BDA sums

Bachelor of Engineering in Information Technology (University of Mumbai)

~~May~~ Dec 2017

Q.1) i) Jaccard Distance between {1,2,3,4} & {2,3,3,7} and {a,a,a,b} & {a,a,b,b,c}

i) $A = \{1,2,3,4\}$ $B = \{2,3,3,7\}$

~~$|A| = 4$~~ ~~$|B| =$~~

$A \cap B = \{2,3\}$ $\therefore |A \cap B| = 2$

$A \cup B = \{1,2,3,4,3,7\}$ $\therefore |A \cup B| = 6$

$\therefore$ Jaccard distance, $d_j(A,B) = 1 - J(A,B)$

$= \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$

$= \frac{6-2}{6} = \frac{4}{6}$

$= 0.667$

ii) $A = \{a,a,a,b\}$, $B = \{a,a,b,b,c\}$

$A \cap B = \{a,a,b\}$ $\therefore |A \cap B| = 3$

$A \cup B = \{a,a,a,b,b,c\}$ $\therefore |A \cup B| = 6$

$\therefore$ Jaccard distance, $d_j(A,B) = 1 - J(A,B)$

$= \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$

$= \frac{6-3}{6} = \frac{3}{6}$

$= 0.5$

2) Hamming distance between 11001 & 01101 and 11001 & 01011

i) $p_1 = 11001$
$p_2 = 01011$

$\therefore d(p_1, p_2) = 2$, because the bit-vectors 1 & 4 differ

ii) ~~$p_1 = 11001$~~ $p_1 = 1100111$
~~$p_2 = 01011$~~ $p_2 = 0101101$

$\therefore d(p_1, p_2) = 3$, because the bit-vectors 1, 4, 5 differs

3) Cosines of the angles between (3, -1, 2) & (-2, 3, 1)

$$\cos\theta = \frac{(3)(-2) + (-1)(3) + (2)(1)}{\sqrt{3^2 + (-1)^2 + 2^2} \cdot \sqrt{(-2)^2 + 3^2 + 1^2}}$$

$$= \frac{-6 - 3 + 2}{\sqrt{14} \times \sqrt{14}} = \frac{-7}{14} = \frac{-1}{2}$$

$\cos\theta = -0.5$

$\therefore \theta = \cos^{-1}(-0.5)$

$\theta = 120^\circ$

May 2017

Q.2] Data stream = 2, 1, 6, 1, 5, 9, 2, 3, 5

a) $h(x) = (2x + 3) \bmod 16$

For 2 :- $h(x) = (2x+3) \bmod 16 = (2\times2+3) \bmod 16 = 7$
For 1 :- $h(x) = (2x+3) \bmod 16 = (2\times1+3) \bmod 16 = 5$
For 6 :- $h(x) = (2x+3) \bmod 16 = (2\times6+3) \bmod 16 = 15$
For 1 :- $h(x) = (2x+3) \bmod 16 = (2\times1+3) \bmod 16 = 5$
For 5 :- $h(x) = (2x+3) \bmod 16 = (2\times5+3) \bmod 16 = 13$
For 9 :- $h(x) = (2x+3) \bmod 16 = (2\times9+3) \bmod 16 = 5$
For 2 :- $h(x) = (2x+3) \bmod 16 = (2\times2+3) \bmod 16 = 7$
For 3 :- $h(x) = (2x+3) \bmod 16 = (2\times3+3) \bmod 16 = 9$
For 5 :- $h(x) = (2x+3) \bmod 16 = (2\times5+3) \bmod 16 = 13$

Datastream	h(x)	binary	r(a)
2	7	0111	0
1	5	0101	0
6	15	1111	0
1	5	0101	0
5	13	1101	0
9	5	0101	0
2	7	0111	0
3	9	1001	0
5	13	1101	0

∴ R = maximax r(a)

R = 0

∴ Number of distinct elements = $2^R = 2^0 = 1$

b) $h(x) = (4x+1) \bmod 16$

For 2 :- $h(x) = (4x+1) \bmod 16 = (4\times2+1) \bmod 16 = 9$
For 1 :- $h(x) = (4x+1) \bmod 16 = (4\times1+1) \bmod 16 = 5$
For 6 :- $h(x) = (4x+1) \bmod 16 = (4\times6+1) \bmod 16 = 9$
For 1 :- $h(x) = (4x+1) \bmod 16 = (4\times1+1) \bmod 16 = 5$
For 5 :- $h(x) = (4x+1) \bmod 16 = (4\times5+1) \bmod 16 = 5$
For 9 :- $h(x) = (4x+1) \bmod 16 = (4\times9+1) \bmod 16 = 5$
For 2 :- $h(x) = (4x+1) \bmod 16 = (4\times2+1) \bmod 16 = 9$
For 3 :- $h(x) = (4x+1) \bmod 16 = (4\times3+1) \bmod 16 = 13$
For 5 :- $h(x) = (4x+1) \bmod 16 = (4\times5+1) \bmod 16 = 5$

Datastream	h(x)	binary	r(a)
2	9	1001	0
1	5	0101	0
6	9	1001	0
1	5	0101	0
5	5	0101	0
9	5	0101	0
2	9	1001	0
3	13	1101	0
5	5	0101	0

∴ R = maximum r(a)
R = 0

∴ Number of distinct elements = $2^R = 2^0 = 1$

c) $h(x) = 5x \bmod 16$

For 2 :- $h(x) = 5x \bmod 16 = 5 \times 2 \bmod 16 = 10$
For 1 :- $h(x) = 5x \bmod 16 = 5 \times 1 \bmod 16 = 5$
For 6 :- $h(x) = 5x \bmod 16 = 5 \times 6 \bmod 16 = 14$
For 1 :- $h(x) = 5x \bmod 16 = 5 \times 1 \bmod 16 = 5$
For 5 :- $h(x) = 5x \bmod 16 = 5 \times 5 \bmod 16 = 9$
For 9 :- $h(x) = 5x \bmod 16 = 5 \times 9 \bmod 16 = 13$
For 2 :- $h(x) = 5x \bmod 16 = 5 \times 2 \bmod 16 = 10$
For 3 :- $h(x) = 5x \bmod 16 = 5 \times 3 \bmod 16 = 15$
For 5 :- $h(x) = 5x \bmod 16 = 5 \times 5 \bmod 16 = 9$

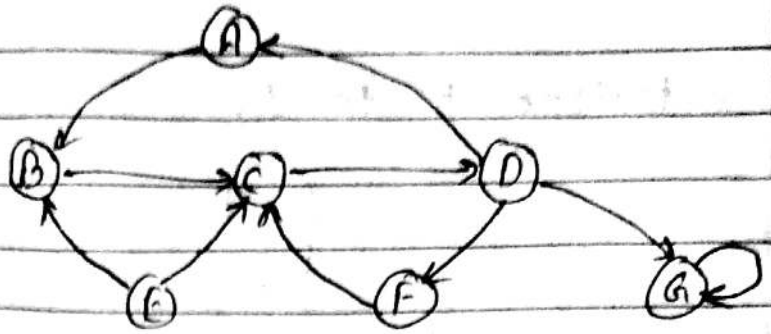
Datastream	$h(x)$	binary	$r(a)$
2	10	1010	1
1	5	0101	0
6	14	1110	1
1	5	0101	0
5	9	1001	0
9	13	1101	0
2	10	1010	1
3	15	1111	0
5	9	1001	0

$\therefore R$ = maximum $r(a)$

$R = 1$

$\therefore$ Number of distinct elements $= 2^R = 2^1 = 2$

Q.3]



Soln:) Matrix for the above graph is,

$$L = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{ccccccc} A & B & C & D & E & F & G \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$L^T = \begin{array}{c} A \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

let, h = hub score
a = authority score

Initialize h to 1,

$$h_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Iteration 1 :-

$$a_1 = L^T h_0$$

$$a_1 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$a_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

For normalization, $= \sqrt{1^2 + 2^2 + 3^2 + 1^2 + 0^2 + 1^2 + 2^2}$

$= 4.472$

$h_1 = L.a_1$

$$h_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 0 \\ 1 \\ 2 \end{bmatrix}$$

$$h_1 = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 5 \\ 3 \\ 2 \end{bmatrix}$$

For normalization of a_1,

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

$$= 4.472 = \sqrt{20}$$

For normalization of h_1,

$$= \sqrt{2^2 + 3^2 + 1^2 + 4^2 + 5^2 + 3^2 + 2^2}$$

$$= 8.246 = \sqrt{68}$$

$$a_1' = \begin{bmatrix} 1/\sqrt{20} \\ 2/\sqrt{20} \\ 0/\sqrt{20} \\ 1/\sqrt{20} \\ 0 \\ 1/\sqrt{20} \\ 2/\sqrt{20} \end{bmatrix} = \begin{bmatrix} 0.223 \\ 0.447 \\ 0.670 \\ 0.223 \\ 0 \\ 0.223 \\ 0.447 \end{bmatrix}$$

$$h_1' = \begin{bmatrix} 2/\sqrt{60} \\ 3/\sqrt{60} \\ 1/\sqrt{60} \\ 4/\sqrt{60} \\ 9/\sqrt{60} \\ 3/\sqrt{60} \\ 2/\sqrt{60} \end{bmatrix} = \begin{bmatrix} 0.242 \\ 0.363 \\ 0.121 \\ 0.485 \\ 0.606 \\ 0.363 \\ 0.242 \end{bmatrix}$$

For Iteration 2:-

$$a_2 = L^T h_1'$$

$$a_2 = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.242 \\ 0.363 \\ 0.121 \\ 0.485 \\ 0.606 \\ 0.363 \\ 0.243 \end{bmatrix}$$

$$a_2 = \begin{bmatrix} 0.485 \\ 0.848 \\ 1.332 \\ 0.121 \\ 0 \\ 0.485 \\ 0.728 \end{bmatrix}$$

$h_2 = t_1 \cdot a_2$

$$h_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.485 \\ 0.848 \\ 1.332 \\ 0.121 \\ 0 \\ 0.485 \\ 0.728 \end{bmatrix}$$

$$h_2 = \begin{bmatrix} 0.848 \\ 1.332 \\ 0.121 \\ 1.698 \\ 2.18 \\ 1.332 \\ 0.728 \end{bmatrix}$$

For normalization of a_2,

$$= \sqrt{0.485^2 + 0.848^2 + 1.332^2 + 0.121^2 + 0^2 + 0.485^2 + 0.728^2}$$

$$= 1.873$$

For normalization of h_2,

$$= \sqrt{0.848^2 + 1.332^2 + 0.121^2 + 1.698^2 + 2.18^2 + 1.332^2 + 0.728^2}$$

$$= 3.52$$

$$a_2' = \begin{bmatrix} 0.485/1.873 \\ 0.848/1.873 \\ 1.332/1.873 \\ 0.121/1.873 \\ 0 \\ 0.485/1.873 \\ 0.728/1.873 \end{bmatrix} = \begin{bmatrix} 0.258 \\ 0.452 \\ 0.711 \\ 0.064 \\ 0 \\ 0.258 \\ 0.388 \end{bmatrix}$$

$$h_2' = \begin{bmatrix} 0.848/3.52 \\ 1.332/3.52 \\ 0.121/3.52 \\ 1.698/3.52 \\ 2.18/3.52 \\ 1.332/3.92 \\ 0.728/3.92 \end{bmatrix} = \begin{bmatrix} 0.240 \\ 0.378 \\ 0.034 \\ 0.482 \\ 0.619 \\ 0.378 \\ 0.206 \end{bmatrix}$$

ii) Spider Trap : Yes

1) A - B - C - D
2) C - D - F
3) G (self-loop)

Dead Ends : No

Because no such nodes are present

Because all the nodes are either pointing/linking to other nodes or to itself.

Q.4] Describe all association rules that have 100% confidence

Soln. Items = {1, 2, 3, ... 100}

Baskets = {B1, B2, ... B100}

Example :- B24 = {1, 2, 3, 4, 6, 8, 12, 24}

Support (A) = Occurence of A in every basket

$$\text{confidence } (A \rightarrow B) = \frac{\text{support } (A \cup B)}{\text{support } A}$$

In above question, the rules can be constructed as :-

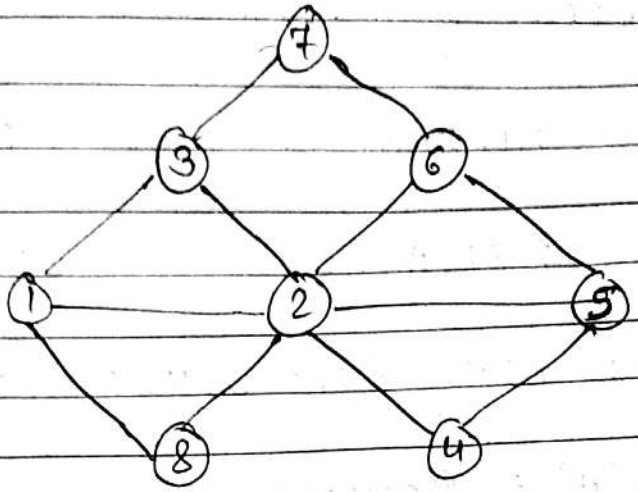
B1 = {1}
B2 = {1, 2}
B3 = {1, 3}
⋮
B100 = {1, 2, 4, 5, 10, 20, 25, 50, 100}

→ By observing, we can conclude that item 1 is the only item that is occuring in every basket.

→ To have 100% confidence, the item has to appear in every basket, which is item 1 only.

→ So any rule for e.g. {3, 5} → 1, {2, 4, 45} → 1 has 100% confidence

Q.3] Clique Percolation Method, find Cliques i.e



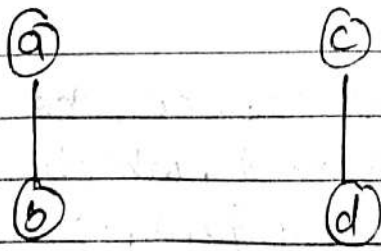
Soln.

Step 1 : All k-cliques present in G are extracted.

∴ Here, we have four 3-cliques present in the graph G,

a = (1, 2, 3)
b = (1, 2, 8)
c = (2, 4, 5)
d = (2, 6, 5)

Step 2 : A new graph, G_C i.e. clique graph is formed where each node represents an identified clique and two vertices (clique) in G_C are connected by an edge, if they have k-1 common vertices.



In the above clique graph, since $k = 3$, therefore we will find $k-1 = 3-1 = 2$ common vertices between the identified cliques.

Clique a and b have vertices 1 and 2 in common, a
Clique c and d have vertices 2 and 3 in common, therefore they will be connected through an edge.

Step 3: Connected components in G_c are identified.

Connected components ~~are~~ in G_c are (a, b) and (c, d) and this form the communities.

Step 4: Each connected component in G_c represents a community.

1) C1 : (1, 2, 3, 8)
2) C2 : (2, 4, 5, 6)

Step 5: Set C be the set of communities formed for G.

Thus, the community set $C = \{C1, C2\}$, where vertex 2 overlaps both the communities. Vertex 7 is not part of any community as it is not a part of any 3-cliques.