2D21503037

ANNA UNIVERSITY (University Departments)
B.E (Full Time) END SEMESTER EXAMINATIONS – April /May 2024
Computer Science and Engineering
VI Semester
CS6001 DATA MINING
(Regulation 2018 - RUSA)

Time: 3 Hours Max. Marks 100

CO 1	Demonstrate the knowledge of the ethical considerations involved in Data Mining
CO 2	Examine data and select suitable methods for data analysis
CO 3	Integrate various classification, clustering, association rule mining techniques on real world data
CO 4	Synthesize the different algorithms and analyze it with the support of tools
CO 5	Interpret the concept of spatial, multimedia and distributed, text and web mining and able to retrieve the data, analyze and make decision.

BL – Bloom's Taxonomy Levels
(L1 - Remembering, L2 - Understanding, L3 - Applying, L4 - Analyzing, L5 - Evaluating, L6 -Creating)

PART- A (10 x 2 = 20 Marks)
(Answer all Questions)

Q. No.	Questions	Marks	CO	BL
1.	List any four characteristics of data ware house.	2	1	L1
2.	Under what conditions does the mean of a dataset coincide with its standard deviation, resulting in both values being equal?	2	2	L3
3.	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): Compute the Manhattan distance between the two objects.	2	2	L2
4.	Write the basic principle of wavelet transform and provide one real-world application for its usage.	5	3	L3
5.	How can the frequency of an itemset (support count) be used to assess confidence in frequent itemset mining?	2	4	L2
6.	List the clustering algorithms which form the spherical shape clusters.	2	3	L3
7.	What is a key advantage of Multilayer Perceptrons (MLPs) compared to simpler linear models like linear regression in machine learning tasks?	2	3	L1
8.	In the context of multi-relational data mining, what is the key advantage of using Inductive Logic Programming (ILP) for multi-relational classification tasks?	2	5	L4
9.	What is the primary distinction between traditional web search engines and techniques used in web mining?	2	5	L4
10.	What for the precision and recall measures used in classification?	2	4	L2

PART- B (8 x 8 = 64 Marks)
Answer any **eight** questions

Q. No.	Questions	Marks	CO	BL
11.	Explain the ETL process and its significance in data warehousing and analytics. Provide examples of each stage of the ETL process and discuss the challenges .	8	1	L2

<table>
<tr><td>12.</td><td>Discuss the methodology and application of a specific technique for image dimensionality reduction, outlining its advantages, limitations, and real-world applications in various domains.</td><td>8</td><td>1</td><td>L1</td></tr>
<tr><td>13.</td><td>Given a transaction database D containing a set of transactions T, a minimum support threshold (minSup), and a minimum confidence threshold (minConf), how can the Apriori algorithm be used to efficiently identify all frequent itemsets and generate strong association rules?</td><td>8</td><td>2</td><td>L3</td></tr>
<tr><td>14.</td><td>The following table gives the training set from an employee database, Status be the class label. Given a data tuple having the values "department = system, age=26….30 and salary=46….50K, what would a naive Bayesian classification of the "status" for the tuple?
<table>
<tr><th>Department</th><th>Status</th><th>Age(yrs)</th><th>Salary in K</th><th>Count</th></tr>
<tr><td>Sales</td><td>Senior</td><td>31…35</td><td>46….50</td><td>30</td></tr>
<tr><td>Sales</td><td>Junior</td><td>26…30</td><td>26….30</td><td>40</td></tr>
<tr><td>Sales</td><td>Junior</td><td>31…35</td><td>31….35</td><td>40</td></tr>
<tr><td>Systems</td><td>Junior</td><td>21…25</td><td>46….50</td><td>20</td></tr>
<tr><td>Systems</td><td>Senior</td><td>31…35</td><td>66….70</td><td>5</td></tr>
<tr><td>Systems</td><td>Junior</td><td>26…30</td><td>46….50</td><td>3</td></tr>
<tr><td>Systems</td><td>Senior</td><td>41…45</td><td>66….70</td><td>3</td></tr>
<tr><td>Marketing</td><td>Senior</td><td>36…40</td><td>46….50</td><td>10</td></tr>
<tr><td>Marketing</td><td>Junior</td><td>31…35</td><td>41….45</td><td>4</td></tr>
<tr><td>Secretary</td><td>Senior</td><td>46…50</td><td>36….40</td><td>4</td></tr>
<tr><td>Secretary</td><td>Junior</td><td>26…30</td><td>26…..30</td><td>6</td></tr>
</table></td><td>8</td><td>2</td><td>L2</td></tr>
<tr><td>15.</td><td>Discuss the key differences between bagging and boosting in terms of training approaches and addressing variance and bias.</td><td>8</td><td>3</td><td>L4</td></tr>
<tr><td>16.</td><td>Prove that in the DBSCAN, for a fixed Minpts value and two neighbourhood thresholds, ε1< ε2, a cluster C with respect to ε1 and Minpts must be a subset of a cluster C' with respect to ε2 and Minpts.</td><td>8</td><td>3</td><td>L5</td></tr>
<tr><td>17.</td><td>Discuss in detail about the how the data-mining task is carried out in multimedia database.</td><td>8</td><td>5</td><td>L2</td></tr>
<tr><td>18.</td><td>Given a social network graph where nodes represent users and edges represent connections between them, how can graph mining techniques be used to identify communities of users with similar interests or characteristics?</td><td>8</td><td>5</td><td>L3</td></tr>
</table>

19.	Consider the following data set consisting of the scores of two variables on each of seven individuals illustrate the k- means algorithm. (k=2) Person ID | A | B 1 | 10 | 10 2 | 15 | 20 3 | 30 | 40 4 | 50 | 70 5 | 35 | 50 6 | 45 | 50 7 | 35 | 45	8	4	L3
20.	Construct a decision tree using ID3 algorithm using the data given in the to predict whether a movie is worth seeing. S.No. | Major Studio? | Cost to make | Genre | Worth seeing 1 | True | 146 | Science | Yes 2 | False | 28 | Horror | Yes 3 | False | 52 | Romance | Yes 4 | False | 74 | Science | No 5 | True | 22 | Romance | No 6 | False | 30 | Science | No	8	4	L2
21.	Discuss two specific applications of data mining in different domains, and then elaborate on two emerging trends in data mining and their potential impact on data analysis.	8	4	L2

PART- C (2 x 8 = 16 Marks)
(Answer all questions)

Q. No.	Questions	Marks	CO	BL
22.	How can data mining techniques be adapted to address the challenges of extracting knowledge from complex data objects and explain any two standard mining principles.	8	4	L4
23.	Suppose in a group of 1500 people, gender was noted. Each person polled their vote for preferred type of reading was fiction or non-fiction. Find the correlation between the two attributes gender and preferred reading (Note the significance level is 10.828) | | Male | Female | Total | Fiction | 250 | 200 | 450 Non fiction | 50 | 1000 | 1050 Total | 300 | 1200 | 1500	8	3	L4