COMP4433 Data Mining & Data Warehousing

Assignment 2 (Due: 23:59, 21 Nov 2023 Tuesday)

Instructions:
- Answer all questions.
- Interpret the questions logically, show your steps and write down your assumption(s) when necessary.
- Please submit your answer to L@PU before the due date.
- Late Submission Policy
  - 3-hour "grace period" is given.
  - 10% off for every 3-hour late
- Plagiarism Policy
  - Both giver and receiver subject to the same penalty below
  - All the students involved will receive 0 marks for this assessment. In addition, they will receive an additional 50% penalty, e.g., 5 marks for a 10-mark assessment.

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.

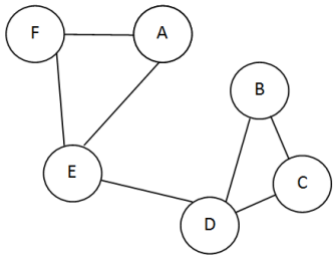


Fig.1 Social network graph with six nodes

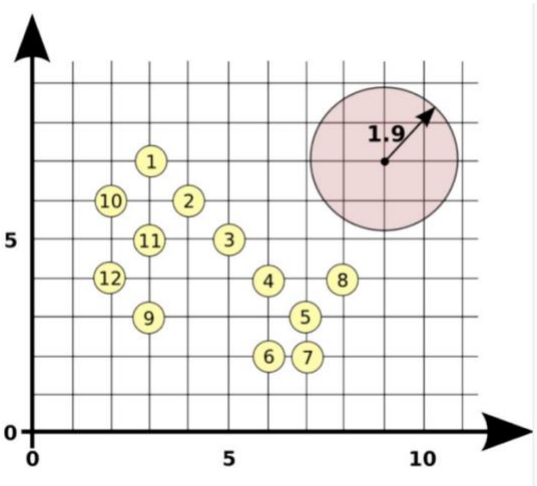
$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}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

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**.

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.

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.

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
BYD	*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.

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

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

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

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

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

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

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

b) Based on your 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). No guessing/inference of the missing attribute data is allowed, i.e., classify the data based on your classifier in part (a). Justify your way to do so.

Today is	2TDB	1TDB	TD	NTD
2 Nov. 2015	*Level*	*Up*		?
3 Nov. 2015	*Up*			?