12. Define Clustering

Ans:

Clustering is the process of grouping a set of instances into clusters, so that instances within a cluster have high similarity

13. Write the applications for Clustering

Ans:

The applications of clustering are

a. Human genetic clustering
b. Medical genetic clustering
c. Market research
d. Pattern recognition

14. Write the requirements of Clustering

Ans.

The requirements of clustering are

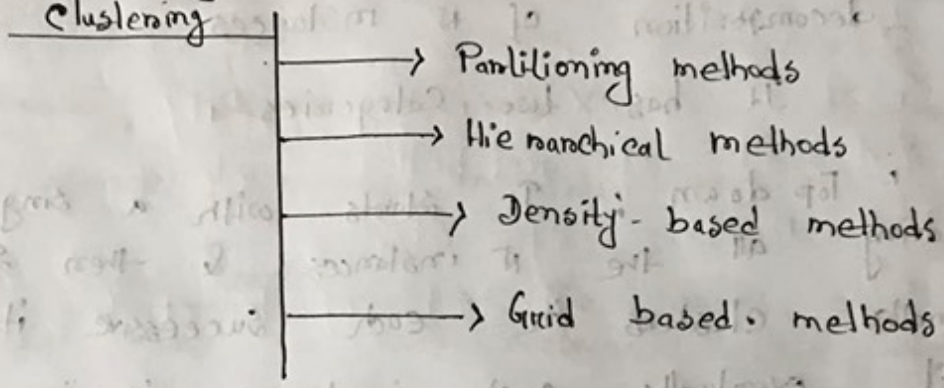
a. Dealing with large data sets containing different types of attributes.

b. find the clusters with ~~arbitra~~ arbitrary shape.

c. Ability to ded with noisy data in data streaming environment.

d. Handling with high dimensional data sets.

⊞ Clustering চার প্রকারের হয়ে থাকে।



15. Describe different clustering methods.

Ans

a. Partitioning method: Construct k Partitio

Where $k <= N$.

Here
k → means number of clusters
N → means number of instances.

It finds mutually exclusive clusters of spherical shape using traditional distance measures. Effective for clustering high-dimensional big data.

Example: k-means clustering, k-medoids

b. Hierarchical methods: It creates a hierarchical decomposition of N instances.

It has two categories

i. Top down: It starts with a single cluster having all the N instances & then split into smaller clusters in each successive iteration until eventually each instance is in one cluster. or termination condition holds.

ii. Bottom up: Each instance form a separate cluster. Then the clusters are merged to one another. It continues until all the clusters are merged into a single cluster. or termination condition holds.

Example: Agg Agnes, BIRCH.

c. Density based method: Clustering instances based on the distance between instances. It finds arbitrarily shaped clusters. It can cluster instances as dense regions in data space.

d. Grid based method: The grid based method user uses a multi resolution grid data structure. It's independent of number of instances.

16. Write the rule of Similarity Measure

Ans:

The similarity rule

a. Each cluster must contain at least one instance.

b. Each instance must belong to exactly one cluster.

17. Write the Property of C-mean

Ans:

High degree of similarity among instances in clusters is obtained.

High degree of dissimilarity among instances in different clusters is achieved simultaneously.

18. Write k-means clustering algorithm.

Ans:

Input: $X = \{x_1, x_2, \ldots, x_N\}$

// A set of unlabelled instances.

k // Number of clusters.

Output: A set of k clusters.

Method:

1. Arbitrarily choose k number of instances, $\{xk_1, xk_2, \ldots, x_{kN}\} \in X$ as the initial k clusters center.

2. Repeat.

3. Re-assign each $x_i \in X \to k$, so mean value $x_m \in k$.

4. Update the k means that is calculate the mean value of the instances for each clusters.

5. Until no change

19. Write the drawbacks of k-means Clustering

Ans.

The drawbacks are

a. Not guaranteed to converge to the global optimum

b. Can not be used in data with nominal features.

c. Not suitable for dis-covering clusters.

d. Fixed number of clusters can make it ~~diff~~ difficult to Predict what k should be.

e. The runtime $O(nkt)$. $n \rightarrow$ Total number of instances. $k \rightarrow$ number of clusters. $t \rightarrow$ Number of iterations.

Table: Height Data

Name	Gender	Height	Output
Kristina	F	1.6 m	Short
Jim	M	2 m	Tall
Maggie	F	1.9 m	Medium
Martha	F	1.88 m	Medium
Stephanie	F	1.7 m	Short
Bob	M	1.85 m	Medium
Kathy	F	1.6 m	Short
Dave	M	1.7 m	Short
Worth	M	2.2 m	Tall
Steven	M	2.1 m	Tall
Debbie	F	1.8 m	Medium
Todd	M	1.95 m	Medium
Kim	F	1.9 m	Medium
Amy	F	1.8 m	Medium
Wynette	F	1.75 m	Medium

Do K means clustering on k=2

Solve:

Step:1 Let

$k = 2$

$c_1 = 1.6$

$c_2 = 2.2$

Serial	Height	Distance from c_1	Distance from c_2
1	1.6	$\sqrt{(1.6-1.6)^2} = 0$	$\sqrt{(1.6-2.2)^2} = 0.6$
2	2	$\sqrt{(2-1.6)^2} = 0.4$	$\sqrt{(2-2.2)^2} = 0.2$
3	1.9	$\sqrt{(1.9-1.6)^2} = 0.3$	$\sqrt{(1.9-2.2)^2} = 0.3$
4	1.88	$\sqrt{(1.88-1.6)^2} = 0.28$	$\sqrt{(1.88-2.2)^2} = 0.32$
5	1.7	$\sqrt{(1.7-1.6)^2} = 0.1$	$\sqrt{(1.7-2.2)^2} = 0.5$
6	1.85	$\sqrt{(1.85-1.6)^2} = 0.25$	$\sqrt{(1.85-2.2)^2} = 0.35$
7	1.6	$\sqrt{(1.6-1.6)^2} = 0$	$\sqrt{(1.6-2.2)^2} = 0.6$
8	1.7	$\sqrt{(1.7-1.6)^2} = 0.1$	$\sqrt{(1.7-2.2)^2} = 0.5$
9	2.2	$\sqrt{(2.2-1.6)^2} = 0.6$	$\sqrt{(2.2-2.2)^2} = 0$
10	2.1	$\sqrt{(2.1-1.6)^2} = 0.5$	$\sqrt{(2.1-2.2)^2} = 0.1$
11	1.8	$\sqrt{(1.8-1.6)^2} = 0.2$	$\sqrt{(1.8-2.2)^2} = 0.4$
12	1.95	$\sqrt{(1.95-1.6)^2} = 0.35$	$\sqrt{(1.95-2.2)^2} = 0.25$
13	1.9	$\sqrt{(1.9-1.6)^2} = 0.3$	$\sqrt{(1.9-2.2)^2} = 0.3$
14	1.8	$\sqrt{(1.8-1.6)^2} = 0.2$	$\sqrt{(1.8-2.2)^2} = 0.4$
15	1.75	$\sqrt{(1.75-1.6)^2} = 0.15$	$\sqrt{(1.75-2.2)^2} = 0.45$

~~$C_1 = \{1.6, 2$~~

$C_1 = \{1, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15\}$

$C_2 = \{2, 9, 10, 12\}$

Step:2

Finding the mean

$$C_1' = \frac{1.6 + 1.9 + 1.88 + 1.7 + 1.85 + 1.6 + 1.7 + 1.8 + 1.9 + 1.8 + 1.75}{11}$$

$$= \frac{19.48}{11}$$

$$= 1.77$$

$$C_2' = \frac{2 + 2.2 + 2.1 + 1.95}{4}$$

$$= \frac{8.25}{4}$$

$$= 2.06$$

3/4 Prog

Serial	Height	Distance from C_1	Distance from C_2
1	1.6	$\sqrt{(1.6-1.77)^2} = 0.17$	$\sqrt{(1.6-2.06)^2} = 0.46$
2	2	$\sqrt{(2-1.77)^2} = 0.23$	$\sqrt{(2-2.06)^2} = 0.06$
3	1.9	$\sqrt{(1.9-1.77)^2} = 0.13$	$\sqrt{(1.9-2.06)^2} = 0.16$
4	1.88	$\sqrt{(1.88-1.77)^2} = 0.11$	$\sqrt{(1.88-2.06)^2} = 0.18$
5	1.7	$\sqrt{(1.7-1.77)^2} = 0.07$	$\sqrt{(1.7-2.06)^2} = 0.36$
6	1.85	$\sqrt{(1.85-1.77)^2} = 0.08$	$\sqrt{(1.85-2.06)^2} = 0.21$
7	1.6	$\sqrt{(1.6-1.77)^2} = 0.17$	$\sqrt{(1.6-2.06)^2} = 0.46$
8	1.7	$\sqrt{(1.7-1.77)^2} = 0.07$	$\sqrt{(1.7-2.06)^2} = 0.36$
9	2.2	$\sqrt{(2.2-1.77)^2} = 0.43$	$\sqrt{(2.2-2.06)^2} = 0.14$
10	2.1	$\sqrt{(2.1-1.77)^2} = 0.33$	$\sqrt{(2.1-2.06)^2} = 0.04$
11	1.8	$\sqrt{(1.8-1.77)^2} = 0.03$	$\sqrt{(1.8-2.06)^2} = 0.26$
12	1.95	$\sqrt{(1.95-1.77)^2} = 0.18$	$\sqrt{(1.95-2.06)^2} = 0.11$
13	1.9	$\sqrt{(1.9-1.77)^2} = 0.13$	$\sqrt{(1.9-2.06)^2} = 0.16$
14	1.8	$\sqrt{(1.8-1.77)^2} = 0.03$	$\sqrt{(1.8-2.06)^2} = 0.26$
15	1.75	$\sqrt{(1.75-1.77)^2} = 0.02$	$\sqrt{(1.75-2.06)^2} = 0.31$

$C_1 = \{1, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14, 15\}$

$C_2 = \{2, 9, 12\}$

Step: 4

Finding Mean

$$c_1' = \frac{1.6+1.9+1.88+1.7+1.85+1.6+1.7+2.1+1.8+1.9+1.8+1.75}{12}$$

$$c_2' = \frac{2+2.2+1.95}{3}$$

So,

$c_1' = 1.79$

$c_2' = 2.05$

Serial	Height	Distance from c_1'	Distance from c_2'
1	1.6	$\sqrt{(1.6-1.79)^2} = 0.19$	$\sqrt{(1.6-2.05)^2} = 0.45$
2	2	$\sqrt{(2-1.79)^2} = 0.21$	$\sqrt{(2-2.05)^2} = 0.05$
3	1.9	$\sqrt{(1.9-1.79)^2} = 0.11$	$\sqrt{(1.9-2.05)^2} = 0.15$
4	1.88	$\sqrt{(1.88-1.79)^2} = 0.09$	$\sqrt{(1.88-2.05)^2} = 0.17$
5	1.7	$\sqrt{(1.7-1.79)^2} = 0.09$	$\sqrt{(1.7-2.05)^2} = 0.35$
6	1.85	$\sqrt{(1.85-1.79)^2} = 0.06$	$\sqrt{(1.85-2.05)^2} = 0.2$
7	1.6	$\sqrt{(1.6-1.79)^2} = 0.19$	$\sqrt{(1.6-2.05)^2} = 0.45$
8	1.7	$\sqrt{(1.7-1.79)^2} = 0.09$	$\sqrt{(1.7-2.05)^2} = 0.35$
9	2.2	$\sqrt{(2.2-1.79)^2} = 0.41$	$\sqrt{(2.2-2.05)^2} = 0.15$
10	2.1	$\sqrt{(2.1-1.79)^2} = 0.31$	$\sqrt{(2.1-2.05)^2} = 0.05$
11	1.8	$\sqrt{(1.8-1.79)^2} = 0.01$	$\sqrt{(1.8-2.05)^2} = 0.25$
12	1.95	$\sqrt{(1.95-1.79)^2} = 0.16$	$\sqrt{(1.95-2.05)^2} = 0.1$
13	1.9	$\sqrt{1.9-1.79)^2} = 0.11$	$\sqrt{(1.9-2.05)^2} = 0.15$
14	1.8	$\sqrt{(1.8-1.79)^2} = 0.01$	$\sqrt{(1.8-2.05)^2} = 0.25$
15	1.75	$\sqrt{(1.75-1.79)^2} = 0.04$	$\sqrt{(1.75-2.05)^2} = 0.3$

Ex 2.5

$C_1 = \{1, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15\}$

$C_2 = \{2, 9\}$