PVP 20

Code: 20IT4501E

III B.Tech - I Semester – Regular Examinations - DECEMBER 2022

DATA MINING
(INFORMATION TECHNOLOGY)

Duration: 3 hours Max. Marks: 70

Note: 1. This paper contains questions from 5 units of Syllabus. Each unit carries 14 marks and have an internal choice of Questions.
2. All parts of Question must be answered in one place.

BL – Blooms Level CO – Course Outcome

			BL	CO	Max. Marks
UNIT-I					
1	a)	Write a brief note on relational databases and data warehouses.	L2	CO1	5 M
	b)	Describe the data mining functionalities, and the kinds of patterns they can discover.	L2	CO1	9 M
OR					
2	a)	Describe the various phases in knowledge discovery process with a neat diagram.	L2	CO1	10 M
	b)	Explain how the evolution of database technology led to data mining.	L2	CO1	4 M
UNIT-II					
3	a)	What is the curse of dimensionality? How to reduce it?	L3	CO2	7 M
	b)	Illustrate binning methods for data smoothing.	L3	CO2	7 M
OR					
4	a)	In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.	L3	CO2	8 M

<table>
<tr><td></td><td>b)</td><td>Discuss issues to consider during data integration.</td><td>L2</td><td>CO2</td><td>6 M</td></tr>
<tr><td colspan="6">UNIT-III</td></tr>
<tr><td rowspan="2">5</td><td>a)</td><td>Can we design a method that mines the complete set of frequent item sets without candidate generation? Explain with example.</td><td>L3</td><td>CO3</td><td>9 M</td></tr>
<tr><td>b)</td><td>Prove that all nonempty subsets of a frequent itemset must also be frequent.</td><td>L4</td><td>CO3</td><td>5 M</td></tr>
<tr><td colspan="6">OR</td></tr>
<tr><td rowspan="2">6</td><td>a)</td><td>How are association rules generated from frequent itemsets? Illustrate.</td><td>L3</td><td>CO3</td><td>7 M</td></tr>
<tr><td>b)</td><td>Apply FP-Growth algorithm to the following transactional data to find frequent itemsets. List all frequent itemsets with their minimum support count of 2 and confidence = 50%.
<table>
<tr><th>TID</th><th>List of Item IDs</th></tr>
<tr><td>1</td><td>i1,i3,i5,i7</td></tr>
<tr><td>2</td><td>i2,i4,i6,i8</td></tr>
<tr><td>3</td><td>i1,i3,i5,i7</td></tr>
<tr><td>4</td><td>i9,i7,i5,i1</td></tr>
<tr><td>5</td><td>i2,i4,i6,i7</td></tr>
<tr><td>6</td><td>i1,i2,i3,i4</td></tr>
<tr><td>7</td><td>i3,i4,i5,i6</td></tr>
<tr><td>8</td><td>i7,i8,i6,i1</td></tr>
<tr><td>9</td><td>i8,i5,i3,i2</td></tr>
<tr><td>10</td><td>i1,i3,i4,i6</td></tr>
</table></td><td>L4</td><td>CO5</td><td>7 M</td></tr>
<tr><td colspan="6">UNIT-IV</td></tr>
<tr><td>7</td><td>a)</td><td>Define information gain and explain its importance in decision tree induction.</td><td>L2</td><td>CO4</td><td>4 M</td></tr>
</table>

	b)	The following table consists of training data from an employee database. The data have been generalized. For example, "31 : : : 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.	L4	CO4	10 M

department	*status*	*age*	*salary*	*count*
sales	senior	31...35	46K...50K	30
sales	junior	26...30	26K...30K	40
sales	junior	31...35	31K...35K	40
systems	junior	21...25	46K...50K	20
systems	senior	31...35	66K...70K	5
systems	junior	26...30	46K...50K	3
systems	senior	41...45	66K...70K	3
marketing	senior	36...40	46K...50K	10
marketing	junior	31...35	41K...45K	4
secretary	senior	46...50	36K...40K	4
secretary	junior	26...30	26K...30K	6

(i) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?
(ii) Use your algorithm to construct a decision tree from the given data.
(iii) Given a data tuple having the values "systems," "26 . . . 30," and "46–50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

OR

8	a)	Consider a school with a total population of 100 persons. These 100 persons can be seen either as 'Students' and 'Teachers' or as a	L4	CO5	7 M

		population of 'Males' and 'Females'. With below tabulation of the 100 people, what is the conditional probability that a certain member of the school is a 'Teacher' given that he is a 'Man'? <table><tr><td></td><td>Female</td><td>Male</td><td>Total</td></tr><tr><td>Teacher</td><td>8</td><td>12</td><td>20</td></tr><tr><td>Student</td><td>32</td><td>48</td><td>80</td></tr><tr><td>Total</td><td>40</td><td>60</td><td>100</td></tr></table>			
	b)	Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?	L2	CO4	7 M
		UNIT-V			
9	a)	Compare k-means with k-medoids algorithms for clustering.	L3	CO4	6 M
	b)	Discuss various evaluation measures used to evaluate clustering algorithms.	L2	CO4	8 M
		OR			
10	a)	Discuss the similarity measures and distance measures frequently used in clustering the data.	L3	CO4	10 M
	b)	Discuss about key issues in Hierarchical clustering.	L3	CO4	4 M