Master of Computer Applications

MCAO 302: Data Mining
Unique Paper Code: 223403302
Semester III
December-2021 (OBE)
Year of admission: 2020

Time: 3 Hours + 1 Hour for uploading **Maximum marks: 70**

Note: Answer any 4 questions. All questions carry equal marks.

1. Consider a retailer that sells items, namely *bread, butter, juice, milk,* and *soda*. He has observed a decline in sales in the recent past because of poor inventory management. He then consulted an inventory management engineer to forecast items to be kept in stock. The engineer then collected the past sales data (Table 1) and started analyzing for the frequent pattern. He decided to use the frequent-pattern (FP) growth approach with minimum support of 3.

Table 1: Transaction Database

TID	List of Items
1	butter, juice, milk
2	bread, milk, soda
3	bread, butter, juice
4	bread
5	bread, butter, juice
6	bread, juice, milk, soda
7	bread, butter, milk
8	bread, butter, juice, milk
9	butter, juice, soda
10	bread, butter

He wants you to parallelly construct the FP-tree and mine the constructed tree for the frequent patterns to verify the calculations. He also wants you to generate all the possible rules with an associated confidence value. Provide a step-by-step description of the construction process.

2. In the game of cricket, players' fitness is one of the major concerns. Most often, players get injured because of bad playing conditions. To overcome this, the club has hired you to forecast the playing conditions based on the weather data. The club also provided you with the weather data for the last seven days (Table 2, Decision attribute: *Play*). You then decided to built a decision tree classifier that is at least 3 levels deep (root at level 1) with information gain as an attribute test condition. Write the step-by-step decision tree construction process.

Table 2: Weather Dataset

Humidity	Outlook	Windy	Play
Low	Overcast	No	Yes
Low	Overcast	Yes	Yes
High	Sunny	No	Yes
High	Sunny	Yes	No
Low	Sunny	No	No
Low	Sunny	Yes	Yes
High	Overcast	No	No

Suppose that the data entry operator has missed entering the first row's value for Humidity. Describe two ways in which you can handle the missing value. Further, generate the rules from the constructed decision tree?

3. Alice, Bob, Georgia, Juliet, and Lily are students attending the class of Data Mining. For one of assignment related to the performance of agglomerative clustering with various cluster similarity measure, namely *single-link, complete-link, average-link*, and *Centroid distance* they have collected their monthly expenses (in thousand) on movie and travel. The data is shown in the table below.

Table 3: Expense Dataset I

Person	movie	travel
Alice	2	4
Bob	8	2
Georgia	9	3
Juliet	1	5
Lily	8.5	1

They decided to use Euclidean distance to calculate the similarity. To verify the assignment, you have to write the step-by-step hierarchy of clustering created by the single-link, complete-link, average-link, and centroid-based cluster similarity measure. Also, write your observation about the pros and cons of the four cluster similarity measures.

4. (a) We are given with the age of sixteen students with the following values:

$$10, 11, 13, 14, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75.$$

Apply the equal width and equal frequency binning to map the values to three bins. (3)

(b) Clustering algorithm can be used for the discretization of continuous variable. TRUE or FALSE? Justify your answer. (3)

(c) Explain predictive and descriptive data mining techniques. Why clustering is called descriptive data mining task. (3)

(d) You have to group eight individuals using the principle of DBSCAN algorithm based on the data related to their expenses (in thousand) over food and cloths (Table 4). Let, *epsilon* and *minpoint*, the two user defined parameters, are given as 2. What are the clusters that DBSCAN would discover? (8.5)

Table 4: Expense Dataset II

Person	Food	Cloth
Aahan	2	10
Parth	2	5
Aakil	8	4
Kiaan	5	8
Ram	7	5
Shankar	6	4
Amit	1	2
Sandeep	4	9

5. (a) Why is it important to consider density when clustering a dataset? Illustrate your argument(s) with examples. (3.5)

(b) Explain how you can address the overfitting problem in decision tree construction. (3.5)

(c) Given a training set with 6+ and 16− examples. What is the entropy value and Gini associated with this data set? (3.5)

(d) Consider a binary class classification problem with class label *Cat*, and *Dog*. Suppose we train a model to predict whether an image is of *Cat* or *Dog*. After training the model, we apply it to a test-set of 100 new images (also labeled) and the model produces the following contingency table.

		Ground Truth	
		Cat	Dog
Predicted	Cat	5	25
Class	Dog	20	50

Compute the *precision, recall, specificity*, and *f-measure* of this model. (3.5)

(e) For two runs of *K-mean* clustering is it expected to get same clustering results? Yes or No? Justify your answer. (3.5)

6. (a) Consider the relationship between support and confidence of association rules, which of the following is true? (1.5)

   (i) Confidence is always larger than or equal to support

   (ii) Support is always larger than or equal to confidence

   (iii) Support and confidence have no necessary relationships

   (iv) Support depends on confidence, but not the other way around

   (v) Confidence depends on support, but not the other way around

   (vi) None of the above

(b) You can list all frequent itemsets and their support of a data set if you know all maximal itemsets of the data set. TRUE or FALSE? Justify your answer. (2)

(c) Let $MFCS = \{\{D, E, F, G, H\}, \{C, E, F, G, H\}, \{B, D\}, \{B, C\}\}$ be the set of maximal frequent candidates (MFCS) and $S_2 = \{\{C, F\}, \{C, H\}\}$ be the set of infrequent itemsets. Find out the updated MFCS. (3.5)

(d) In predictive classification, the quality of the classifier is based solely on how well it classifies the training data. TRUE or FALSE? Justify your answer. (3.5)

(e) Discuss issues that are important to consider when employing a Decision Tree based classification algorithm. (3.5)

(f) Can we use custering algorithm for dimensionality reduction task? Yes or No? Justify your answer. (3.5)