Birla Institute of Technology & Science-Pilani, Hyderabad Campus
1st Semester 2011-2012
Data Mining (CS / IS C415)
Test II (Regular)

Type: Closed Book **Time: 50 mins**
Wt.age: 20 %(40 Marks) **Date: 17/11/2011**

---

1. Construct an FP-tree of the following transactional data. **(5 Marks)**

Transaction ID	Items
T_1	{A,B,C}
T_2	{C,D}
T_3	{E}
T_4	{B,D}
T_5	{A,C,D,E}
T_6	{E}
T_7	{A,B,D,E}

Transaction ID	Items
T_8	{A,C,D,E}
T_9	{E}
T_{10}	{B,C}
T_{11}	{B,D}
T_{12}	{ C,D,E}
T_{13}	{A,B,C}
T_{14}	{E}
T_{15}	{A,C,D,E}

2. Consider the following customer sequence dataset: **(6+2 Marks)**

Sequence ID	Sequence
1	< {1 5} {2} {3} {4} >
2	< {1} {3} {4} {3 5} >
3	< {1} {2} {3} {4} >
4	< {1} {3} {5} >
5	< {4} {5} >

(a) Apply the GSP algorithm to the dataset in the table using minimum support S = 33% to determine all large sequences.
(b) Identify the maximal sequence patterns.

3. Imagine we would like to cluster houses around India without using their exact addresses. For each house, we map properties of the house to a numeric value. For instance, the house's location is mapped as Alwal = 0, Bharathnagar = 1, Shameerpet = 2, etc., the exterior material is brick = 0, aluminum = 1, wood = 2, etc., the kitchen color is white = 0, green = 1, tan = 2, etc. We have 50 such features so each house can be represented as a vector in R 50. **(4 Marks)**

Which of the three clustering algorithms learned in class (hierarchical clustering, k-means, BRICH, CURE models) would be most appropriate for this task? Explain briefly why do you prefer this algorithm.

4. In the figure below shows the sitting arrangement of students A. . .O in a lecture hall. The lecturer clusters the students based on where they are sitting in the classroom using the DBSCAN algorithm. He is using the Manhattan distance measure, an EPS of 2.1 and MinPts of 4 (Note that the MinPts value includes the point itself and the Manhattan distance used is **MANHATTAN DISTANCE(X,Y) = ($|X_1-Y_1| + |X_2-Y_2| + \cdots + |X_N-Y_N|$)**. The lecturer finds that there are two clusters. **(1+3+3 Marks)**

1				**A**					
2		**B**		**C**					**D**
3				**E**	**F**		**G**		
4				**H**	**I**				
5									
6	**J**								
7		**K**	**L**		**M**	**N**			
8									
9			**O**						
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

(a) Which students are core points?
(b) Student P enters the room. He feels connected to both clusters of students. Where should he sit if he wants to belong to both clusters, but he does not want the clusters to merge into one?
(c) Where should Student P sit if he wants there to be only one cluster?

5. Assume that the following 3 instances belong to the same cluster and that the Euclidean distance is used to measure the distance between two data instances over all the four attributes: car-name, cylinders, model-year, and mpg. **(2+3 = 5 Marks)**

	Car Name	**Cylinders**	**Model – Year**	**Mpg**
A1	Chevrolet	8	70	18
A2	Chevrolet	4	82	27
A3	Toyota	4	70	24

a. Calculate the centroid of this cluster of 3 instances. Show your work.
b. Calculate the medoid of this cluster of 3 instances. Show your work. For this, you need to calculate the Euclidian distances:

6. The doctor of a school has measured the height and weight of pupils in a 5th grade class as shown in the table below. Detect the outlier using Grubb's Test. **(5 Marks)**

Student ID	Height	Weight
S1	130	37
S2	132	40
S3	138	39
S4	136	40.5
S5	131	42
S6	153	51
S7	131	41.5
S8	133	39
S9	129	41
S10	133	30

7. Consider a very small collection C that consists in the following three documents: **(1+1+2+2 = 6 Marks)**
d1: “times of india”
d2: “indian postal service”
d3: “hyderabad times”
Given the following query: “times times india”,
Construct the following and find the cosine similarity between the query and the documents using the Term-document incidence

1. The Term-document incidence matrix based on binary values.
2. The Raw term frequency matrix
3. Normalized Term frequency