

5 Years Integrated MCA, Utkal University
End Semester Examination, March 2021

Semester: 9th
Full Mark:70

Sub: DM & DW (9.1)
Time: 3Hours

Answer all Questions

1	a)	Explain the components and architecture of Data Mining system along with the basic steps of KDD.	10												
	b)	In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.	4												
		OR													
	c)	Explain the different functionalities of Data Mining System.	7												
	d)	What is noisy data? Explain the different smoothing methods to handle the noisy data during data cleaning with suitable examples.	7												
2	a)	What is OLTP? Explain the differences between OLAP and OLTP.	7												
	b)	Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. i) Draw a schema diagram for the above data warehouse using star and snowflake schema. ii) Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2019?	7												
		OR													
	c)	Explain the three-tier architecture of Data warehouse.	7												
	d)	Write short notes on the followings: i) Features of data warehouse ii) Concept hierarchy	7												
3		What is Apriori property? A database has five transactions. Let min-support=60% and min-confidence=80%. <table border="1" style="margin-left: 20px;"> <thead> <tr> <th>TID</th> <th>Items-bought</th> </tr> </thead> <tbody> <tr> <td>T1</td> <td>{P, Q, R, S, T, V}</td> </tr> <tr> <td>T2</td> <td>{W, Q, R, S, T, V}</td> </tr> <tr> <td>T3</td> <td>{P, A, S, T}</td> </tr> <tr> <td>T4</td> <td>{P, U, E, S, V}</td> </tr> <tr> <td>T5</td> <td>{E, Q, Q, I, T}</td> </tr> </tbody> </table> Find all frequent item sets using Apriori Algorithm and determine the strong rules from the frequent item sets	TID	Items-bought	T1	{P, Q, R, S, T, V}	T2	{W, Q, R, S, T, V}	T3	{P, A, S, T}	T4	{P, U, E, S, V}	T5	{E, Q, Q, I, T}	14
TID	Items-bought														
T1	{P, Q, R, S, T, V}														
T2	{W, Q, R, S, T, V}														
T3	{P, A, S, T}														
T4	{P, U, E, S, V}														
T5	{E, Q, Q, I, T}														
		OR													
	a)	Explain the concept of support and confidence of an association rule with suitable examples.	4												
	b)	Write the drawbacks associated with Apriori algorithm. Find the frequent	10												

		item sets from the dataset given in Q.3 using FP-Growth algorithm.																																													
4	a)	What is classification? Explain the steps of designing a classification model.	4																																												
	b)	Explain the back propagation learning algorithm used for classification.	10																																												
		OR																																													
	c)	Explain the different criteria for comparing the classification and prediction methods.	4																																												
	d)	<p>The following table consists of training data. Construct a decision tree from these data using the basic tree induction algorithm. Classify the records with “status” attribute. Write down the rules that are generated from this tree.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Department</th> <th>Avg_range</th> <th>Salary_class</th> <th>Status</th> </tr> </thead> <tbody> <tr> <td>sales</td> <td>middle-aged</td> <td>high</td> <td>senior</td> </tr> <tr> <td>sales</td> <td>young</td> <td>low</td> <td>junior</td> </tr> <tr> <td>sales</td> <td>middle-aged</td> <td>low</td> <td>junior</td> </tr> <tr> <td>systems</td> <td>young</td> <td>high</td> <td>junior</td> </tr> <tr> <td>systems</td> <td>middle-aged</td> <td>high</td> <td>senior</td> </tr> <tr> <td>systems</td> <td>senior</td> <td>high</td> <td>Senior</td> </tr> <tr> <td>marketing</td> <td>middle-aged</td> <td>high</td> <td>Senior</td> </tr> <tr> <td>marketing</td> <td>middle-aged</td> <td>average</td> <td>Junior</td> </tr> <tr> <td>secretary</td> <td>senior</td> <td>average</td> <td>Senior</td> </tr> <tr> <td>secretary</td> <td>young</td> <td>low</td> <td>junior</td> </tr> </tbody> </table>	Department	Avg_range	Salary_class	Status	sales	middle-aged	high	senior	sales	young	low	junior	sales	middle-aged	low	junior	systems	young	high	junior	systems	middle-aged	high	senior	systems	senior	high	Senior	marketing	middle-aged	high	Senior	marketing	middle-aged	average	Junior	secretary	senior	average	Senior	secretary	young	low	junior	10
Department	Avg_range	Salary_class	Status																																												
sales	middle-aged	high	senior																																												
sales	young	low	junior																																												
sales	middle-aged	low	junior																																												
systems	young	high	junior																																												
systems	middle-aged	high	senior																																												
systems	senior	high	Senior																																												
marketing	middle-aged	high	Senior																																												
marketing	middle-aged	average	Junior																																												
secretary	senior	average	Senior																																												
secretary	young	low	junior																																												
5	a)	<p>Suppose the data mining task is to cluster the following eight points into three clusters:</p> <p>A1(2, 10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9)</p> <p>The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 As the center of each cluster respectively. Write the K-means algorithm and use it to show only the three cluster centers after the first round execution.</p>	10																																												
	b)	Give a broad classification of different clustering algorithms.	4																																												

		OR	
	c)	Write the strengths and weaknesses of k-means clustering algorithm. Explain the k-medoid clustering algorithm.	10
	d)	Describe briefly about binary variables.	4