12/01/21 Lec-07 ✓5. Proximity measures for ordinal attributes
✓6. Proximity measures for mixed-type "
✓7. Cosine similarity

5) Ordinal attributes :

Step 1) Let ordinal attributes be ranked as :

< fair , good, excellent >

∴ M_f represents total no. of ordered states

Here it is 03.

step 2 : Replace each ordinal data by a rank.

< fair : 01 good: 02 excellent : 03 >

step 3 : Normalize the rankings

$$Z_{if} = \frac{r_{if} - 1}{m_f - 1}$$

$$\text{fair } (1) = \frac{1-1}{3-1} = 0$$

$$\text{Good } (2) = \frac{2-1}{3-1} = 0.5$$

$$\text{Excellent } (3) = \frac{3-1}{3-1} = 01$$

Eg:

obj-id	Test
1	Excellent (01)
2	fair (0)
3	good (0.5)
4	excellent (01)

$d(2,1) = |0-1| = 1$

$d(3,1) = |0.5-1| = 0.5$

$d(3,2) = 0.5-0 = 0.5$

$d(4,1) = 0$

$d(4,2) = 1$

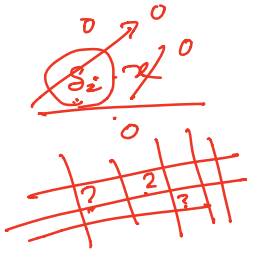
$d(4,3) = 0.5$

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$$

6) Proximity measure for mixed type attributes:

Obj id	Test-1 Nominal	Test-2 ordinal	Test-3 Numerical
1	code A	Excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28



1. dissimilarity matrix for nominal: $d(i,j) = \frac{p-m}{p}$

where p = # attributes

m = # match b/w i & j

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

2. dissimilarity matrix for ordinal attribute:

$\left(z_{if} = \frac{r_{if} - 1}{M_f - 1} \right)$

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0.5	0

3. dissimilarity matrix for numerical data:

To normalize, the data is mapped in the range of [0,1].

$$d(i,j) = \frac{|x_{if} - x_{jf}|}{max - min}$$

i q

1 → 45

2 → 22

3 → 64

4 → 28

	1	2	3	4
1	0			
2	0.55	0		
3	0.45	01	0	
4	0.4	0.14	0.86	0

$$d(2,1) = \frac{|45-22|}{64-22} = 0.55$$

$$d(3,1) = \frac{|64-45|}{64-22} = 0.45$$

$$d(3,2) = \frac{|64-22|}{64-22} = 01$$

6) formula for the mixed attribute :

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij} d_{ij}}{\sum_{f=1}^{p} \delta_{ij}}$$

where $\delta_{ij} = 0$ if x_{if} or x_{jf} is missing or $x_{if} = x_{jf} = 0$ and f is asey. binary
else $\delta_{ij} = 1$

	1	2	3	4
1	0			
2	0.85	0		
3	0.65	0.83	0	
4	0.13	0.71	0.79	0

$$d(2,1) = \frac{(1 \times 1) + (1 \times 1) + (1 \times 0.55)}{1+1+1} = \frac{2.55}{03} = 0.85$$

7) Cosine Similarity : & (cosine distance) :

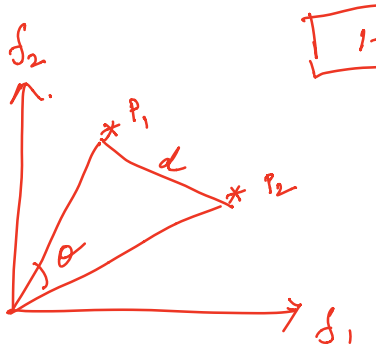
→ widely used in recommendation system (NLP)

P1 ● ○ P2

dis ↑ sim ↓

dis ↓ sim ↑

1 - cos sim = cos - distance

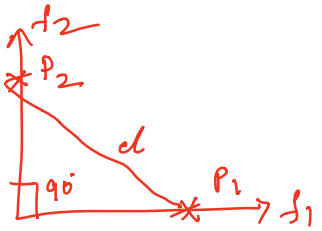


If $\theta = 45°$

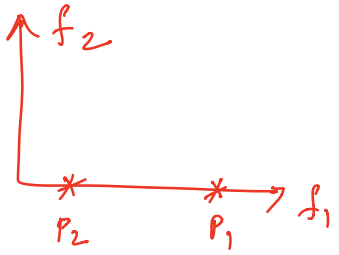
$\therefore \cos\theta = \cos 45° = 0.53$

(This means P_1 & P_2 are 53% similar based on angle)

cos-dis = 1 - cos sim
= 1 - 0.53
= 0.47



$\cos(90°) = 0$



$\cos(0^\circ) = 1$

2 points are similar as the angle is 0° and similarity is 100%.

Term - frequency vector :

→ Each document is an object represented by a term-frequency vector.

→ Term frequency are very long and sparse (ie too many 0 values).

→ Cosine similarity is a measure of similarity that can be used to compare documents or give a ranking of documents w.r.t a given vector of query words.

sim fun: $$\boxed{sim(x, y) = \frac{x \cdot y}{||x||\ ||y||}}$$

where $||x||$ is the Euclidean norm of vector $(x_1, x_2, \ldots x_p)$ defined as $||x|| = \sqrt{x_1^2 + x_2^2 + \ldots x_p^2}$. Conceptually, it is the length of the vector.

→ A cosine value 0 means the 2 vectors are 90° to each other (orthogonal) hence don't match.

→ The closer the cosine value is to 1, the smaller is the angle and greater is the match b/w 2 vectors.

Eg: Term frequency vector :

doc	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
1	5	0	3	0	2	0	0	2	0	0
2	3	0	2	0	1	1	0	1	0	1
3	0	7	0	2	1	0	0	3	0	0
4	0	1	0	0	1	2	2	0	3	0

Let x = doc 1 & y = doc 2

find similarity b/w x & y.

Soln: $sim(x,y) = \dfrac{x \cdot y}{||x||\ ||y||}$

$x = \langle 5\ 0\ 3\ 0\ 2\ 0\ 0\ 2\ 0\ 0 \rangle$

$y = \langle 3\ 0\ 2\ 0\ 1\ 1\ 0\ 1\ 0\ 1 \rangle$

$x \cdot y = (5\times3) + (0) + (3\times2) + (0) + (2\times1) + (0) + (0) + (2\times1) + 0 + 0 = 25$

$||x|| = \sqrt{5^2+0^2+3^2+0^2+2^2+2^2} = 6.48$

$||y|| = \sqrt{3^2+0^2+2^2+0^2+1+1+1+1} = 4.12$

$sim(x,y) = \dfrac{25}{6.48 \times 4.12} = 0.94$

$\therefore$ 2 documents x & y are 94% similar.