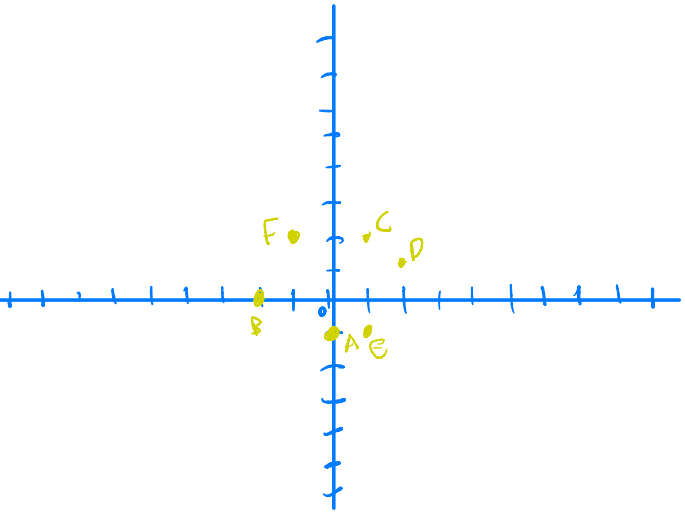
CZ/CE 4032
Data Analytics & Mining

Tutorial – Week 2
Clustering – basic methods

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Questions

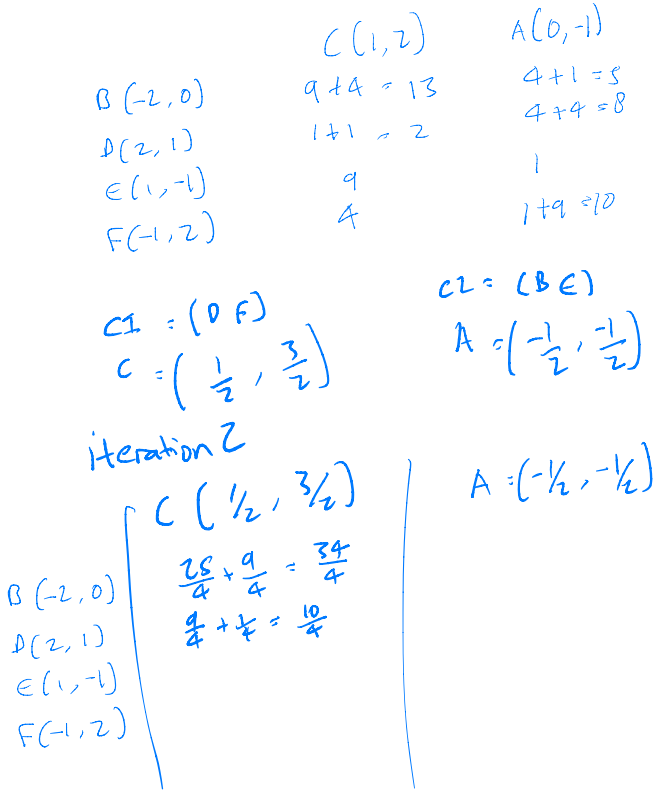
- Q1:
- Perform K-means on the following 6 data points for 2 iterations. The number of clusters is 2. The initial centroids are at data point E and F.



A	[0, -1]
B	[-2, 0]
C	[1, 2]
D	[2, 1]
E	[1, -1]
F	[-1, 2]

Questions

- Q2:
- Given the 6 data points in Q1, perform K-means for 2 iterations. The number of clusters is 2. The initial centroids are at data point C and A.



Questions

- Q3:
- Given the 6 data points in Q1, perform hierarchical clustering (agglomerative) until there is only 1 cluster. Use Single Linkage method.

Iteration 1
step 1 = Cluster assignment
1) Calculate Squared L2 distance

	E (1,-1)	F (-1,2)
A (0,-1)	$(0-1)^2+(-1-(-1))^2=1$	$1+9=10$
B (-2,0)	$(-2-1)^2+(0-(-1))^2=10$	$1+9=5$
C (1,2)	$(1-1)^2+(2-(-1))^2=9$	$4+0=4$
D (2,1)	$(2-1)^2+(1-(-1))^2=5$	$9+1=10$
E (1,-1)	0	$4+9=13$
F (-1,2)	$4+9<13$	0

2) assign each point to its nearest centroid

C1 : (A, D, E) C2 = (B, C, F)

$E = \left(\frac{0+2+1}{3}, \frac{-1+1-1}{3}\right) = \left(1, -\frac{1}{3}\right)$

$F = \left(\frac{-2+1+-1}{3}, \frac{0+2+2}{3}\right) = \left(-\frac{2}{3}, \frac{4}{3}\right)$

iteration 1

	$(1, -1/3)$	$(-2/3, 4/3)$
A (0,-1)	$1+\frac{4}{9}=\frac{13}{9}=1.44$	$\frac{4}{9}+\frac{49}{9}=\frac{53}{9}=5.89$
B (-2,0)	$9+\frac{1}{9}=\frac{82}{9}=9.11$	$\frac{16}{9}+\frac{16}{9}=\frac{32}{9}=3.56$
C (1,2)	$0+\frac{81}{9}=5.44$	$\frac{25}{9}+\frac{4}{9}=\frac{29}{9}=3.22$
D (2,1)	2.78	$\frac{64}{9}+\frac{1}{9}=\frac{65}{9}=7.22$
E (1,-1)	$0+\frac{4}{9}=0.44$	8.22
F (-1,2)	$0+\frac{49}{9}=5.44$	0.56

2)

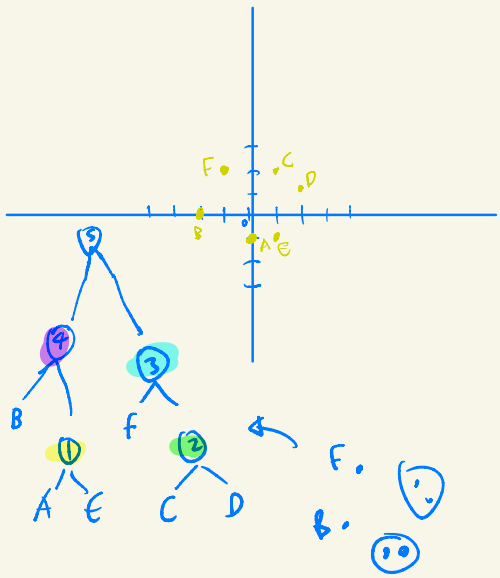
	C (1,2)	A (0,-1)
A (0,-1)	10	0
B (-2,0)	13	5
C (1,2)	0	10
D (2,1)	2	8
E (1,-1)	9	1
F (-1,2)	4	10

$C_1 (C, D, F) = \left(\frac{1+2-1}{3}, \frac{2+1+2}{3}\right) = \left(\frac{2}{3}, \frac{5}{3}\right)$

$C_2 (A, B, E) = \left(\frac{0-2+1}{3}, \frac{-1+0+1}{3}\right) = \left(-\frac{1}{3}, -\frac{2}{3}\right)$

	$(2/3, 5/3)$	$(-1/3, -2/3)$
A (0,-1)	7.56	0.22
B (-2,0)	9.89	3.22
C (1,2)	0.22	8.89
D (2,1)	2.22	8.22
E (1,-1)	7.22	1.89
F (-1,2)	2.89	7.56

3)



	A(0,-1)	B(-2,0)	C(1,2)	D(2,1)	E(1,-1)	F(-1,2)
A(0,-1)	0	4+1=5	1+9=10	4+4=8	1+0=1	1+9=10
B(-2,0)		0	9+4=13	16+1=17	9+1=10	1+4=5
C(1,2)			0	1+1=2	0+9=9	4+0=4
D(2,1)				0	1+4=5	9+1=10
E(1,-1)					0	4+9=13
F(-1,2)						0

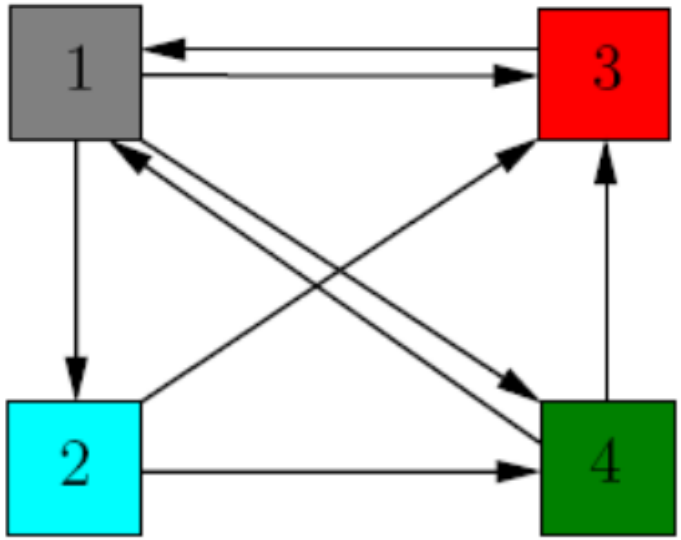
	B	AE	CDF
B (-2,0)	0	5	5
AE		0	5
CDF			0

! we have 3 equal distances, we can randomly select

Link Analysis: PageRank Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

- Q1. Given the graph below, calculate the page rank score for each node using the following two methods:
  - 1) directly solving the flow equations;
  - 2) using the power iterative method for 2 iterations.



1)

blue	in	out
1	2	3
2	1	2
3	3	1
4	2	2

$r_1 + r_2 + r_3 + r_4 = 1$ ⓞ

$r_1 = r_3 + \frac{1}{2} r_4$ ①

$r_2 = \frac{1}{3} r_1$ ②

$r_3 = \frac{1}{3} r_1 + \frac{1}{2} r_2 + \frac{1}{2} r_4$ ③

$r_4 = \frac{1}{3} r_1 + \frac{1}{2} r_2$ ④

use ⑥ ⑤ ② in ⓞ

$r_1 + (\frac{1}{3} r_1) + (\frac{3}{4} r_1) + (\frac{1}{2} r_1) = 1$

$\frac{12+4+9+6}{12} r_1 = 1$

$r_1 = \frac{12}{31}$

$r_2 = 4/31$

$r_3 = 9/31$

$r_4 = 6/31$

use ① in ④

$r_4 = \frac{1}{3} r_1 + \frac{1}{2}(\frac{1}{3} r_1) = \frac{1}{2} r_1$ ⑤

use ⑤ in ①

$r_1 = r_3 + \frac{1}{2}(\frac{1}{2} r_1) \equiv r_1 - \frac{1}{4} r_1 = r_3$

$r_3 = \frac{3}{4} r_1$ ⑥

in → out, Ly ToT is out ↓ in

Transition matrix

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

$r = M \cdot r$

② Power iteration

$r_1 = M r$

Transition matrix

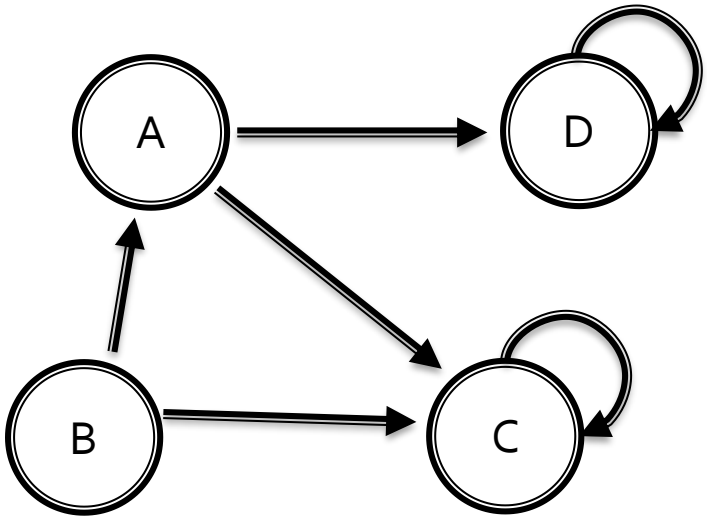
$$\begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

$r_0 = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$

$$r_1 = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{pmatrix} 1/4 + 1/8 \\ 1/12 \\ 1/12 + 1/8 + 1/8 \\ 1/12 + 1/8 \end{pmatrix} = \begin{pmatrix} 3/8 \\ 1/12 \\ 1/3 \\ 5/24 \end{pmatrix}$$

$$r_2 = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 3/8 \\ 1/12 \\ 1/3 \\ 5/24 \end{pmatrix} = \begin{pmatrix} 1/8 + 5/48 \\ 3/24 \\ 3/24 + 1/24 + 5/48 \\ 3/24 + 1/24 \end{pmatrix} = \begin{pmatrix} 7/16 \\ 1/8 \\ 17/48 \\ 1/6 \end{pmatrix} \begin{matrix} 0.44 \\ 0.125 \\ 0.291 \\ 0.167 \end{matrix}$$

- Q2. A graph is given below.
  - a) calculate the page rank score for each node using the power iteration method for 3 iterations.
  - b) identify one spider trap group if there are any.



2)

	in	out
A	1	2
B	0	2
C	3	1
D	2	1

Matrix =

	A	B	C	D
A	0	1/2	0	0
B	0	0	0	0
C	1/2	1/2	1	0
D	1/2	0	0	1

$r_0 = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$

$r_{t+1} = M \cdot r_t$

$r_1 = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix} = \begin{matrix} 1/8 \\ 0 \\ 1/8 + 1/8 + 1/4 \\ 1/8 + 1/4 \end{matrix} = \begin{matrix} 1/8 \\ 0 \\ 4/8 + 1/2 \\ 3/8 \end{matrix}$

$r_2 = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1/8 \\ 0 \\ 1/2 \\ 3/8 \end{pmatrix} = \begin{matrix} 0 \\ 0 \\ 1/16 + 1/2 = 9/16 \\ 1/16 + 3/8 = 7/16 \end{matrix}$

$r_3 = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 1 & 0 \\ 1/2 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 9/16 \\ 7/16 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 9/16 \\ 7/16 \end{pmatrix}$

Only Spider trap node have Values

(3) Spider trap = no out link node

C, D

Spider trap groups =
= C - A, C, D Ac!
= D - C, D Bc!

Spider trap is not dead node

!! dead node is totaly no out-link

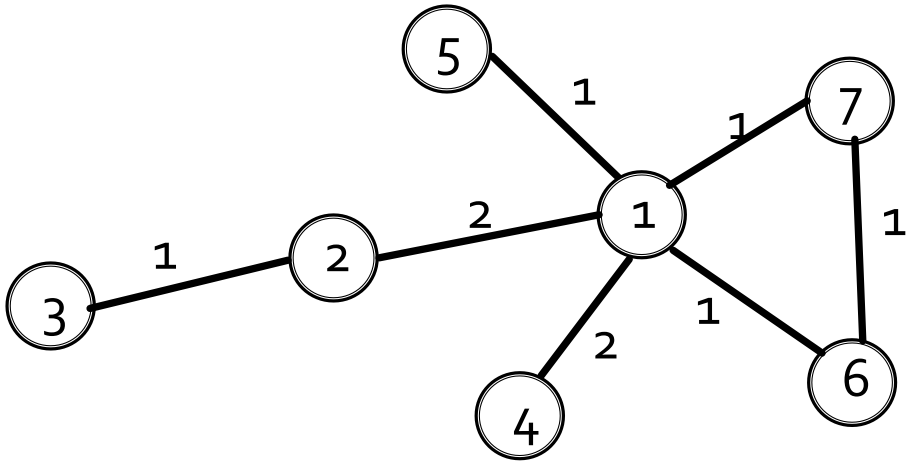
Spider trap might exist with node only linking to itself

Graph Neural Network Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Question:

Given a graph below (next page), the task is to do node-wise classification using a 2-layer graph convolutional network (GCN) and a cross-entropy loss.



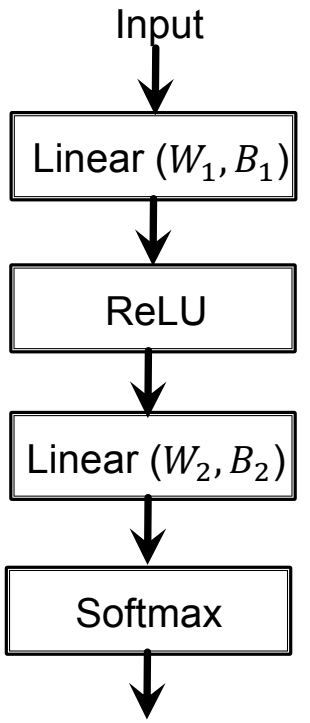
Input graph

need calculate node 3 and 7

(continued from last page)

The initial node features and the GCN network architecture are given below:

$$x_1 = [0, -1]^T,$$
$$x_2 = [0, 1]^T,$$
$$x_3 = [-1, 0]^T,$$
$$x_4 = [1, \ 0]^T,$$
$$x_5 = [-1, -1]^T,$$
$$x_6 = [2, 1]^T,$$
$$x_7 = [-1, 1]^T.$$



GCN architecture

(continued from last page)

The initial GCN weight parameters given below (W_k and B_k are the weight matrices for neighborhood aggregation and self transformation, respectively, for the k-th layer).

Calculate the prediction of **node 3** by performing one forward pass.

$$W_1 = \begin{bmatrix} 0 & -0.1 \\ 0.1 & 0 \end{bmatrix} \quad B_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & -0.1 \end{bmatrix}$$

$$W_2 = \begin{bmatrix} -0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \quad B_2 = \begin{bmatrix} 0 & 0.1 \\ -0.1 & 0 \end{bmatrix}$$

1) input → linear → ReLU → Linear → softmax

$$\hat{h}_i^{(k)} = W_k \sum_{j \in N_i} a'_{ij} h_j^{(k-1)} + B_k h_i^{(k-1)}$$

$$\max(0, h)$$

row-normalized affinity matrix =

$$\begin{pmatrix} 0&2&0&2&1&1&1 \\ 2&0&1&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 2&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 1&0&0&0&0&0&1 \\ 1&0&0&0&0&1&0 \end{pmatrix} \rightarrow \begin{pmatrix} 0&2/7&0&2/7&1/7&1/7&1/7 \\ 2/3&0&1/3&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 1/2&0&0&0&0&0&1/2 \\ 1/2&0&0&0&0&1/2&0 \end{pmatrix}$$

node 3 by FW pass

node 1

$$a'_{12}h_2 + a'_{14}h_4 + a'_{15}h_5 + a''_{16}h_6 + a'_{17}h_7$$

$$\begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}-1\\-1\end{pmatrix} + \begin{pmatrix}2\\1\end{pmatrix} + \begin{pmatrix}-1\\1\end{pmatrix} =$$

$$W_1 \sum a_{1j} h_j^{(k-1)} = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

node 2

$$a'_{21}h_1 + a'_{23}h_3$$

$$\begin{pmatrix}0\\-1\end{pmatrix} + \begin{pmatrix}-1\\0\end{pmatrix} =$$

$$W_1 \sum a_{2j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

node 3

$$a'_{32}h_2 = 0$$

$$1 \begin{pmatrix}0\\1\end{pmatrix} = 0$$

$$W_1 \sum a_{3j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 0 \end{pmatrix}$$

$$B_1 h_3 = \begin{pmatrix} 0.1 & 0 \\ 0 & -0.1 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 0 \end{pmatrix}$$

node 4

$$a_{41}h_1$$

$$\cdot \begin{pmatrix}0\\-1\end{pmatrix}$$

$$W_1 \sum a_{4j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

node 5

$$a_{51}h_1 \quad W_1 \sum a_{5j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

$$1 \cdot \begin{pmatrix}0\\-1\end{pmatrix}$$

node 6

$$a_{61}h_1 + a_{67}h_7$$

$$\begin{pmatrix}0\\-1\end{pmatrix} + \begin{pmatrix}-1\\1\end{pmatrix} =$$

$$W_1 \sum a_{6j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

node 7

$$a_{71}h_1 + a_{76}h_6$$

$$\begin{pmatrix}0\\-1\end{pmatrix} + \begin{pmatrix}2\\1\end{pmatrix} =$$

$$W_1 \sum a_{7j} h_j = \begin{pmatrix} 0 & -0.1 \\ 0.1 & 0 \end{pmatrix} =$$

Linear $= \begin{pmatrix} -0.1 \\ 0 \end{pmatrix} h \begin{pmatrix} 0 \\ 0.1 \end{pmatrix} = \begin{pmatrix} -0.2 \\ 0 \end{pmatrix}$

Relu $= \max\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} -0.2 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ or $h_3^{(1)}$

↑ might be diff

Step 2 calculate $h_2^{(1)}$

Relu $= \begin{pmatrix} 0.067 \\ 0 \end{pmatrix}$ or $h_2^{(1)}$

GCN layer 2 ! !

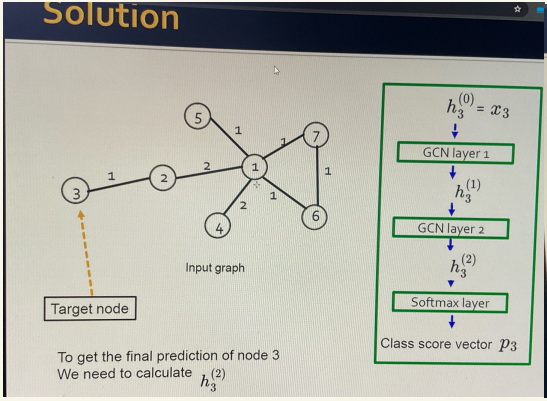
$h_3^{(2)}$

$W_2 \sum a_{32} h_2 = \begin{pmatrix} -0.1 & 0 \\ 0 & 0.1 \end{pmatrix} \begin{pmatrix} 1.0067 \\ 0 \end{pmatrix} = \begin{pmatrix} -0.0067 \\ 0 \end{pmatrix}$

$b_2 h_3 = \begin{pmatrix} 0 & 0.1 \\ -0.1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

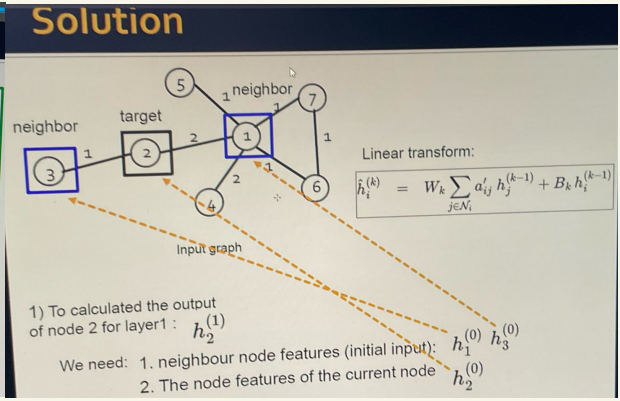
$h_3^{(2)} = \begin{pmatrix} -0.0067 \\ 0 \end{pmatrix} \rightarrow$ Soft max $\rightarrow \begin{pmatrix} 0.478 \\ 0.512 \end{pmatrix}$

Solution



To get the final prediction of node 3
We need to calculate $h_3^{(2)}$

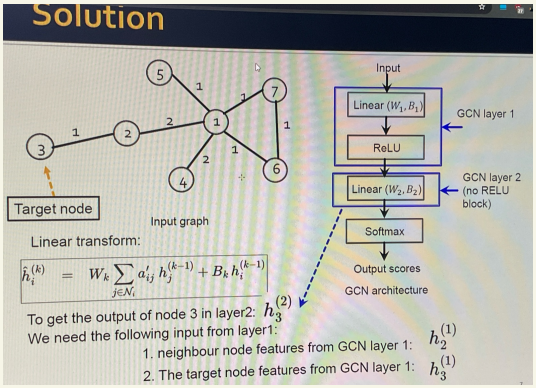
Solution



1) To calculated the output of node 2 for layer1 : $h_2^{(1)}$

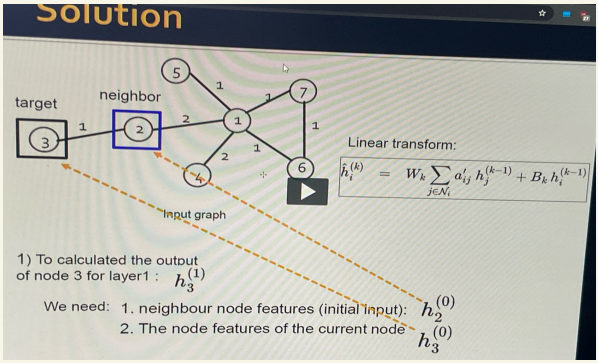
We need: 1. neighbour node features (initial input): $h_1^{(0)}$ $h_3^{(0)}$
2. The node features of the current node $h_2^{(0)}$

Solution



To get the output of node 3 in layer2: $h_3^{(2)}$
We need the following input from layer1:
1. neighbour node features from GCN layer 1: $h_2^{(1)}$
2. The target node features from GCN layer 1: $h_3^{(1)}$

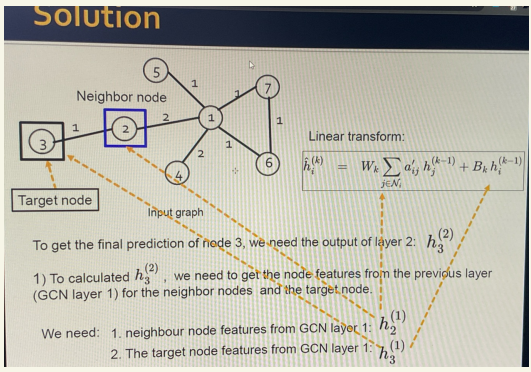
Solution



1) To calculated the output of node 3 for layer1 : $h_3^{(1)}$

We need: 1. neighbour node features (initial input): $h_2^{(0)}$
2. The node features of the current node $h_3^{(0)}$

Solution

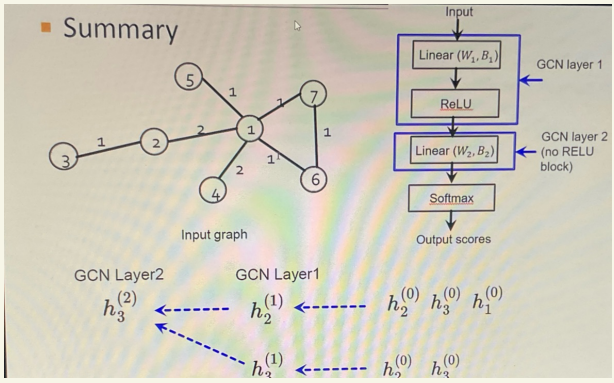


To get the final prediction of node 3, we need the output of layer 2: $h_3^{(2)}$

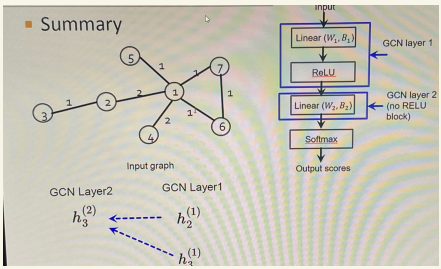
1) To calculated $h_3^{(2)}$, we need to get the node features from the previous layer (GCN layer 1) for the neighbor nodes and the target node.

We need: 1. neighbour node features from GCN layer 1: $h_2^{(1)}$
2. The target node features from GCN layer 1: $h_3^{(1)}$

Summary



Summary



Similarity Search Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Q1 LSH

- **Q1**: given the query sample and the samples in the database below, find the top 2 nearest neighbours in the database. Use LSH-random projection with the hash functions given below.

Query sample:

A	[0, -1]

Database samples (5):

B	[-2, 0]
C	[1, 2]
D	[2, 1]
E	[1, -1]
F	[-1, 2]

w_1 =[-1 1] ^{T};
w_2 =[-1 0] ^{T};
w_3 =[0 1] ^{T};
w_4 =[1 -1] ^{T};
w_5 =[1 0] ^{T};
w_6 =[-1 -1] ^{T};

Note: the question is modified from the LSH example in the lecture class by adding two additional hashing functions (w_5 and w_6)

Query sample:

A	[0, -1]

Database samples (5):

B	[-2, 0]
C	[1, 2]
D	[2, 1]
E	[1, -1]
F	[-1, 2]

$w_1 = [-1\ 1]^T$;
$w_2 = [-1\ 0]^T$;
$w_3 = [0\ 1]^T$;
$w_4 = [1\ -1]^T$;
$w_5 = [1\ 0]^T$;
$w_6 = [-1\ -1]^T$;

A $h_1 = w_1^T x_A = \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = -1$
$h_2 = w_2^T x_A = \begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = 0$
$h_3 = w_3^T x_A = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = -1$
$h_4 = w_4^T x_A = \begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = 1$
$h_5 = w_5^T x_A = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = 0$
$h_6 = w_6^T x_A = \begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = 1$

$h_1(x_A) = 0$
$h_2(x_A) = 0$
$h_3(x_A) = 0$
$h_4(x_A) = 1$
$h_5(x_A) = 0$
$h_6(x_A) = 1$

B $h_1 = w_1^T x_B = 2 = 2$
$h_2 =$
$h_3 =$
$h_4 =$
$h_5 =$
$h_6 =$

$h_1(x_B) = 1$
$h_2(x_B) = 1$
$h_3(x_B) = 0$
$h_4(x_B) = 0$
$h_5(x_B) = 0$
$h_6(x_B) = 1$

C $h_1 = \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = -1+2 = 1$
$h_2 = \begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = -1 \to 0$
$h_3 = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = 2 \to 1$
$h_4 = \begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = 1-2 = -1 \to 0$
$h_5 = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = 1+0 = 1$
$h_6 = \begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix} = -3 \to 0$

D $h_1 = \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = -1 \to 0$
$h_2 = \begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = -2 \to 0$
$h_3 = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = 1$
$h_4 = \begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = 1$
$h_5 = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = 2 \to 1$
$h_6 = \begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix} = -3 \to 0$

E $h_1 = \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = -2 \to 0$
$h_2 = \begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = -1 \to 0$
$h_3 = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = -1 \to 0$
$h_4 = \begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = 2 \to 1$
$h_5 = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix} = 1$
$h_6 = \begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}0\\-1\end{pmatrix} = -2 \to 0$

F $h_1 = \begin{pmatrix}-1\\1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = 1+2 = 3 \to 1$
$h_2 = \begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = 1$
$h_3 = \begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = 2 \to 0$
$h_4 = \begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = -1-2 = -3 \to 0$
$h_5 = \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = -1 \to 0$
$h_6 = \begin{pmatrix}-1\\-1\end{pmatrix}\begin{pmatrix}-1\\2\end{pmatrix} = 1-2 \to 0$

A	0	0	0	1	0	1
B	1	1	0	0	0	1
C	1	0	1	0	1	0
D	0	0	1	1	1	0
E	0	0	0	1	1	0
F	1	1	0	0	0	0

	B	C	D	E	F
A	3	5	3	2	4

L2 distances

A(0,-1]	B(-2,0)	C(1,2)	D(2,1)	E(1,-1)	F(-1,2)
	4+1=5	1+9=10	4+4=8	1+0=1	1+9=10

Q2 PQ

- **Q2. Product Quantization (PQ)**

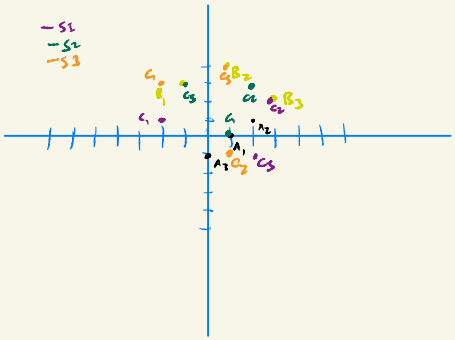
! study asymmetry !

Two input vectors are given below:
A: [1, 0, 2, 1, -1, 0]
B: [-1, 3, 1, 4, 3, 2]

Define 3 Subspaces. Each subspace has 3 centroids:
Subspace 1: C1: [-2, 1], C2: [3, 2], C3: [2, -1]
Subspace 2: C1: [1, 0], C2: [2, 3], C3: [-1, 3]
Subspace 3: C1: [-2, 3], C2: [1, -1], C3: [1, 4]

a): Use PQ to compute the compressed vectors of A and B

b): Construct distance lookup tables and calculate the approximate (symmetric case) squared-L2 distance of A and B using the compressed vectors.



Squared L2 distance

	A(1,0)	A(2,1)	A(1,0)	B(-1,3)	B(1,4)	B(3,2)
S1 C(-2,1)	9+1=10			1+9=5		
C(3,2)	4+4=8	C_3		16+1=17	C_1	
C(2,-1)	1+1=2			9+16=25		
S2 C(1,0)		1+1=2	C_1		0+16=16	
C(2,3)		0+4=4			1+1=2	C_2
C(-1,3)		9+4=13			4+1=5	
S3 C(-2,2)			1+9=10			25+1=26
C(1,-1)			4+1=5	C_2		4+9=13
C(1,4)			4+16=25			4+4=8

$A = [3, 1, 2]$

$B = [1, 2, 3]$

Symmetric distance

S1 $d(C_3, C_1) = 16+4=20$

S2 $d(C_1, C_2) = 1+9=10$

S3 $d(C_2, C_3) = 0+25=25$

$D(A, B) = 20 + 10 + 25 = 55$

Clustering - DBSCAN Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Question 1

Q: Given the data points below and the parameters: *epsilon=3, minPoints*=4, write down which of the points are core points, border points and noise points, respectively.

A(2,2) B(4,3) C(3,5) D(1,5)
E(-5,3) F(5,1) G(-4,6) H(-3,4) I(-2,6)

eps = 3
minpoints = 4

Start from A

A

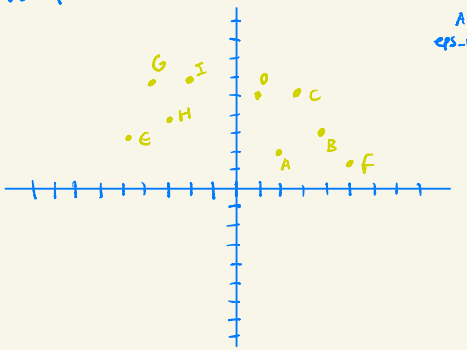
A is core

eps_radius: 5) = minPts(4)

q = A

↓

q = B C D F

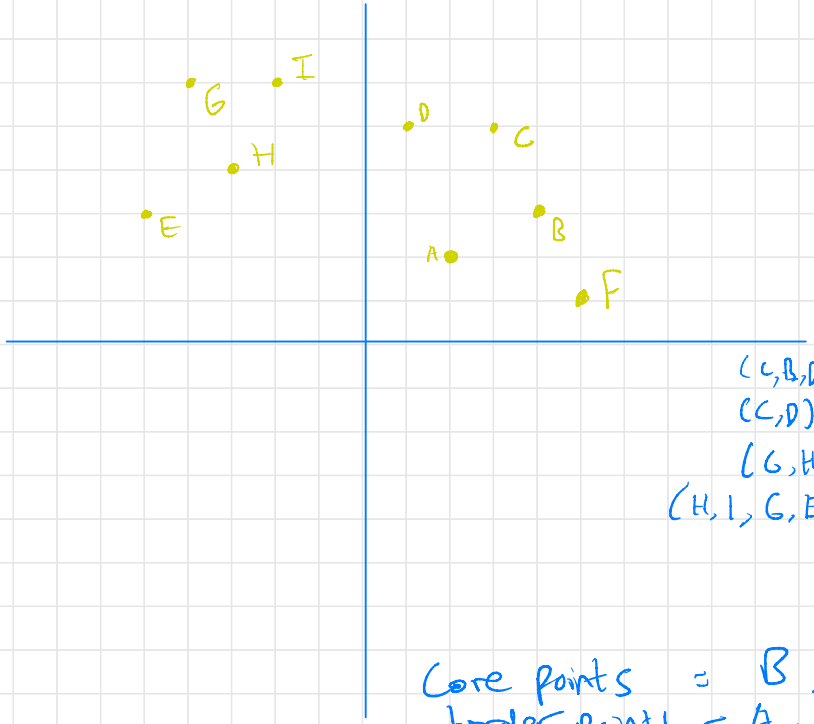


Solution

DBSCAN parameters: *epsilon*=3, *minPoints*=4

To simplify calculation, here we use squared L2 distance, so we need to compared the squared distance with squared epsilon: $\epsilon^2 = 3^2 = 9$

	A	B	C	D	E	F	G	H	I	#points within epsilon
A	0	5	10	10	50	10	52	29	32	2
B	5	0	5	13	81	5	73	50	45	4
C	10	5	0	4	68	20	50	37	26	3
D	10	13	4	0	40	32	26	17	10	2
E	50	81	68	40	0	104	10	5	18	2
F	10	5	20	32	104	0	106	73	74	2
G	52	73	50	26	10	106	0	5	4	3
H	29	50	37	17	5	73	5	0	5	4
I	32	45	26	10	18	74	4	5	0	3



A is not core point
is Border Point
(reach from B)

B is core point
eps_rad : 4 >= minPoints(4) (B, A, C, F)

F is border point to b
(C, B, D) C is border point eps_rad : 3 <= minPoints(4)
(C, D) D is noise point eps_rad : 2 <= minPoints(4)
(G, H) I is border point eps_rad 3 <= minPoints(4)
(H, I, G, E) H is core point eps_rad 4 >= minPoints(4)
G is border point
E is border point

Core points = B, H
border points = A, C, E, F, G, I
Noise points = D

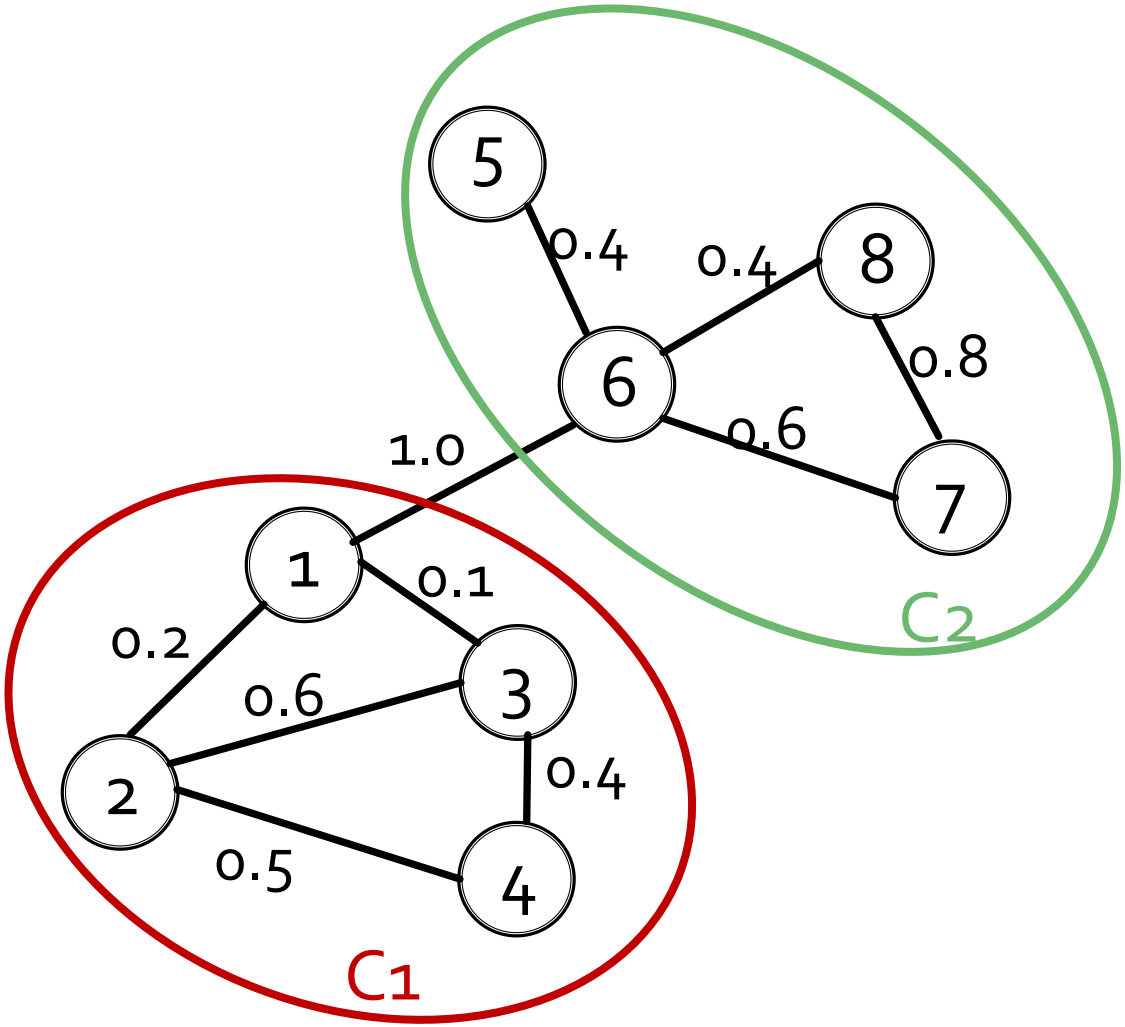
Graph community detection Tutorial

Lin Guosheng
School of Computer Science and Engineering
Nanyang Technological University

Question 1

Q: A graph is given below.
The current community assignment is:
C1={1, 2, 3, 4},
C2={5,6,7,8}.

Process node #1 to update the communities using Louvain Algorithm.


5
8
6
7
1
3
2
4
0.4
0.4
0.8
0.6
1.0
0.1
0.2
0.6
0.4
0.5
C2
C1

Process node #1

$\Delta Q(1 \to \{5,6,7,8\}) = Q_{after} - Q_{before}$

$= [Q(\{1,5,6,7,8\}) + Q(\{2,3,4\})] - [Q(\{5,6,7,8\}) + Q(\{1,2,3,4\})]$

$\Delta Q(1 \to 2) = Q_{after} - Q_{before}$

$= Q(1+2) - [Q(1) + Q(2)]$

$M = 12+1+2+2 = 5$

$Q(1) = -\left(\frac{k_1}{2m}\right)^2 = -\left(\frac{1.3}{2(5)}\right)^2 = -0.0169$

$Q(2) = -\left(\frac{k_1}{2m}\right)^2 = -\left(\frac{1.3}{2(5)}\right)^2 = -0.0169$

$\Sigma_{in} =$

$0.04 - 0.0676 = -0.0276$

$Q(\{1,2,3\}) = \frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m}\right)^2 = \frac{2.0.2}{2.5} - \left(\frac{2.6}{2.5}\right)^2 =$

$\Delta Q(1 \to 2) = Q_{After} - Q_{before}$

$= -0.0276 - [-0.0169 - 0.0169]$

$= 0.0062$

Lec

affinity matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix} \quad A' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.25 & 0 & 0.25 & 0.5 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 0 \end{pmatrix}$$

node 2

CZ4032 Data Analytics and Mining

Tutorial for Week 7: association rules

Q1 The following is a pseudocode of Apriori algorithm

1) Generate frequent itemsets of length 1
2) Repeat until no new frequent itemsets are identified
    a) Generate length (k+1) candidate itemsets from length k frequent itemsets
    b) Prune candidate itemsets containing subsets of length k that are infrequent
    c) Count the support of each candidate by scanning the DB
    d) Eliminate candidates that are infrequent, leaving only those that are frequent

Explain 2 b) with an example.

Q2 Illustrating Apriori Principle with an example.

Q3 In the lecture, we introduce how to generate length (k+1) candidate itemsets from length k frequent itemsets. Explain this with an example. Can give another way of generating candidates?

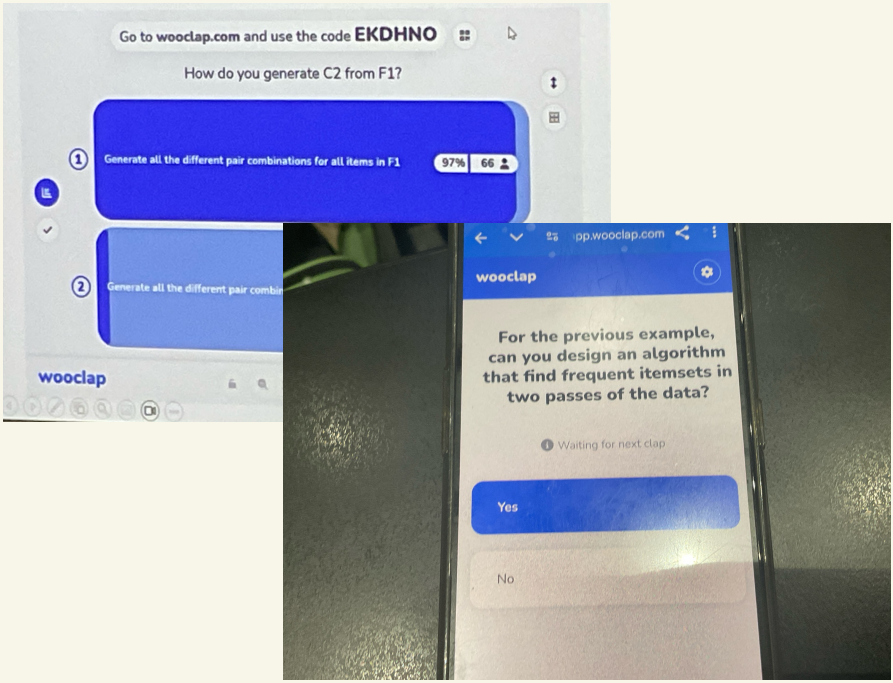
Q4 A dataset with 4 records

1: a, c,d
2, b, c,e
3, a, b, c, e
4, b, e,

Suppose minimum support is 2. Find frequent itemsets step-by-step.

Q5 Suppose {B,C,D} is a frequent itemset. Enumerate the candidate rules:

Q6 Consider the observation: If A,B,C→D is below confidence, so is A,B→C,D. Can we design an efficient order of generating rules based on the observation?


Go to wooclap.com and use the code EKDHNO
How do you generate C2 from F1?
1
Generate all the different pair combinations for all items in F1
97%
66
2
Generate all the different pair combi
wooclap
pp.wooclap.com
wooclap
For the previous example, can you design an algorithm that find frequent itemsets in two passes of the data?
Waiting for next clap
Yes
No

CZ4032 Data Analytics and Mining

Association Rule Mining: Tutorial 2

Q1 Explain the following observation for PCY algorithm

a. If a bucket contains a frequent pair, then the bucket is surely frequent
b. However, even without any frequent pair, a bucket can still be frequent

Q2 Given a dataset, minsup threshold, which of the following has the largest number of itemsets? Which has the smallest number of itemsets?

2) Frequent itemsets
3) Maximal frequent itemsets
4) Closed frequent itemsets

Q3 Discuss the impact of the following characteristics of a transaction table on the use of the FP tree to mine frequent itemsets from the table:

(a) Number of unique items in table

(b) Average number of items in a transaction

(c) Number of transactions in table

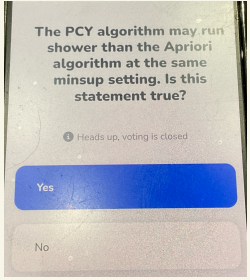
Q4 A database has four transactions. Let *min_sup* = 60% and *min_conf* = 80%.

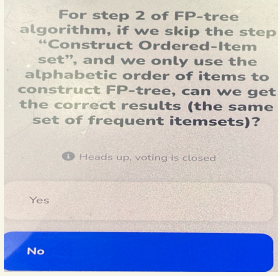
TID	Date	Items_bought
T100	20006-01-01	{*K*, *A*, *D*, *B*}
T200	20006-01-01	{*D*, *A*, *C*, *E*, *B*}
T300	20006-01-01	{*C*, *A*, *B*, *E*}
T400	20006-01-01	{*B*, *A*, *D* }

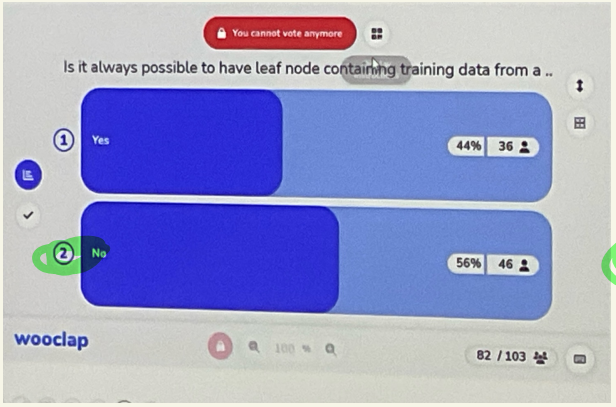
(a) Find all frequent itemsets using FP-growth.

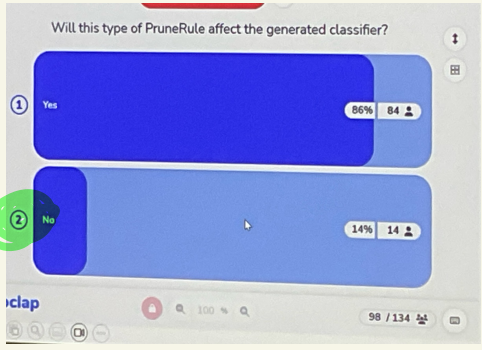
Q5: for sequence pattern mining, answer the following questions

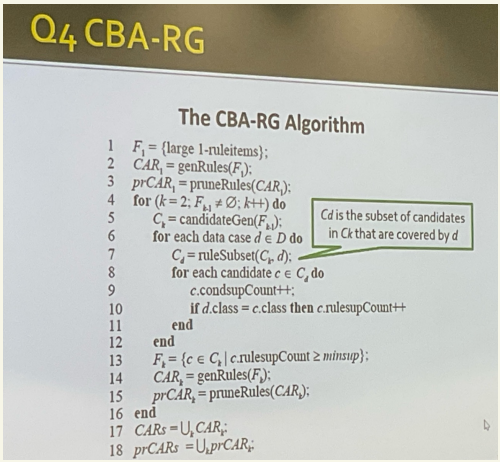
- Can <{a},{b},{c}> merge with <{b},{c},{f}> ?
- Can <{a},{b},{c}> merge with <{b,c},{f}>?
- Can <{a},{b},{c}> merge with <{b},{c,f}>?
- Can <{a,b},{c}> merge with <{b},{c,f}> ?
- Can <{a,b,c}> merge with <{b,c,f}>?
- Can <{a}{b}{a}> merge with <{b}{a}{b}>?
- Can <{b}{a}{b}> merge with <{a}{b}{a}> ?

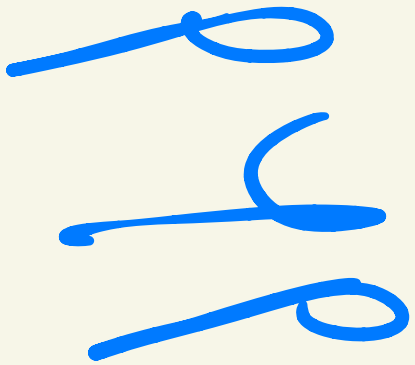

The PCY algorithm may run shower than the Apriori algorithm at the same minsup setting. Is this statement true?
Heads up, voting is closed
Yes
No


For step 2 of FP-tree algorithm, if we skip the step "Construct Ordered-Item set", and we only use the alphabetic order of items to construct FP-tree, can we get the correct results (the same set of frequent itemsets)?
Heads up, voting is closed
Yes
No


You cannot vote anymore
Is it always possible to have leaf node containing training data from a ..
1 Yes 44% 36
2 No 56% 46
wooclap
82 / 103


Will this type of PruneRule affect the generated classifier?
1 Yes 86% 84
2 No 14% 14
98 / 134


Q4 CBA-RG
The CBA-RG Algorithm
1 F_1 = {large 1-ruleitems};
2 CAR_1 = genRules(F_1);
3 $prCAR_1$ = pruneRules(CAR_1);
4 for (k = 2; $F_{k-1} \neq \emptyset$; k++) do
5 C_k = candidateGen(F_{k-1});
6 for each data case $d \in D$ do
7 C_d = ruleSubset(C_k, d);
8 for each candidate $c \in C_d$ do
9 c.condsupCount++;
10 if d.class = c.class then c.rulesupCount++
11 end
12 end
13 F_k = {$c \in C_k$ | c.rulesupCount ≥ $minsup$};
14 CAR_k = genRules(F_k);
15 $prCAR_k$ = pruneRules(CAR_k);
16 end
17 $CARs = \bigcup_k CAR_k$;
18 $prCARs = \bigcup_k prCAR_k$;
Cd is the subset of candidates in Ck that are covered by d



S1 2022-2023

1) a) - need to specify the initial

b) $h_{1a} = 1.0 + 2.1 + 2 - 1 = 0$

$h_{1b} = -2\ 0 + 1.1 + 4.-1 = -3 \rightarrow 0$

$h_{2a} = 1\ 1 + 2\ 0 + 2\ 0 = 1 \rightarrow$ diff from a

$h_{2b} = -2\ 1 + 1\ 0 + 4\ 0 = -2 \rightarrow 0$

hamming dist = 1

c)

	Case	C	W	default	error
r_2	3	3		Y	0+3=3
r_1	3	3		Y	0+3=3
r_5	2	1	1	Y	1+2=3
r_4	3	2	1	N	1+2=3
r_3					

A	B	C	Label
2	4	2	N
4	2	1	Y
4	4	2	Y
2	2	1	Y
1	2	1	N
2	3	3	N
4	1	4	N
4	3	3	N
3	4	2	Y
3	1	4	Y
3	3	4	N
3	2	2	N
2	4	1	Y
2	1	1	Y