Artificial Intelligence

By

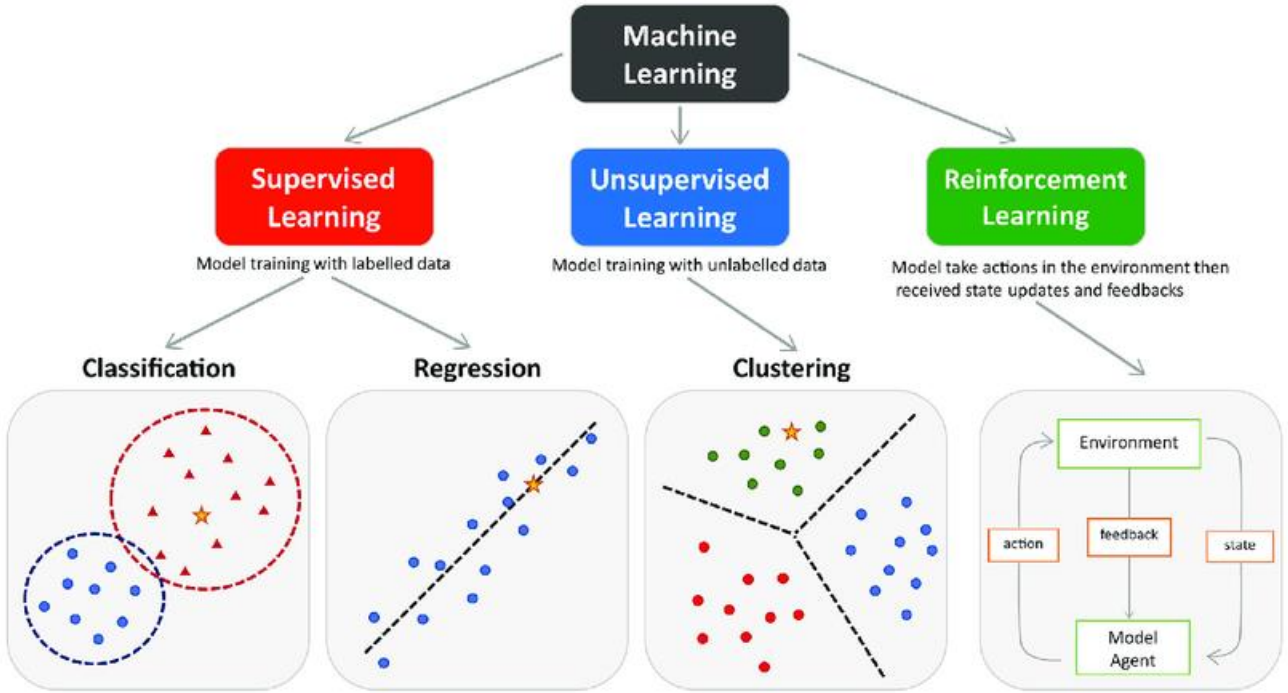
Dr. Manoj Kumar

University School of Automation and Robotics

GGSIP University, East Campus, Delhi, India

Clustering : Introduction

- Clustering is an important data analysis tool that provides improved data understanding.
- It plays a key role in searching for structures in data.
- Cluster analysis is aimed at dividing data whose identity is not known in advance, into homogeneous groups or clusters such that similar data objects belong to the same cluster and dissimilar data objects to different clusters.

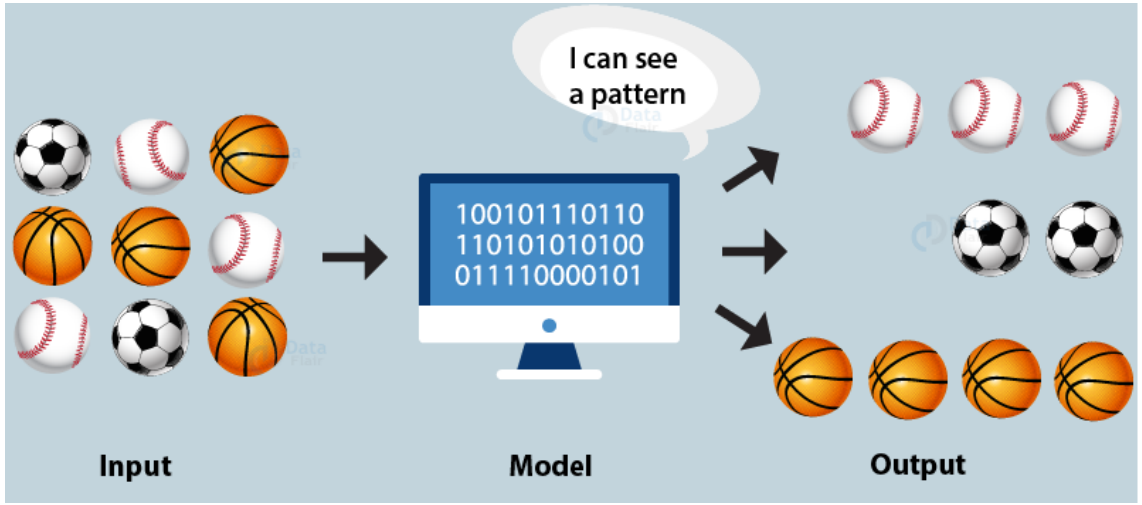


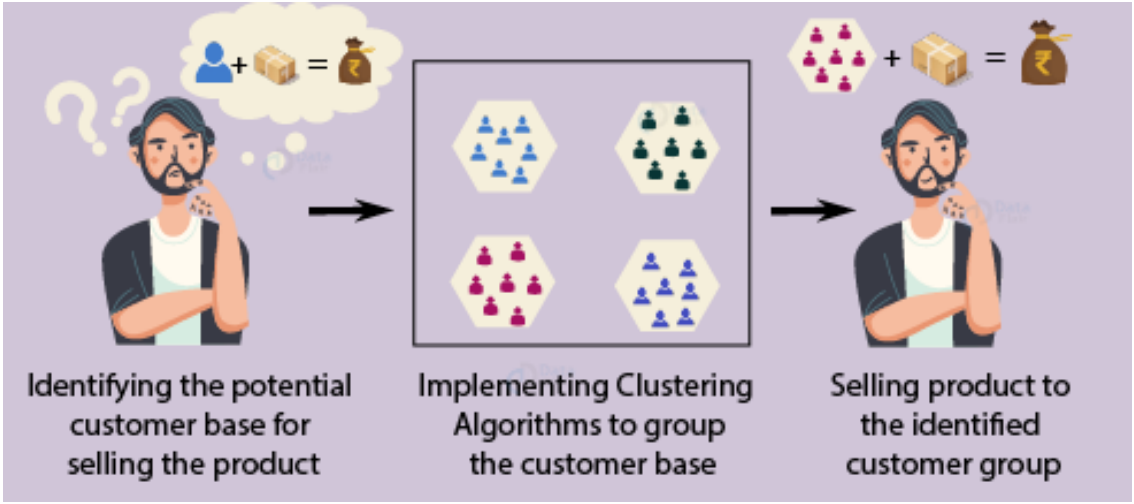
Clustering is a distance based unsupervised machine learning algorithm where data points that are close to each other are grouped in a given number of clusters/groups.

- A cluster is, therefore, a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. In other words, given a finite set of data, X , the problem of clustering in X is to find several cluster centers that are required to form a partition of X such that the degree of association is strong for data within blocks of the partition and weak for data in different blocks.

What is Clustering?

- Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait – often proximity according to some defined distance measure.





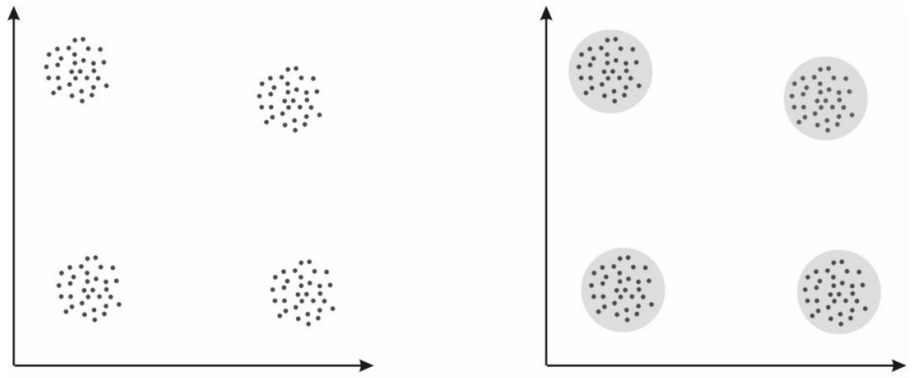


Fig. 1: An example of clusters.

- Clustering deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is, therefore, a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Uses of clustering

Computational Biology and Bioinformatics – In transcriptomics, clustering is used to build groups of genes with related expression patterns. Often such groups contain functionally related proteins, such as enzymes for a specific pathway, or genes that are co-regulated. High throughput experiments using expressed sequence tags (ESTs) or DNA microarrays can be a powerful tool for genome annotation, a general aspect of genomics.
In sequence analysis, clustering is used to group homologous sequences into gene families. This is a very important concept in bioinformatics, and evolutionary biology in general. See evolution by gene duplication.

Plant and Animal Ecology – In the fields of plant and animal ecology, clustering is used to describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments; it is also used in plant systematic to generate artificial phylogenies or clusters of organisms (individuals) at the species, genus or higher level that share a number of attributes.

Marketing Research – Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationship between different groups of consumers/potential customers for the following purposes.

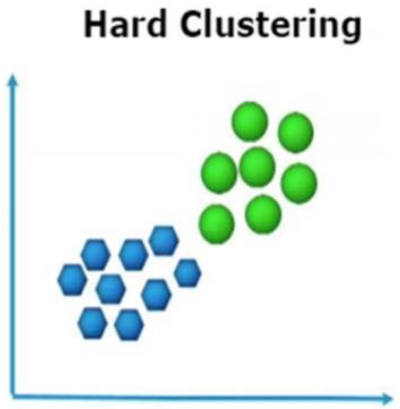
- o Segmenting the Market and determining target markets
- o Product positioning
- o New Product development
- o Selecting test markets

uses of clustering

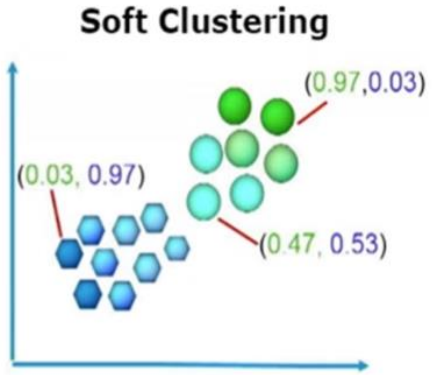
- **Insurance** – Insurance companies use clustering to identify policy holders with high average policy claims; identifying frauds.
- **City-Planning** – Identifying groups of houses according to their house type, value and geographical location.
- **Earthquake Studies** – Past data on earthquake is clustered to study the nature of continental faults.
- **Document Clustering** – Clustering proves to be one of the most promising techniques to reign in the huge unwieldy collection of documents on electronic media. Clustering similar documents together and representing them with a cluster center helps users get a glimpse of contents of a whole collection together without the tedious task of reading all of them.
- **Data Mining** – Many data mining applications involve partitioning data items into related subsets; the marketing applications discussed above represent some examples. Another common application is the division of documents, such as World Wide Web pages, into genres.
- **Social Network Analysis** – In the study of social network, clustering may be used to recognize communities within large groups of people.
- **Image Segmentation** – Clustering can be used to divide a digital image into distinct regions for border detection or object recognition.
- **WWW** – Document classification; clustering web-log data to discover groups of similar access patterns.

Clustering : Introduction

In hard clustering each datapoint is assigned only a single cluster.



On the other hand in soft clustering each data point belongs to a cluster with a certain probability also known as Membership Value.



Definition 1 (Hard Partition): Let X be a set of data, and x_1 be an element of X. A partition $P = \{C_1, C_2, \ldots, C_k\}$ of X is "hard" if and only if the following two condition hold.

(i) for all $x_i \in X$ there exists a partition $C_i \in P$ such that $x_i \in C_j$.

(ii) for all $x_i \in X$, $x_i \in C_j \Rightarrow x_i \notin C_k$ where $k \neq j$, $C_k C_j \in P$.

The first condition in the definition assures that the partition covers all data points in X, whereas the second condition assures that all clusters in the partition are mutually exclusive.

Definition 2 (Soft Partition): Let X be a set of data, and x_i be an element of X. A partition $P = \{C_1, C_2, \ldots, C_k\}$ of X is "soft" if and only if the following two conditions hold.

(i) for all $x_i \in X$ and for all $C_j \in P$, $0 \leq \mu_{cj}(x_i) \leq 1$ such that $x_i \in C_j$.

(ii) for all $x_i \in X$ there exists $C_j \in P$ such that $\mu_{cj}(x_i) > 0$.

where $\mu_{cj}(x_i)$ denote clustering of special interest to which x_i belongs to cluster C_j.

A type of fuzzy clustering of special interest is one that ensures the membership degree of a point x in all clusters adding up to one, i.e.,

$$\sum_j \mu_{cj}(x_i) = 1 \qquad \forall x_i \in X$$

A soft partition that satisfies this additional condition is called a *constrained soft partition*. The fuzzy c-means algorithm, which is best known fuzzy clustering algorithm, produces a constrained partition.

FUZZY CLUSTER ANALYSIS

- Cluster analysis divides data into groups (clusters) such that similar data objects belong to the same cluster and dissimilar data objects to different clusters.

- The resulting data partition improves data understanding and reveals its internal structure.

  - Partitional clustering algorithms divide up a data set into clusters or classes, where similar data objects are assigned to the same cluster whereas dissimilar data objects should belong to different clusters.

  - In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data.

  - Membership degree between zero and one are used in fuzzy clustering instead of crisp assignments of the data to clusters. The most prominent fuzzy clustering algorithm is the fuzzy c-means, a fuzzification of k-means.

Fuzzy C Means Clustering Algorithm

Fuzzy C Means Clustering Algorithm is an example of soft clustering.

- **Step 1:** Given the data points based on the number of clusters required initialize the membership table with random values.
- Suppose the given data points are **{(1, 3), (2, 5), (6, 8), (7, 9)}**

Or each data point have two features

Dividing this data into two clusters

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

- **Step 2: Find out the centroid.**

n is number of data points

- The formula for finding out the centroid (V) is:

- $V_{ij} = \frac{\sum_{k=1}^{n} \gamma_{ik}^{m} * x_k}{\sum_{k=1}^{n} \gamma_{ik}^{m}}$

  - γ: Fuzzy membership value
  - m: Fuzziness parameter generally taken as 2 and
  - x_k is the data point

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

- $V_{11} = \frac{(0.8^2 * 1 + 0.7^2 * 2 + 0.2^2 * 4 + 0.1^2 * 7)}{(0.8^2 + 0.7^2 + 0.2^2 + 0.1^2)} = 1.568$
- $V_{12} = \frac{(0.8^2 * 3 + 0.7^2 * 5 + 0.2^2 * 8 + 0.1^2 * 9)}{(0.8^2 + 0.7^2 + 0.2^2 + 0.1^2)} = 4.051$
- $V_{21} = \frac{(0.2^2 * 1 + 0.3^2 * 2 + 0.8^2 * 4 + 0.9^2 * 7)}{(0.2^2 + 0.3^2 + 0.8^2 + 0.9^2)} = 5.35$
- $V_{22} = \frac{(0.2^2 * 3 + 0.3^2 * 5 + 0.8^2 * 8 + 0.9^2 * 9)}{(0.2^2 + 0.3^2 + 0.8^2 + 0.9^2)} = 8.215$

Centroid are: (1.568, 4.051) and (5.35, 8.215)

- **Step 3: Find out the distance of each point from the centroid.**

- $D_{11} = \sqrt{(1 - 1.568)^2 + (3 - 4.051)^2} = 1.2$
- $D_{12} = \sqrt{(1 - 5.35)^2 + (3 - 8.215)^2} = 6.79$
- $D_{21} = \sqrt{(2 - 1.568)^2 + (5 - 4.051)^2} = 1.04$
- $D_{22} = \sqrt{(2 - 5.35)^2 + (5 - 8.215)^2} = 4.64$
- $D_{31} = \sqrt{(4 - 1.568)^2 + (8 - 4.051)^2} = 4.63$
- $D_{32} = \sqrt{(4 - 5.35)^2 + (8 - 8.215)^2} = 1.36$
- $D_{31} = \sqrt{(7 - 1.568)^2 + (9 - 4.051)^2} = 7.34$
- $D_{32} = \sqrt{(7 - 5.35)^2 + (9 - 8.215)^2} = 1.82$

Centroids are:

(1.568, 4.051) and

(5.35, 8.215)

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9
	1	1	2	2

- **Step 4: Updating membership values.**

n is number of clusters here, n=2

- $\gamma_{ki} = \left(\sum_{j=1}^{n} \left\{ \frac{d_{ki}^2}{d_{kj}^2} \right\}^{\left(\frac{1}{(m-1)}\right)} \right)^{-1}$

k is a data point

- For point 1 new membership values are:

- $\gamma_{11} = \left(\left\{ \frac{(1.2)^2}{(1.2)^2} + \frac{(1.2)^2}{(6.79)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.97$

- $\gamma_{12} = \left(\left\{ \frac{(6.79)^2}{(1.2)^2} + \frac{(6.79)^2}{(6.79)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.03$

$D_{11} = 1.2, \quad D_{12} = 6.79$

$D_{21} = 1.04, \quad D_{22} = 4.64$

$D_{31} = 4.63, \quad D_{32} = 1.36$

$D_{31} = 7.34, \quad D_{32} = 1.82$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	**0.97**	0.7	0.2	0.1
2	**0.03**	0.3	0.8	0.9

- For point 2 new membership values are:

- $\gamma_{21} = \left(\left\{ \frac{(1.04)^2}{(1.04)^2} + \frac{(1.04)^2}{(4.64)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.95$
- $\gamma_{22} = \left(\left\{ \frac{(4.64)^2}{(1.04)^2} + \frac{(4.64)^2}{(4.64)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.05$

- For point 3 new membership values are:

- $\gamma_{31} = \left(\left\{ \frac{(4.63)^2}{(4.63)^2} + \frac{(4.63)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.08$
- $\gamma_{32} = \left(\left\{ \frac{(1.36)^2}{(4.63)^2} + \frac{(1.36)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.92$

- For point 4 new membership values are:

- $\gamma_{41} = \left(\left\{ \frac{(7.34)^2}{(7.34)^2} + \frac{(7.34)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.06$
- $\gamma_{42} = \left(\left\{ \frac{(1.82)^2}{(7.34)^2} + \frac{(1.82)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.94$

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.97	0.95	0.08	0.06
2	0.03	0.05	0.92	0.94

- **Step 5:** Repeat the steps (2-4) until the constant values are obtained for the membership values or the difference is less than the tolerance value

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Tolerance (t) = 0.01

Cluster	(1, 3)	(2, 5)	(4, 8)	(7, 9)
1	0.97	0.95	0.08	0.06
2	0.03	0.05	0.92	0.94