DA Endsems Questions 2025

Question 1

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

(1.1) Visual data mining is useful for pre-processing.

(1.2) Outliers are different from noise data.

(1.3) To improve efficiency, in the BUC algorithm, dimensions should be processed in the order of increasing cardinality.

(1.4) As compared to k-means algorithm, k-medoid algorithm is robust to outliers.

(1.5) Must-link constraints and cannot-link constraints are the part of background knowledge.

(1.6) Divisive hierarchical clustering is relatively easier than agglomerative hierarchical clustering.

(1.7) Distant supervision handles the issue of scarcity of labels.

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

(1.9) Bitmap 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.

Question 2

Answer the following briefly $[10 \times 3 = 30]$

(2.1) How CLARANS improve performance over PAM and CLARA? [3]

(2.2) Suppose a group of 12 sales price records has been sorted as follows:

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

Partition them into three bins using: (i) Equal-frequency (equal-depth) partitioning (ii) Equal-width partitioning (iii) Clustering

(2.4) Explain the basic idea of the BUC algorithm. Discuss the issues of BUC algorithm as compared to top-down approaches. [3]

(2.5) What are the challenges of classifying stream data? Explain how ensemble method is effective in classifying stream data. [3]

(2.6) In evaluating the quality of association rules discuss the pitfalls of confidence measure through example. Briefly discuss about one better measure to resolve the problem of confidence measures. [3]

(2.7) Explain how CLIQUE (Clustering in QUEst) exploits Apriori property for finding density-based clusters in subspaces. [3]

(2.8) Explain how error-correcting codes can be used to improve the accuracy of multiclass classification. [3]

(2.9) Explain the steps of the backpropagation algorithm by considering a two-layer neural network. [3]

(2.10) Explain the importance of Transaction-weighted Downward Closure Property in extracting utility patterns. [3]

Question 3

Find the equations about the complexity of clustering algorithms. Justify the equations based on the steps of the algorithm. [5]

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

Question 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." [3]

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

Question 5

At a minimum support of 60%, find all frequent itemsets using the Apriori algorithm. [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

Question 6

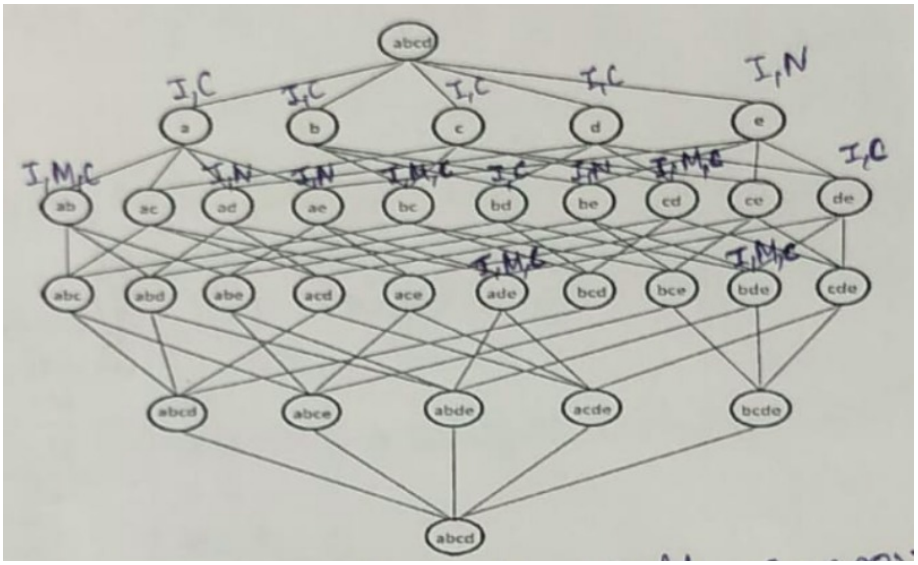
6. Given the following lattice structure and the transactions, label each node with the following letters: [5]

- **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
- **I** if it is infrequent

Assume minimum support threshold as 30%.

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



Question 7

Assume that you are a senior officer in the 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 the ID3 algorithm. [5]

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:

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

Question 8

Consider the database of Question No. 7. Predict the "Evade" for the attribute values

$$\text{Govt Employee} = \text{No}, \quad \text{Marital Status} = \text{Married}, \quad \text{Taxable Income} = 700\text{K}$$

using a naive bayesian classifier. [5] **Hint:**

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right)$$

Question 9

Given a set of 5-dimensional categorical samples:

$$A = 10110,$$
$$B = 11010,$$
$$C = 00110,$$
$$D = 01010,$$
$$E = 10101,$$
$$F = 01101.$$

Apply the agglomerative clustering algorithm using Single-link and Complete-link methods. You may employ appropriate similarity measure by giving justification. Draw the corresponding dendrograms. [5]

Question 10

The following contingency table summarizes supermarket transaction data, where *"hot dogs"* refers to transactions containing hot dogs, $\overline{\text{hot dogs}}$ refers to transactions that do not contain hot dogs, *"hamburgers"* refers to transactions containing hamburgers, and $\overline{\text{hamburgers}}$ refers to transactions that do not contain hamburgers. Suppose that the association rule

$$\text{hot dogs} \rightarrow \text{hamburgers}$$

is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong? [5]

	hot dogs	$\overline{\textbf{hot dogs}}$	$\sum$ **row**
hamburgers	2000	500	2500
$\overline{\textbf{hamburgers}}$	1000	1500	2500
$\sum$ **col**	3000	2000	5000