b) Find the frequent item sets with support count = 2, using apriori method for following transactional database : 8

TID	List of items
T100	M, P
T200	K, N
T300	K, C
T400	M, K, N
T500	M, C
T600	K, C
T700	M, C
T800	M, K, C, P
T900	M, K, C

c) Explain how we can improve the efficiency of apriori algorithm. 6

4. a) Explain the approaches to mining multilevel association rules. 6

b) What is the DMQL syntax for the following : 6

i) Association ii) Characterization iii) Concept hierarchy ?

c) Explain decision tree induction. 8

MODULE – III

5. a) What are spatial databases ? How spatial cube is constructed ? 9

b) Explain k-medoid method with an example. 6

c) Explain optics technique. 5

6. a) Explain any one grid based method. 6

b) Explain any two density based method used for cluster analysis. 10

c) Differentiate between classification and clustering. 4

MODULE – IV

7. a) Write a short note on : 10

i) Social impacts of data mining ii) Trends in data mining.

b) Explain any 4 multidimensional features used to choose a data mining system. 10

8. a) Explain how data mining can be used for biomedical and DNA analysis. 10

b) Discuss the privacy protection principles in the context of data mining. 10

B.E. (IT) (Sem

DISTR

Duration : 3 Hours

Instructions : i)

ii

iii

1. a) Explain the basic o

b) What do you mean of transparency. W highest degree of

c) Differentiate betw microkernel shou

2. a) Explain the follo principal choices

i) Reliability

b) In near future dis of such system s

c) Describe the fu

d) Explain the des

3. a) Explain the app

b) Explain Logic variations in th How does this system difficu algorithm to h

IT 7 EL 1 (1)

B.E. (IT) (Semester – VII) Examination, December 2009
DATA MINING AND WAREHOUSES (E-I)

ation : 3 Hours Total Marks : 100

***Instruction** : Attempt **any five** questions such that at least **one** question from **each** Module is **selected**.*

MODULE – I

a) What is the main difference between the data warehouse and operational database system ? With an example explain the three steps in the process of knowledge discovery in databases. 10

b) Explain the following types of databases : 10

i) Relational database

ii) Transactional database

iii) Object oriented database

iv) Object relational database

v) Text database and multimedia database.

a) Do we prefer to use the OLAP operations in the multimedia data model ? If yes then what are those operations and if no then why are they not preferred ? 6

b) Explain how data mining is considered as a step in the process of knowledge discovery. 8

c) Discuss the methods for filling in the missing values for the attributes in data cleaning. 6

MODULE – II

a) Do we prefer to use the neural network in data mining ? What are the limitations and consequences of choosing neural network in decision support systems ? 6

P.T.O.

4. a) Write the main steps in the algorithm for classification using prediction. 5

b) Consider a two dimensional database D with the records : R1 (2, 2), R2 (2, 4), R3 (4, 2), R4 (4, 4), R5 (3, 6), R6 (7, 6), R7 (9, 6), R8 (5, 10), R9 (8, 10), R10 (10, 10). Show the results of the k-means algorithm at each step, assuming that you start with two clusters (k = 2) with centers C1 = (6, 6) and C2 = (9, 7). Write the main steps you follow. 8

c) Write short notes on 'Grid Based Clustering'. 7

MODULE – III

5. a) Why is outlier mining important ? Briefly describe the following different approaches used for outlier analysis. 12

i) Statistical-based outlier detection.

ii) Distance-based outlier detection.

b) With the help of the neat labeled diagram explain the data warehouse architecture. 8

6. a) What do you understand by slice and dice ? Give an example. 5

b) What is a data warehouse ? Consider a cube used to analyze sales by location, item and date (weekly). Assume data is available for 1000 locations, 10,000 items for 100 weeks. Discuss an efficient scheme to compute the resulting data cuboid. 9

c) How does a snowflake scheme differ from a STAR schema ? Explain with an example. Name any two disadvantages of snowflake schema. 6

MODULE – IV

7. a) Explain the following types of parallelism used for parallel execution of the tasks within SQL statements. 4

i) Horizontal parallelism

ii) Vertical parallelism

b) With the help of a neat labeled diagram explain the Shared Memory Architecture for parallel processing. 8

c) With the help of a diagram compare the MOLAP v/s ROLAP architectures. 8

8. a) Write short notes on : 16

i) Metadata Interchange Initiative

ii) Multidimensional data model.

b) How is mining for user behavior on the web carried out ? 4

MODULE – II

3. a) Clarify the difference between classification and clustering.

b) Explain the significance of 'Naive' used by the Naïve Bayesian Classifiers. Briefly describe the difference between Naive Bayes Classification and Bayesian Belief Networks.

c) Consider the following training data set in Table 1 :

Table 1

Attribute 1	Attribute 2	Attribute 3	Class
A	70	True	Class 1
A	90	True	Class 2
A	85	False	Class 2
A	95	False	Class 2
A	70	False	Class 1
B	90	True	Class 1
B	78	False	Class 1
B	65	True	Class 1
B	75	False	Class 1
C	80	True	Class 2
C	70	True	Class 2
C	80	False	Class 1
C	80	False	Class 1
C	96	False	Class 1

i) Calculate the gain on Attribute 1 as Gain (x_1)

ii) Explain the main steps in construction of the decision tree using the information in i).

IT 7 (EI) 1 (RC)

B.E. (IT) (Semester – VII) (RC) Examination, Nov./Dec. 2014

DATA MINING AND WAREHOUSES

Elective – I

Max. Marks : 100

ıration : 3 Hours

Instructions : 1) Attempt **any five** questions by selecting atleast **one** from **each** Module.

2) Assume **suitable** data **if** necessary.

MODULE – I

1. a) Define :

i) Data Mining

ii) Knowledge Discovery

State the difference between the two. Draw a complete labeled diagram of a typical data mining system. 8

b) The age values for the data tuples are : 6

20, 20, 21, 22, 22, 25, 25, 25, 25, 13, 15, 16, 16, 19, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Use min-max normalization to transform the values 20, 30, 40 and 70 in the range [0.0, 1.0].

c) Write the four factors which test the interestingness of the patterns. 2

d) What is Iceberg Query ? Give its general syntax. 4

2. a) What is association Rule Mining ? What are the two steps in the process ? Explain. 6

b) Suppose that we have sales data given by Address and the Address fields include House Number, Street Name, City, State, Pincode and Country. Write a DMQL statement for expressing the concept hierarchy. 4

c) Suppose a group of 12 sales price records has been sorted as follows : 8

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.

i) Equidepth partitioning ii) Equi-width partitioning.

d) What is numerosity reduction using regression ? 2

P.T.O.

IT 7 (EI) 1 (RC)

B.E. (IT) (Semester – VII) (RC) Examination, Nov./Dec. 2014

DATA MINING AND WAREHOUSES

Elective – I

Max. Marks : 100

ıration : 3 Hours

Instructions : 1) Attempt **any five** questions by selecting atleast **one** from **each** Module.

2) Assume **suitable** data **if** necessary.

MODULE – I

1. a) Define :

i) Data Mining

ii) Knowledge Discovery

State the difference between the two. Draw a complete labeled diagram of a typical data mining system. 8

b) The age values for the data tuples are : 6

20, 20, 21, 22, 22, 25, 25, 25, 25, 13, 15, 16, 16, 19, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Use min-max normalization to transform the values 20, 30, 40 and 70 in the range [0.0, 1.0].

c) Write the four factors which test the interestingness of the patterns. 2

d) What is Iceberg Query ? Give its general syntax. 4

2. a) What is association Rule Mining ? What are the two steps in the process ? Explain. 6

b) Suppose that we have sales data given by Address and the Address fields include House Number, Street Name, City, State, Pincode and Country. Write a DMQL statement for expressing the concept hierarchy. 4

c) Suppose a group of 12 sales price records has been sorted as follows : 8

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into three bins by each of the following methods.

i) Equidepth partitioning ii) Equi-width partitioning.

d) What is numerosity reduction using regression ? 2

P.T.O.

i) Classify the tuple X = (age = Youth, Income= Medium, Student= Yes, Credit Rating = Fair); using Naïve Bayesian Classifier. 3

c) Define Accuracy, Precision and Recall; metrics for evaluating Classifier performance. 8

4. a) Using the data given in Question [3b)]; classify the data sample X = (Age = Youth, Income = High, Student = Yes, Credit Rating = Fair); using L_2-norm and L_1-norm for 3NN classifier. 7

b) Give the k-mediods Algorithm (PAM) for Clustering. With an example bring out the drawback of k-means algorithm. 5

c) Which clustering technique is the algorithm DBSCAN based on ? How does the DBSCAN find clusters ?

MODULE – 3

5. a) Is outlier detection import in the study of Data Mining. What are the some of the challenges of outlier detection ? 6

b) Find the outlier among the listed temperatures using the mean-variance method, assuming that it is normal distributed. T_{MIN} = {18.7, 18.6, 18.3, 18.5, 18.0, 17.9, 18.3, 18.4, 18.5, 18.6, 18.7, 18.6}. 6

c) Give the DB(r, π) – outlier detection Algorithm. How does it compute outliers ? 8

6. a) Using an example explain atleast four OLAP operations on multi-dimensional data. 6

b) Differentiate between OLTP and Data Warehousing. 6

c) Explain the different steps in building a Data Warehouse. 8

MODULE – 4

7. a) With the help of a neat diagram explain the Shared Disk Architecture for parallel Database processing. 8

b) Explain the following : 6

i) HOLAP

ii) ROLAP

iii) MOLAP.

c) Differentiate between multidimensional OLAP and multi-relational OLAP. 6

8. a) Write short note on techniques for Web Usage Mining. 10

b) Write notes on any two applications of Data Mining techniques. 10

IT 7 – (E-I) 1(RC)

B.E. (IT) Semester – VII (RC) Examination, Nov./Dec. 2015

DATA MINING AND WARE HOUSING

Duration : 3 Hours

Total Marks : 100

Instruction : *Answer **any five** questions by selecting at least **one** question from **each** Module.*

MODULE – 1

1. a) What is Data Mining ? Explain the different steps involved in Knowledge Discovery from Data. 8

   b) Given the vectors x = (12, –12, 10, 20, 10, –30) and y = (–10, 10, –10, 10, 10, –10), calculate the proximity between them using the following measures : 6

      i) Cosine

      ii) L_2 norm

      iii) Tanimoto coefficient.

   c) Given the following data :

      4, 18, 15, 21, 22, 24, 24, 25, 25, 27, 28, 34.

      Try "smoothing by bin medians" and "smoothing by bin boundaries", on above data with bin size of 4. 6

2. a) Explain the following types of databases : 8

      i) Multimedia database

      ii) Relational database

      iii) Object oriented database

      iv) Transactional database.

   b) Differential between Data Mart and Data Warehouse. 4

P.T.O.

c) Find the frequent item sets with minimum support count of 3 for the following transactional data :

T_ID	Bread	Milk	Detergents	Shampoo	Eggs	Soft Drink
1	1	1	1	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

MODULE – 2

3. a) Differentiate between Clustering and Classification. Give a brief application for each.

b) Consider the following data :

TID	AGE	Income	Student	Credit Rating	Class : Buys Comp
1	Youth	High	No	Fair	No
2	Youth	High	No	Excel	N
3	Mid_Age	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	No
6	Senior	Low	Yes	Excel	No
7	Mid_age	Low	Yes	Excel	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excel	Yes
12	Mid_Age	Medium	No	Excel	Yes
13	Mid_Age	High	Yes	Fair	Yes
14	Senior	Medium	No	Excel	No

i) Classify the tuple X = Credit Rating = Fair};

c) Define Accuracy, Precisi performance.

4. a) Using the data given in Qu Income = High, Student L_1-norm for 3NN classifi

b) Give the k-mediods Algo out the drawback of k-m

c) Which clustering techni the DBSCAN find cluste

5. a) Is outlier detection imp the challenges of outlie

b) Find the outlier among th assuming that it is nor 17.9, 18.3, 18.4, 18.5,

c) Give the DB(r, π) – ou outliers ?

6. a) Using an example exp data.

b) Differentiate between

c) Explain the different s

7. a) With the help of a ne parallel Database pr

b) Explain the following
i) HOLAP
ii) ROLAP
iii) MOLAP.

c) Differentiate betwee

8. a) Write short note on

b) Write notes on any

MODULE – II

3. a) Why is tree pruning useful in decision tree induction ? How does tree pruning work ?
   b) Given two objects $A_1(3, 10)$ and $A_2(2, 5)$
      i) Compute the Euclidean distance between the two objects.
      ii) Compute the Manhattan distance between the two objects.
      iii) Compute the Minkowski distance between the two objects (use $p = 3$).
   c) Design an efficient method that performs effective naive Bayesian classification over an infinite data stream.
   d) What are the characteristics for agglomerative and divisive hierarchial clustering ?

4. a) Explain the K-means algorithm for clustering. Differentiate between K-means and K-medoids.
   b) Differentiate between Bayesian belief networks and Naive Bayesian classifier.
   c) Write short notes on :
      i) Density based methods.
      ii) Hierarchial methods.

(5+5)

MODULE – III

5. a) Differentiate between OLTP and Data Warehousing.
   b) What is anamoly detection ? List and explain anomaly detection methods.
   c) Mention the need, functions and applications of data warehouses in the field of data mining.

6. a) With the help of a neat labelled diagram, explain the data warehouse architecture.
   b) Write note on the following :
      i) Star Schema
      ii) Snowflake Scheme
      iii) Fact Constellation Schema
      iv) Efficient computation of data cube.

MODULE – IV

7. a) What is web usage mining ? Explain. 5

b) Explain in brief the following :

i) Intraquery parallelism. 6

ii) Interquery parallelism.

c) Compare ROLAP, MOLAP and hybrid servers with its architecture. 9

8. a) With the help of a neat diagram explain the shared disk architecture for parallel processing. 8

b) Write short notes on : 10

i) Web content Mining.

ii) OLAP tools and internet. 2

c) List the four types of data partitioning techniques.

IT 7 – (E-I) 1(RC)

B.E. (IT) (Semester – VII) (RC) Examination, Nov./Dec. 2016

DATA MINING AND WAREHOUSING

Duration : 3 Hours

Total Marks : 100

Instructions : 1) Attempt **any five** questions, by selecting atleast **one** question from **each** Module.
2) Assume necessary data if **required**.

MODULE – I

1. a) Explain how the data transformation and integration steps are carried out in data mining. 10

   b) Explain concept hierarchy generation for numerical data. 5

   c) The age values for the data tuples are
   20, 20, 21, 22, 22, 25, 25, 25, 25, 13, 15,16, 16, 19, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   Suggest a method of data smoothing that can be used on the above data. Show the working. 5

2. a) How is data warehouse different from a database ? How are they similar ? 4

   b) List and elaborate the major issues in data mining. 6

   c) Consider the following data : 10

T_ID	Items Purchased
101	a, b, e
102	b, d
103	b, c
104	a, b, d
105	a, c
106	b, c
107	a, c
108	a, b, c, e
109	a, b, e

Using apriori algorithm, find candidate itemsets and frequent itemsets. (Assume minimum support count = 2).

P.T.O.

MODULE – III

5. a) Describe major applications of data warehouse. 6
   b) Write short note on : 6
      i) Deviation based technique
      ii) Statistical Based Techniques.
   c) Explain with suitable example following data warehouse schema : 8
      i) Star Schema 6
      ii) Snowflake schema.

6. a) Differentiate between OLTP and OLAP. 6
   b) Explain different categories of measures in data warehouse. 8
   c) Explain Multi tier architecture of data warehouse.

MODULE – IV

7. a) Write short note on : 8
      i) Shared memory architecture
      ii) Shared nothing architecture. 6
   b) What is significance of user behavior Mining ? 6
   c) Give some important features of OLAP servers. 8

8. a) Write short note on :
      i) HOLAP
      ii) MOLAP. 6
   b) Explain different data partitioning schemes. 6
   c) Explain in details different query and reporting tools.

---

b) List and explain different data mining stores on which data mining can be performed.

c) Explain pattern interestingness measure as one of the data mining primitives.

d) Write short note on Data Mining Query language.

MODULE – II

3. a) Write short note on following clustering techniques :
   i) Density based methods
   ii) Grid based methods.

b) Draw decision tree for following data set. Explain steps.

CustID	Gender	Car Type	Income	Class
1	M	Family	Average	C0
2	F	Sports	High	C1
3	M	Luxury	High	C1
4	M	Family	Low	C0
5	M	Sports	Average	C0
6	F	Luxury	High	C1
7	F	Luxury	High	C1
8	M	Family	Low	C1
9	F	Luxury	High	C0
10	M	Sports	High	C1

c) Explain K nearest neighbor classification with suitable example. 6

4. a) Consider following dataset 1, 2, 6, 7, 8, 10, 15, 17, 20. From 3 cluster using k-medoid algorithm and considering 6, 7, 8 as 3 medoids of 3 clusters respectively. 8

b) What are the requirements of clustering in data mining ? 4

c) List and explain different applications of classification and prediction. 5

d) Compare the advantages and disadvantages of eager classification versus lazy classification. 3

b) Compute Association rules using Apriori Algorithm for the following set of t... support =3 and Confidence=80%

TID	Item List
100	A,D,E
200	A,B,C,E
300	A,B,D,E
400	A,C,D,E
500	B,C,E
600	B,D,E
700	C,D
800	A,B,C,
900	A,D,E
1000	A,B,E

c) Write a short note on Data Mining Query Language.

d) How is data warehouse different from a database? How are they similar?

Module – II

Q3 a) Briefly outline the major steps of decision tree classification.

b) Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):
   a) Compute the Euclidean Distance between the two objects.
   b) Compute the Manhattan Distance between the two objects.
   c) Compute the Minkowski Distance between the two objects, using q=3.

c) Explain two types of data in cluster analysis.

Q.4a) Differentiate between Clustering and Classification. Give a brief application for each.

b) Consider the following data:

TID	Home Owner	Marital Status	Annual Income	Class: Loan Defaulter
1	Y	S	125	N
2	N	M	100	N
3	N	S	70	N
4	Y	M	120	N
5	N	D	95	Y
6	N	M	60	N
7	Y	D	220	N
8	N	S	85	Y
9	N	M	75	N
10	N	S	90	Y

i. Calculate the best splits using Shannon's Entropy measure and GINI index.
ii. Show the resulting decision trees using both the measures calculated above.

Differentiate between Bayesian Belief Networks and Naïve Bayesian Classifier. 3

Module – III

Write short notes on the following anomaly detection methods: 9
i. Supervised methods.
ii. Semi-supervised methods.
iii. Unsupervised methods.

With the help of a neat labelled diagram explain the data warehouse architecture. 8

Are data warehouses a pre-requisite for data mining? Why/Why not? 3
Explain with suitable example the following data warehouse schema: 6
a) Star Schema
b) Snowflake Schema

) Differentiate between OLTP and OLAP. 6

:) Explain multidimensional data model and its use. 6

d) What is Data Cube? Given an Example. 2

Module – 4

7 a) What is metadata with regard to Databases? What are the contents of the metadata in data warehousing? 6

b) Explain the following architecture for parallel processing: 8
i. Shared memory architecture
ii. Shared nothing architecture

c) Write a short note on the following: 6
a) MOLAP Architecture
b) ROLAP Architecture

Q.8 a) List and briefly write the significance of 4 types of data partitioning techniques. 4

b) Write short notes on : 10
a) Web Content Mining
b) Metadata Interchange Initiative

c) Using a neat diagram, explain web processing model. 6

c) Explain Min-Max Normalization

Module II

Q.3 a) Construct a decision tree for the data give. Use entropy and information gain.

Age	Income	Student	Credit_rating	Buys_computer
<= 30	High	N	Fair	No
<= 30	High	N	Excellent	No
31 ... 40	High	N	Fair	Yes
> 40	Medium	N	Fair	Yes
> 40	Low	Y	Fair	Yes
> 40	Low	Y	Excellent	No
31 ... 40	Low	Y	Excellent	Yes
<= 30	Medium	N	Fair	No
<= 30	Low	Y	Fair	Yes
> 40	Medium	Y	Fair	Yes
<= 30	Medium	Y	Excellent	Yes
31 ... 40	Medium	N	Excellent	Yes
31 ... 40	High	Y	Fair	Yes
> 40	Medium	N	Excellent	No

b) Write a short note on DBSCAN. 08

Q.4 a) Consider the objects given below. Assume $C_1 = x_2$ and $C_2 = x_8$ perform K-medoid using Manhattan distance. Note down the Error criterion for the next iteratori choose a random object x_7 as medoid. Note down the inference. 10

$x_1(2,6)$ $x_2(3,4)$ $x_3(3,8)$ $x_4(4,7)$ $x_5(6,2)$ $x_6(6,4)$ $x_7(7,3)$ $x_8(7,4)$ $x_9(8,5)$ $x_{10}(7,6)$

b) Jack, Mary and Jin are subjected to tests for certain illness they are suffering from. The results are as follow. 10

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	Male	Yes	N	P	N	N	N
Mary	Female	Yes	N	P	N	P	N
Jin	Male	Yes	P	N	N	N	N

i) Justify whether Jack and Mary suffer from same illness
ii) Justify whether Jack and Jin suffer from same illness
iii) Justify whether Jin and Mary suffer from same illness.

BE831

Paper / Subject Code: BE832 / Elective-I (1) Data Mining & Warehousing.

BE832

Total No. of Printed Pages:3

B.E. (Information Technology) Semester- VII (Revised Course 2007-08)
EXAMINATION NOV/DEC 2019
Elective-I
Data Mining & Warehousing

[Duration : Three Hours] **[Total Marks :100]**

Instructions:
1) Answer **any five**questions with at least **one** from **each Module.**
2) **All** questions carry **equal** marks.
3) Make suitable assumptions **wherever** necessary.

Module – 1

Q.1 a) What is Data Mining? How is data mining a step in the process of knowledge discovery? 7

b) Suppose the data for the analysis include the attribute age. The age values for the data tuples are: 8
20, 20, 21, 22, 22, 25, 25, 25, 25, 13, 15, 16, 16, 19, 30, 33, 33, 35,
35, 35, 35, 36, 40, 45, 46, 52, 70.

Evaluate the following:

i. Use Min-Max Normalization to transform the values 20, 30, 40 and 70 in the range [0.0, 1.0]

ii. If the Standard Deviation of age is 12.94 years, what are the Z-Scores corresponding to 20, 30, 40 and 70 years.

iii. Use Smoothing by Bin Means to smooth the data, using a Bin Depth of 3.

c) Explain briefly the techniques for dimensionality reduction. 5

Q.2 a) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. 5

Module III

Q.5
a) With the help of a neat diagram explain data warehouse architecture. 10
b) Write short notes on 10
i) Deviation Based Technique
ii) Statistical Based Technique

Q.6
a) Write short notes on [5 × 4 = 20]
i) Star Schema
ii) Snowflake schema
iii) Fact constellation schema
iv) Data cube

Module IV

Q.7
a) Write short notes on 20
i) Web content mining
ii) ROLAP
iii) MOLAP
iv) Metadata Interchange Initiative

Q.8
a) What is the significance of user behavior mining? 06
b) Discuss privacy protection technique with regard to data Mining. 10
c) Explain the need for OLAP. 04

Paper / Subject Code: BE832 / Elective-I (1) Data Mining & Warehousing.

BE832

Total No. of Printed Pages:03

B.E. (Information Technology) Semester- VII (Revised Course 2007-08)
EXAMINATION MAY/JUNE 2019
Elective-I (1) Data Mining & Warehousing.

[Max. Marks :100]

[Duration : 3 Hours]

Instructions:

a) Assume data whenever necessary.
b) Draw neat labeled diagram using pencil and rules
c) Answer any five questions by selecting at least one from each module

Module - I

Q.1 a) You are given a transaction data as shown below from a fast food restaurant. For simplicity we assign the meal items short names [M1- M5]. 10

Meal- Item	List of item	Meal- Item	List of Items
Order 1	{M1, M2, M5}	Order 6	{M2, M3}
Order 2	{M2, M4}	Order 7	{M1, M3}
Order 3	{M2, M3}	Order 8	{M1, M2, M3, M5}
Order 4	{M1, M2, M4}	Order 9	{M1, M2, M3}
Order 5	{M1, M3}		

For all the min_sup 2/9 and min_conf=7/9. Apply Apriori and identify all k-frequent itemsets. Find all the strong association rules.

b) Construct the FP-tree for the database above. Consider min_sup=2. 10

Q.2 a) Define maximal and closed frequent itemset identify the above from the database given as 10

Transaction ID	Items
T1	{A, C, T, W}
T2	{C, D, W}
T3	{A, C, T, W}
T4	{A, C, D, W}
T5	{A, C, D, T, W}
T6	{C, D, T}

b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are 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. 06
Draw the box plot and indicate the (5-No. summary) five number summary. Also indicate the possible outliers

Module III

a) Write short notes on:- [5 × 4 = 20mks]
   i) Graphical based anomaly detection techniques
   ii) Statistical based anomaly detection technique
   iii) Distance based anomaly detection technique
   iv) Model based anomaly detection technique

a) Explain basic functions of a data warehouse. [6 mks]
b) With the help of a neat diagram explain overall architecture of a data ware house. [8mks]
c) Are the data ware houses a pre-requisite for data mining? Why/Why not? [6 mks]

Module IV

a) How can mining be used to identify the users behavior on the web? [4 mks]
b) Explain web content mining. [8 mks]
c) Discuss the benefits of data mining for the financial data analysis. [8 mks]

a) Write short notes on:- [20 mks]
   i) Advantages of DB Mines.
   ii) Visual and Audio Data Mining
   iii) Query and Reporting Tools
   iv) Data Partitioning scheme

Module II

3. a) Construct a decision tree for the data given below. Generate the rules from the constructed decision tree. [12 mks]

Outlook	Temperature	Humidity	Windy	Play-Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Use Entropy and Information gain.

b) Perform clustering [hierarchical] for the matrix given below and draw the dendrogram. [8 mks]

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

4. a) Use k-means algorithm to cluster the data into 3-cluster. [8 mks]
{3,5,11,12,4,21,32,12,28}
b) Explain k NN with the help of an example. [8 mks]
c) Explain Nominal and Ordinal data. [4 mks]

Q.3 a) Construct first level of decision tree for classification utilizing the following class labelled data, using information Gain as Attribute Selection Method. Target Class Buys_ Computer ="Yes" Or "No"

No	Age	Income	Student	Credit_ Rating	Class: Buys_Computer
1	Youth	High	N	Fair	No
2	Youth	High	N	Excellent	No
3	Middle_Aged	High	N	Fair	Yes
4	Senior	Medium	N	Fair	Yes
5	Senior	Low	Y	Fair	Yes
6	Senior	Low	Y	Excellent	No
7	Middle_Aged	Low	Y	Excellent	Yes
8	Youth	Medium	N	Fair	No
9	Youth	Low	Y	Fair	Yes
10	Senior	Medium	Y	Fair	Yes
11	Youth	Medium	Y	Excellent	Yes
12	Middle_Aged	Medium	N	Excellent	Yes
13	Middle_Aged	High	Y	Fair	Yes
14	Senior	Medium	N	Excellent	No

b) Why is "Naïve Bayes Classifier" called Naïve? (4 Marks)

c) From the following data predict value of price for a distance value of 50 using liner regression.

Distance (kms)	40	42	45	48	52
Price (INR)	8500	8250	8000	7750	6500

PART B

Answer **any two** questions from the following.

Q.4 a) What is Outlier Analysis/ Anomaly Detection? Explain different variations of Anomaly Detection Problems. (6 Marks

b) Explain Graphical/ Visual approaches of anomaly detection. List their limitations. (6 Marks)

c) Compare Distance Based and Density based approaches of outlier detection. (8 Marks)

Q.5 a) Explain why Graph Mining is important (6 Marks)

b) Explain briefly how social networks behave. (6 Marks)

c) What is Spatial Data Mining? (8 Marks)

a) What is a Data Warehouse? (4 Marks)

b) List and explain differences between Operational Database Systems and Data Warehouses. (8 Marks)

c) Explain in detail various steps of ETL process. (8 Marks)

PART C

Answer **anyone** questions from the following. (8 Marks)

a) List and explain any 3 Major Clustering Approaches (6 Marks)

b) Explain How K-Means Clustering algorithm work. (6 Marks)

c) Explain how DBSCAN Density –Based Spatial Clustering algorithm works. (8 Marks)

a) Explain different types of OLAP Servers. (6 Marks)

b) With a diagram explain Star Schema for Data Warehouse design. (6 Marks)

c) Explain OLAP operations of Roll-up and Drill-down. (6 Marks)

Paper / Subject Code: BE832 / Elective-I (1) Data Mining & Warehousing.

BE832

tal No. of Printed Pages:03

B.E. (Information Technology) Semester-VII (Revised Course 2007-08)
EXAMINATION Aug/Sept 2019
Elective-I
Data Mining & Warehousing

[Max. Marks : 100]

uration : Three Hours]

structions:

1. Assume data wherever required.
2. Draw neat labelled diagrams using pencil and ruler.
3. Answer any five questions by selecting at least one from each module.

Module I

1. a) Consider the transactional data base given as

TID	Items	TID	Items
1	{a,b}	6	{a,b,c,d}
2	{b,c,d}	7	{a}
3	{a,c,d,e}	8	{a,b,c}
4	{a,d,e}	9	{a,b,d}
5	{a,b,c}	10	{b,c,e}

i) Mine the association rules using A-Priori for frequent k-itemsets. **[10 mks]**

ii) Construct FP-Tree for frequent itemsets. **[10 mks]**

Assume min_sup=2

[6 mks]

a) Consider the data for the feature price:
4,8,15,21,21,24,25,28,34
Suppose n=3 i.e. size of bin.
i) Perform smoothing by bin means
ii) Perform smoothing by bin median
iii) Perform smoothing by bin boundaries. **[6 mks]**

b) Perform correlation analysis for the data given.

A	B
1	1
2	2
3	5
4	3

[8 mks]

c) Explain the following:-
i) Concept Hierarchy Generation
ii) Antimonotone Property
iii) Data Preprocessing
iv) Data Integration and Transformation

Total No. of Printed Pages: 03

T.E. (Information Technology) (Sem-VI) (Revised Course 2016-2017)
EXAMINATION Nov/Dec 2019
Data Mining

[Duration : Three Hours] **[Total Marks : 100]**

Instruction:

1. Figures to the right indicate full marks.
2. Answer **any five** questions by selecting **two** questions from **Part A** and **two** from **Part B** and one question from **Part C.**
3. Make suitable assumptions if required.

PART A

Answer **any two** questions from the following.

Q.1 a) What is Data Mining? Explain briefly. **(4 Marks)**

b) List and explain any two major issues in Data Mining? **(8 Marks)**

c) Use Apriori Algorithm to find frequent itemsets, where minimum support count is 2. **(8 Marks)**

Transaction Id	List of Item Ids
T1	I1,I2,I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I3,I3,I5
T9	I1,I2,I3,I6

Q.2 a) Explain different ways in which missing values in Data can be handled. **(6 Marks)**

b) Find Mean, Mode, Median, and Standard Deviation for Temperature and Sales Data. **(10 Marks)**
Also find covariance and correlation between the Temperature and Sales.

Temperature Deg. Celsius	27	28	32	35	39	34	27	28	31	33
Sales(INR)	800	900	600	300	250	600	900	1000	600	350

c) Smoothen following data using bin size of 3 and using smoothing by bin means. **(4 Marks)**
10, 12, 14, 16, 18, 20, 21, 21, 23, 24, 24, 24