TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2076 Chaitra

Exam.	Regular		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt All questions.*
- ✓ *The figures in the margin indicate Full Marks.*
- ✓ *Assume suitable data if necessary.*

1. Find the principal components and the proportion of the total variance explained by each when the covariance matrix of the three random variables X_1, X_2, and X_3 is: [4]

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

2. (a) Given the following points compute the distance matrix using the Manhattan and the Supremum distance. [2+1+2]

Points	X	Y
P1	6	3
P2	2	2
P3	3	4

(b) Given the following two vectors compute the Cosine similarity between them.
D1 = [4 0 2 0 1]
D2 = [2 0 0 2 2]

(c) Given the following two binary vectors compute the Jaccard similarity and Simple Matching Coefficient.
P = [0 0 1 1 0 1]
Q = [1 1 1 1 0 1]

3. Suppose that a data warehouse for a sales company consists of five dimensions: *time*, *location*, *supplier*, *brand*, and *product*, and two measures: *count* and *price*. [3+3]
(a) Draw a *snowflake schema* diagram for the data warehouse.
(b) Starting with the base cuboid [*time*, *location*, *supplier*, *brand*, *product*], what specific OLAP operations should one perform in order to list the total *count* for a certain *brand* for each *state* per *year* (assume *location* has three levels: *country*, *state*, *city*; and assume *time* has three levels: *year*, *month*, *day*)?

4. Why is a conflict resolution strategy often necessary for rule-based classifiers? Describe the common conflict resolution strategies for rule-based classifiers. [2+4]

5. The following dataset will be used to train a decision tree for predicting whether a mushroom is edible or not based on its shape, color and odor. [2+5]

Shape	Color	Odor	Edible
C	B	1	Yes
D	B	1	Yes
D	W	1	Yes
D	W	2	Yes

C	B	2	Yes
D	B	2	No
D	G	2	No
C	U	2	No
C	B	3	No
D	W	3	No

(a) Which attribute would the ID-3 algorithm choose to use for the root of the decision tree?

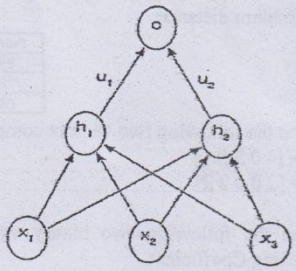
(b) Draw the full decision tree that would be learned for the given data.

6. Consider the multi-layer feed-forward neural network shown in the following figure. This neural network has three inputs (x_1), (x_2) and (x_3) connected to a hidden layer consisting of two nodes (h_1) and (h_2). The weight of the edge connecting (x_i) to (h_j) is (w_{ji}). The two hidden nodes are connected to the output node (o). The weight of the edge connecting the hidden node (h_i) to the output node (o) is (u_i). The activation functions at hidden and output layers is set to sigmoid function defined as follows: [2+3+4]

$$\sigma(\theta) = \frac{1}{1 + exp(-\theta)}$$

Using the target output (t), the squared error is used as the loss function at the output node, and is defined as:

$$E(o, t) = \frac{1}{2}(o - t)^2$$



(a) Using the symbols given above, compute the activation at (h_1).

(b) Compute the gradient of the loss with respect to the output (o).

(c) Compute the gradient of the loss with respect to the weight (w_{12}).

7. Consider the transaction data shown in the following table from a fast food restaurant. [5+3]

Meal Item	List of Item IDs
Order:1	M1, M2, M5
Order:2	M2, M4
Order:3	M2, M3
Order:4	M1, M2, M4
Order:5	M1, M3
Order:6	M2, M3
Order:7	M1, M3

Order:8	M1, M2, M3, M5
Order:9	M1, M2, M3

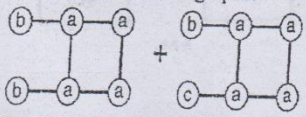
There are 9 distinct transactions (Order: 1 – Order: 9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity, the meal items have been assigned short names (M1-M5). Assume that the minimum support is 2/9 and the minimum confidence is 7/9.

(a) Apply the Apriori algorithm to the dataset of transactions and identify all frequent k-itemsets.

(b) Find all strong association rules of the form: $X \wedge Y \rightarrow Z$ and note their confidence values.

8. (a) List all the 4-subsequences contained in the data sequence: < {1,3} {2} {2,3} {4} >

(b) Draw all candidate sub-graphs obtained from joining the pair of graphs shown below using edge-growing method to expand the sub-graphs. [3+3]



9. Given the matrix (X) whose rows represent different data points, perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here (K) is chosen as 3. The center of the 3 clusters are initialized as red (6.2, 3.2), green (6.6, 3.7) and blue (6.5, 3.0). Provide the final cluster centers and comment on the number of iterations required for the clusters to converge. [8]

$$X = \begin{bmatrix} 5.9 & 3.2 \\ 4.6 & 2.9 \\ 6.2 & 2.8 \\ 4.7 & 3.2 \\ 5.5 & 4.2 \\ 5.0 & 3.0 \\ 4.9 & 3.1 \\ 6.7 & 3.1 \\ 5.1 & 3.8 \\ 6.0 & 3.0 \end{bmatrix}$$

10. The table below is a distance matrix for six objects: [4+4]

	A	B	C	D	E	F
A	0					
B	0.12	0				
C	0.51	0.25	0			
D	0.84	0.16	0.14	0		
E	0.28	0.77	0.70	0.45	0	
F	0.34	0.61	0.93	0.20	0.67	0

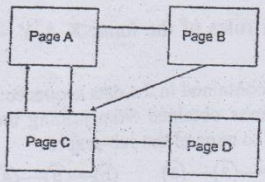
(a) Show the final result of hierarchical clustering with single-link by drawing a dendrogram.

(b) Show the final result of hierarchical clustering with complete-link by drawing a dendrogram.

11. (a) Discuss the issues related to anomaly detection. [2]

(b) If the probability that a normal object is classified as an anomaly is 0.01 and the probability that an anomalous object is classified as anomalous is 0.99, then what is the false alarm rate and detection rate if 99% of the objects are normal? [3]

12. Consider the following subset of pages and their links. Apply the PageRank algorithm using a damping factor of 0.85. A minimum of five iterations are required. Assume initial page rank of all pages is 0.25. [8]



***

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2076 Ashwin

Exam.	Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. What are the fundamental differences between Data Mining and Data Warehousing? Describe the steps of KDD for data mining. [3+7]
2. What do you mean by dimensional data? What are base & apex cuboid? Slicing & Dicing? Roll Down and Roll UP operations? Give example. [2+3+3+3]
3. How do you measure the accuracy of classifiers? How do you select best root attribute in decision tree? Explain. [4+6]
4. What are prior and posterior probabilities? Explain the algorithmic steps of Bayesian classifier and write its strengths. [3+7]
5. For the transactions given below, consider confidence=60% and minimum support=30%. Identify large itemsets (L-Itemset) at L=3 with possible associations using A-priori algorithm and generate F-List using FP-Growth algorithm. [12]

Transactions	Items description
T1	A, B, C, T, M, P, D, K
T2	A, B, T, P, D, K
T3	B, C, T, D, M, A, P
T4	A, C, T, M, D,
T5	A,C, D, K, M
T6	B, C, T

6. How DBSCAN algorithm works? How do we avoid the issues of DBSCAN? [8+2]
7. Explain web mining taxonomy. [8]
8. Write short notes on (**Any Three**) [3+3+3]
   a. Data smoothing techniques
   b. Clustering and its application in anomaly detection
   c. AprioriALL: Sequential pattern mining algorithm
   d. Various similarity measures between data tuples.

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2075 Chaitra

Exam.	Regular / Back		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective I) (CT72502)*

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. Explain Data Warehouse architecture with its analytical processing. [8]
2. Why data preprocessing is necessary? Explain the methods for data preprocessing to maintain data quality. [4+4]
3. Define Decision Tree Classifier with Gini-Index with suitable example. How can you handle overfitting in Decision Tree? [6+4]
4. What do you mean by frequent Pattern growth, draw FP-tree with given tabular data. [4+4]

TID	Items
01	f, a, c, d, g, i, m, p
02	a, b, c, f, l, m, o
03	b, f, h, j, o, w
04	b, c, k, s, p
05	A, f, c, e, l, p, m, n

5. How ANN works? Explain with Algorithm. [8]
6. What is the application of clustering in data mining? Explain K-means clustering with example. [2+6]
7. How DBSCAN clustering is used for handling noise in data? [8]
8. What is outlier? Explain the distance base approaches for the anomaly detection. [5]
9. What are the challenges of web mining? Explain about time series data mining with an example. [5]
10. Write short notes on: (Any three) [4+4+4]
    a) Market Basket Analysis
    b) Visual Data Mining
    c) OLAP and OLTP
    d) Data Normalization

***