UNIT - I

1. Explain kinds of data can be mined? Give examples.

2. Differentiate Operational database systems and data warehousing.

3. Explain the star schema and fact constellation schemas.

4. List and describe the five primitives for specifying a data mining task.

5. What are the differences between the three main types of data warehouse usage: information processing, analytical processing, and data mining? Discuss the motivation behind OLAP mining (OLAM).

6. Explain what kinds of patterns can be mined? Give examples.

7. State why, for the integration of multiple heterogeneous information sources, many companies in industry prefer the update-driven approach, rather than the query-driven approach.

8. Suppose that a data warehouse consists of the 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. Enumerate three classes of schemas that are popularly used for modeling data warehouses and explain.

9. Briefly compare the following concepts using examples Discovery-driven cube, multi feature cube, virtual warehouse

10. Briefly compare the following concepts with example Snowflake schema, fact constellation, starnet query model.

11. Explain technologies used for data mining?

12. Suppose that a data warehouse consists of the 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. Draw a schema diagram for the above data warehouse.

13. Present an example where data mining is crucial to the success of a business. What data mining functionalities does this business need? Can such patterns be generated alternatively by data query processing or simple statistical analysis?

14. Explain which kinds of applications are targeted for data mining?

15. Suppose that a data warehouse consists of the 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. 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 2010? Explain the concepts involved.

16. A data warehouse can be modeled by either a star schema or a snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another.

17. Explain the difference and similarity between discrimination and classification, between characterization and clustering, and between classification and regression.

18. Draw an architecture of typical data mining system with a neat sketch and explain its components.

19. What are the various OLAP operations are used in the multidimensional data model? Explain them in detail with an example.

20. Explain the data warehouse implementation.

21. Briefly describe data mining functionalities.

22. Discuss the major issues in data mining.

23. Briefly explain about efficient computation of Data Cubes.

24. Write the difference between OLTP vs OLAP

25. Explain the basic elements of Data warehouse with a neat sketch

26. Explain ROLAP, MOLAP and HOLAP.

27. What are various schemas for multidimensional data models?

28. Can you list the characteristic differences between OLAP and OLTP?

29. With an example, describe snowflake and fact constellations

30. What are the OLAP operations? Explain.

31. Why data mining functionalities are used? Explain with an example data characterization and data discrimination.

32. Illustrate on what kinds of patterns can be mined.

33. Differentiate operational database systems and data warehousing.

UNIT - II

1. Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. Partition them into three bins by each of the following methods:

(i) equal-frequency (equal-depth) partitioning (ii) equal-width partitioning (iii) clustering

2. Write short notes on the following:

(i) Data Preprocessing (ii) Data Discretization (iii) Concept Hierarchy

3. What are the value ranges of the following normalization methods?

(i) min-max normalization (ii) z-score normalization

(iii) z-score normalization using the mean absolute deviation instead of standard deviation

(iv) normalization by decimal scaling

4. Explain in detail about data pre-processing.

5. What is data integration and discuss issues to consider during data integration.

6. Given the following data (in increasing order) for the attribute age: 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.

(i) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].

(ii) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

(iii) Use normalization by decimal scaling to transform the value 35 for age.

7. What is the need of dimensionality reduction? Explain any two techniques for dimensionality reduction

8. Suppose a group of 12 sales price records has been sorted as follows: 5,10,11,13,15,35,50,55,72,92,204,215. Partition them into three bins by each of the following methods:

(i) equal-frequency (equal-depth) partitioning (ii) equal-width partitioning (iii) clustering

9. Use these methods to normalize the following group of data: 200,300,400,600,1000

(i) min-max normalization by setting min = 0 and max = 1 (ii) z-score normalization

(iii) z-score normalization using the mean absolute deviation instead of standard deviation.

10. What is data consolidation? In detail discuss various techniques used to consolidate data.

11. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	23	23	27	27	39	41	47	49	50
Fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
Fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

(i) Normalize the two attributes based on z-score normalization.

(ii) Calculate the correlation coefficient (Pearson's product moment coefficient). Are these two attributes positively or negatively correlated? Compute their covariance.

12. Describe the problem of data quality with some examples. Explain the usage of feature subset selection in data preprocessing

13. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

14. Discuss in detail about data transformation with suitable examples.

15. Explain various data pre-processing methods with appropriate examples.

16. Give the following data (in increasing order) for the attribute age: 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. How might you determine outliers in the data? Relate it with data cleaning.

17. Briefly discuss the forms of Data preprocessing with neat diagram.

18. What is the use of data discretization? Explain entropy based data discretization?

19. What is data integration? Discuss the issues to be considered for data integration.

20. Differentiate between data reduction and dimensionality reduction for data discretization

21. Explain about concept hierarchy generation for categorical data.

22. Describe Data Transformation & Data Discretization.

23. What is data reduction? Discuss about dimensionality reduction.

24. What is Numerosity Reduction? What are the available techniques for numerosity reduction?

25.How can we smooth out noise in data cleaning process? Explain

26. Normalize the following group of data by using the following techniques. 200, 300, 400, 600, 1000

i) min-max normalization technique ii)z-score normalization iii)Decimal scaling.

Write your observations on the above techniques.

27. What is data reduction? What is dimensionality reduction? What is lossless and lossy dimensionality reduction?

28.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.

i) Use min-max normalization to transform the value35for age onto the range [0:0; 1:0].

Ii)Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. Iii)Use normalization by decimal scaling to transform the value 35 for age.

29. Why preprocessing of data is needed? Discuss about data integration in detail.

30. In real world data tuples with missing values for some attributes are common occurrence. Describe various methods for handling this problem.

31. What is redundancy? Why correlation analysis is useful? Describe how correlation coefficient is computed?

32. b) What are the value ranges of the following normalization methods?

(i) min-max normalization (ii) z-score normalization (iii) normalization by decimal scaling

UNIT - III

1. Use the C4.5 algorithm to build a decision tree for classifying the followingobjects:

Class	Size	Color	Shape
A	Small	Yellow	Round
A	Big	Red	Round
A	Big	Red	Round
A	Small	Black	Round
B	Small	Black	Round
B	Big	Black	Cube
B	Big	Yellow	Cube
B	Big	Black	Round
B	Small	Yellow	Cube

2. Why information gain is considered as attribute selection measure? Illustrate with an example.

3. Explain the decision tree induction algorithm with appropriate examples. Discuss the disadvantages of this approach? What is over fitting, and how can it be prevented for decision trees?

4. What is visual mining? Explain the application of decision tree induction algorithm in it.

5. What are the new features of C4.5 algorithm comparing with original Quinlan's ID3 algorithm for decision-tree generation?

6. What is attribute selection measure? Briefly describe the attribute selection measures for decision tree induction.

7. What are the new features of C4.5 algorithm comparing with original Quinlan's ID3 algorithm for decision-tree generation?

8. What is attribute selection measure? Briefly describe the attribute selection measures for decision tree induction.

9. Given a training data set Y:

A	B	C	Class
15	1	A	C1
20	3	B	C2
25	2	A	C1
30	4	A	C1
35	2	B	C2
25	4	A	C1
15	2	B	C2
20	3	B	C2

Find the best split point for decision tree for attribute A.

10. Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

11. Make a decision tree for the following database using Gini Index. Indicate all intermediate steps.

Example	Colour	Shape	Size	Class
1	Red	Square	Big	+
2	Blue	Square	Big	+
3	Red	Circle	Big	+
4	Red	Circle	Small	-
5	Green	Square	Small	-
6	Green	Square	Big	-

12. Given data set, D, the number of attributes, n, and the number of training tuples, |D|, show that the computational cost of growing a tree is at most $n \times |D| \times \log(|D|)$

13. Describe the classification task in induction and deduction phases. Explain with example classification tasks.

14. Calculate the gain in the Gini Index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

A	B	CLASS LABEL
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

15. Why information gain is considered as attribute selection measure? Illustrate with an example.

16. Identify the attribute that will act as the root node of a decision tree to predict golf play for following database with Gini Index. Indicate all the intermediate steps.

Outlook	Wind	Play Golf
Rain	Strong	No
Sunny	Weak	Yes

Overcast	Weak	Yes
Rain	Weak	Yes
Sunny	Strong	Yes
Rain	Strong	No
Overcast	Strong	No

17. What measures are used to find best split in Decision Tree Induction algorithm? How Can we improve the scalability in Decision Tree Induction algorithm?

18. Explain in detail about Attribute Selection methods in Classification

19. What is the role of data preprocessing in classification? Give examples of common data preprocessing tasks and explain how they impact the performance of a classification model.

20. Discuss about contusion matrix in detail.

21.Compare and contrast post-pruning and pre-pruning techniques in the context of data mining. How do these techniques help in addressing over fitting, and what are the potential consequences of not pruning a decision tree?

22. Explain the classification by decision tree induction.

23. Explain the purpose of "Attribute selection measures" in classification by decision tree induction? How we can use the "Tree pruning" in classification?

24. What is the Basic Concept of Classification?

25. Discuss in detail visual mining for decision tree induction.

26. Explain the techniques and strategies used to improve the scalability of decision tree induction in data mining, such as parallelization and distributed computing.

27. Describe the importance of visual mining in data mining, with a focus on decision tree induction. How can visual representations of decision trees aid in model interpretation and decision support?

28. How tree pruning in decision tree induction is useful? Explain various methods for pruning decision trees.

29. Write a note on attribute selection measures.

30. Explain decision tree induction algorithm for classifying data tuples and with suitable example

31. Explain the concepts of class labels, features, and training data in a classification problem. How are these components used to train and evaluate classification models?

UNIT - IV

1. Why is the process of discovering association rules relatively simple compared to generating large item sets in transactional databases?

2. Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, explain it with the following table:

Tid	List of Items
001	milk, dal, sugar, bread
002	Dal, sugar, wheat,jam
003	Milk, bread, curd, paneer
004	Wheat, paneer, dal, sugar
005	Milk, paneer, bread
006	Wheat, dal, paneer, bread

3. Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved?

4. Consider the transaction data-set:

Trans ID	Items
T1	{a,b}
T2	{b,c,d}
T3	{a,c,d,e}
T4	{a,d,e}
T5	{a,b,c}
T6	{a,b,c,d}
T7	{a}
T8	{a,b,c}
T9	{a,b,d}
T10	{b,c,e}

Construct the FP tree by showing the trees separately after reading each transaction.

5. Given a simple transactional database X:

TID	Items
T01	A,B,C,D
T02	A,C,D,F
T03	C,D,E,G,A
T04	A,D,F,B
T05	B,C,G
T06	D,F,G
T07	A,B,G
T08	C,D,F,G

Using the threshold values support = 25% and confidence = 60%, find all large item sets in database X

6. How association rules mined from large databases? Explain.

7. What is a frequent item set? How to find frequent item sets for a transactional database? Explain any one approach with illustrations.

8. Find frequent item sets for the following table using FP-Growth algorithm. Assume relevant thresholds.

TID	List of Items
T1	I1,I3,I5
T2	I2,I4,I5
T3	I1,I2,I3,I4
T4	I5,I3,I2
T5	I1,I2,I5
T6	I3,I4,I5

9. What are the frequent item sets with a minimum support of 3 for the given set of transactions?

TID	Items
101	A,B,C,D,E
102	A,C,D
103	D,E
104	B,C,E
105	A,B,D,E
106	A,B
107	B,D,E
108	A,B,D
109	A,D
110	D,E

10. Write the algorithm to discover frequent item sets without candidate generation and explain it with an example.

11. Consider the following table to find frequent item sets using vertical data format. Support threshold 30%

TID	List of Items
T01	Milk, biscuits, surf powder, teabags
T02	Teabags, sugar, soap
T03	Milk, sugar, bread, soap
T04	Bread, teabags, biscuits
T05	Chocolates, milk, biscuits
T06	Milk, teabags, bread

T07	Bread, biscuits, chocolate
T08	Milk, surf powder, bread

12. Discuss Apriori Algorithm with a suitable example and explain how its efficiency can be improved?

13. Assume 5 transactions and explain the two-step approach to generate frequent item sets and to mine association rules using Apriori algorithm.

14. A database has four transactions. Let min_sup=60% and min_conf=80%

TID	Date	Items_bought
100	10/15/2022	{K,A,B,D}
200	10/15/2022	{D,A,C,E,B}
300	10/19/2022	{C,A,B,E}
400	10/22/2022	{B,A,D}

Find all frequent items using Apriori& FP-growth, respectively. Compare the efficiency of the two-meaning process.

15.Write the algorithm to discover frequent item sets without candidate generation and explain it with an example.

16. Make a comparison of Apriori and FP-Growth algorithms for frequent item set mining in transactional databases. Apply these algorithms to the following data:

TID	List of Items
1	Bread, Milk, Sugar, TeaPowder, Cheese, Tomato
2	Onion, Tomato, Chillies, Sugar, Milk
3	Milk, Cake, Biscuits, Cheese, Onion
4	Chillies, Potato, Milk, Cake, Sugar, Bread
5	Bread, Jam, Mik, Butter, Chilles
6	Butter, Cheese, Paneer, Curd, Milk, Biscuits
7	Onion, Paneer, Chilies, Garlic, Milk
8	Bread, Jam, Cake, Biscuits, Tomato

17. Discuss about basic concepts of frequent item set mining.

18. What are the drawbacks of Apriori Algorithm? Explain.

19. What are the advantages of FP-Growth algorithm?

20. Write about basic concept in Association Rule mining. How many association rules can be generated for a given transactional database?

21. Consider the following transactional data for a commercial shop.

TID	List of items with ids
T1	I2,i4
T2	I1,i2,i5
T3	i2, i3
T4	i1, i3
T5	i1, i2, i4
T6	i2, i3
T7	i1, i3
T8	i1, i2, i3
T9	i1, i2, i3, i5

Generate all the frequent itemsets using Apriori algorithm. Consider the minimum support count is 2. Clearly show your computational steps.

22. Explain Mining Frequent Patterns using FP-Growth.

23. What are the various Constraints in Constraint based Association rule mining? Explain.

24. Explain how confident-based pruning is used in rule generation. What is the purpose of pruning rules, and how does it impact the quality and interpretability of the rules?

25. Apply Apriori Algorithm in tracing all the frequent item datsets

Transactions	Itemset
T100	1 2 3
T200	2 3 5
T300	1 2 3 5
T400	2 5
T500	1 3 5

#Hint: Consider appropriate minimal support and minimal configure values for generating the rules.

26. Explain the closed item set and maximal item set concepts in the context of compact representation. What distinguishes closed item sets from maximal item sets, and why are they useful?

27. Explain in detail the candidate generation procedures.

28. With an example, explain the Fp-growth algorithm?

29. Discuss about closed frequent itemsets in detail.

30. How can you find frequent itemsets using candidate generation?

31. Write and explain the APRIORI algorithm with an example.

32. Give examples of real-world applications where association analysis is commonly used. How does identifying associations between items or events benefit these applications?

UNIT - V

1. Consider five points {X1 , X2 , X3 , X4 , X5 } with the following coordinates as a two dimensional sample for clustering : X1 = (0.5,2.5); X2 = (0,0); X3 = (1.5,1); X4 = (5,1); X5 = (6,2) Illustrate the K-means partitioning algorithms using the above data set.

2. What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables.

3. Describe k-means clustering algorithms in terms of the following criteria:

(i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and

(iii) limitations.

4. What is Cluster Analysis? What are some typical applications of clustering? What are some typical
requirements of clustering in data mining?

5. Given the samples X1 = {1, 0}, X2 = {0, 1}, X3 = {2, 1}, and X4 = {3, 3}, suppose that the samples are randomly clustered into two clusters C1 = {X1, X3} and C2 = {X2, X4}. Apply one iteration of the K-means partitional clustering algorithm, and find a new distribution of samples in clusters.

6. Describe how categorization of major clustering methods is being done?

7. Suppose that the data-mining task is to cluster the following eight points (representing location) into three clusters: A1 (2;10) ; A2 (2;5) ; A3 (8;4) ; B1 (5;8) ; B2 (7;5) ; B3 (6;4) ; C1 (1;2) ; C2 (4;9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to determine: the three cluster centers after the first round of execution.

8. What are the advantages and disadvantages of k-means clustering against model-based clustering?

9. Cluster the following data into three clusters, using the k-means method.

X	Y
10.9	12.6
2.3	8.4
8.4	12.6
12.1	16.2
7.3	8.9
23.4	11.3
19.7	18.5
17.1	17.2
3.2	3.4
1.3	22.8

2.4	6.9
2.4	7.1
3.1	8.3
2.9	6.9
11.2	4.4
8.3	8.7

10. Describe K means clustering with an example.

11. Consider five points {X1 , X2 , X3 , X4 , X5 } with the following coordinates as a two dimensional sample for clustering : X1 = (0.5,2.5); X2 = (0,0); X3 = (1.5,1); X4 = (5,1); X5 = (6,2) Illustrate the K-means partitioning algorithms using the above data set.

12. What is cluster analysis? Describe the dissimilarity measures for interval-scaled variables and binary variables.

13. Suppose that the data mining task is to cluster points into three clusters, where the points are A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only the three cluster centers after the first round of execution

14. What are the requirements for cluster analysis? Explain briefly.

15. Given the points x1 = {1, 0}, x2 = {0,1}, x3={2, 1}, and x4 = {3, 3}. Suppose that these points are randomly clustered into two clusters: C1 = {x1, x3} and C2 = {x2, x4}. Apply one iteration of Kmeans partitional-clustering algorithm and find new distribution of elements in clusters. What is the change in a total square error?

16. Use an example to show why the k-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation

17. What is meant by clustering? Explain the partitioning methods with an example.

18. What is the drawback of k-means algorithm? How can we modify the algorithm to diminish? That problem?

19. Explain K Means clustering method

20. What is cluster analysis? Describe the types of data in cluster analysis

21. Classify various Clustering methods.

22. Write partitioning around mediods algorithm.

23. Describe the concept of bi-secting K-means as an extension of the K-means algorithm. How does this technique aim to address some of the limitations of standard K-means?

24. What is clustering and what is conceptual clustering? Describe dimensions and measures in a spatial data cube.

25. What are the main factors to consider when choosing the number of clusters (k) in K-means? Discuss common methods for determining the optimal value of k.

26. Discuss in detail additional issues of K- Means algorithm?

27. Describe density-based clustering and its main idea. How does it identify clusters based on the density of data points rather than distance measures?

28. Define spherical, ellipsoidal, and arbitrary-shaped clusters. How do the shape and distribution of data points impact the choice of clustering techniques?

29. Discuss the importance of cluster analysis in real-world applications. Provide examples of industries or fields where cluster analysis is a valuable tool for gaining insights and making decisions.

30. Compare and contrast the advantages and disadvantages of K-means and bisecting K-means in terms of performance and quality of clustering results.

31. Explain the key goals of cluster analysis. How does it help in discovering hidden patterns in data and making data-driven decisions?

32. Explain the influence of the initial seed point selection on K-means clustering results. What strategies can be used to improve the reliability of K-means?