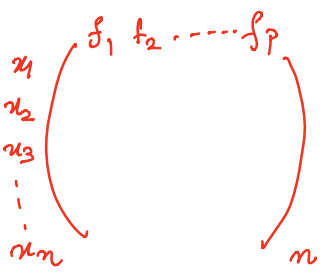
08/01/21 Lec-06

Topics to be discussed :

1. data matrix vs. dissimilarity matrix
2. proximity measure for nominal attribute
3. " " " binary "
4. " " " numeric "
5. " " " ordinal "
6. " " " mixed "
7. cosine similarity

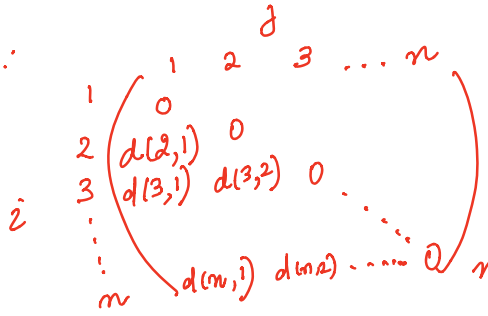
data matrix :

2-mode matrix



n → no. of objects (persons)
p → " " attributes (age, ht, wt, sal, gender etc.)

dissimilarity :



→ In general, d(i,j) is a non negative no. that is close to 0 when objects i and j are highly similar or near to each other, and become larger when they differ.

→ d(i,i) = 0 (difference between object and itself is 0)

→ d(i,j) = d(j,i) (symmetric)

$$\boxed{sim(i,j) = 1 - d(i,j)}$$ (for nominal data)

2) Proximity measure for nominal attribute :

Nominal attributes : Color code | Grading

Color code	Grading
Blue	A
red	B
green	C
yellow	A

dissimilarity b/w i & j :

$$d(i,j) = \frac{p - m}{p}$$

where p = no. of attributes
m = no. of match between i & j.

Eg :

ID	Test Res
1	code A
2	code B
3	code C
4	code A

f_1 (p=1)

$$d(2,1) = \frac{1-0}{1} = 1$$

$$d(3,1) = \frac{1-0}{1} = 1$$

$$d(3,2) = \frac{1-0}{1} = 1$$

$$d(4,1) = \frac{1-1}{1} = 0$$

$$d(4,2) = 1$$

$$d(4,3) = 1$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{pmatrix} \end{array} \quad 4 \times 4$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

2) Proximity measure for binary attributes :

- sym. binary (Gender : M/F)
- asym. binary (Medical Test Outcome)
- dissimilarity
  - sym. Binary : $d(i,j) = \frac{r+s}{q+r+s+t}$
  - asym. binary : $d(i,j) = \frac{r+s}{q+r+s}$
- Similarity
  - sym. binary
  - asym. binary ✗

Simple matching coefficient

$$Sim(i,j) = \frac{M_{11} + M_{00}}{M_{11} + M_{10} + M_{01} + M_{00}}$$

* Jaccard coefficient

$$Sim(i,j) = \frac{q}{q + r + s}$$

$$= \frac{M_{11}}{M_{11} + M_{10} + M_{01}} = 1 - d(i,j)$$

Contingency Table :

		(j) 1	(j) 0
(i)	1	q^{11}	r^{10}
	0	s^{01}	t^{00}

q : # attributes that equal 01 for obj i & j

r : # " " " 01 " i & 0 for j

s : # " " " 0 " i & 1 " "

t : # " " " 0 " " " 0 " j

Total no. of attributes $= p = q + r + s + t$

Eg :

	T_1	T_2	T_3	T_4	T_5	T_6
Jack	1	0	1	0	0	0
Jim	1	0	1	0	1	0
Mary	1	1	0	0	0	0

Case 1: Asymm. Binary : $d(i,j) = \frac{M_{10} + M_{01}}{M_{11} + M_{10} + M_{01}}$

	Jack	Jim	Mary
Jack	0		
Jim	0.33	0	

Mary \ 0.66 0.75 0 /

$$d(\underset{i}{\text{Jim}}, \underset{j}{\text{Jack}}) = \frac{1+0}{2+1+0} = \frac{1}{3} = 0.33$$

d (Mary, Jack) = 0.66

d (Mary, Jim) = 0.75

Case 2 : for symm. binary :

	Jack	Jim	Mary
Jack	0		
Jim	1/6	0	
Mary	1/3	1/2	0

4) Proximity measure for numeric attribute :

3 most commonly used measures are :

1. Minkowski distance
2. Euclidean distance
3. Manhattan distance
4. Supremum distance

1. Minkowski distance :

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where h is a real no s.t. $h \geq 1$. — (1)

2. Euclidean distance :

when $h = 2$, Eqn (1) becomes L_2 norm i.e Euclidean distance.

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \quad — (2)$$

where i & j are 2 objects having "p" no. of attributes.

3. **Manhattan distance** : / or (city block distance)

In eqn. (1), when $h = 1$, L1 norm or Manhattan distance is derived.

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}| \quad \text{— (3)}$$

4) **Supremum distance** :

This is also referred as L_{max}, L_∞ norm and Chebyshev distance which is a generalization of Minkowski distance for $h \to \infty$.

→ To compute it, find the attribute f that gives the maximum distance in values b/w 2 objects.

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{1/h}$$

$$= \max_{f}^{p} |x_{if} - x_{jf}|$$

L^∞ norm is also known as uniform norm.

Eg 1)

	f_1	f_2
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

i) **Euclidean** :

$d(P_2, P_1) = \sqrt{(2-0)^2 + (0-2)^2}$

$= \sqrt{4+4} = \sqrt{8}$

	P_1	P_2	P_3	P_4
P_1	0			
P_2	2.8	0		
P_3	3.2	1.4	0	
P_4	5.1	3.2	2	0

(ii) **Manhattan** :

$d(P_2, P_1) = |2-0| + |0-2|$

$= 2+2$

$= 4$

	P_1	P_2	P_3	P_4
P_1	0			
P_2	4	0		
P_3	4	2	0	
P_4	6	4	2	0

(iii) Supremum distance :

$$d(P_2, P_1) = \max\{|2-0|, |0-2|\}$$

$$= \max\{2, 2\}$$

$$= 2$$

	P_1	P_2	P_3	P_4
P_1	0			
P_2	2	0		
P_3	3	1	0	
P_4	5	3	2	0