COMP4433 Data Mining & Data Warehousing Applications

Assignment 2 (Reference Answers)

1. Social network data is usually modelled as a graph with nodes depicting users and edges showing the relationship between them. For the simplest relationship, a binary link (i.e. 0 or 1) is typically used to denote the existence of a relationship between users. For the graph depicted in Fig.1, users A, B, C and F have 2 friends individually, user D has friends B, C and E while user E has friends A, D and F. Such relationship can be represented by the adjacency matrix shown in Fig.2.

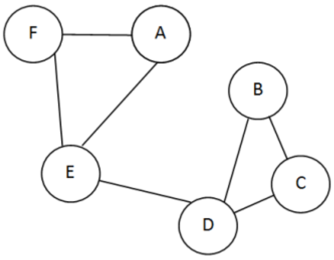


Fig.1 Social network graph with six nodes

$$\begin{array}{c|c} & \begin{matrix} A & B & C & D & E & F \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \end{matrix} & \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Fig.2 Adjacency matrix for the social network graph in Fig.1

a) Propose a dissimilarity metric for clustering the nodes in Fig.1 and prepare the corresponding dissimilarity matrix.

Ans. **[15 marks]**

There exist many solutions here. One can treat Fig.1 or Fig.2 as transactional data and use the dissimilarity function:

$$Dissim(T_1, T_2) = 1 - \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$$

On the other hand, one may treat the graph as binary attributed dataset and using "City Block Distance" to measure the dissimilarity, e.g.

$$d(A, B) = |0-0| + |0-0| + |0-1| + |0-1| + |1-0| + |1-0| = 4.$$

Note here that the City Block distance measures the difference between data objects and treats "common friend" (|1-1|) the same as "common unfriend".

With respect to Fig.2, we have

$$d_1(i,j) = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & & & & & \\ B & 4 & 0 & & & & \\ C & 4 & 2 & 0 & & & \\ D & 3 & 3 & 3 & 0 & & \\ E & 3 & 3 & 3 & 6 & 0 & \\ F & 2 & 4 & 4 & 3 & 3 & 0 \end{array}$$

b) Based on the dissimilarity matrix in part(a), use the **single linkage agglomerative hierarchical clustering** algorithm to cluster the six social network users. Show your steps.

Ans. **[20 marks]**
Based on the dissimilarity matrix $d_1(i,j)$ shown in part (a),
we can either merge B and C or A and F, e.g., BC

$$d_2(i,j) = \begin{array}{c|ccccc} & A & BC & D & E & F \\ \hline A & 0 & & & & \\ BC & 4 & 0 & & & \\ D & 3 & 3 & 0 & & \\ E & 3 & 3 & 6 & 0 & \\ F & 2 & 4 & 3 & 3 & 0 \end{array}$$

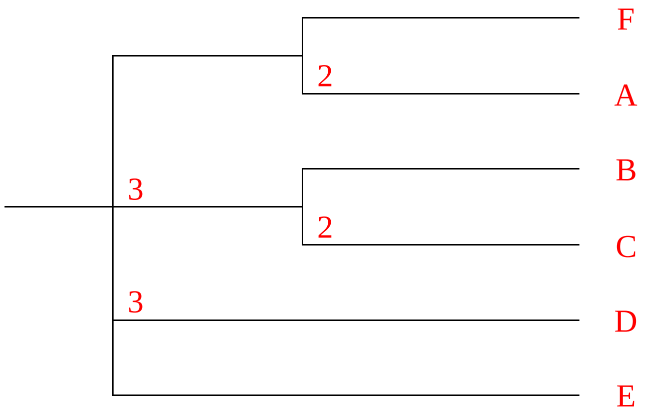
And then merge A and F to have

$$d_3(i,j) = \begin{array}{c|cccc} & AF & BC & D & E \\ \hline AF & 0 & & & \\ BC & 4 & 0 & & \\ D & 3 & 3 & 0 & \\ E & 3 & 3 & 6 & 0 \end{array}$$

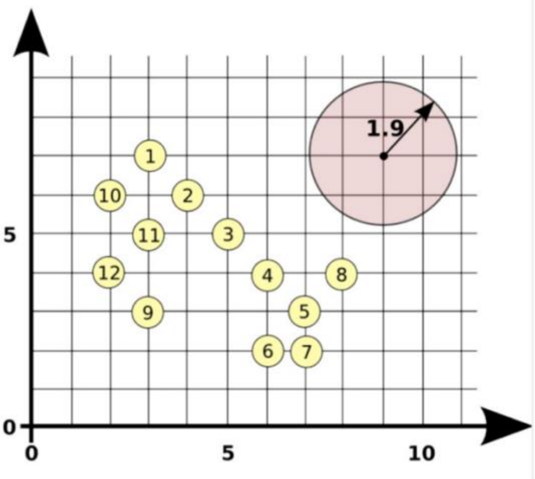
We continue to merge AF and D (or other options wrt $d_3(i,j)=3$) to have

$$d_4(i,j) = \begin{array}{c|ccc} & AFD & BC & E \\ \hline AFD & 0 & & \\ BC & 3 & 0 & \\ E & 3 & 3 & 0 \end{array}$$

And we can see that eventually all the three clusters/nodes AFD, BC and E will be merged at dissimilar value=3. The corresponding dendrogram is depicted as follows.



2. Following the tutorial question on spatial clustering, i.e. applying DBSCAN with e=1.9 and MinPts=4 to the data recapped below, answer the following questions.

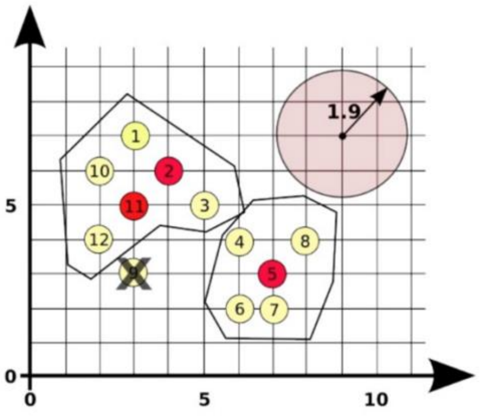


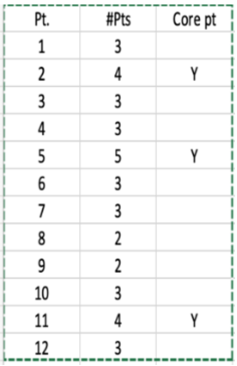
a) Add one more data point (e.g. point 13) so that only one cluster is formed.
b) Following (a), update the list of core points, border points and noise points accordingly.
c) For MinPts=4 specified above, specify a change of *e* value so that there will have NO noise point.
d) For *e* =1.9, specify a change of MinPts value so that there will have only ONE cluster.
e) For *e* =1.9 and MinPts=4 specified above, add one more point so that there will have NO noise point.

Ans. **[Total 25 marks; 5 marks for each part]**

a) Consider the time we form a cluster for core point 2:
Point 3 is not a core point and hence the growing process to the right stopped. So, if it is a core point together with having point 4 as another core point, the 2 original core points 2 and 5 will be density connected and one large cluster will be form. In order to do so, we can add a data point in between 3 and 4, say with coordinate (5.5, 4.5), and points 3 and 4 will become core.
b) The change is minimal, with points 3 and 4 changed from border point to core point. All others remain the same.
c) One simple change is to use a larger neighborhood radius, e.g. ε=2.1.
d) Decrease MinPts to 3 or 2.
e) There exists one noise data point 9 here. So, add a point so that point 9 can be density reachable to, e.g., a point at (3,4).

Note that the answer is not unique.
You may refer to our tutorial answer recapped below.





Pt.	#Pts	Core pt
1	3	
2	4	Y
3	3	
4	3	
5	5	Y
6	3	
7	3	
8	2	
9	2	
10	3	
11	4	Y
12	3	

3. Your R&D team has been assigned a project to carry out price movement classification on the stock data. After pre-processing the collected numeric data, the following database is given:

Stock Price Movement Database

Stock	Price Movement from 19 Oct. – 30 Oct., 2015									
	19 Oct	20 Oct	21 Oct	22 Oct	23 Oct	26 Oct	27 Oct	28 Oct	29 Oct	30 Oct
PCCW	*Up*	*Up*	*Level*	*Down*	*Level*	*Up*	*Up*	*Down*	*Level*	*Up*

where the movement labels *Up*, *Down* & *Level* denote the stock price going up, down and level respectively in the corresponding trading day. In order to classify next trading day's price movement, the stock data above is extracted as follows.

Extracted Stock Price Movement Database for Classification

Today is	Price Movement of PCCW for			
	2 Trading Day before (2TDB)	1 Trading Day before (1TDB)	Today (TD)	Next Trading Day (NTD)
21 Oct	*Up*	*Up*	*Level*	*Down*
22 Oct	*Up*	*Level*	*Down*	*Level*
23 Oct	*Level*	*Down*	*Level*	*Up*
26 Oct	*Down*	*Level*	*Up*	*Up*
27 Oct	*Level*	*Up*	*Up*	*Down*
28 Oct	*Up*	*Up*	*Down*	*Level*
29 Oct	*Up*	*Down*	*Level*	*Up*

with the last column NTD as the **class attribute**.

a) Suppose you are asked to adopt the decision tree to classify the given stock data. Show how the last two rows (i.e., Today is 28 Oct. & 29 Oct. resp.) are classified when all the seven data records above are used for training.

$$\log_2 x = \log_{10} x / \log_{10} 2 \cong \log_{10} x / 0.30103$$

$$I(c_1, c_2, c_3) = -\frac{c_1}{c_1 + c_2 + c_3}\log_2 \frac{c_1}{c_1 + c_2 + c_3} - \frac{c_2}{c_1 + c_2 + c_3}\log_2 \frac{c_2}{c_1 + c_2 + c_3} - \frac{c_3}{c_1 + c_2 + c_3}\log_2 \frac{c_3}{c_1 + c_2 + c_3}$$

$$I(1,0,0) = I(2,0,0) = I(0,2,0) = 0$$

$$I(1,1,0) = I(1,0,1) = 1$$

$$I(1,2,0) = I(0,1,2) \cong 0.918$$

$$I(1,2,1) = 1.5$$

$$I(2,3,2) \cong 1.557$$

Ans. [Total 40 marks; Part(a): 25 marks, Part(b): 15 marks]

Determining the root attribute:

I(3,2,2)=1.557

Entropy for 2TDB

2TDB	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	I($\#_{Up}$, $\#_{Level}$, $\#_{Down}$)
Up	1	2	1	1.5
Level	1	0	1	1
Down	1	0	0	0

Entropy(2TDB)=(4/7)*I(1,2,1) + (2/7)*I(1,0,1) + (1/7)*I(1,0,0)≈1.143

Information_Gain(2TDB)=1.557-1.143=0.414

Entropy for 1TDB

1TDB	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	I($\#_{Up}$, $\#_{Level}$, $\#_{Down}$)
Up	0	1	2	0.918
Level	1	1	0	1
Down	2	0	0	0

Entropy(1TDB)=(3/7)*I(0,1,2) + (2/7)*I(1,1,0) + (2/7)*I(2,0,0)≈0.679

Information_Gain(1TDB)=1.557-0.679=0.878

Entropy for TD

TD	$\#_{Up}$	$\#_{Level}$	$\#_{Down}$	I($\#_{Up}$, $\#_{Level}$, $\#_{Down}$)
Up	1	0	1	1
Level	2	0	1	0.918
Down	0	2	0	0

Entropy(TD)=(2/7)*I(1,0,1) + (3/7)*I(2,0,1) + (2/7)*I(0,2,0)≈ 0.679

Information_Gain(TD)=1.557-0.679=0.878

Hence, either TD or 1TDB can be chosen as the root node.

Let TD be the root node.

The “Down” branch will terminate.

For the “Up” branch, we have I(1,1,0)=1

Obviously, any of 2TDB and 1TDB can be used to differentiate the 2 samples for NTD=Up and NTD=Down

For the “Level” branch, we have I(2,0,1)=0.918

By inspection, we can easily conclude that we should choose 1TDB as the intermediate node because when 1TDB=”Down”, the two samples are NTD=”Up” and when 1TDB=”Up”, the only one sample is NTD=”Down”.

Hence, we have 5 rules generated. One version is:

R1: IF TD=”Down” THEN NTD=”Level” (2 samples)

R2: IF TD=”Up” AND 1TDB=“Level” THEN NTD=”Up” (1 samples)

R3: IF TD=”Up” AND 1TDB=“Up” THEN NTD=”Down” (1 samples)

R4: IF TD=”Level” AND 1TDB=“Down” THEN NTD=”Up” (2 samples)

R5: IF TD=”Level” AND 1TDB=“Up” THEN NTD=”Down” (1 samples)

To classify the last two rows for Today is 28 Oct. & 29 Oct., we have

28 Oct.: matching the rule IF TD=”Down” THEN NTD=”Level”

Hence, the predicted class is “Level” which is correct.

29 Oct.: matching the rule IF TD=”Level” AND 1TDB=“Down” THEN NTD=”Up”

Hence, the predicted class is “Up” which is correct.

(25 marks)

b) Based on the classifier in part (a), think about and show how the following two cases can be properly classified, i.e., to classify next trading day’s price movement with missing data (empty boxes). Justify your way to do so.

Today is	2 Trading Day before	1 Trading Day before	Today	Next Trading Day
2 Nov. 2015	*Level*	*Up*		?
3 Nov. 2015	*Up*			?

Ans.:

For the first record (Today is 2 Nov. 2015), all rules satisfying

1. 2TDB=”Level”, 1TDB=”Up” and TD=”Up”
2. 2TDB=”Level”, 1TDB=”Up” and TD=”Level”
3. 2TDB=”Level”, 1TDB=”Up” and TD=”Down”

will be matched and hence R1, R3 and R5 are selected and hence NTD=”Down” by majority rule.

For the second record (Today is 3 Nov. 2015), all rules satisfying

2TDB=”Up”, 1TDB=Up/Level/Down and TD= Up/Level/Down

will be matched and hence for our R1-R5 all rules will be matched. Thus, we may classify the records as NTD=”Up” by majority rule (3 Up, 2 Level and 2 Down).

Also, you may want to use your predicted class label of 2 Nov. 2015, i.e., “Down” to substitute 1TDB of 3 Nov. 2015 to carry out the classification. So, with TD=Down/Level/Up and 1TDB=Down, R1 and R4 will be matched:

R1: IF TD=”Down” THEN NTD=”Level” (2 samples)

R4: IF TD=”Level” AND 1TDB=“Down” THEN NTD=”Up” (2 samples)

So, the prediction should be Level/Up

We may have another conclusion for a different decision tree formed in part(a), i.e., using 1TDB as root node.

(15 marks)