Question Bank

Year: 2016- 2017 **Semester:** First and Second

Subject Dept: IT **Subject Name:** Data Mining and Data Warehousing

……

Q1) What is data warehouse?

Q2) What is the benefits of data warehouse?

Q3) What is the difference between OLTP and OLAP?

Q4) Briefly state different between data ware house & data mart?

Q5) What are the difference among dependent data mart, independent data mart and hybrid data mart?

Q6) List some properties of data Marts.

Q7) List the Reasons for Creating a Data Mart

Q8) What are the characteristics of data warehouse?

Q9) Differentiate between Data Warehouse versus Operational DBMS

Q10) multiple choice questions

1. Operational database is

Q11) Give some alternative terms for data mining.

Q12) What is KDD.

Q13) What are the steps involved in KDD process.

Q14) What are the benefits of data mining?

Q15) Define prediction and description models

Q16) Describe challenges to Separate Data Warehouse regarding performance and functions issues.

Q17). Why Do We Need Data Warehouses?

Q18). How is a Data Warehouse different from an Operational DBMS.

Q19). Define an **operational system**

Q20) What is Decision tree?

Q21) What is meant by Pattern?

Q22) What are the different Applications of Data Mining

Q23) Explain the following Major Tasks in Data Processing

Q24) Explain the knowledge discovery phases.

Q25) Define the following terms:

Q26). What is Descriptive and predictive data mining?

Q27) **Write the preprocessing steps that may be applied to the data for prediction.**

Q28) **Why do you need data warehouse life cycle process? and what are the steps in the life cycle approach?**

Q29) **Explain OLAP and OLTP?**

Q30) **Write the main steps of Data cleaning tasks**

Q31) **Why Do We Need Data Warehouses?**

Q32) **Name some of the data mining applications?**

Q33) **Why is data quality so important in a data warehouse environment?**

Q34) **How can data visualization help in decision-making?**

Q35) **Explain the different types of data repositories on which mining can be performed?**

Q36) Explain the concept of Data Mining Classification Schemes

Q37) Define data warehouse. Draw the architecture of data warehouse and explain the three tiers in detail.

Q38) Draw and explain block diagram of Online Transaction Processing Cycle.

Q39) Consider the sales market transactions shown in table below, what is the Multidimensional OLAP Cube that can be derived from this data set.

Q40) What are the main advantages and disadvantages of MOLAP cube.

……

Q41) Cluster analysis is said to be a collection of objects. It is used in various application in the real world. Enumerate the applications of cluster analysis, in details.

……..

Q42) Explain in details each one of these steps.

……

Q43) Consider a database, ***D***, consisting of 9 transactions. Suppose ***min.support*** count required is 2 and let ***min.confidence*** required is 70%. Use the apriori algorithm to generate all the frequent candidate itemsets ***Ci*** and frequent itemsets ***Li***.

Then, generate the association rules from frequent itemsets using min. support & min. confidence take this case only ***l*** = {I1,I2,I5}.

TID	List of Items
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

……

Q44) Draw a flowchart to show how the K-Mean Clustering algorithm works?

……

Q45) Clustering technique is used in various fields of our real life enumerate five of the Clustering Applications.

……..

Q46) Explain in details each one of these steps.

1. **Decision Support System,**
2. **Market-Basket Data**
3. **Association rules**
4. **The Apriori algorithm Key Concepts**

……

Q47) Consider a database, ***D***, consisting of 5 transactions. Use this table to show the implementation of k-means algorithm together with Euclidean distance function. Use K=2 and suppose **A** and **C** are selected as the initial means.

i	X_1	X_2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

.....

Q48) When we can say the association rules are interesting?

Q49) Explain Association rule in mathematical notations.

Q50) Define support and confidence in Association rule mining.

Q51) Suppose that we have the following table of a database of transactions D, depending on these transactions determine Support and Confidence values for the following items I.

$X \Rightarrow Y$
Bread $\Rightarrow$ PeanutButter
PeanutButter $\Rightarrow$ Bread
Beer $\Rightarrow$ Bread
PeanutButter $\Rightarrow$ Jelly
Jelly $\Rightarrow$ PeanutButter
Jelly $\Rightarrow$ Milk

items I

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

a database of transactions D

Q52) Use PAM to Cluster the following data set of ten objects into two clusters i.e. k = 2. Consider a data set of ten objects as follows.

X_1	2	6
X_2	3	4
X_3	3	8
X_4	4	7
X_5	6	2
X_6	6	4
X_7	7	3
X_8	7	4
X_9	8	5
X_{10}	7	6

Let us assume x2 and x8 are selected as medoids, so the centers are
c1 = (3,4) and c2 = (7,4)

……

Q53) How are association rules mined from large databases?

Q54) What is the purpose of Apriori Algorithm?

Q55) How to generate association rules from frequent item sets?

Q56) Define the concept of classification and explain the main steps.

Q57) What is Decision tree?

Q58) What is Hierarchical method?

Q59) Explain the General Steps of Hierarchical Clustering

Q60) Explain the Methods of Hierarchical Clustering and give example for each one.

Q61) Differentiate between Agglomerative and Divisive Hierarchical Clustering Algorithm?

Q62) Explain with Examples the various applications of Classification Algorithm.

Q63) Explain classification by Decision tree induction?

……..

Q64) Discuss (shortly) whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their profitability.

(b) Computing the total sales of a company.

(c) Sorting a student database based on student identification numbers.

(d) Predicting the outcomes of tossing a (fair) pair of dice.

(e) Predicting the future stock price of a company using historical records.

(f) Monitoring the heart rate of a patient for abnormalities.

(g) Monitoring seismic waves for earthquake activities.

(h) Extracting the frequencies of a sound wave.

……..

Q65) What are the main advantages and disadvantages of Decision Tree classification algorithms?

Q66) Explain the partitioning method of clustering.

Q67) State the categories of clustering methods?

Q68) Name some specific applications of cases where the data analysis task is Classification.

Q69) What are the essential steps in the process of making a decision approach?

Q70) How can data visualization help in decision-making?

Q71) Explain classification techniques by rule-based classifier and decision tree induction with example for each case?

Q72) Visual classification: an interactive approach to decision tree construction. Draw a flowchart to explain in detail all the main steps of visual classification.

Q73) What is descriptive and predictive data mining, name three examples for each one?

Q74) Multiple Choice Questions. Please choose the best answer for the following questions:-

1. Which of the following is the most important when deciding on the data structure of a data mart?
(a) XML data exchange standards
(b) Data access tools to be used
(c) Metadata naming conventions
(d) Extract, Transform, and Load (ETL) tool to be used

2. Which of the following is not a data mining functionality?
A) Characterization and Discrimination
B) Classification and regression
C) Selection and interpretation
D) Clustering and Analysis

3. Strategic value of data mining is .....
A) cost-sensitive
B) work-sensitive
C) time-sensitive
D) technical-sensitive

4. The output of KDD is .....
A) Data
B) Information
C) Query
D) Useful information

5. Which one manages both current and historic transactions?
(a) OLTP
(b) OLAP

(c) Spread sheet
(d) XML

6. Which of the following process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evolution and knowledge presentation?
(a) KDD process
(b) ETL process
(c) KTL process
(d) MDX process

7. Data modeling technique used for data marts is
(a) Dimensional modeling
(b) ER – model
(c) Extended ER – model
(d) Physical model

8. Which of the following tools a business intelligence system will have?
(a) OLAP tool
(b) Data mining tool
(c) Reporting tool
(d) Both(a) and (b) above

9. Which of the following is not a kind of data warehouse application?
A) Information processing
B) Analytical processing
C) Data mining
D) Transaction processing

10. The data is stored, retrieved and updated in .....
A) OLAP
B) OLTP

C) Data Mart
D) FTP

11. The ..... allows the selection of the relevant information necessary for the data warehouse.

A) top-down view
B) data warehouse view
C) data source view
D) business query view

Q75) What is ETL?

Q76) What is data warehouse architectures: conceptual view, explain in details.

Q77) Decision support systems are used for

a. Management decision making
b. Providing tactical information to management
c. Providing strategic information to management
d. Better operation of an organization

Q78) Decision support systems are essential for

A. Day–to-day operation of an organization.
B. Providing statutory information.
C. Top level strategic decision making.
D. Ensuring that organizations are profitable.

Q79) Multiple Choice Questions. Please choose the best answer for the following questions:-

1. Data mining is best described as the process of
a. identifying patterns in data.
b. deducing relationships in data.
c. representing data.
d. simulating trends in data.

2. Data used to build a data mining model.
   a. validation data
   b. training data
   c. test data
   d. hidden data

3. Classification problems are distinguished from estimation problems in that
a. classification problems require the output attribute to be numeric.
b. classification problems require the output attribute to be categorical.
c. classification problems do not allow an output attribute.
d. classification problems are designed to predict future outcome.

4. This approach is best when we are interested in finding all possible interactions among a set of attributes.
a. decision tree
b. association rules
c. K-Means algorithm
d. genetic learning

5. This step of the KDD process model deals with noisy data.
a. Creating a target dataset
b. data preprocessing
c. data transformation
d. data mining

6. This clustering algorithm initially assumes that each data instance represents a single cluster.

a. agglomerative clustering
b. conceptual clustering
c. K-Means clustering
d. expectation maximization

Q80) Construct a decision tree with root node *Type* from the data in the table below. The first row contains attribute names. Each row after the first represents the values for one data instance. The output attribute is *Class*.

Scale	Type	Shade	Texture	Class
One	One	Light	Thin	A
Two	One	Light	Thin	A
Two	Two	Light	Thin	B
Two	Two	Dark	Thin	B
Two	One	Dark	Thin	C
One	One	Dark	Thin	C
One	Two	Light	Thin	C

Q81) What is outlier analysis?

Q82) Explain each one

1. What is data cleaning?

2. What is data integration?

3. What is data transformation?

Q83) What are the two steps in data classification?

Q84) What is the difference between classification and clustering?

Q85) What are hierarchical methods, and give example for each one?

Q86) List out some clustering methods.

Q87) Define the following terms: Data Mart, MOLAP, OLTP,

Q88) The data classification process includes two steps of training set and test set explain these steps with example.

Q89) List the general steps of hierarchical clustering algorithm.

Q90) Briefly discuss the hierarchical agglomerative algorithm and convert the following scenario to agglomerative.
In this scenario, after the second step of the agglomerative algorithm will yield clusters, {a} {b c} {d e} {f}. In the third step will yield clusters {a} {b c} {d e f}, which is a clustering, in the fourth step will give a small number but larger clusters that are {a} {b c d e f} and finally will yield cluster of {a b c d e f}.

Q91) Explain each one with example.

1. Traditional Hierarchical Clustering
2. Traditional Dendrogram
3. Non-traditional Hierarchical Clustering
4. Non-traditional Dendrogram

Q92) Explain the market basket analysis problem

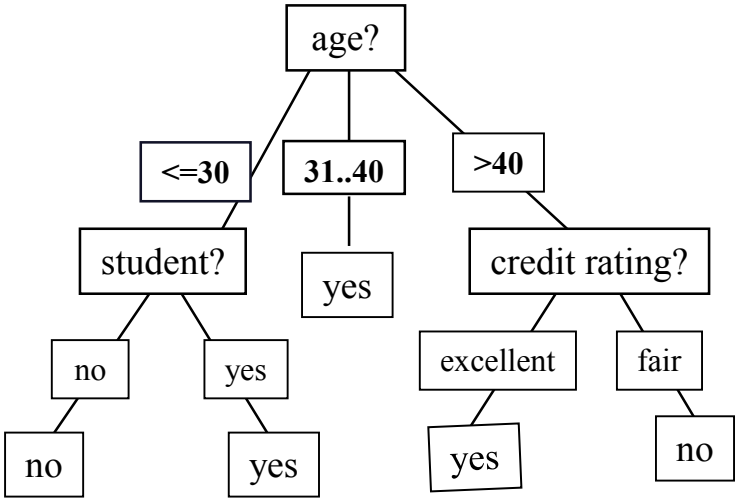
Q93) Consider a database, ***D***, consisting of 7 transactions. Use this table to show the implementation of k-means algorithm together with Euclidean

distance function. Use K=2 and suppose **1** and **4** are selected as the initial means.

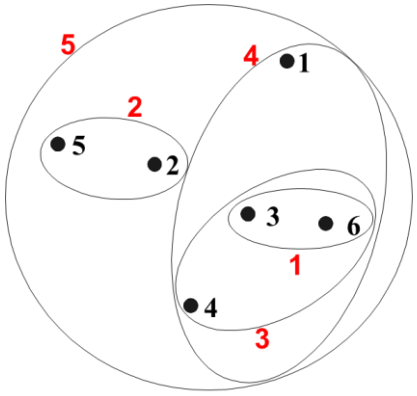
Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Q94) List the basic steps to develop a clustering task

- Q95) Extract a rule based system from a decision tree given bellow, use rule-based ordering technique.



Q96) Explain the dendrogram of the hierarchical technique and convert the numbers of the figure below of nested clusters to a dendrogram.



Q97) State the advantages of the decision tree approach over other approaches for performing classification.

Q98) Explain in detail the coverage of a rule and accuracy of a rule methods of a data mining classification with example for each one.

Q99) What do you mean by hierarchical cluster analysis.

Q100) What do you mean by the Apriori algorithm Key Concepts.

Q101) Consider a database, D, consisting of 4 transactions. Suppose min.support count required is 2 and let min.confidence required is 70%. Use

the apriori algorithm to generate all the frequent candidate itemsets Ci and frequent itemsets Li.

Data base D

TID	Items
10	a, c, d
20	b, c, e
30	a, b, c, e
40	b, e

……

Q102) Consider a database, D, consisting of 4 transactions. Suppose min.support count required is 2 and let min.confidence required is 70%. Use the apriori algorithm to generate all the frequent candidate itemsets Ci and frequent itemsets Li.

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

……