CSA16-DATA WAREHOUSING AND DATA MINING

UNIT1-INTRODUCTION TO DATA MINING

2 Marks

1. Define Data Mining
2. Define the following data mining functionalities: characterization, discrimination
3. What are the steps involved in knowledge discovery from data?
4. How to classify the datamining system.
5. Why Preprocess the Data?
6. List out the Strategies for data transformation.
7. Define outlier analysis with an example
8. How to integrate the data mining system with a database or Datawarehouse
9. Define KDD process

5 Marks

10. List and describe the five primitives for specifying a data mining task.
11. The data for analysis contains the values 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. Use smoothing by bin means, by bin medians and by bin boundaries to smooth the above data, using a bin depth of 3. Illustrate your steps.
12. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
13. What are the strategies for data reduction? Explain them in detail.
14. Outliers are often discarded as noise. However, one person's garbage could be another's treasure. Propose two methods that can be used to detect outliers and discuss which one is more reliable.

10 Marks

15. Define each of the following data mining functionalities: characterization, discrimination, association and correlation analysis, classification, prediction, clustering, and evolution analysis. Give examples (applications) of each data mining functionality, using a real-life database that you are familiar with.
16. The fifteen data values for analysis are given as
    {20, 20, 21, 22, 22, 25, 25, 25, 25, 30,33, 33, 35, 35, 35}.
    (i) What is the mean, median and mode of the data?
    (ii) smooth out the data using binning method. Use bin depth of 3 and bin means to replace values.
17. Illustrate the architecture of a data mining system with a neat diagram.

18. Suppose your task as a software engineer at Big-University is to design a data mining system to examine their university course database, which contains the following information: the name, address, and status (e.g., undergraduate) of each student, the courses taken, and their cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?
19. An Electronics company is planning to build a data mining system for its production forecast. Explain each of the steps involved in developing a system to forecast production required for the next year as knowledge discovery process with a neat diagram.

UNIT 2-INTRODUCTION TO DATAWAREHOUSING

2 MARKS

1 Define Datawarehouse
2 Differentiate Operational Database Systems and DataWarehouse
3 Define datamart
4 Define meta data repository .
5 What is data cube?
6 Define multidimensional Datamodel
7 Define the modelling schemes for Datawarehouse .
8 How are concept hierarchies useful in OLAP?
9 How many cuboids are there in n-dimensional data cube?
10 How many cuboids in an n-dimensional cube with L levels?
11 Define concept hierarchy

5 MARKS

12 Differentiate Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP)

13 Briefly explain about the Starnet Query Model**?**

14 The data warehouse for a 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 snowflake schema diagram for the data warehouse.

b. Starting with the base cuboid [student; course; semester; instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each University student.

15 Briefly compare the following concepts. You may use an example to explain your point(s).

(a) Snowflake schema, fact constellation,
(b) Data cleaning, data transformation, refresh
(c) Enterprise warehouse, data mart, virtual warehouse

16 Suppose that a data warehouse consists of 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.

a. Draw a Star schema diagram for the above data warehouse

b. 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 2004?

17 What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining?

10 MARKS

18 What type of architecture does the data warehouse adopt? Briefly explain the layers of the architecture?

19 Given a multidimensional database with four dimensions [item, customer, time and location], Show the various OLAP operations sketching them on the multidimensional data cube.

20 Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate.

(a) Draw a star schema diagram for the data warehouse.

(b) Starting with the base cuboid [date; spectator; location; game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators.

21 What is data warehouse? What are the applications of data warehouse? Explain the steps involved in the design process of data warehouse.

23 The data warehouse can be modeled by either a star schema or a snowflake schema.

a) Briefly describe the similarities and the differences of the two models(5)

b) Analyze their advantages and disadvantages with regard to one another. (5)

UNIT – 3

CLASSIFICATION AND PREDICTION

2 Marks

20. Define Classification

21. What is Supervised Learning and Unsupervised Learning?

22. Define Information Gain

23. What is Tree Pruning?

24. Define Bayes Theorem

25. What is feed Forward Neural Network

26. What is Confusion Matrix?

27. List the methods for Evaluating the Accuracy of Classifiers

28. Given a decision tree, you have the option of

    (a) converting the decision tree to rules and then pruning the resulting rules, or

    (b) pruning the decision tree and then converting the pruned tree to rules.

    What advantage does (a) have over (b)?

29. Define Classification by back propagation.

5 marks

30. Given the confusion matrix for positive and negative tuples as shown in table.

Actual class	Predicted class		
		C1	C2
	C1	t_pos (654)	f_neg (46)
	C2	f_pos (98)	t_neg (586)

Confusion Matrix

a. Calculate the sensitivity and specificity

b. Show that accuracy is a function of sensitivity and specificity

31. Explain the cross-validation and bootstrap methods of evaluating accuracy of classifier or predictor.

32. How does the value 63.2%, come from for the bootstrap method for accuracy estimation? Briefly explain the method

33. State how the tree pruning is useful in decision tree induction?

34. Briefly outline the major steps of decision tree classification.

35. Explain the methods for evaluating the accuracy of a Classifier

36. What is confusion matrix? How will you find out the sensitivity, specificity, precision, and accuracy of the classifier using the confusion matrix.

10 Marks

37. A hospital has historical data for the classification of its patients as to whether they suffer from heart disease or not ('Heart' is the class label attribute). construct a decision tree using the data set as given in the table 2 training set.

Age	Trestbps	Choi	Gender	Heart
< 50	<120	<200	Male	No
< 50	<120	<200	Female	No
< 70	<120	<200	Male	Yes
< 60	< 140	<200	Male	Yes
< 60	<160	>200	Male	Yes
< 60	<160	>200	Female	No
< 70	<160	>200	Female	Yes
< 50	<140	<200	Male	No

Table:

< 50	<160	>200	Male	Yes
<60	<140	>200	Male	Yes
<50	<140	>200	female	Yes
<70	< 140	<200	female	Yes
<70	<120	>200	Male	Yes
<60	<140	<200	female	No

Training Set

39 The following table consists of Class-labelled training tuples from thecustomer database.

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
6	Senior	Low	Yes	excellent	no
7	middle aged	Low	Yes	excellent	yes
8	Youth	Medium	No	excellant	YES
9	Youth	Low	Yes	fair	yes
10	Senior	Medium	No	faie	NO
11	Youth	Medium	No	excellant	yes
12	middle aged	Medium	No	fair	yes
13	middle aged	High	Yes	fair	yes
14	Senior	Medium	No	excellant	no

a. Given a tuple ***X*** = (*age = youth, income = medium, student = yes, credit rating = fair*), Predict the class label using naïve Bayesian classification.

b. State and explain Bayes theorem and Naive Bayesian classification.

40 Discuss the major ideas of Naive Bayesian classification with an example

UNIT4-CLUSTERING, MINING FREQUENT PATTERNS,ASSOCIATIONS

2 Marks

38. Define frequent patterns with example

39. What are the two steps in association rule mining.

40. What is a frequent -itemset?

41. Define Association Rule with measures Support and Confidence

42. Define Apriori Property

43. FP-growth Algorithm is more efficient than AprioriAlgorithm,Why?

44. Define Clustering

45. List out the Different methods of Clustering Algorithm

46. What is Dendogram

47. Define Agglomorative and Divisive Method of hierarchial Clustering

5 Marks

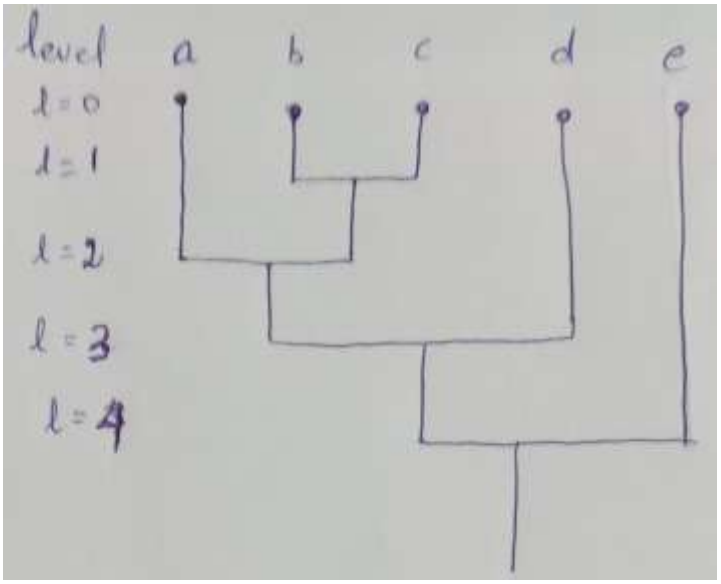
48. Explain K-Means Clustering(partion method) Algorithm

49. Given two objects represented by the tuples (13, 4, 7, 7) and (6, 4, 6, 4):

    (a) Compute the Euclidean distance between the two objects.

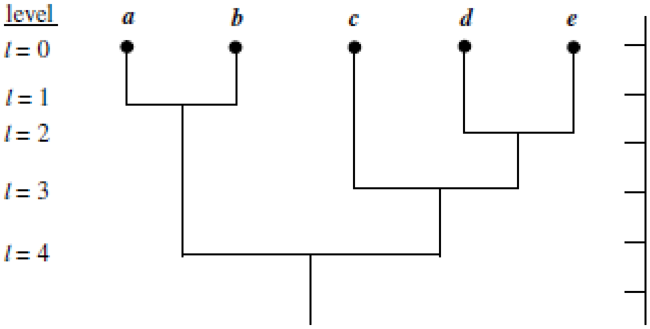
    (b) Compute the Manhattan distance between the two objects.

50. The dendrogram representation for hierarchical clustering of data objects {a,b,c,d,e} is given below.



Construct the Agglomerative and divisive hierarchical clustering on data objects {a,b,c,d,e} with the above dendrogram and provide suitable explanation for the samev

51. The dendrogram representation for hierarchical clustering of data objects {a,b,c,d,e} is given below.



Construct the Agglomerative and divisive hierarchical clustering on data objects {a,b,c,d,e} with the above dendrogram and provide suitable explanation for the same.

10 Marks

52. A database (Refer Table1) has 4 transactions. Let min sup = 2 and min conf = 60%. Determine all frequent itemsets using Apriori Algorithm

Tid	**Items**
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Table 1database

53. Using FP-Growth algorithm find the frequent itemset for the transactional Data of Supermarket as given in the Table:

TID	**List of item IDS**
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Table : Supermarket Transaction Database

54. Using Apriori algorithm find the frequent itemset for the transactional Data of Supermarket as given in the Table:

TID	**List of item IDS**
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Table : Supermarket Transaction Database

55. Describe in detail on how the partitioning method is used to constructs several partitions of the given dataset, with each partition representing a cluster.

56. A database (Refer Table) has five transactions. Let min sup = 60% and min conf = 80%.

TID	*items bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find all frequent itemsets using FP-Growth Algorithm

57. Apply the Apriori algorithm with minimum support of 30% and minimum confidence of 70%, and find all the frequent itemset in the data set given in Transaction Table 2. Show the step by step process of the problem.

TID	**Items**
1	A,B,C,D
2	A,B,C,D,E,G
3	A,C,G,H,K
4	B,C,D,E,K
5	D,E,F,H,L
6	A,B,C,D,L
7	B,I,E,K,L
8	A,B,D,E,K
9	A,E,F,H,L
10	B,C,D,F

Table: Transaction Table

58. What is clustering? Illustrate both methods of hierarchical clustering for a data set of five objects {M,N,O,P,Q}

59. Explain Agglomerative and divisive methods of hierarchical algorithm with an example.

60. Describe Expectation-Maximization and Hierarchical algorithm in detail.

61. Explain FP-growth algorithm for frequent itemset mining in detail.

UNIT – 5 APPLICATIONS

2 marks

62. What are knowledge that can be mined from Text?

63. List the Applications of text mining

List the Applications of spatial data mining

What are the challenges in web mining?

64. What is Text Mining:

5 marks

65. Explain in detail about document ranking method in Information Retrival System

66. Briefly describe the following advanced database systems and applications: object-relational databases, spatial databases, text databases, multimedia databases, the World Wide Web.

67. Analyse and give your suggestions of how data warehousing could be used in telecommunication industry.

68. State the significance of text mining and list some application areas of text mining.

10 marks

69. Explain in details about text mining with a neat diagram

70. Explain in details about web mining with a neat diagram

71. Explain in details about spatial data mining with a neat diagram