1. Identify whether the following statements are TRUE or FALSE. If the statement is FALSE, correct it and justify the corrected sentence. If the statement is TRUE, justify it. Restrict the justification to a few (less than five) sentences. [10*1=10]

1.1 To improve efficiency, in BUC algorithm, dimensions should be processed in the order of increasing cardinality

1.2 It is not possible covert convertible constraints as monotonic or antimonotonic

1.3 Subpace clustering addresses the curse of dimensionality

1.4 Must-link constraints and cannot-link constraints are the part of background knowledge

1.5 Divisive hierarchical clustering is relatively easier than agglomerative hierarchical clustering

1.6 Distant supervision handles the issue of scarcity of labels

1.7 The rules which have high lift value are different from the rules which have high confidence value

1.8 "k-anonymity" method is employed to improve privacy preservation in data mining.

1.9 Bit-map index is better than hash and tree indexes in OLAP

1.10 Distributive and algebraic data cube measures are easier to compute than holistic data cube measures.

2. Answer the following briefly [10*3=30]

2.1 Why data should be standardized? Explain the corresponding procedure.

2.3 How CLARANS improves performance over PAM and CLARA.

2.2 How ensemble classifiers improve accuracy? Give an outline of Adaboost algorithm.

2.4 In data warehousing technology, a multi dimensional view can be represented by a ROLAP (Relational OLAP), MOLAP (Multi-dimensional OLAP), and HOLAP (hybrid OLAP). Explain which technique you implement and why?

2.5 Explain basic idea of the BUC algorithm? Discuss the issues with the BUC algorithm as compared to top-down approaches.

2.6 What are the challenges of classifying stream data? Explain how ensemble method is effective in classifying stream data?

2.7 In evaluating the quality of association rules discuss the pitfalls and appropriateness of confidence measure through example. Briefly discuss about one better measure to resolve the problem of confidence measure.

2.8 In the last section of the paper, it is mentioned that "BIRCH is significantly superior to CLARANS in terms of quality, speed and order-sensitivity". Discuss how and why?

2.9 Explain how error- correcting codes can be used to improve the accuracy of multiclass classification.

2.10 What is the role of sigmoid function in the backpropagation algorithm?

3. Find the equations for to computing complexity of clustering algorithms. Provide justification. [5]

- K-means: Time: O(I*K*m*n); Space: O((m+K)n) where, "K" denotes number of clusters, "m" is number of items and "n" is number of attributes, "I" is number of iterations.
- Hierarchical algorithm: Time: $O(m^2 \log m)$; Space: $O(m^2)$, where "m" is the number of data items
- Time complexity of Chameleon algorithm: $O(mp + m \log m + p^2 \log p)$, where "m" is the number of items and "p" is number of partitions.

4. (a) Explain the following regarding DBSCAN: "directly density-reachable is symmetric for pairs of core points. In general, however, it is not symmetric if one core point and one border point are involved".

(b) In DBSCAN algorithm, explain the heuristic to determine the Eps and MinPts. [5]

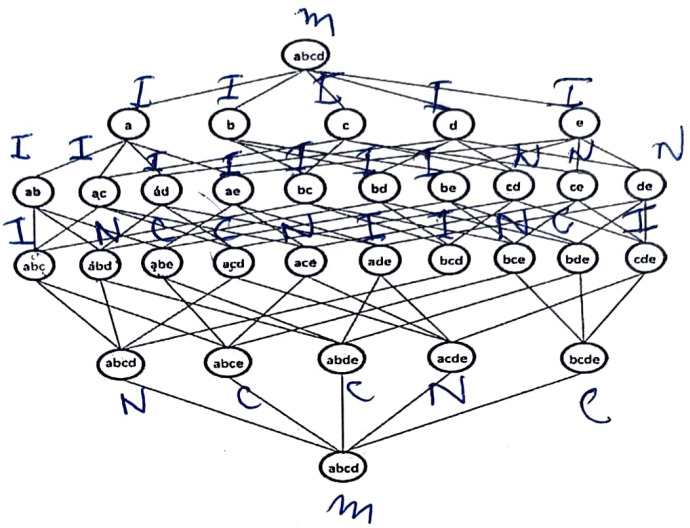
5. Consider the data set given in the following table. At a minimum support of 60%, find all frequent itemsets using a vertical data format. [5]

TID	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Cola
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Cola

6. Given the following lattice stricture and the transactions, label each node with the following letters. [5]
   - M if it is a maximal frequent itemset
   - C if it is a closed frequent itemset
   - N if it is frequent but neither maximal nor closed, and
   - I if it is frequent

   Assume minimum support threshold as 30%.

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



7. Assume that you are a senior officer in Income TAX department of India. You are given a representative sample of past records of people who are supposed to pay taxes. Derive a decision tree for alerting task force using ID3 algorithm. [5]

Data set for question # 7				
Tid	Govt Employee	Marital status	Taxable income	Evade
1	YES	Single	1250K	NO
2	NO	Married	1000K	NO
3	NO	Single	700K	NO
4	YES	Married	1200K	NO
5	NO	Divorced	950K	YES
6	NO	Married	600K	NO
7	YES	Divorced	2200K	NO
8	NO	Single	950K	YES
9	NO	Married	750K	NO
10	NO	Single	90K	NO

$$Hint: Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

8. Consider the database of question #7. **Predict the** "Evade" for the attribute values <Govt Employee: *NO*, Marital status: *Married* , Taxable Income: *700K*> using a naïve Bayesian classifier. [5]

$$Hint:\ P(A_i|c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

9. Given a set of 5-dimensional categorical samples: A=10110; B=11010; C=00110; D=01010; E=10101; F=01101. Apply agglomerative clustering algorithm using single-link and complete-link methods. You can employ appropriate similarity measure by giving justification. Draw the corresponding dendrograms. [5]

10. One of the definition of data mining is "Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. In this course, you have learned several data mining functionalities, like attribute oriented induction, data cube, association rule mining, classification and clustering. List one example application for each functionality and the kind of interesting information (refer the definition) the functionality could help you to extract to improve the performance of the corresponding application. [5]