithms that are insensitive to the order of input. Some clustering algorithms cannot incorporate newly inserted data
- **High dimensionality:** Finding clusters of data objects in high dimensional space is challenging,
- **Constraint-based clustering:** Real-world applications may need to perform clustering under various kinds of constraints. The challenging task is to find groups of data with good clustering behavior that satisfy specified constraints.
- **Interpretability and usability**: Users expect clustering results to be interpretable, comprehensible,and usable.

APPROACHES TO CLUSTER ANALYSIS

- There are a number of clustering methods:
- **Hierarchical methods**
  - Agglomerative methods, in which subjects start in their own separate and end, with the optimum number of clusters .
  - Divisive methods, in which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster
- **Non-hierarchical methods** (often known as k-means clustering methods)

Clustering Algorithms

BELGIUM CAMPUS iTversity
It's the way we're wired

- Clustering algorithms may be classified as listed below:

- **Exclusive Clustering>** data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster
- **Overlapping Clustering>**uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership
- **Hierarchical Clustering>**algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.
- **Probabilistic Clustering>** uses a completely probabilistic approach.

Clustering Algorithms

BELGIUM CAMPUS iTversity
It's the way we're wired

- Each of these algorithms belongs to one of the clustering types :
    1. **K-means** is an exclusive clustering algorithm,
    2. **Fuzzy C-means** is an overlapping clustering algorithm,
    3. **Hierarchical** clustering is **Hierarchical Clustering**
    4. **Mixture of Gaussian** is a probabilistic clustering algorithm

Types of Data in Cluster Analysis

- **Data matrix (or *object-by-variable structure*):**objects, such as persons, with p variables (also called measurements or attributes), such as age, height, weight, gender, and so on.
- **Dissimilarity matrix (or *object-by-object structure*):** This stores a collection of proximities that are available for all pairs of *n* objects. often represented by an n-by-n table:

TYPES OF DATA AND MEASURES OF DISTANCE

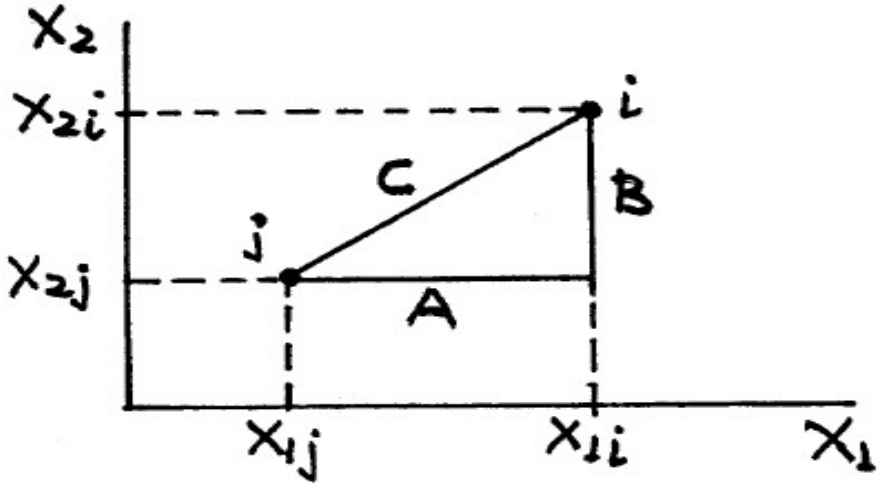
- The data used in cluster analysis can be interval, ordinal or categorical.
- Having a mixture of different types of variable will make the analysis more complicated.
- In cluster analysis we need to have some way of measuring the distance between observations and the type of measure used will depend on what type of data you have.
- A number of different measures have been proposed to measure 'distance' for binary and categorical data

TYPES OF DATA AND MEASURES OF DISTANCE

- Clustering methods require a more precise definition of “similarity" (“closeness", “proximity") of observations and clusters.
- When the grouping is based on variables, it is natural to employ the familiar concept of distance.
- For interval data the most common distance measure used is the Euclidean distance.

The Euclidean distance between the two points is the hypotenuse of the triangle ABC:

$$D(i,j) = \sqrt{A^2 + B^2} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2}.$$



TYPES OF DATA AND MEASURES OF DISTANCE

BELGIUM CAMPUS iTversity
It's the way we're wired

- An alternative measure is the squared Euclidean distance.
- The squared distance between the two points i and j is

$$D_2(i,j) = A^2 + B^2 = (X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2.$$

- Another measure is the city block distance, defined as

$$D_3(i,j) = |A| + |B| = |X_{1i} - X_{1j}| + |X_{2i} - X_{2j}|.$$

Distance you travel if the points i and j were located at opposite corners of a city block

Distance measures can be extended to more than two variables

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}.$$

• TYPES OF DATA AND MEASURES OF DISTANCE (hierarchical agglomerative clustering methods

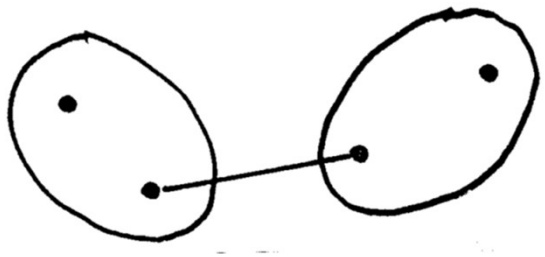
BELGIUM CAMPUS iTversity
It's the way we're wired

- Given a distance measure, a reasonable procedure for grouping n observations proceeds in the following steps:

1. Begin with as many clusters as there are observations, that is, with each observation forming a separate cluster
2. Merge that pair of observations that are nearest one another, leaving n -1 clusters for the next step.
3. Next, merge into one cluster that pair of clusters that are nearest one another, leaving n - 2 clusters for the next step.
4. Continue in this fashion, reducing the number of clusters by one at each step, until a single cluster is formed consisting of all n observations.
5. At each step, keep track of the distance at which the clusters are formed.

Euclidean distance

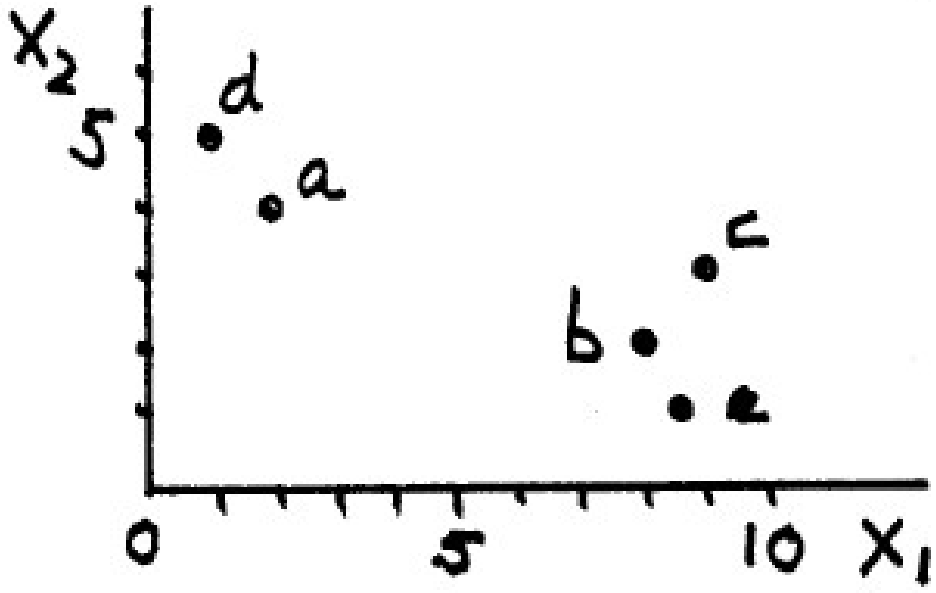
- The problem with this procedure is how to measure the distance between clusters consisting of two or more observations.
- the simplest method is to treat the distance between the two nearest observations, one from each cluster, as the distance between the two clusters.
- This is known as the nearest neighbor (or single linkage) method



Euclidean distance is the appropriate measure of proximity.

Euclidean distance

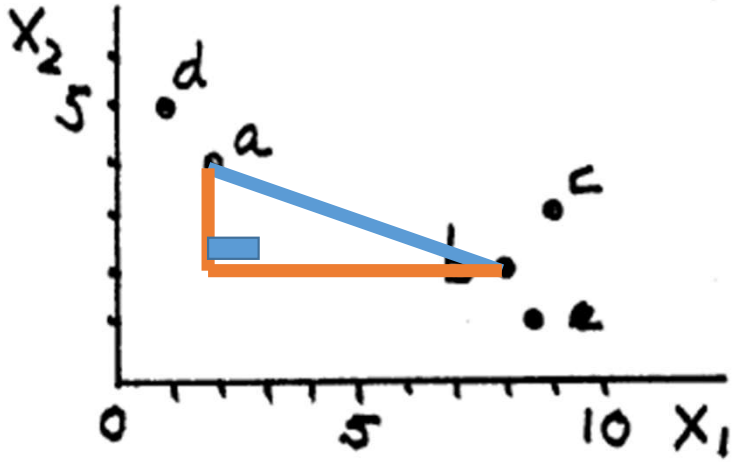
Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1



Euclidean distance

- to determine the number of clusters, consider the step(s) at which the merging distance is relatively large.

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

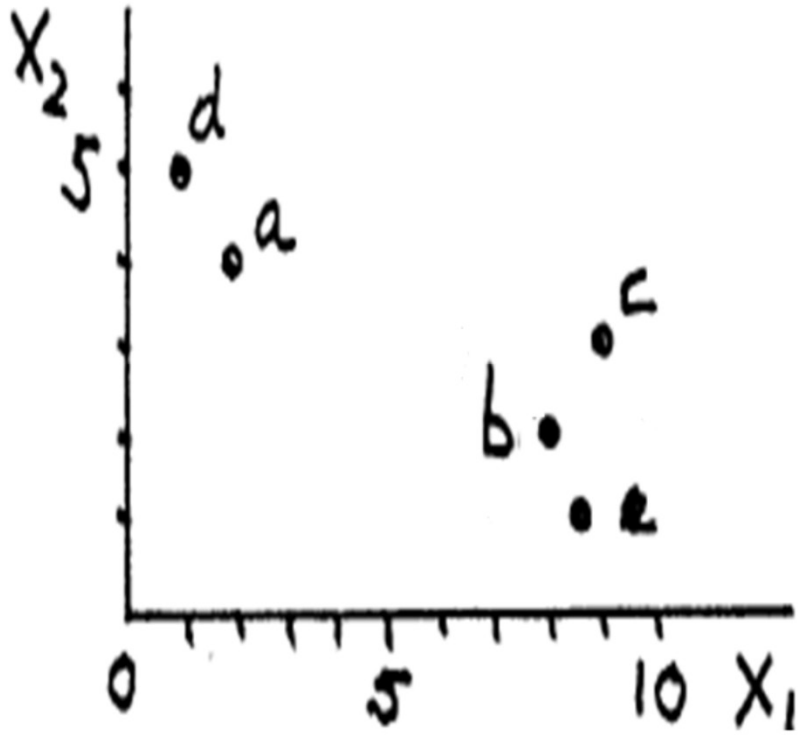


distance between a and b is

$$\sqrt{(2-8)^2+(4-2)^2} = \sqrt{36+4} = 6.325.$$

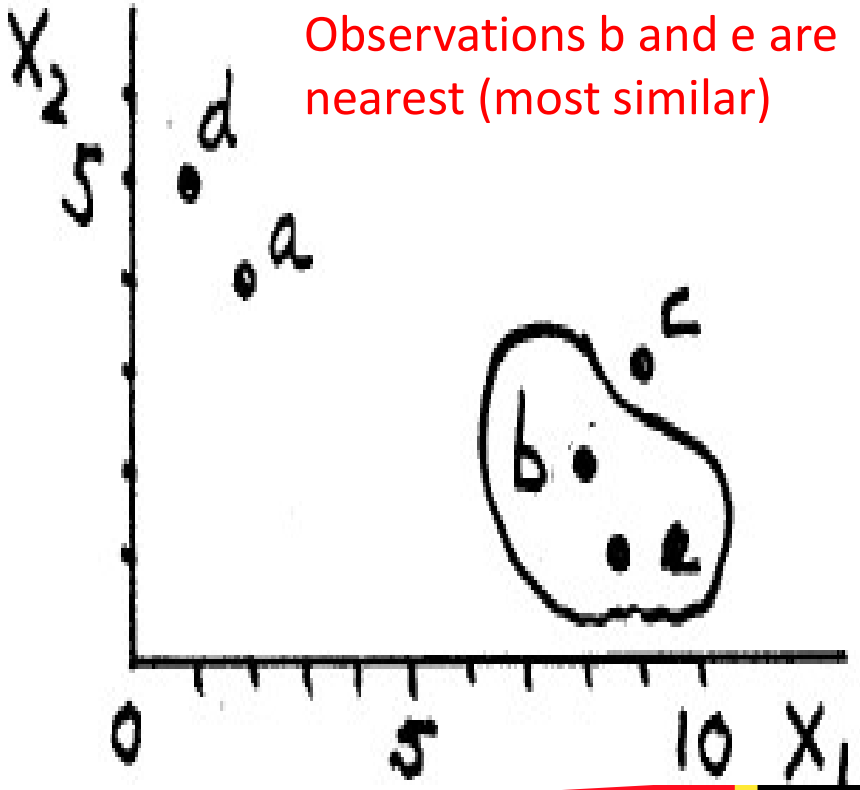
FILL IN THE STARS

Cluster	a	b	c	d	e
a	0	6.325	☆	1.414	☆
b		0	1.414	7.616	☆
c			0	☆	2.062
d				0	☆
e					0



Euclidean distance

Cluster	a	b	c	d	e
a	0	6.325	7.071	1.414	7.159
b		0	1.414	7.616	1.118
c			0	8.246	2.062
d				0	8.500
e					0



Observations b and e are nearest (most similar)

Nearest neighbor method

- Assuming the nearest neighbor method is used, the distance between
- The cluster (be) and another observation is the smaller of the distances between that observation, on the one hand, and b and e, on the other
- Two pairs of clusters are closest to one another at distance 1.414; these are (ad) and (bce).

$$D(be, a) = \min\{D(b, a), D(e, a)\} = \min\{6.325, 7.159\} = 6.325.$$

Nearest neighbor method

- select (a;d) as the new cluster

The distance between (be) and (ad) is

$$D(be, ad) = \min\{D(be, a), D(be, d)\} = \min\{6.325, 7.616\} = 6.325,$$

while that between c and (ad) is

$$D(c, ad) = \min\{D(c, a), D(c, d)\} = \min\{7.071, 8.246\} = 7.071.$$

Nearest neighbor method

, between c and (ad) is

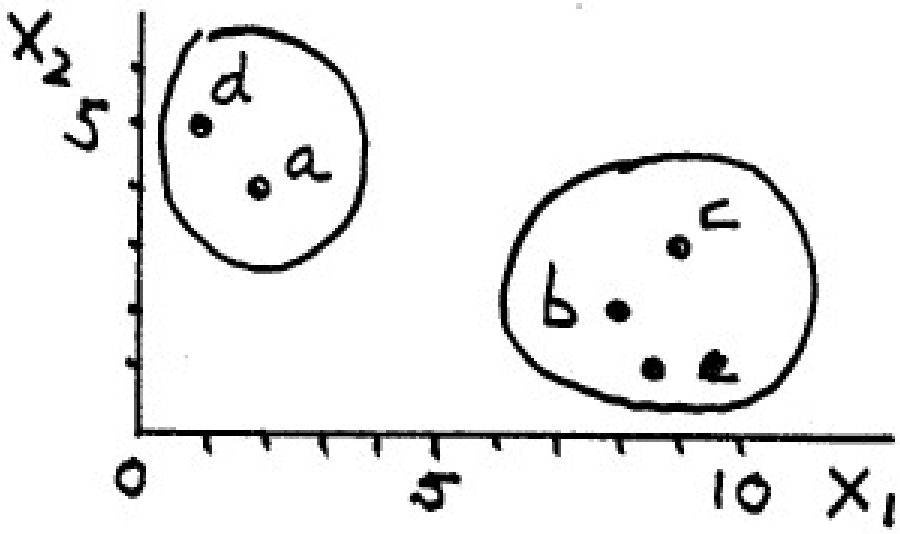
$$D(c, ad) = \min\{D(c, a), D(c, d)\} = \min\{7.071, 8.246\} = 7.071.$$

The distance between the two remaining clusters is

$$D(ad, bce) = \min\{D(ad, be), D(ad, c)\} = \min\{6.325, 7.071\} = 6.325$$

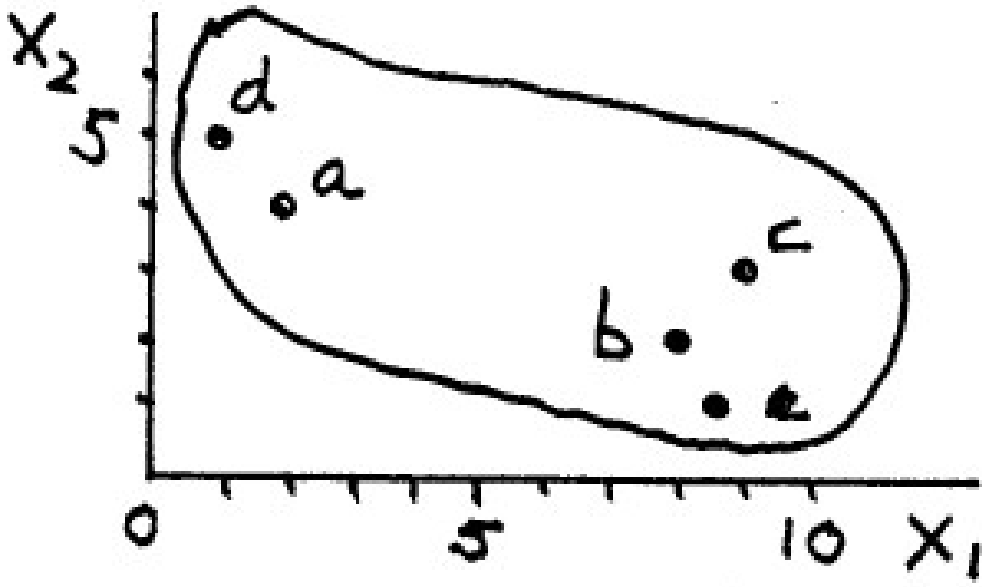
Nearest neighbor method

Cluster	(be)	(ad)	c
(be)	0	6.325	1.414
(ad)		0	7.071
c			0



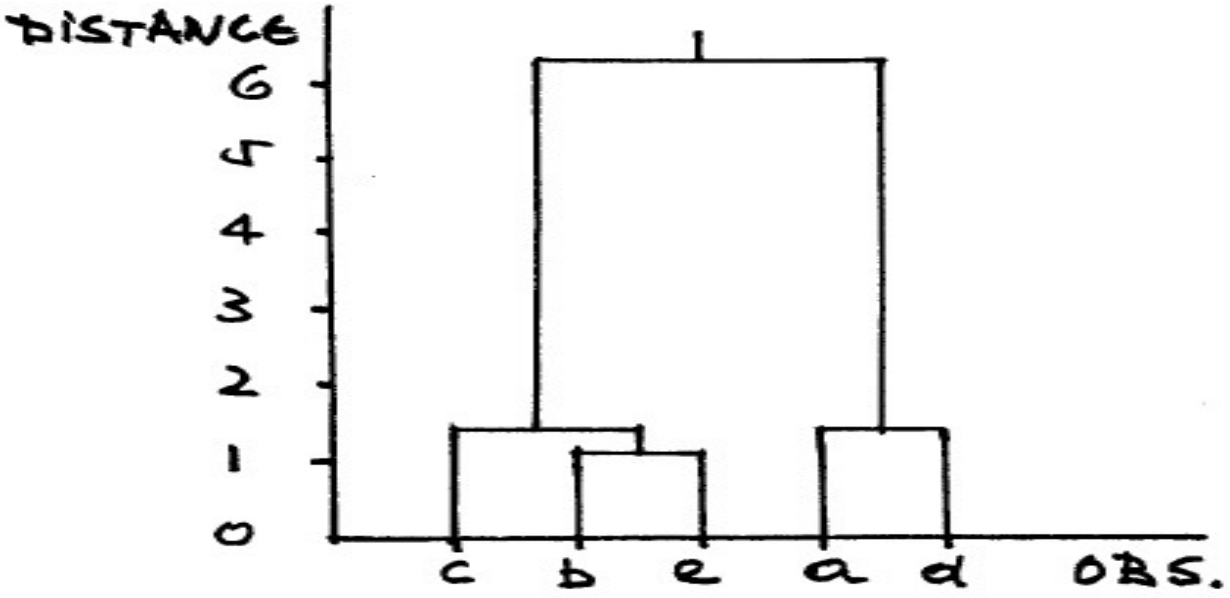
Nearest neighbor method

Cluster	(bce)	(ad)
(bce)	0	6.325
(ad)		0

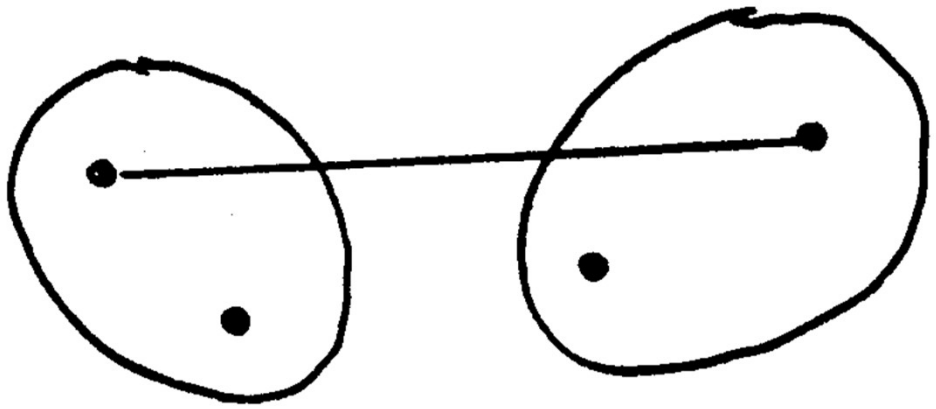


Nearest neighbor method, dendrogram

- The groupings and the distance at which these took place are also shown in the tree diagram (dendrogram)



Cluster distance, furthest neighbor method



- The nearest neighbor is not the only method for measuring the distance between clusters.
- Under the furthest neighbor (or complete linkage) method, the distance between two clusters is the distance between their two most distant members.

Cluster distance, furthest neighbor method

- The furthest neighbor method also calls for grouping b and e at step 1.
- However, the distances between (be), on the one hand, and the clusters (a), (c), and (d), on the other, are different:

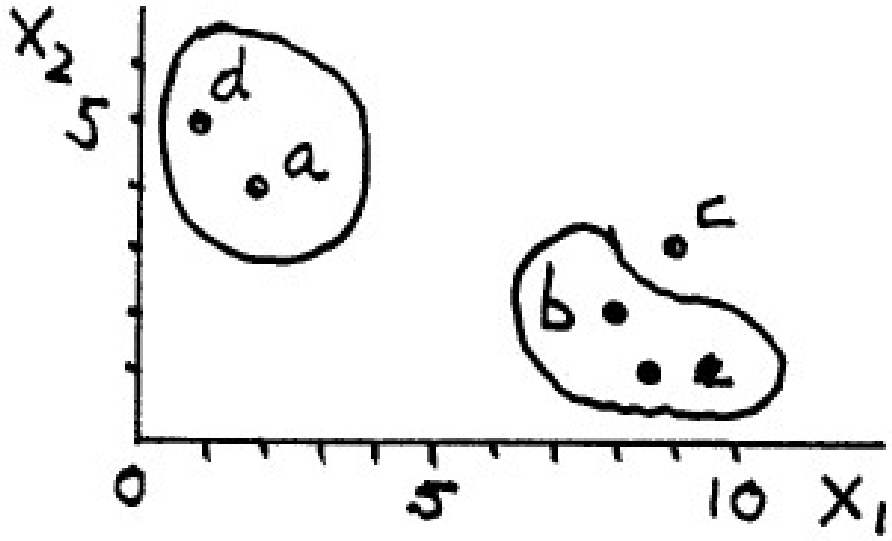
$$D(be, a) = \max\{D(b, a), D(e, a)\} = \max\{6.325, 7.159\} = 7.159$$
$$D(be, c) = \max\{D(b, c), D(e, c)\} = \max\{1.414, 2.062\} = 2.062$$
$$D(be, d) = \max\{D(b, d), D(e, d)\} = \max\{7.616, 8.500\} = 8.500$$

CLUSTER DISTANCE, FURTHEST NEIGHBOR METHOD

BELGIUM CAMPUS iTversity
It's the way we're wired

Cluster	(be)	a	c	d
(be)	0	7.159	2.062	8.500
a		0	7.071	1.414
c			0	8.246
d				0

(a)



(b)

CLUSTER DISTANCE, FURTHEST NEIGHBOR METHOD

BELGIUM CAMPUS iTversity
It's the way we're wired

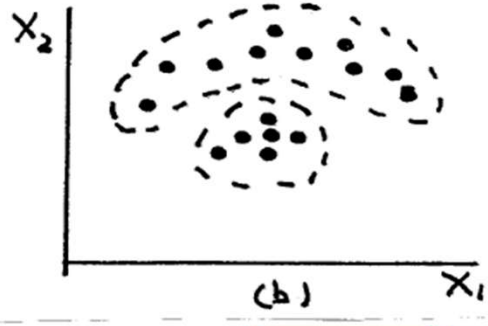
- The nearest clusters are (a) and (d), which are now grouped into the cluster (ad).
- the nearest and furthest neighbor methods produced the same results in this example.
- In other cases, however, the two methods may not agree.

CLUSTER DISTANCE, FURTHEST NEIGHBOR METHOD

BELGIUM CAMPUS iTversity
It's the way we're wired

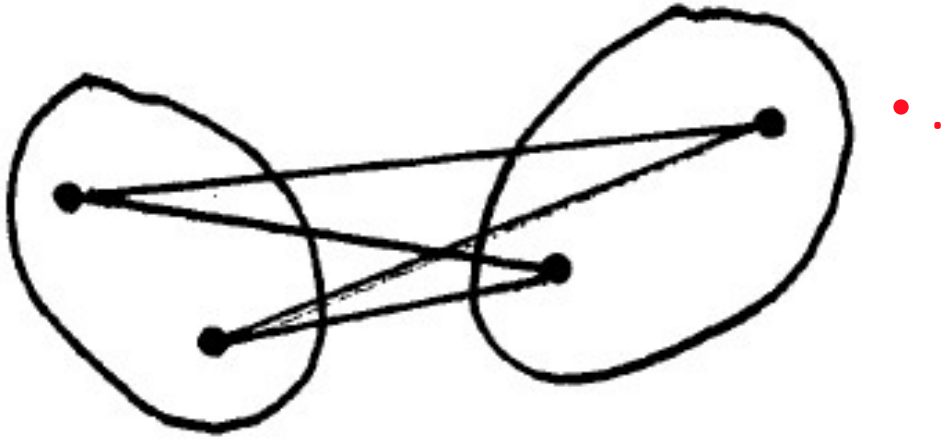
- The nearest neighbor method will probably not form the two groups percived by the naked eye.
- This is so because at some intermediate step the method will probably merge the two “closest" points joined into the same cluster, and proceed to string along the remaining points in chain-link fashion
- The furthest neighbor method, will probably identify the two clusters because it tends to resist merging clusters the elements of which vary substantially in distance from those of the other cluster.

On the other hand, the nearest neighbor method will probably succeedd in forming the two groups marked



Average Linkage

- A compromise method is average linkage, under which the distance between two clusters is the average of the distances of all pairs of observations, one observation in the pair taken from the first cluster and the other from the second cluster



Hierarchical divisive method

- This method follows the reverse procedure in that it begins with a single cluster consisting of all observations, forms next 2, 3, etc. clusters, and ends with as many clusters as there are observations

Non-hierarchical clustering method, (k-means method)

- Step 1. Specify the number of clusters and, arbitrarily or deliberately, the members of each cluster.
- Step 2. Calculate each cluster's "centroid" and the distances between each observation and centroid. If an observation is nearer the centroid of a cluster other than the one to which it currently belongs, re-assign it to the nearer cluster.
- Step 3. Repeat Step 2 until all observations are nearest the centroid of the cluster to which they belong.
- Step 4. If the number of clusters cannot be specified with confidence in advance, repeat Steps 1 to 3 with a different number of clusters and evaluate the results

(k-means method)

Person	X_1	X_2
a	2	4
b	8	2
c	9	3
d	1	5
e	8.5	1

Suppose two clusters are to be formed
First assign a, b and d to Cluster 1,
Next assign c and e to Cluster 2.
Calculate cluster centroids

The cluster centroid is the point with coordinates equal to the average values of the variables for the observations in that cluster

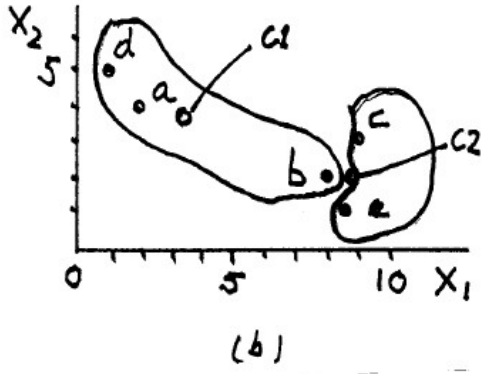
Centroid of Cluster 1 is the point (X1 = 3:67, X2 = 3:67),

(k-means method)

- The cluster centroid is the point with coordinates equal to the average values of the variables for the observations in that cluster.

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	c	9	3
b	8	2	e	8.5	1
d	1	5			
Ave.	3.67	3.67	Ave.	8.75	2

(a)



(b)

calculate the distance between **a** and the two centroids:

$$D(a, abd) = \sqrt{(2-3.67)^2 + (4-3.67)^2} = 1.702,$$

$$D(a, ce) = \sqrt{(2-8.75)^2 + (4-2)^2} = 7.040.$$

a is closer to the centroid of Cluster 1, to which it is currently assigned. a is not reassigned

Next, calculate the distance between b and the two cluster centroids:

$$D(b, abd) = \sqrt{(8-3.67)^2 + (2-3.67)^2} = 4.641$$

$$D(b, ce) = \sqrt{(8-8.75)^2 + (2-2)^2} = 0.750.$$

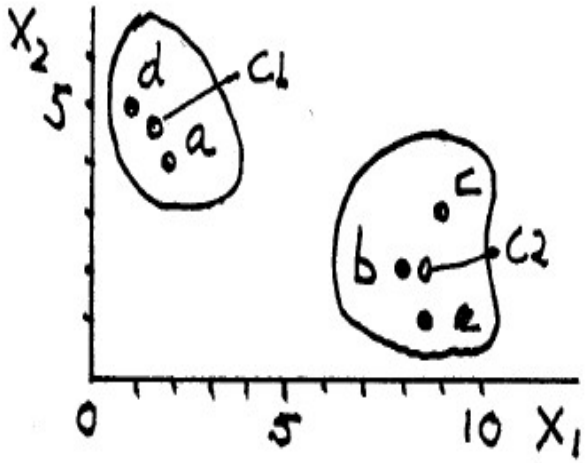
Since b is closer to Cluster 2's centroid than to that of Cluster 1, it is reassigned to Cluster 2.

(k-means method)

BELGIUM CAMPUS iTversity
It's the way we're wired

Cluster 1			Cluster 2		
Obs.	X_1	X_2	Obs.	X_1	X_2
a	2	4	c	9	3
d	1	5	e	8.5	1
			b	8	2
Ave.	1.5	4.5	Ave.	8.5	2

(a)



(b)

Every observation belongs to the cluster to the centroid of which it is nearest, and the k-means method stops

New cluster centroids

Obs.	Distance from Cluster 1	Distance from Cluster 2
a	0.707*	6.801
b	6.964	0.500*
c	7.649	1.118*
d	0.707*	8.078
e	7.826	1.000*

DISTANCE MEASURES FOR ATTRIBUTES

description of four persons according to marital status (single, married, divorced, other) and gender (male,female):

Obs.	Marital status	Gender
a	Single	Female
b	Married	Male
c	Other	Male
d	Single	Female

A reasonable measure of the similarity of two observations is the ratio of the number of matches (identical categories) to the number of attributes.

$$D_a(i,j) = 1 - \frac{\text{Number of matches}}{\text{Number of attributes}}$$

DISTANCE MEASURES FOR ATTRIBUTES

BELGIUM CAMPUS iTversity
It's the way we're wired

- Since **a** and **d** are both single and female, the similarity measure is 2/2 or 1; b and c do not have the same marital status but are both male, so the similarity measure is 1/2.
- The distances between all pairs of observations in our example are as follows:

Obs.	a	b	c	d
a	0	1	1	0
b		0	0.5	1
c			0	1
d				0

Any of the clustering methods described earlier can be applied to the above distances. nearest neighbor, furthest neighbor, or complete linkage methods a and d would be grouped to form the first cluster.

DISTANCE MEASURES FOR ATTRIBUTES

BELGIUM CAMPUS iTversity
It's the way we're wired

- When the grouping is to be based on variables and attributes the simplest approach is to convert the variables to attributes and then apply the measure $D_a(i; j)$ to the distance between any pair of observations
- Say the four observations will be grouped according to marital status, gender, and age:

Obs.	Marital status	Gender	Age (years)	Age category
a	Single	Female	15	Y
b	Married	Male	30	M
c	Other	Male	60	O
d	Single	Female	32	M

DISTANCE MEASURES FOR ATTRIBUTES

- If we make age an attribute with, three categories: Y (under 25 years old), M (25 to 50), and O (more than 50 years old).
- The “distance" between b and c, is

$$D_a(b, c) = 1 - \frac{1}{3} = \frac{2}{3}.$$

What will be the distances between all pairs of observations?

Obs.	Marital status	Gender	Age (years)	Age category
a	Single	Female	15	Y
b	Married	Male	30	M
c	Other	Male	60	O
d	Single	Female	32	M

DISTANCE MEASURES FOR ATTRIBUTES

Obs.	a	b	c	d
a	0	1	1	1/3
b		0	2/3	2/3
c			0	1
d				0

QUESTION

BELGIUM CAMPUS iTversity
It's the way we're wired

- Six observations on two variables are available, as shown in the following table:

Obs.	X_1	X_2
a	3	2
b	4	1
c	2	5
d	5	2
e	1	6
f	4	2

a. Plot the observations in a scatter diagram. How many groups would you say there are, and what are their members?
b. Apply the nearest neighbour method and the squared Euclidean distance as a measure of dissimilarity. Use a dendrogram to arrive at the number of groups and their membership.
c. Apply the k-means method, assuming that the observations belong to two groups and that one of these groups consists of a and e.