

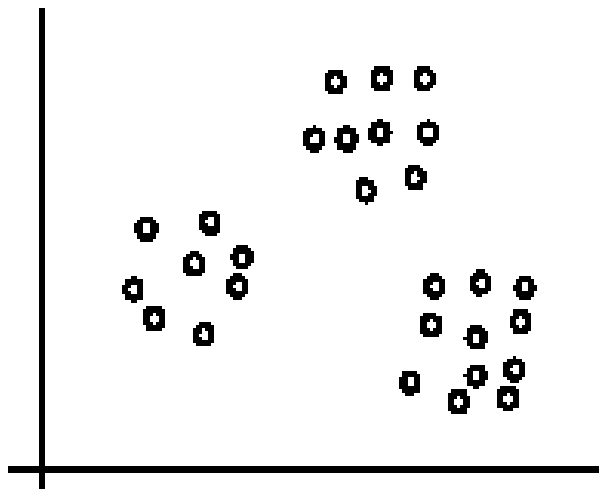
Clustering

Clustering

- Clustering is a technique for finding similar groups in data, called **clusters**.
- It groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- Clustering is often called an **unsupervised learning** task as no class values denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning i.e., only data is known but labels are unknown.
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning.
 - In fact, association rule mining is also unsupervised

An illustration

The data set has three natural groups of data points, i.e., 3 natural clusters.



What is clustering for?

Let us see some real-life examples

Example 1: groups people of similar sizes together to make small, medium and large T-shirts.

- ❖ Tailor-made for each person: too expensive
- ❖ One-size-fits-all: does not fit all.

Example 2: In marketing, segment customers according to their similarities

- ❖ To do targeted marketing.

Example 3: Given a collection of text documents, we want to organize them according to their content similarities,

- ❖ To produce a topic hierarchy

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

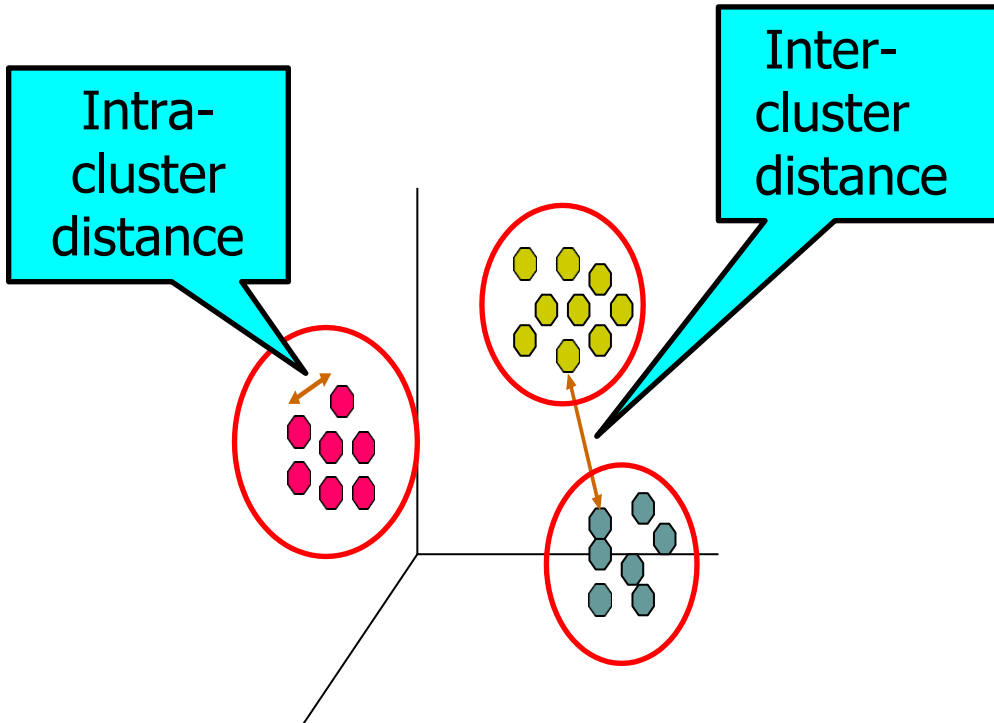
- ❖ It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.
- ❖ In recent years, due to the rapid increase of online documents, text clustering becomes important.

Aspects of clustering

- A clustering algorithm (Types of Clustering)
 - Partitional clustering: A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.
 - Hierarchical clustering: A set of nested clusters organized as a hierarchical tree
 - Agglomerative (Bottom-Up)
 - Divisive (Top-Down)
 - Fuzzy C-Means (FCM) Clustering
 - Density-based clustering
- Generally a distance (similarity e.g. K-Means Clustering, or dissimilarity e.g. Agglomerative) function is used to classify clusters.
- Clustering quality
 - Inter-clusters distance \Rightarrow maximized
 - Intra-clusters distance \Rightarrow minimized
- The *quality* of a clustering result depends on the algorithm, the distance function and the application.

What is Cluster Analysis?

Finding groups of objects in data such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Cluster Analysis



How many clusters do you expect?



starshado7671ickr



starshado7671ickr



K-Means Clustering

K-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.

K is positive integer number.

The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

K-Means Clustering Algorithm

Step 1: Begin with a decision on the value of K = number of clusters

Step 2: Put any initial partition that classifies the data into k clusters. Assign training samples as the following:

- a. Take the first k training sample as single-element clusters.
- b. Assign each of the remaining (N-K) training samples to the cluster with the nearest centroid.
- c. After each assignment, re-compute the centroid of the gaining cluster.

Step 3: Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing it.

Step 4: Repeat step until convergence is achieved, that is until a pass through the training sample causes no new assignments.

A simple examples showing the implementation of K-means algorithm using (K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

Initialization: Randomly we choose following two centroids. In this case, The 2 centroids are: $\mu_1 = (1.0, 1.0)$ and $\mu_2 = (5.0, 7.0)$.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

$$\text{Euclidian distance} = \sqrt{|\mu_{1x} - x_i|^2 + |\mu_{2y} - y_i|^2}$$

$$d(\mu_1, 1) = \sqrt{|1.0 - 1.0|^2 + |1.0 - 1.0|^2} = 0$$

$$d(\mu_2, 1) = \sqrt{|5.0 - 1.0|^2 + |7.0 - 1.0|^2} = 7.21$$

Individual	Centroid1	Variable1	Centroid2	Variable2
1	0.00	1	7.21	1
2	1.12	1.5	6.10	2
3	3.61	3	3.61	4
4	7.21	5	0.00	7
5	4.72	3.5	2.50	5
6	5.32	4.5	2.06	5
7	4.30	3.5	2.92	4.5

Step 2:

Thus, we obtain two clusters containing: {1, 2, 3} and {4, 5, 6, 7}

Their new centroids are:

$$\mu_1 = \left\{ \frac{1}{3}(1.0, +1.5 + 3.0), \frac{1}{3}(1.0, +2.0 + 4.0) \right\} = \{1.83, 2.33\}$$

$$\mu_2 = \left\{ \frac{1}{4}(5.0, +3.5 + 4.5 + 3.5), \frac{1}{4}(7.0, +5.0 + 5.0 + 4.5) \right\} = \{4.12, 5.38\}$$

$$d(\mu_1, 1) = \sqrt{|1.83 - 1.0|^2 + |2.33 - 1.0|^2} = 1.57$$

$$d(\mu_2, 1) = \sqrt{|4.12 - 1.0|^2 + |5.38 - 1.0|^2} = 5.38$$

Individual	Centroid1	Variable1	Centroid2	Variable2
1	1.57	1	5.38	1
2	0.47	1.5	4.28	2
3	2.04	3	1.78	4
4	5.64	5	1.84	7
5	3.15	3.5	0.73	5
6	3.78	4.5	0.54	5
7	2.74	3.5	1.08	4.5

Step 3:

Now using these two centroids we compute the Euclidean distance of each object as shown in table.

Therefore, the new clusters are {1, 2} and {3, 4, 5, 6, 7}

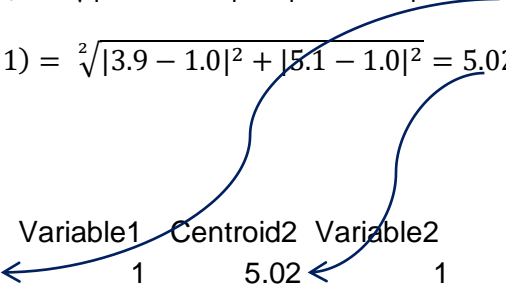
Next centroids are:

$$\mu_1 = \left\{ \frac{1}{2}(1.0, +1.5), \frac{1}{2}(1.0, +2.0) \right\} = \{1.25, 1.5\}$$

$$\begin{aligned} \mu_2 &= \left\{ \frac{1}{5}(3.0 + 5.0, +3.5 + 4.5 + 3.5), \frac{1}{5}(4.0 + 7.0, +5.0 + 5.0 + 4.5) \right\} \\ &= \{3.9, 5.1\} \end{aligned}$$

$$d(\mu_1, 1) = \sqrt{|1.25 - 1.0|^2 + |1.5 - 1.0|^2} = 0.56$$

$$d(\mu_2, 1) = \sqrt{|3.9 - 1.0|^2 + |5.1 - 1.0|^2} = 5.02$$



Individual	Centroid1	Variable1	Centroid2	Variable2
1	0.56	1	5.02	1
2	0.47	1.5	3.92	2
3	2.04	3	1.42	4
4	5.64	5	2.20	7
5	3.15	3.5	0.41	5
6	3.78	4.5	0.61	5
7	2.74	3.5	0.72	4.5

Step 4:

Since, there is no change in the cluster.

Thus, the algorithm comes to a halt here and final result consists of 2 clusters {1, 2} and {3, 4, 5, 6, 7}.