WELCOME TO

TECH I.S.

Introduction to Clustering

Cluster and Clustering

Cluster: a collection of data objects

Clustering: Grouping a set of data objects into clusters

Clustering is unsupervised classification: no predefined target class are given.

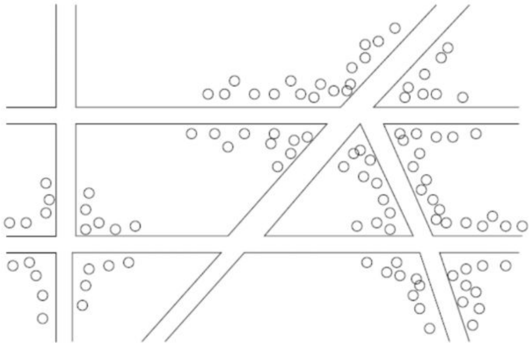
What is clustering used for?

Examples:

- Groups customers to make “small”, “medium” and “large” T-Shirts.
- Given a collection of text documents, we want to compare content similarities To check copyrights.

In fact, clustering is one of the most utilized data mining techniques.

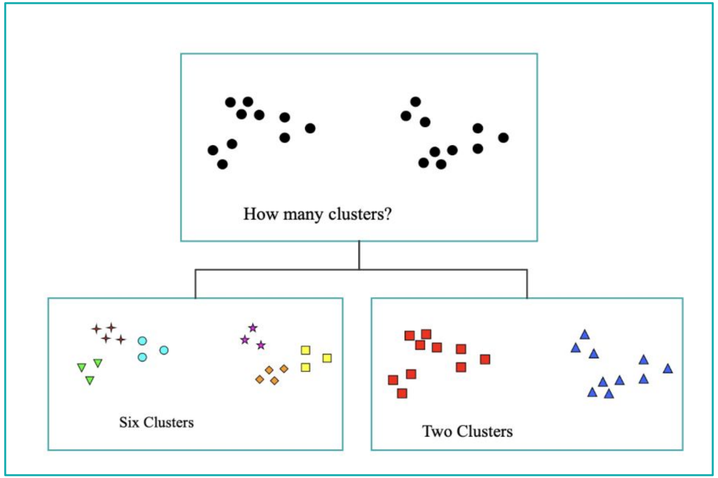
John Snow, London 1854

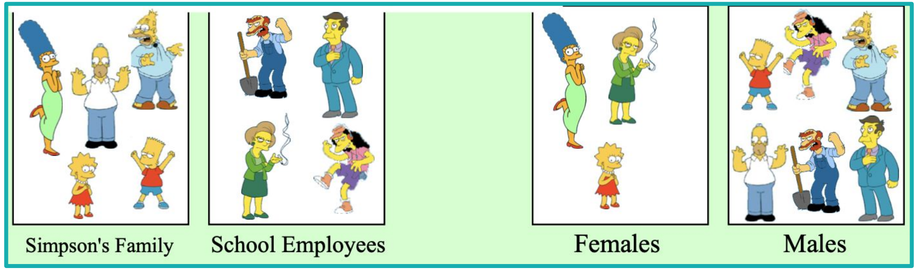


Plotting cholera cases on a map of London

Clustering - Ambiguous and Subjective

The following objects and subjects can be divided in more than one way:



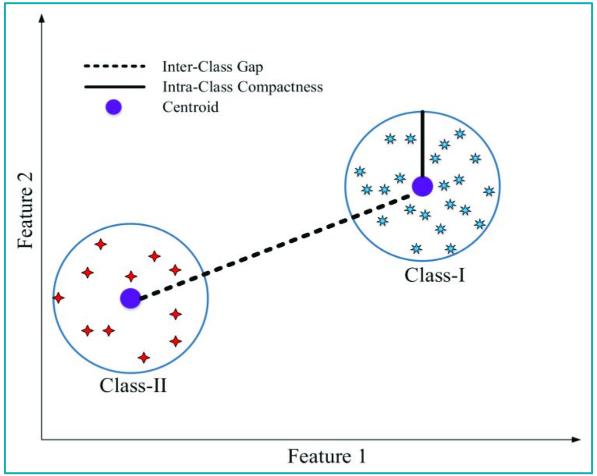


Quality of Clustering

The quality of a clustering result depends on <u>similarity measure</u> used and its implementation.

<u>A good clustering method will produce high quality clusters with</u>

- **<u>High INTRA class similarity</u>**
- **<u>Low INTER class similarity</u>**



Similarity and Dissimilarity for objects

If p and q are the attribute values for two data objects then d is the similarity accoridngly:

Attribute Type	Dissimilarity	Similarity
Discrete	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Continuous	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Data as Distance Matrix

Represents pairwise distance in n objects

- An n by n matrix
- d(i,j): distance or dissimilarity between objects i and j – Nonnegative
- Close to 0: similar

	s 1	s 2	s 3	s 4	…
g 1	0.13	0.72	0.1	0.57	
g 2	0.34	1.58	1.05	1.15	
g 3	0.43	1.1	0.97	1	
g 4	1.22	0.97	1	0.85	
g 5	-0.89	1.21	1.29	1.08	
g 6	1.1	1.45	1.44	1.12	
g 7	0.83	1.15	1.1	1	
g 8	0.87	1.32	1.35	1.13	
g 9	-0.33	1.01	1.38	1.21	
g 10	0.10	0.85	1.03	1	
…					

Original Data Matrix

	g 1	g 2	g 3	g 4	…
g 1	0	*d*(1,2)	*d*(1,3)	*d*(1,4)	
g 2		0	*d*(2,3)	*d*(2,4)	
g 3			0	*d*(3,4)	
g 4				0	
…					

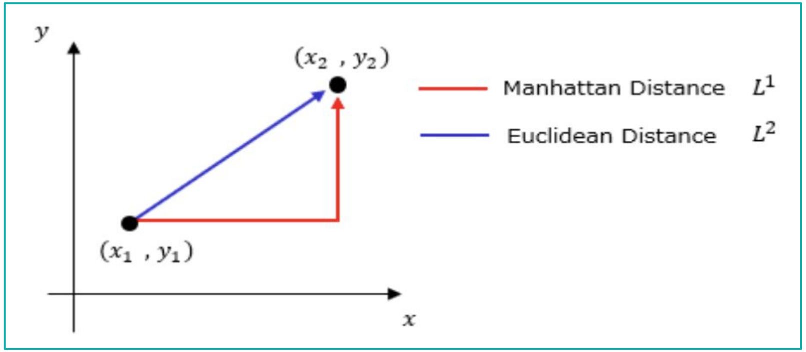
Distance Matrix

Distance as Similarity Measure

Lesser the distance, more is the similarity.

Euclidean Distance (Norm 2) → $L^2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Manhattan Distance (Norm 1) → $L^1 = |x_2 - x_1| + |y_2 - y_1|$



Document and Cosine Similarity

Each document can be represented as a vector,

Each word can be a component of the vector representing the number of times that term occurs in the document and their dot product represents the similarity between them.

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$$

d_1 = 3 2 0 5 0 0 0 2 0 0
d_2 = 1 0 0 0 0 0 0 1 0 2

$\cos(d_1, d_2) = .3150$

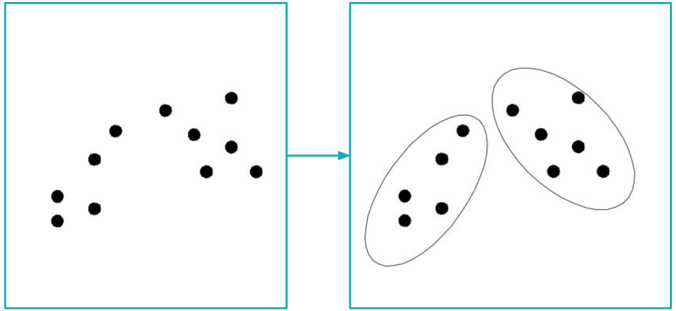
Two Types of Clustering

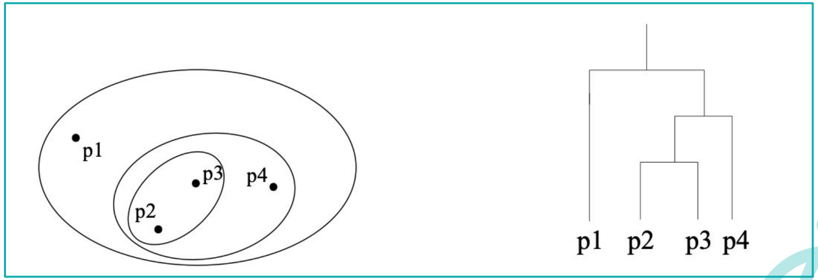
Partitional algorithms:

A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

Hierarchical algorithms:

Creating a set of nested clusters organized as a hierarchical tree.





Clustering as an optimization problem

Clustering algorithm finds clusters such that it minimize or maximize an objective function.

Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function.

- Global objective function: Typically used in partitional clustering.
- Local objective function: Used by Hierarchical & Density-based clustering algorithms

Desirable Properties of a Clustering Algorithm

Clustering Scalability: In order to handle extensive databases, the clustering algorithm should be scalable.

High Dimensionality: The algorithm should be able to handle high dimensional space.

Flexibility: Algorithm Usability with multiple data kinds.

Dealing with unstructured data: Ability to handle missing values, and noisy or erroneous data.

Interpretability: The clustering outcomes should be interpretable, comprehensible, and usable.

Much obliged.

TECH I.S.