MDS 601-2081✡

Tribhuvan University

Institute of Science and Technology

2081

✡

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. Define research. Describe the main characteristics of scientific research.

2. Describe the nature and purpose of historical research.

3. Describe the APA method of citation for single author and multiple auuthor with example.

4. Illustrate Nominal scale and Ratio Scale.

5. Why research problem is started in a study? What are the criteria for a good research problem?

Group B **[5×6=30]**

6. Define the term "Validity". Describe the various types of validity.

7. Illustrate the major steps involved in scientific research process.

8. Mention the principal steps in a sample survey.

9. Under what condition a laboratory experiment be better than field experiment.

OR

What is sampling error? How sampling error can be reduced during the process of data collection.

10. What library skills are required for thorough survey of literature related to a research topic in Data science.

OR

In Measuring reactions time, a psychologist estimates that the standard deviation is 0.05 seconds. How large a sample of measurement must be taken in order to be 99% confident that the error of his estimate will not exceed 0.01 seconds?

IOST,TU

Tribhuvan University

Institute of Science and Technology

2081

✡

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. List the differences between OLAP and OLTP.
2. What is data pre-processing? List the data pre-processing techniques useful for numeric data.
3. Explain the senario where Ensemble methods are very useful.
4. What is association analysis?
5. Explain Fuzzy Clustering technique.

Group B (5×6=30)

6. What is the kernel function in SVM? Explain Gaussian Kernel Radial Basis Function (RBF) with examples.

OR

What is the frequent pattern growth approach for Association analysis? Discuss how this solves the issue of Apriori algorithm.

7. Explain different types of data visualization techniques.

OR

Explain the evaluation techniques that can be used for evaluating clustering models.

8. Explain the concept of Knowledge Discovery in Databases (KDD) and its importance in data mining. Discuss the key steps involved in the KDD process and provide examples to illustrate each step.
9. What is a decision tree? How can we construct a decision tree based classification model? Explain with appropriate examples.
10. A telecommunications company, Telecom Co, wants to identify unusual call patterns within its large dataset of customer call records. They have decided to use clustering techniques for outlier detection. Define what an outlier is in this context and explain how TelecomCo can use a clustering based approach to detect these outliers.

MDS603-2081✡

Tribhuvan University

Institute of Science and Technology

2081

✡

Master Level / Second Year /Third Semesters
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. Explain about the different characteristics of big data.
2. How the functional programming of MapReduce different from imperative programming?
3. What are the core components of Hadoop?
4. Explain about the different types of NoSQL databases.
5. Explain about the different configuration modes of Apache Pig.

Group B **[5×6 = 30]**

6. Explain about the anatomy of a MapReduce job run, failures, shuffle and sort with reference to word frequency count example.

7. Explain about the anatomy of write operation ofthe Hadoop Distributed File System (HDFS).

OR

Explain about the HDFS architecture and HDFS commands.

8. Explain about the CAP theorem. What do you mean by "Eventual Consistency" in NoSQL databases?

OR

Write short notes on:

a) Resilient Distributed Datasets
b) Spark Streaming

9. Write mongodb query for the following database:

➢ db.order.find({})

```
[{ cust_id: "ID1", "ord_date": ISODate("2018-05-04"), "price": 400, "status": "A", "item
qty": 20},
{ cust_id: "ID1", "ord_date": ISODate("2020-05-04"), "price": 400, "status": "not A",
"item qty": 20},
{ cust_id: "ID2", "ord_date": ISODate("2015-05-04"), "price": 200, "status": "not A",
"item qty": 10},
... ]
```

a) For each unique cust_id, sum the price field and sort it by sum.

b) For each unique cust_id, and unique day, month and year, sum the price field.

c) For each unique cust_id with status A, sum the price field and return only those records where the sum is either 250 or 300.

10. Explain about the architecture of Apache Hive. Also, compare Apache Hive with Apache Spark and Apache Pig.

MDS606-2081✡

Tribhuvan University

Institute of Science and Technology

2081

✡

Master Level / Second Year /Third Semester
Full Marks: 45

Data Science (MDS 606)
Pass Marks: 22.5

(Decision Analysis)
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 Risk Profile, Risk Appetite and Risk tolerance.

2. Explain about decision making Conditions in decision analysis.

3. A trader has two investment opportunities, A and B available to him but does not have enough capital to invest in the both. The probability of success on A is 0.70 while that on B is 0.40. Both the investments require an initial capital of Rs.200000 and both return nothing if the venture is not successful. Investment A returns Rs.300000 over cost if it is successful, whereas the successful completions of B will return Rs.500000 over cost. Using EMV criterion decide the best strategies the trader should adopt.?

4. The payoff matrix of a two-person zero sum game is

Player A	Player B		
	B_1	B_2	B_3
A_1	1	2	1
A_2	0	-4	-1
A_3	1	3	-2

Determine the number of saddle points and the corresponding optimal solutions. Find also the best strategy for each player.

5. A company manufactures two products radio and transistors which must be processed through assembly and finishing departments. Assembly has 90 hours available; finishing can handle up to 72 hours of work. Manufacturing one radio requires 6 hours in assembly and 3 hours in finishing. Each transition requires 2 hours in assembly and 4 hours in finishing. If profit is 1.20 per radio and 1.90 per transistor, determine the best combination of radios and transistors to realize a maximum profit of 2000. Formulate a problem as a GP problem.

Group B (5×6=30)

6. What is decision analysis? Describe the various steps of decision-making processes.

OR

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

7. A physician purchases a particular vaccine on Monday each week. The vaccine must be used within the week following, otherwise it becomes worthless. The vaccine costs Rs.2 per dose and the physician charges Rs.4 per dose. In the past 50 weeks, the physician has administered the vaccine in the following quantities.

Doses per week	20	25	50	60
Number of weeks	5	15	25	5

a) Determine how many doses the physician should buy every week.

b) Compute EVPI.

c) A physician is thinking of spending on a small market survey to obtain additional information regarding the demand levels. How much should he be willing to spend on such a survey?

OR

A distribution of past daily sales of a commodity is as follows:

Daily sales (units)	500	600	700	800	900
Probability	0.05	0.15	0.35	0.30	0.15

If selling price per unit is Rs.40 and cost price per unit is Rs.25 and salvage price per unit is Rs.5, what is

a) optimum quantity?

b) maximum expected profit?

c) expected values for perfect information?

8. A newspaper vendor buys a new started local paper at the rate of Rs.5 and sells it at the rate of Rs.10. The unsold papers do not have any value. The vendor knows that he cannot sell more than 20 papers in a day and the minimum sale would not be less than 18. How many papers should he buy based on (a) Maximax criterion (b) maximin criterion (c) minimax regret criterion?

9. Use dominance rule to reduce the size of the following game to 2×2 game and hence find the optimal strategies and the value of the game.

Player A	Player B		
	B_1	B_2	B_3
A_1	6	12	7
A_2	11	7	12
A_3	10	6	11

10. Solve the following Goal Programming model by simplex method:

Minimize $Z = d_1^-$

Subject to the constraints

$200x_1 + 100x_2 \leq 600$

$100x_1 + 100x_2 \leq 400$

$400x_1 + 800x_2 + d_1^- - d_1^+ = 10000$

and $x_1, x_2, d_1^-, d_1^+ \geq 0$, where x's represents the decision variables and d's represents deviational variables.

MDS 607-2081✡

Tribhuvan University

Institute of Science and Technology

2081

✡

Master Level / Second Year /Third Semester | Full Marks: 45

Data Science (MDS 607) | Pass Marks: 22.5

(Monte Carlo Methods) | 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. How does Bayesian inferences differ from other approaches of Statistics? Explain.

2. What are good random numbers? Describe.

3. Distinguish the random walks in 1D and 2D.

4. Write down the meaning of transition probabilities with examples.

5. Explain the meaning of "continuous state space".

Group B **(5×6=30)**

6. Discuss Metropolis Hasting Markov Chain Monte Carlo method with suitable examples.

OR

Explore 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. How do you distinguish "stationary distribution" from other distributions. Explain an algorithm to get "stationary distribution" using Python code.

9. "Monte Carlo method is the only way to carry on multi dimensional integration". Justify this statement by examples.

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 whether 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 3 defective ones. Calculate the posterior distribution of θ (you may use table of binomial distribution):
$f(x|\theta = 0.05) = F(3|\theta = 0.05) = 0.9841$; $F(2|\theta = 0.05) = 0.9245$
and $f(x|\theta = 0.10) = F(3|\theta = 0.10) = 0.8670$; $F(2|\theta = 0.10) = 0.6769$.

IOST,TU