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?

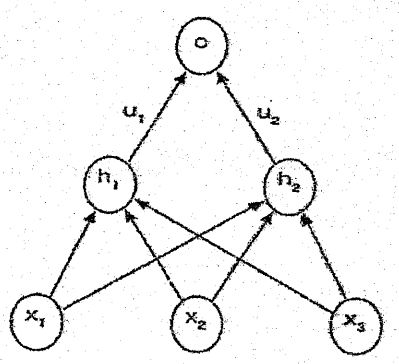
(b) Draw the full decision tree that would be learned for the given data.

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

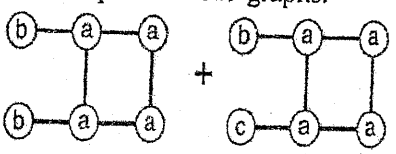
There are 9 distinct transactions (Order: 1 – Order: 9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity, the meal items have been assigned short names (M1-M5). Assume that the minimum support is 2/9 and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

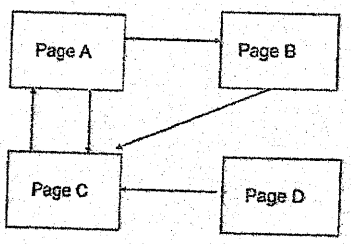
(a) Show the final result of hierarchical clustering with single-link by drawing a dendrogram.

(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

***

37B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2073 Shrawan

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BE, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

Subject: - Data Mining *(Elective II) (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. "The world is data rich but information is poor". Justify with your own words. [8]
2. What are the measuring elements of data Quality? Explain different data transformation by normalization methods with an example. [2+6]
3. What is a decision tree and how information gain is used for attribute selection? Explain with example. [8]
4. Explain ROC. Using the following data, calculate TPR, FPR, precision for given confusion matrix. [1+3+6]

	A	B
A	20	5
B	10	40

Classify, A = Yes, B = No

5. What is FP Tree? How FP-growth algorithm eliminate the problem of Apriori algorithm? Construct the FP tree and find association rules for the following transaction database using FG- Growth algorithm. Support = 30% and confidence = 75%. [10]

Transaction ID	Items
1	P,R,S
2	R,S,T
3	P,Q,R
4	P,R,S,T
5	P,S,T
6	P,Q,T
7	Q,S,T
8	Q,R,T

6. What are Categorical data? What are the possible issues arrives when using Categorical data? How can you handle such issues? [2+3+3]
7. What is the application of clustering in data mining? Explain the k-means algorithm with example. [8]
8. What is anamoly detection? Explain distance based method for anamoly detection. [8]
9. Write short notes on: [4×3]

   i) Data transformation
   ii) Web mining
   iii) OLAP

***

34B TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2072 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 II) (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 is data mining? Explain all the steps of knowledge discovery. [2+6]
2. How do you perform analysis of multidimensional data? Explain with the concept of OLAP. [10]
3. Predict Class label using naive Bayesian classifier for X = (age = youth, income = medium, student = yes, credit-rating = fair) using the following data set. [10]

RID	Age	Income	Student	Credit-rating	Class Buy computer
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle-age	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle-age	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle-age	Medium	No	Excellent	Yes
13	Middle-age	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

4. The confusion matrix for a classifier is given as follows: [10]

		actual class	
		class1	class2
predicted class	class1	21	6
	class2	7	41

calculate a. accuracy
b. sensitivity
c. specificity
d. precision
e. recall

5. What is the importance of SUPPORT and COFIDENCE during association analysis? Explain FP-Growth method with example. [10]
6. What are the types of clustering methods? Explain DBSCAN method of clustering with an example. [10]
7. What is the use of Apriori Algorithm in market basket analysis? Explain with suitable example. [10]
8. Write short notes on: [4×3]
   i) Time series Data mining
   ii) Issues in anomaly/Fraud detection
   iii) Categorical data and related issues

***

24C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2071 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.*
- ✓ ***All** questions carry equal marks.*
- ✓ *Assume suitable data if necessary.*

1. What is a Data Mining? Explain its application.
2. Explain the properties that a Distance Metric needs to support with respect to Minkowski's distance.
3. What is a decision tree? Explain Gini Index with suitable example.
4. Explain a Bayes classifier. In what cases can Naive Bayes and Bayesian Belief Network be used?
5. Why is a clustering an unsupervised learning? How can hierarchical clusters be generated using Bisecting K-means algorithm?
6. Explain the different measures of cluster validity.
7. How does Apriori Algorithm optimize the brute force approach for frequent item set generation?
8. What is an Anomaly Detection? Explain few distance based approaches that can be used for Anomaly Detection.

***

27C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2072 Kartik

Exam.	New Back (2066 & Later Batch)		
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 is a data mining? Explain general steps in brief. [4]
2. Why data preprocessing is required in the data mining? Explain some of approaches of data clearing. [5+5]
3. Write about Hunt's Algorithm for Decision Tree induction. Explain the test conditions that can be used for different attribute types. [10]
4. What is an ANN classifier? Explain its general consideration that required for the classifier. [2+6]
5. What is an association analysis? Explain its importance in market-basket analysis. [2+5]
6. What is a Frequent item set? Explain FP growth method with example. [1+8]
7. What is a cluster analysis? How it is different from classification? [5]
8. Explain a DBSCAN algorithm with example. [7]
9. What is an Anomaly detection? Discuss its importance in security. [5]
10. Explain Time series data mining in brief. [6]
11. Write short notes on: [3×3]
    i) Data transformation
    ii) Sequential pattern
    iii) Cluster evaluation

***

27C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2071 Shawan

Exam.	New Back (2066 & Later Batch)		
Level	BE	Full Marks	80
Programme	BEX, BCT	Pass Marks	32
Year / Part	IV / I	Time	3 hrs.

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

- ✓ Candidates are required to give their answers in their own words as far as practicable.
- ✓ *Attempt **All** questions.*
- ✓ ***All** questions carry equal marks.*
- ✓ *Assume suitable data if necessary.*

1. What is data mining? Explain different data types of attributes in a dataset.
2. How can principle component analysis be used for dimentionality reduction?
3. Why is classification a super vised learning method? Explain different impurity measures used in decision tree classifier.
4. Explain Naive Bayes classifier. How can over fitting problem be solved in case of classification?
5. Explain FP-growth algorithm in detail.
6. What are association rules? How can spriori algorithm be used to generate association rules.
7. What is contiguous cluster? Explain an algorithm that can be used to generate contiguous clusters.
8. Explain K-means clustering with limitation Use k-means clustering to cluster the following dataset.

A	B
1.0	1.0
1.5	2.0
3.0	4.0
5.0	7.0
3.5	5.0
4.5	5.0
3.5	4.5

9. How can Nearest-Neighbor algorithm be used for anomaly defection?
10. Write short notes on:
    a) Time-series data mining
    b) Data warehouse and data mart

***

37C TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
Examination Control Division
2070 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) (CT725)*

- ✓ 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. a) What is "curve of Dimensionality"? How can it be avoided? [5]
   b) Discuss the impact of noisy data in data mining? [5]
2. Explain rule based classifier? How can CN2 Algorithm be used for rule based classification? Define "Accuracy" and "Laplace" measures used for rule evaluation. [9]
3. An input sequence "A A B B B A A A B B" was used for classification. The Classifier 'X' predicted the sequences as: "A A B B B A A A B B" where as the Classifier 'Y'predicted the sequences as: "A A A A B B A A A B". Develop the corresponding confusion matrix for the classifiers and find their corresponding. [10]
   i) Accuracy
   ii) Precision
   iii) True Positive Rate
   iv) False Positive Rate
4. Explain Apriori algorithm. Use Apriori to generate frequent item sets with support of 50% for the following transaction database. [10]

TID	Items
1	ACD
2	BD
3	ABCE
4	BDF

5. Why is pattern evaluation important in association rule mining? Explain with example the statistical based measures used for measuring interestingness of association rules. [8]
6. What is a density based cluster. Explain an algorithm that can be used to generate density based clusters. [8]
7. What is Hierarchical Clustering? Differentiate between agglomerative and divisive approach of hierarchical clustering. Augment your answer with appropriate illustrative examples. [10]
8. Write short notes on: [15]
   i) Data ware house and Data mart
   ii) Base Rate Fallacy
   iii) Web mining
   iv) Anomaly Detection
   v) Convex Hull Method

***