Code No: 137BQ

R16

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY HYDERABAD
B. Tech IV Year I Semester Examinations, December - 2019
DATA MINING
(Common to CSE, IT)

Time: 3 Hours **Max. Marks: 75**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 25 marks. Answer all questions in Part A. Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks and may have a, b as sub questions.

PART – A
(25 Marks)

1.a)	Define data mining.	[2]
b)	List the methods of filling missing values.	[3]
c)	Define closed frequent itemset.	[2]
d)	What is the need of confidence measure in association rule mining?	[3]
e)	List the measures for selecting best split in decision tree construction.	[2]
f)	Quote an example for Bayesian belief network.	[3]
g)	What are the limitations of single linkage algorithm?	[2]
h)	List the typical requirements of clustering in data mining.	[3]
i)	What is meant by stop words?	[2]
j)	Give the taxonomy of web mining	[3]

PART – B
(50 Marks)

2. Discuss data mining as a step in knowledge discovery process and various challenges associated. [10]

OR

3. Use a flowchart to summarize the following procedures for attribute subset selection:
a) Stepwise forward selection
b) Stepwise backward elimination. [10]

4. Classify frequent pattern mining methods and explain the criteria followed for classification. [10]

OR

5. Apply apriori algorithm to find frequent itemsets from the following transactional database. Let min_sup = 30%. [10]

TID	Items_bought
1	Pen, notebook, ruler
2	Pencil, eraser, sharpener
3	Pen, ruler, chart, sharpener
4	Pencil, clip, eraser
5	Ruler, pin, story book, pen
6	Marker, chart, sketchpens

6. State classification problem and briefly explain general approaches to solve it. [10]

OR

7. Apply Naïve-Bayesian classifier to identify class label(campus_placement) to the new sample/student < 7 to 8, 'Fair', 'Excellent','No'>. [10]

SID	CGPA	Coding Skills	Soft Skills	Hackathon Participation	Campus_placement
1	7 to 8	Excellent	Fair	Yes	Yes
2	8 to 9	Fair	Excellent	Yes	Yes
3	9 to10	Poor	Fair	No	Yes
4	5 to 6	Poor	Excellent	No	No
5	7 to 8	Excellent	Poor	No	No
6	8 to 9	Fair	Fair	Yes	Yes
7	9 to 10	Poor	Poor	No	No

8. Suppose that the data mining task is to cluster the following eight students into three clusters, the distance function is Manhattan.Assign record 1,2,3 as the centroid of each cluster respectively.Use the k-means algorithm to show the final three clusters. [10]

RecordID	Height(cms)	Weight(kgs)
1	145	35
2	165	55
3	170	90
4	135	60
5	140	50
6	160	75
7	150	40
8	155	65

OR

9. Appraise the importance of outlier detection and its application. Explain any one approach for outlier detection. [10]

10. Discuss various kinds of patterns to be mined from web/server logs in web usage mining. [10]

OR

11. Compare and contrast text mining with web content mining using lucid examples. [10]

--ooOoo—