Lec-28 19/03/21.

Introduction to clustering

- Basic concept
- distance measures
- K-means clustering algo.

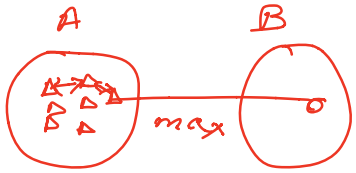
defn:

The process of grouping a set of physical or abstract objects into classes of similar objects is called grouping.

→ clustering is also called data segmentation because clustering partitions large data set into groups based on their similarity.

→ clustering can be used for outlier detection.

→ In machine learning, clustering is an example of unsupervised learning.



Distance Measures:

(a) Geometric distance measure (dealing with numeric data)

(b) Percent disagreement (dealing with categorical data)

a) Geometric distance measure:

i) Minkowski distance:

$$d(i,j) = \left(|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p \right)^{1/p}$$

where p is a positive integer. Such a distance is called Lp norm.

2) when $p = 1$ (i.e L_1 norm) → It represents Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$$

3) when $p = 2$ (i.e L_2 norm) → It represents Euclidean distance

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$$

4) Chebychev distance :

$$\text{dis}(A,B) = \max |x_i - y_i|$$

5) Power distance :

$$\text{powder dis}(A,B) = \left(\sum_i |x_i - y_i|^p \right)^{1/r}$$

where p and r are parameters defined by user.

(b) Percent disagreement :

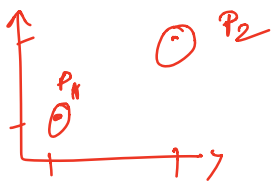
$$\text{Percent disagreement} = (\text{no. of } x_i \neq y_i)/i$$

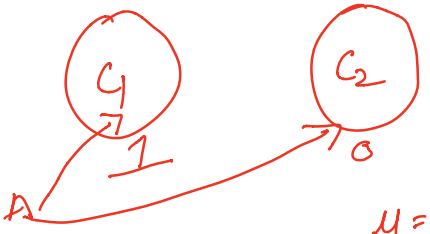
G1:	A	C (x)	T	G (x)	T	G (x)	G (x)	G	C	A
G2:	A	T	T	C	T	A	A	G	C	A

$$\text{Percent disagreement} = \left(\frac{4}{10}\right) \times 100\% \quad (i = 10)$$

$$= 40\%$$

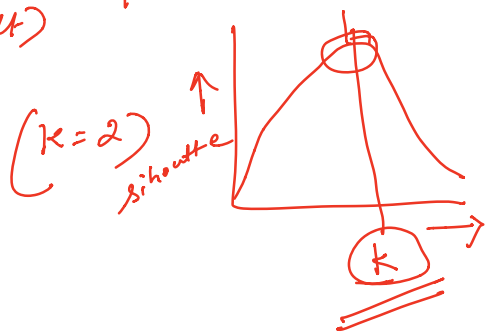
K-means clustering (Partitioning method)





K mean: (hard) $\mu = 0.3$

Fuzzy c-means (soft) $\mu = 0.7$



P_1 (1, 1)	P_5 (5, 2)
P_2 (6, 7)	P_6 (2, 3)
P_3 (4, 6)	P_7 (1, 2)
P_4 (5, 7)	P_8 (3, 1)

Item 1 :

Randomly choose 2 points as cluster centroid.

Here it- P_5 (5, 2) and P_7 (1, 2)

P_1 < P_5 = 4.123, P_7 = 1 ✓

P_2 < P_5 = 5.099 ✓, P_7 = 7.071

P_3 < P_5 = 4.123 ✓, P_7 = 5

P_4 < P_5 = 5 ✓, P_7 = 6.403

P_5 < P_5 = 0 ✓, P_7 = 4

P_6 < P_5 = 3.162, P_7 = 1.414 ✓

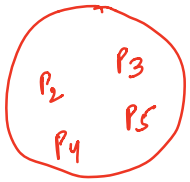
P_7 < P_5 = 4.0, P_7 = 0 ✓

P_8 < P_5 = 2.236, P_7 = 2.236

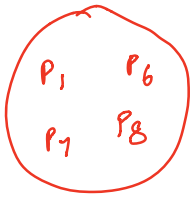
arbitrary break the tie by assigning P_8 to P_7 cluster

C1 (P_5)	C2 (P_7)
$\{P_2, P_3, P_4, P_5\}$	$\{P_1, P_6, P_7, P_8\}$

Item. 2 :



C1



C2

$C_1' = x' = \frac{5+5+6+4}{4} = 5$

$y' = \frac{7+6+7+2}{4} = 5.5$

$C_2' = x' = 1.75$

$y' = 1.75$

P_1 → $C_2' = 1.06$ ✓

P_1 → $C_1' = 6.02$

P_2 → $C_2' = 6.75$

P_2 → $C_1' = 1.80$ ✓

P_3 → $C_2' = 4.808$

P_3 → $C_1' = 1.118$ ✓

P_4 → $C_2' = 6.174$

P_4 → $C_1' = 1.5$ ✓

P_5 → $C_1' = 3.5$

P_5 → $C_2' = 3.259$ ✓

P_6 → $C_1' = 3.905$

P_6 → $C_2' = 1.274$ ✓

P_7 → $C_1' = 5.315$

P_7 → $C_2' = 0.790$ ✓

P_8 → $C_1' = 4.924$

P_8 → $C_2' = 1.457$ ✓

$C_1' = \{ P_2 \ P_3 \ P_4 \}$

$C_2' = \{ P_1 \ P_5 \ P_6 \ P_7 \ P_8 \}$

Item 3

C_1'' $x = \frac{5+4+6}{3} = 5$

$y = \frac{7+6+7}{3} = 6.66$

C_2'' $x = \frac{1+1+2+3+5}{5} = 2.4$

$y = \frac{1+2+3+1+2}{5} = 1.8$

o/p of 3rd iterns :

$C_1^{11} = \{ P_2 \; P_3 \; P_4 \}$ $\quad$ $C_2^{11} = \{ P_1 \; P_5 \; P_6 \; P_7 \; P_8 \}$

The points inside cluster don't change in two successive iterations. Thus, K-means is converged.

Exit