Master of Computer Applications

MCAO 302: Data Mining
Unique Paper Code: 223403302
Semester III
December-2022
Year of admission: 2021

Time: Three Hours **Maximum marks: 70**

Note: Answer all the questions.

1. Consider a movie dataset in which movies belong to one or more genres, as illustrated in the following table. The last column of the table contains feedback provided by a particular user. '?' denotes that the user has not yet watched the movie.

MID	Movie Genres	Feedback
M1	Comedy, Romance	Dislike
M2	Comedy, Drama, Romance, Action	Dislike
M3	Comedy, Drama	Dislike
M4	Thriller, Action	Like
M5	Drama, Thriller, Action, Horror	Like
M6	Action, Horror	Like
M7	Thriller, Horror	?
M8	Drama, Romance	?

$\frac{4}{10} \times 6$

2.4

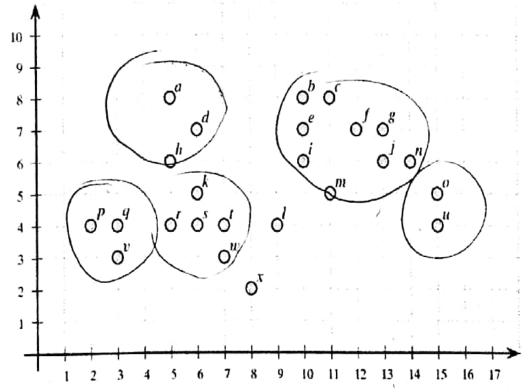
$\frac{\cancel{75}^{3}}{\cancel{100}} \times \cancel{6}^{3} \times 2$

$\frac{9}{2}$

(a) Mine all the rules using the Apriori algorithm with at least 40% support and 75% confidence. Based on these rules, would you recommend the movie M7 or M8 to the user? [8]

(b) Explain whether or not the running time can be further improved by replacing Apriori algorithm with a Dynamic Itemset Counting algorithm that stops after every 3 transaction to add more itemsets. [6]

2. (a) (i) Show the density-based clusters and the noise points over the data given in figure below. Use Manhattan distance ($Distance([x_1, y_1)], [x_2, y_2]) = |x_1 - x_2| + |y_1 - y_2|$) to calculate the distance between two points, $\epsilon = 2$ and $minpts = 4$. [5]



(ii) Is o density reachable from i? Show the intermediate points on the chain or the point where the chain breaks. [1]

(iii) List all the core points. [2]

(b) Describe each of the following clustering algorithms in terms of: shapes of clusters that can be determined; input parameters that must be specified; and limitations. [6]

(i) k-medoids

(ii) CLARA

(iii) BIRCH

3. (a) Explain why data analysts need to normalize their numeric variables. Use the methods below to normalize the following group of data. [7]

$$200, 300, 400, 600, 1000$$

(i) min-max normalization by setting $min = 0$ and $max = 1$

(ii) z-score normalization

(iii) normalization by decimal scaling

(b) Consider the 1-dimensional data set with 10 data points {1, 2, 3,... 10}. Show three iterations of the k-means algorithms when k = 2, and the random seeds are initialized to {1, 2}. Does the algorithm discover the correct clusters if the initial centers are changed to {2, 9}? What does this tell you about the running time? [7]

4. (a) True or False: Two different decision trees (constructed using different methods) that both correctly classify all the examples in a given training set will also classify any other testing example in the same way (i.e., both trees will predict the same class for any other example). [1]

(b) Suppose we would like to convert a categorical attribute X with 4 values to a data table with only binary variables. How many new attributes are needed? [1]

(c) Suppose there are eight items, A, B, . . . , H, and the following are the maximal frequent itemsets: {A, B}, {B, C}, {A, C}, {A, D}, {E}, and {F}. Find the negative border. [2]

(d) Suppose ABC is a frequent itemset and $BCDE$ is NOT a frequent itemset. Given this information, we can be sure that certain other itemsets are frequent and sure that certain itemsets are NOT frequent. Others may be either frequent or not. Which of the following statement(s) is/are correct. [2]

(i) ACD is certainly not frequent

(ii) ABCD can be either frequent or not frequent

(iii) BCE is certainly not frequent

(iv) BC is certainly frequent

(e) You have a friend who only does one of four things on every Saturday afternoon: go shopping, watch a movie, play tennis, or just stay in. You have observed your friend's behavior over 11 different weekends. On each of these weekends you have noted the weather (sunny, windy, or rainy), whether her parents visit (visit or no-visit), whether she has drawn cash from an ATM machine (rich or poor), and whether she had an exam during the coming week (exam or no-exam). You have to build a decision tree to predict your friend's activity. What feature you will select at the root level for both *Entropy* and *Gini Index* splitting criterion? [8]

# ex.	Weather	Parents	Cash	Exam	Decision
1	sunny	visit	rich	yes	cinema
2	sunny	no-visit	rich	no	tennis
3	windy	visit	rich	no	cinema
4	rainy	visit	poor	yes	cinema
5	rainy	no-visit	rich	no	stay-in
6	rainy	visit	poor	no	cinema
7	windy	no-visit	poor	yes	cinema
8	windy	no-visit	rich	yes	shopping
9	windy	visit	rich	no	cinema
10	sunny	no-visit	rich	no	tennis
11	sunny	no-visit	poor	yes	tennis

c, T, S-i, S

5. (a) Consider this training data set. Examples are A-E, and the single attribute is X.

Example	A	B	C	D	E
Attribute Value (X)	0.1	0.6	0.8	2.0	3.0

Draw the dendogram that results from applying hierarchical agglomerative clustering to this data. The similarity between two clusters is measured using centroid distance. [3]

(b) What are the steps of knowledge discovery in databases? [4]

(c) What is the basic principle of FP-Tree algorithm? Describe its important step. Construct FP-Tree for the data given below. [7] Support count = 2

TID	Items_bought
T100	umbrella, detergent, milk, cheese, bread, diaper
T101	bread, water, cheese, umbrella
T102	water, beer, umbrella, detergent
T103	water, milk, detergent
T104	water, bread, umbrella, diaper, milk, detergent
T105	bread, detergent
T106	milk, diaper, cheese, bread
T107	detergent, bread, umbrella
T108	beer, diaper, detergent
T109	bread

De 9
Br 7
U 5
M 4
Di 4
W 4
C 3
Be 2