Tribhuvan University
Institute of Science and Technology
SCHOOL OF MATHEMATICAL SCIENCES
First Assessment 2079

Full Marks: 45
Pass Marks: 22.5
Time: 2hrs

Subject: **Advanced Data Mining**
Course No: **MDS 602**
Level: **MDS /II Year /III Semester**

Candidates are required to give their answers in their own words as far as practicable.

Attempt All Questions.

Group A [3 ×5 = 15]

1. What is data mining? Describe about Major issues in Data mining?
2. Write down the steps to find the principal component of a data set.
3. What is data warehouse? Compare OLTP and OLAP.
4. What are categorical data? How can you handle the categorical data in association mining?
5. Explain Rule based classifier. How it is different than decision tree classifier.

Group B [5 × 6 = 30]

6. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity (Example: Age in years. Answer: Discrete, quantitative, ratio).
   a) Brightness as measured by a light meter.
   b) Angles as measured in degrees between 0 ∘ and 360 ∘.
   c) Bronze, Silver, and gold medals as awarded at the Olympics.
   d) Number of patients in a hospital.
   e) Ability to pass light in terms of the following values: opaque, translucent, transparent.
   f) Military rank.
7. Given the following data (in increasing order) for the attribute age: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. (b) How might you determine outliers in the data?
8. Consider the data set shown in Table

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

Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each transaction ID as a market basket.

Use the results in part (a) to compute the confidence for the association rules {b, d} ⟶ {e} and {e} ⟶ {b, d}. Is confidence a symmetric measure?

Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

…………… ………OR

CS CamScanner

Construct the FP tree from the above transactional data table with minimum support of 33%.

9. Explain ID3 Algorithm with an example.

10. Describe Confusion Matrix with example. Define Accuracy, Precision, TPR, TNR, FPR, FNR of the classifier model.

OR

Given the following confusion matrix, determine Accuracy, Precision, TPR, TNR, FPR, FNR of the classifier model.

N=165	Predicted: No	Predicted: Yes	
Actual: No	50	10	60
Actual: Yes	5	100	105
	55	110	

***

C 10 or 2

5

Tribhuvan University
Institute of Science and Technology
SCHOOL OF MATHEMATICAL SCIENCES
Second assessment 2079

Subject: **Advanced Data Mining**
Course No: **MDS 602**
Level: **MDS/II Year/III Semester**

Full Marks: **45**
Pass Marks: **22.5**
Time: **2hrs**

Candidates are required to give their answer in their own words as far as practicable.

Attempt ALL questions.

Group A [5×3=15]

1. What is a super vector machine (SVM)? Why does SVM work fast and well?
2. Why does ensemble method classifier give better result? Explain main types of ensemble method.
3. How is the cluster quality measured? What are the factors affecting cluster quality. Explain with an example.
4. What is anomaly? Why it is important? Explain different types of outliers that occur in a dataset.
5. Explain distance-based outlier detection approach.

Group B [5×6=30]

6. Consider the 5 datapoints shown below: P1:(1,2,3), P2:(0,1,2), P3:(3,0,5), P4:(4,1,3), P5:(5,0,1). Apply the K-means clustering algorithm, to group those data points into 2 clusters, using the Manhattan distance. Also calculate the SSE (Sum of Square Error) Suppose that the initial centroids are C1:(1,0,0) and C2:(0,1,1).

OR

Cluster the following data and cluster the data using fuzzy C-means clustering algorithms. Assume k=2 and P=2.

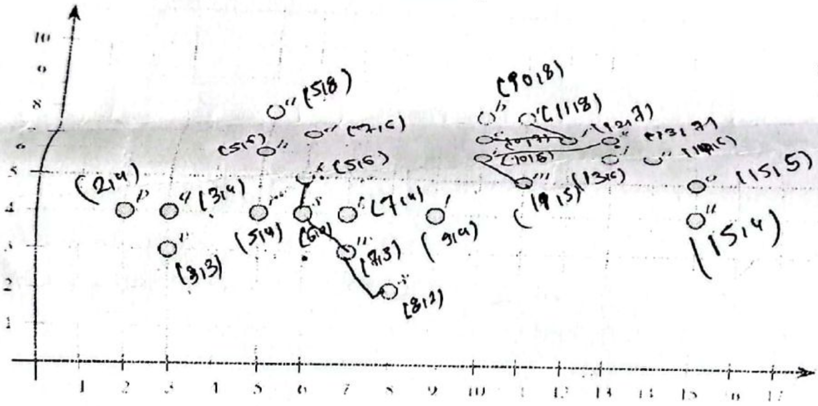
X1	1	2
X2	2	3
X3	9	4
X4	10	1

CS CamScanner

7. Assume that database D is given by the table below. Follow single link technique to find clusters in D. Use any distance measure method

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

8. Consider the figure below and answer the following questions, we use the Euclidean distance between points, and the Ɛ=2 and minpts=3.



a) List all the core points.

b) Is **a** directly reachable from **d**?

c) Is **o** density reachable from **i**? Show the intermediate points on the chain or the point where the chain breaks.

9. Explain different statistical based clustering method.

10. What are the approaches for density-based outlier detection. Explain Local Outlier Factor approach for outlier detection.

OR

A (0,0), B (1,0), C (1,1) and D (0,3) and K=2. Use LOF to detect one outlier among these 4 points. Use Manhattan distance method to measure the distance between the points.

***

MDS 602-2079✡

Tribhuvan University

Institute of Science and Technology

2079

✡

Master Level / Second Year /Third Semester/ Science
Data Science (MDS 602)
(Advanced Data Mining)

Full Marks: 45
Pass Marks: 22.5
Time: 2 hours

Candidates are required to give their answers in their own words as for as practicable.

Attempt All Questions

Group A **[5 × 3 = 15]**

1. What is data mining? Explain about data preprocessing.
2. What is sequential pattern mining? How Apriori is different than FP-Tree algorithm.
3. Explain different method for estimating a classifier's accuracy.
4. Explain K-means clustering algorithm.
5. What are the contextual and collective outliers in attribute of credit card company (name, age, job, address, annual-income, annual-expense, average-balance, credit-limit) and why?

Group B **(5×6=30)**

6. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.
7. What is a rule-based classifier? Briefly discuss different types of rule-based classifier.

OR

The following table consists of training data from an employee database. The data have been generalized. For example, "31 . . . 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

Department	Status	Age	Salary	Count
sales	senior	31-35	46k-50k	30
sales	junior	26-30	26k-30k	40
sales	junior	31-35	31k-35k	40
systems	junior	21-25	46k-50k	20

CamScanner

systems	senior	31-35	66k-70k	5
systems	junior	26-30	46k-50k	3
systems	senior	41-45	66k-70k	3
marketing	senior	36-40	46k-50k	10
marketing	junior	31-35	41k-45k	4
secretary	senior	46-50	36k-40k	4
secretary	junior	26-30	26k-30k	6

Given a data tuple having the values "systems," "26-30," and "46K-50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be?

8. A database has 5 transactions as below. Let min sup = 60% and min conf = 80%

$\frac{60 \times 5}{100}$

TID	items bought
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Find complete set of frequent itemsets using Apriori algorithm. Also find the top two strong association rules between the items sets.

9. Assume that database D is given below. Follow complete link technique to find clusters in D. Also show the dendrogram.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

OR

Write an algorithm for DBSCAN and prove that in DBSCAN, for a fixed MinPts value and two neighborhood thresholds $\in_1 < \in_2$, a cluster C with respect to $\in_1$ and MinPts must be a subset of a cluster C′ with respect to $\in_2$ and MinPts.

10. Explain proximity-based outlier detection approaches.

CS CamScanner