Duration: 3hrs **[Max Marks: 80]**

N.B. : (1) Question No 1 is Compulsory.
(2) Attempt any three questions out of the remaining five.
(3) All questions carry equal marks.
(4) Assume suitable data, if required and state it clearly.

1 Attempt any **FOUR** **[20]**

a Elucidate Market Basket analysis with an example. **[5]**

b A dimension table is wide, the fact table deep. Explain **[5]**

C Differentiate between OLTP and OLAP. **[5]**

d In real-world data, tuples with *missing values* for some attributes are common occurrence. Describe various methods for handling this problem. **[5]**

e Define initial load, incremental load and full refresh. **[5]**

2 a A database has five transactions. Let min sup count =3 and min conf =70% **[10]**

TID	Items
10	1,3,4
20	2,3,5
30	1,2,3,5
40	2,5
50	1,3,5

b 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 **[10]**
(i)Draw a star schema diagram for the above data warehouse.
(ii) 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?
(iii)To obtain the same list, write an SQL query assuming the data are stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).

3 a Explain Data mining as a step in KDD. Give the architecture of typical data mining. **[10]**

b Why is entity – relationship modeling technique is not suitable for data warehouse? How is dimensional modeling different? **[10]**

4 a Develop a model to predict the salary of college graduates with 10 years of work experience using linear regression. **[10]**

Year of experience (x)	Salary in $100 (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

b Show the dendogrm created by the complete link clustering algorithm for given set of points. **[10]**

	A	B
P1	2	4
P2	8	2
P3	9	3
P4	1	5
P5	8.5	1

5 a The college wants to record the marks for the courses completed by students using the dimensions a) Course b) Student c) Time and measure of aggregate marks. **[10]**
Create a cube and describe following operations
1) Roll up 2) Drill Down 3) slice 4) Dice

b Compare and contrast linear and logistic regression. **[10]**

6 a Demonstrate Multidimensional association Rule mining with suitable example. **[10]**

b 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) **[10]**
The distance function is Euclidean distance. Suppose initially we assign A1,B1,C1, as a center of each cluster, respectively. Use K-means algorithm to show only
i) the three cluster centers after the first round of execution
ii) the final three clusters.

****