GURU GHASIDAS VISHWAVIDYALAY, CENTRAL UNIVERSITY, BILASPUR

BSC VI END SEMESTER EXAM 2022 (Online)

INSTRUCTION : On the top of the answer sheet write Date, your Name, Class, Semester, Name of Subject, University Roll Number. Total pages of the answer sheet strictly limited to 4sheets only.

One mark will be deducted per minute in case of late submission

Class : BSc-VI SEM Mark : 70

Subject: Data Mining Time : 10am-12noon

Answer any fourteen(14). Calculator is allowed.

<table>
<tr><th>Q. No.</th><th></th><th>Mark</th></tr>
<tr><td>1</td><td>Differentiate between cluster analysis, outlier analysis and evolution analysis.</td><td>5</td></tr>
<tr><td>2</td><td>What are the major challenges of mining a large amount of data in comparison with mining a small amount of data?</td><td>5</td></tr>
<tr><td>3</td><td>Suppose that a data warehouse for GGU University consists of the following four dimensions: student, course, semester and instructor and two measures count and avg_grade. When at the lowest conceptual level the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.
(a)Draw a snow-flake schema for the data warehouse</td><td>5</td></tr>
<tr><td>4</td><td>For the above drawn schema of Q. 3 write the DMQL</td><td>5</td></tr>
<tr><td>5</td><td>Explain the indexing techniques used in data warehouse with suitable examples.</td><td>5</td></tr>
<tr><td>6</td><td>Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70
(a)Use min-max normalization to transform the value 35 for age onto the range [0.0 1.0]
(b)Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94.</td><td>5</td></tr>
<tr><td>7</td><td>Find the IDFT value of $X(k) = \{2, -j, 0, j\}$</td><td>5</td></tr>
<tr><td>8</td><td>How dimensions of data can be reduced using transform based techniques?</td><td>5</td></tr>
<tr><td>9</td><td>A 2x2 contingency table is given as follows :
Are gender and preferred_reading correlated? Check by Chi-square test. Given
$value\ of\ \chi^{2} = 10.828\ for\ \alpha = 0.001\ and\ \deg ree\ of\ \ freedom = 1$
<table>
<tr><th></th><th>Male</th><th>Female</th><th>total</th></tr>
<tr><td>Fiction</td><td>50</td><td>1000</td><td>1050</td></tr>
<tr><td>Non-fiction</td><td>250</td><td>200</td><td>450</td></tr>
<tr><td>Total</td><td>300</td><td>1200</td><td>1500</td></tr>
</table></td><td>5</td></tr>
<tr><td>10</td><td>Find the dissimilarity matrix for the following mixed types variables</td><td>5</td></tr>
</table>

Object	Test-1 (Categorical)	Test-2 (Ordinal)	Test-3 (Ratio scaled)
1	Code-A	Excellent	445
2	Code-B	Fair	22
3	Code –C	Good	164
4	Code-A	Excellent	1210

11 What are the different training methods available for classification or prediction? 5

12 Explain the k-mean clustering algorithm for IRIS data set. 5

13 Plot the dendogram using agglomerative algorithm of hierarchical clustering with single linkage for the following points: 5

A	3	3.5
B	4	4
C	3	4
D	5	5
E	1.5	1.5
F	1	1

14 Write the APRIORI Algorithm of association rule mining. 5

15 Explain the artificial neural network based time series prediction. 5

16 How to choose the Root node in decision tree based classifier? 5

17 Using the below given table and Naïve Bayesian Classification algorithm find out the class of the new tuple, 5

X={age=senior, income=medium, student=yes, credit rating =fair}

RID	Age	Income	Student	Credit rating	Class : Buys computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes

18 Suppose that the data mining task is to cluster the following nine points into three clusters. 5

A_1(2, 10), A_2(2, 5), A_3(8,4), B_1(5, 8), B_2(7, 5), B_3(6, 4), C_1(1, 2), C_2(4, 9), C_3(3,4)

The distance function is Euclidean distance.

Suppose initially we assign A_2, B_2 and C_2 as the center of each cluster respectively. Use the k-means algorithm to show the three cluster centers after the first round of execution