MDS 601-2079 ✡

MDS 2nd Year III Semester

Tribhuvan University

Institute of Science and Technology

2079

✡

Master Level / Second Year /Third Semester/ Science

Data Science (MDS 601)

(Research Methodology)

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 applied research? How is it different from basic research?
2. What are the characteristics of scientific research? Is it possible to have good research that is not generalizable? Why or why not?
3. Why research problem is stated in a study? What are the criteria for a good problem statement?
4. What are the goals of a literature review? What distinguish a strong from a weak literature review?
5. Describe the nature of descriptive research. What is the value of descriptive research in data science?

Group B

[5×6=30]

6. What is scientific research? Describe the major steps involved in the scientific research process?
7. What type of experiment do you think it would be possible to conduct in the following situation? Describe the different steps that you would take in conducting the experiment. "A company has designed a new product. You think that the consumer will like this product more than other similar products. You wish to demonstrate by scientific experiment, that this is the case."
8. "A valid instrument is always reliable, but a reliable instrument may not always be valid" comment on this statement.
9. What is non-response error? Describe how non response error can occur? What can be done to rescue the non-response errors associated with mail questionnaire?

OR

Describe sampling. Why is sampling used in research? List the steps involved in the process of sampling.

10. What factors should be kept in mind while selecting a research topic? Explain the attributes of a good research topic.

OR

What is a research report? What purposes does it serve? Why is reporting of research study important?

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

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%

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 $\epsilon_1 < \epsilon_2$, a cluster C with respect to ϵ_1 and MinPts must be a subset of a cluster C' with respect to ϵ_2 and MinPts.

10. Explain proximity-based outlier detection approaches.

Tribhuvan University
Institute of Science and Technology
2079

Master Level / Second Year /Third Semester/ Science
Data Science (MDS 603)
(Techniques for Big Data)

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 are the different characteristics of big data?
2. How is functional programming language different than imperative programming language?
3. What are the core components of Hadoop Ecosystem?
4. What are structured, unstructured and semi structured data?
5. How Spark is different than Map Reduce? Compare Hive QL with SQL.

Group B [5×6 = 30]

6. Explain the scopes and applications of big data analytics in the different sectors with example.
7. What is map reduce? Explain its execution overview with reference to word frequency count example.
8. What are the different daemons running in a Hadoop cluster? Explain how these daemons work in Master/Slave Architecture.

   OR

   What are the different configuration modes to setup the Hadoop? Which setup mode is preferred for development and why?
9. Explain the limited taxonomy of NoSQL data base with example.

   OR

   What are the limitations of distributed databases like NoSQL? Explain it with reference to CAP Theorem.
10. Explain about resilient distributed dataset and data frames in spark.

Tribhuvan University
Institute of Science and Technology
2079
✡

Master Level / Second Year /Third Semester/ Science
Data Science (MDS 606)
(Decision Analysis)

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. Explain different types of decisions in brief.

2. A man has the choice of running either a hot -snack stall or an ice-cream stall at a sea side resort during the summer season. If it is a fairly cool summer, he should make Rs.5000 by running the hot-snack stall, but if the summer is quite hot, he can only expect to make Rs.1000. On the other hand, if he operates the ice-creams stall, his profit is estimated at Rs.6500 if the summer is hot, but only Rs.1000 if it is cool. There is a 40% chance of the summer being hot. What should be his decision to maximize expected profit by using EMV criterion?

3. Determine the saddle point solution, the associated pure strategies and the value of the game for the following game. The payoff matrix for player A is given by

Player A	Player B		
	B_1	B_2	B_3
A_1	15	2	3
A_2	6	5	7
A_3	-7	4	0

4. Define goal programming. Distinguish between Linear programming problem and goal programming.

5. Write down current trends in Enterprise Risk Management (ERM).

Group B (5×6=30)

6. Define group decision making. Describe the techniques of group decision making.

OR

Describe about the different types of decision theories with suitable examples.

7. A distribution of past sales of a commodity for ABC Enterprises is as follows:

Quantities buyer's bought	20 units	25 units	40 units	60 units
Probability	0.10	0.30	0.50	0.10

ABC Enterprises buys these for Rs.6 and sells them for Rs.10.

a) What quantities should be bought to maximize expected profits?

b) What is the Expected Profit with Perfect Information (EPPI)?

c) What is the Expected Value with Perfect Information (EVPI)?

OR

The captain's table is a mail-order distributor of fresh lobsters. The company buys these for Rs.4 per pound and sells them for Rs.7.50 per pound. The per week shipment distribution is as follows:

Shipments per week, pound	3000	5000	8000	12000	18000
Probability of occurrence	0.05	0.20	0.20	0.40	0.15

The company has been approached by a consulting of firm specializing in sales forecasting. The firm has offered to provide the captain's table with a sales-forecasting model, which will increase the distributor's present profit by matching purchases with sales. The cost of buying and running this model will be Rs.7500 a week. Should the company buy it?

8. The following matrix gives the payoff of different strategies (alternatives) S_1, S_2, S_3 against states of nature (events) D_1, D_2, D_3, and D_4

Strategies	States of nature			
	D_1 (Rs.)	D_2 (Rs.)	D_3 (Rs.)	D_4 (Rs.)
S_1	4000	-100	6000	18000
S_2	20000	5000	400	0
S_3	20000	15000	-2000	1000

Indicate the decision taken under the following approach:

a) Pessimistic criterion b) Optimistic criterion c) Minimax regret criterion

d) Laplace criterion e) Hurwitz criterion.

Assume that the coefficient of optimism (α) is 0.60.

9. In a game of matching coins with two players A and B, suppose A wins 2 units of value when there are two heads, wins nothing when there are two tails and losses 1 unit of value when there is one head and one tail. Determine the payoff matrix, the best strategies for each player and the value of the game to A.

10. Solve the following Goal Programming (GP) by simplex method, Minimize $Z = d_1^-$

Subject to the constraints $2x_1 + x_2 \leq 6$, $x_1 + x_2 \leq 4$, $4x_1 + 8x_2 + d_1^- - d_1^+ = 100$ and $x_1, x_2, d_1^-, d_1^+ \geq 0$.

Tribhuvan University
Institute of Science and Technology
2079
✡

Master Level / Second Year /Third Semester/ Science
Data Science (MDS 607)
(Monte Carlo Methods)

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. Distinguish between frequentists and Bayesian inferences.
2. Explain the "random numbers" with a suitable example.
3. What is " random walk"? Explain.
4. Write down the meaning of transition probabilities with examples.
5. Point out main differences between Metropolis and Metropolis-Hastings algorithm.

Group B (5×6=30)

6. Discuss necessary theory, algorithm and applications of Gibbs sampling to solve any simple problem like bivariate normal distributions.

   OR

   Discuss convergence criteria in Gibbs sampling.

7. What is the significance of Importance Sampling over Simple Sampling? Illustrate by considering one of the examples to evaluate integration of a function.

   OR

   What are main characteristics of random numbers? Also discuss criteria to check the quality of random number generators.

8. Discuss Metropolis algorithm. Also explain why Hastings realize it to modify and what were his modifications
9. Explore the significance of Markov Chain Monte Carlo considering one example.
10. A manufacturer claims that the shipment contains only 5% of defective items, but the inspector feels that in fact it is 10%. We have to decide wheather to accept or to reject the shipment based on θ, the proportion of defective parts. Before we see the real data. Let's

assign a 50-50 chance to both suggested values of θ i.e. $\pi(0.05) = \pi(0.10) = 0.5$. A random sample of 20 parts has defective ones. Calculate the posterior distribution of θ (you may use table of binomial distribution):

$f(x|\theta = 0.05) = F(3|\theta = 0.05) = 0.9841;\quad F(2|\theta = 0.05) = 0.9245$

and $f(x|\theta = 0.10) = F(3|\theta = 0.10) = 0.8670;\quad F(2|\theta = 0.10) = 0.6769$