BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
HYDERABAD CAMPUS
SECOND SEMESTER 2018 – 2019
Data Mining (CS F415) - MID SEMESTER EXAM (REGULAR)

Date: 15.03.2019 **Weightage:30%(60M)**

Duration: 90min. **Type: Closed Book**

Note: Answer all parts of a question sequentially

No of pages in the question paper: 2

Q1. Data Preprocessing **[6 M]**

For each of the following three attributes identify its type as binary, discrete or continuous and further classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio)

a) seasons of a year **b)** decay in a radioactive element **c)** pair-wise distances between cities

Q2. Association Rule Mining

a) Consider the transactional database in the **Table-1** given below **[4+3+3=10 M]**

tid	items
1	a,b,c,d
2	b,c,e,f
3	a,d,e,f
4	a,e,f
5	b,d,f

Table-1

i. Apply Apriori algorithm and generate frequent itemsets with a min support 60% and confidence as 50%.

ii. Determine the maximal frequent and closed frequent itemsets.

iii. Generate the association rules using 3 frequent item sets.

b) Consider the following set of candidate 3-itemsets:

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}

i) Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

Condition 1: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.

Condition 2: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than maxsize. Assume maxsize = 2 for this question.

Condition 3: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to maxsize, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node. **[4 M]**

ii) Consider a transaction that contains the following items: {1, 2, 3, 5}. **[3 M]**

Using the hash tree constructed in part (i), which leaf nodes will be checked against the transaction? What are the candidate 3-itemsets contained in the transaction?

c) Consider the following frequent 3-sequences:

< {1, 2, 3} >, < {1, 2}{3} >, < {1}{2, 3} >, < {1, 2}{4} >, < {1, 3}{4} >,
< {1, 2, 4} >, < {2, 3}{3} >, < {2, 3}{4} >, < {2}{3}{3} >, and < {2}{3}{4} >.

i) List all the candidate 4-sequences produced by the candidate generation step of the Generalized Sequential Patterns (GSP) algorithm. **[3 M]**

ii) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints). **[2 M]**

iii) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming maxgap = 1). **[3 M]**

Q3. Clustering **[29]**

a) In a wireless sensor network (WSN) the senor nodes are scattered and they transmit their data to a base station. If every node transmits its data periodically to the base station the sensor's battery loses its life and the lifetime of the WSN is reduced. Suggest a clustering technique to this problem such that the lifetime can be increased. **[6 M]**

b) For each of the following two **Figures 1.a and 1.b** of data points Which among the following clustering algorithms will perform well in accurately clustering the given data and why?

i) K-means **ii)** Single-link hierarchical **iii)** Complete-link hierarchical **iv)** DBSCAN

Note: Just writing the name of the algorithm will not be awarded any marks. **[4 M]**

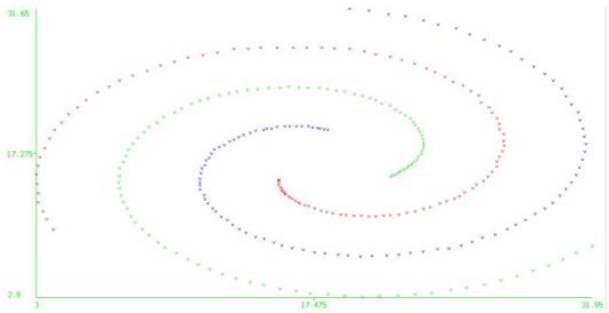

Figure 1.a

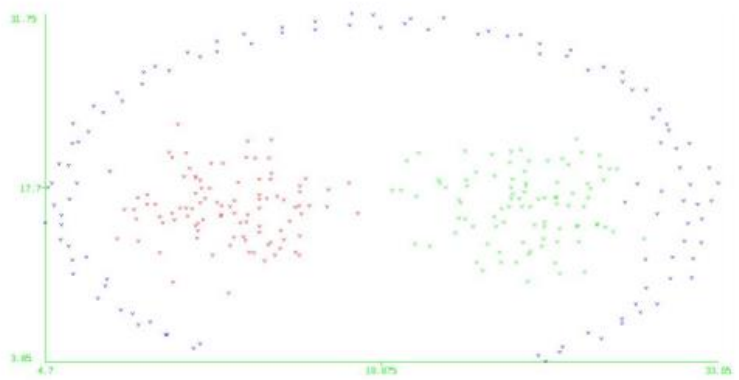

Figure 1.b

c) Given the dissimilarity matrix (**Table-2**) of 5 objects, show the working of Divisive Analysis(DIANA) algorithm using the Min link. **[5 M]**

	A	B	C	D	E
A	0.0	2.0	6.0	10.0	9.0
B	2.0	0.0	5.0	9.0	8.0
C	6.0	5.0	0.0	4.0	5.0
D	10.0	9.0	4.0	0.0	3.0
E	9.0	8.0	5.0	3.0	0.0

Table-2

d) Consider the given dataset (**Table-3**) which is about *what to do in the evening*:

Attribute Tuples	Deadline?	Is there a party?	Lazy?	Activity
1	Urgent	Yes	Yes	Party
2	Urgent	No	Yes	Study
3	Near	Yes	Yes	Party
4	None	Yes	No	Party
5	None	No	Yes	Pub
6	None	Yes	No	Party
7	Near	No	No	Study
8	Near	No	Yes	TV
9	Near	Yes	Yes	Party
10	Urgent	No	No	Study

Table-3

Using DBSCAN find clusters with ε=0.25 and min_pts=1. **[5 M]**

e) i) Consider the following 8 points in R^3 and build the cluster feature tree using the branching factor as 2 and threshold radius as 2 units. Use **L1 norm** to compute the distance between a pair of points.

$(1, 1, 1)^t$, $(1, 1, 2)^t$, $(1, 3, 2)^t$, $(2, 1, 1)^t$ $(6, 3, 1)^t$, $(6, 4, 4)^t$, $(6, 6, 6)^t$, $(6, 5, 7)^t$ **[4 M]**

ii) Since BIRCH builds the CF Tree on partial data what factors contribute to its success? **[2 M]**

f) In the CURE clustering algorithm once we find the representative points they are moved by a fraction of their distance from original location towards the centroid of the cluster. Would it make more sense to move them all a fixed distance towards the centroid instead? Why or why not? **[3 M]**

********** *That's all folks* **********