MDS 501-2080✡

Tribhuvan University
Institute of Science and Technology
2080
✡

Subject: **Fundamental of Data Science**
Course No: **MDS 501**
Level: **MDS /I Year /I Semester**

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

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

Group A [5 × 3 = 15]

1. Is CRISP-DM waterfall or Agile? Support your answer with suitable example.
2. What are the different commonly used data formats used across data science project?
3. What do you mean by regression and classification? Are they supervised or unsupervised technique? Justify.
4. What do you mean by backpropagation? Conceptually, how do they differ from forward propagation in neural network?
5. How group unaware selection and adjusted group thresholds are useful for addressing bias?

Group B [5 × 6 = 30]

6. Explain the different phase of OSEMN lifecycle for data science projects.
   OR
   What do you mean by TDSP? How does it differ from other frameworks?
7. Explain time series analytics with its types.
8. What do you mean by missing data? How do you tackle them in any data science project? Explain with example of your own.
9. Elaborate the concept behind Naïve Bayes algorithm for classification task.
10. Apply map-reduce to the following set of data:

    Data, Science, Engineering
    Engineering, Data, Analytics
    Analytics, Intelligence, Science

    OR

    What is Hadoop? Explain the different components of Hadoop.

Tribhuvan University
Institute of Science and Technology
2080

✡

Subject: **Data Structures and Algorithms**
Course No: **MDS 502**
Level: **MDS /I Year /I Semester**

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

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

Attempt All questions.

Group A [5 × 3 = 15]

1. Define asymptotic notation. Explain big-oh (O) notation with example. (1 + 2)
2. How do you implement enqueue and dequeue operations in linear queue? (3)
3. What is precondition for binary search? Explain binary search algorithm. (1 + 2)
4. Explain preorder traversal with example. (3)
5. What is spanning tree? Explain minimum spanning tree in brief. (1 + 2)

Group B [5 × 6 = 30]

6. Explain algorithm for converting an infix expression to postfix using stack. Use this algorithm to convert (A + B - C) * D to postfix. (4 + 2)

   OR

   List some applications of stack. Explain algorithm for evaluating a postfix expression using stack with suitable example. (1.5 + 4.5)

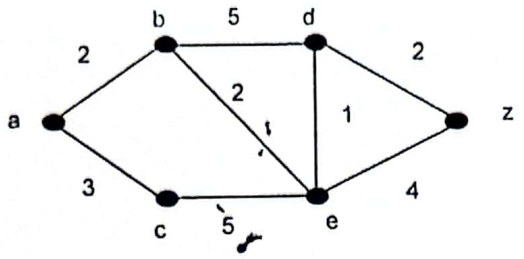
7. Compare linked list with array. How do you insert and remove nodes in singly linked list? (2 + 4)

   OR

   What is circular linked list? How do you implement stack using linked list? (1.5 + 4.5)

8. Explain quick sort along with its time complexity. Trace the execution of quick sort algorithm with the array of numbers 30, 20, 15, 37, 45, 9, 23, 15, and 3. (2 + 4)
9. Define AVL tree. Construct AVL tree for the sequence 27, 66, 80, 9, 4, 14, 28, 8, and 6. (1 + 5)

10. Use Dijkstra's shortest path algorithm to find the shortest path between the vertices a and z in the graph given below. (6)



Tribhuvan University
Institute of Science and Technology
2080

✡

Subject: **Database Management Systems**
Course No: **MDS 505**
Level: **MDS /I Year/I Semester**

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

Candidates are required to give their answer in their own words as far as practicable. The figures in the margin indicate full marks.
Attempt ALL questions.

Group A [5 × 3 = 15]

1. How problems related to data integrity and automicity occur in flatfile system?
2. Describe relationship type and relationship set in ER Model.
3. Define parallel database with its possible architectures.
4. What are the security threats to databases?
5. State the concept of information rule and non-subversion rule in Codd's Rule.

Group B [5 × 6 = 30]

6. Create a relation containing trivial and non-trivial functional dependencies. When a relation is in 3NF? [3+3]

OR

What is the stored procedure? Create a stored procedure of your choice containing input and output parameters. [2+4]

7. Describe how incorrect summary and unrepeatable read problems occur in concurrent execution of transactions? Illustrate with examples. [6]

OR

Consider any two transactions T1, T2 performing read and write operations over data items X, Y, Z. Now show how Two-Phase Locking Protocol ensures serializability in concurrent execution of T1 and T2. [6]

8. How fixed length attributes and variable length attributes are represented in variable length record approach in file organization? What is a slotted page structure? [4+2]
9. What is embedded SQL? How is heuristic optimization of query trees done? Illustrate with an example. [2+4]
10. Consider the following relations defining exam management system. [6]
Answersheet(*Copyno, Symbolno, Subject_name, Marks, Eid, Sid*)

Examiner(*Eid, Ename, Qualification*)
Scrutinizer(*Sid, Sname*)
Scrutinydetail(*Eid, Sid, Scrutiny_date*)

Write the **SQL and Relational Algebra** statements for following;

a) Find the names of the subjects having marks greater than 40.
b) Find the number of copies which are examined and scrutinized.
c) Find the names of the examiners and scrutinizers with the scrutinizing date July 1.
d) Find the names of examiners who have not examined any answer sheet.

Tribhuvan University

Institute of Science and Technology

2080

✡

Subject: **Mathematics for Data Science** *Full Marks*:45
Course No: **MDS 504** *Pass Marks*:22.5
Level: **MDS /I Year /I Semester** *Time*:2 hrs

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

Group A [5 × 3 = 15]

1. Are the following sets form the subspace of R^2? Justify.

   a) The set $S = \{\begin{pmatrix} x \\ y \\ z \end{pmatrix} : y + z = 2 \text{ and } x, y, z \in R\}$ in the vector space R^3.

   b) The closed L_2 ball, $B(0,3)=\{X=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in R^2 : \|X\| \leq 9\}$. **[1.5+1.5]**

2. What is rank of a matrix? Reduce the matrix $A = \begin{pmatrix} 3 & 1 & 3 & 8 \\ 2 & 2 & 6 & -1 \\ 10 & 3 & 9 & 7 \\ 8 & 4 & 11 & 17 \end{pmatrix}$ to Echelon form and hence find the rank. **[1+2]**

3. Let $T: R^2 \to R^2$ be defined by $T(1, 0) = (1, 3)$, $T(1, 2) = (0, -2)$. If T is linear, find the formula for $T(x, y)$. What is the matrix represented by T relative to the standard basis? **[1.5+1.5]**

4. What is meant by linear independence of a set of vectors. Let V and W be two vector spaces over the field F and $T: V \to W$ be a linear transformation. Then prove that the kernel of T is a subspace of V and the image of T is a subspace of W. **[1+1+1]**

5. What is quadratic form? Let S be a 3 ×3 square matrix with a quadratic form in 3 variables. Then there exists a 3 ×3 symmetric matrix T such that $X^TSX = X^TTX, \forall X \in R^3$. **[1+2]**

Group B [5 ×6 = 30]

6. What is L_∞ norm on a vector n-space R^n? Write any two properties. Let $\|.\|$ be the Euclidean norm , X and Y be two vectors in R^n. State and prove the Triangle Inequality and Parallelogram Law. Verify Cauchy -Schwarz inequality for $X = (1, 3)$ and $Y = (2, 1)$. [1+2+2+1]

7. Define Kernel and Image of a linear transformation T: $V \to W$. If $v_1, v_2, \ldots, v_n$ are linearly independent vectors in V and $KerT = \{0_V\}$.Is the set $\{ T(v_1), T(v_2), \ldots, T(v_n)\}$ forms linearly independent vectors in W? Justify. Also, show that the set {(0, 1), (1, 1)} of vectors span R^2. [1+2.5+2.5]

OR

What is quadratic form. Let A be a 3 ×3 square matrix with a quadratic form in 3 variables. Then there exists a 3 ×3 symmetric matrix B such that $X^TAX = X^TBX, \forall X \in R^3$. Further, express the quadratic form $x_1^2 + x_1x_2 - 4x_3x_1 + 2x_2x_3 - 4x_3^2$ as the difference of squares. [1+2.5+2.5]

8. Let $A = \begin{pmatrix} 3 & 2 \\ 2 & 1 \end{pmatrix}$ be a symmetric matrix. Find the orthogonal matrix D such that $D^{-1}AD$ is a diagonal matrix.Let $u_1, u_2, \ldots, u_n$ be the eigenvectors associated with the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of a $n \times n$ symmetric matrix B respectively, then prove that $B = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \ldots + \lambda_n u_n u_n^T$. . [3+3]

9. Explain the role of basic linear algebra techniques that are useful in the study of data science. Describe how the various concepts of Vector Spaces are applied in Machine Learning. [2+4]

OR

Define four fundamental subspaces of a matrix A. Determine the values of the constants a and b for which the system $3x - 2y + z = b$, $5x - 8y + 9z = 3$, $2x + y + az = -2$ has a unique solution, no solution and infinitely many solutions. [2+4]

10. Discuss in brief about singular value decomposition? Find the singular value decomposition of the matrix $\begin{pmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{pmatrix}$. [1+5]

)S503-2080 ✡

Tribhuvan University
Institute of Science and Technology
2080

✡

)ject: **Statistical Computing with R**
urse No: **MDS 503**
)el: **MDS /I Year /I Semester**

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

ndidates are required to write answers with examples for answering question numbers 1-5in the answer sheet and use top foranswering question numbers 6-10 with R scripts. R scripts must be knitted as PDFwith the puts/interpretation of question number 6-10 and it must be saved in a folder with the R script and the name/exam roll nber for grading.

tempt ALL Questions.

Group A [5 × 3 = 15]

1. Explain these terms with examples for R:
   a) Getting multi-way table with array
   b) Creating class intervals of continuous variable
   c) Missingness vs nothingness

2. Explain following concepts with focus on R software:
   a) Raw data
   b) Data wrangling
   c) Tidy data

3. Explain the followings with examples for R:
   a) Reference range based on mean
   b) Reference range based on median
   c) Outliers and extreme values

4. Explain the following concepts with focus on R software:
   a) Test of normality
   b) Parametric tests
   c) Residual analysis

5. Describe decision tree classification model with focus on:
   a) Bagging
   b) Improved bagging
   c) Boosting

Group B [5 × 6 = 30]

6. Do the following in R Studio with R script:
   a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with 150 random cases of each variable. Your exam roll number must be used to set the random seed.
   b) Show a sub-divided bar diagram of body mass index variable by sex and socio-economic variables separately with interpretations.

c) Show multiple bar diagram of age variable withsexand educational level variablesand interpret carefully
d) Show boxplots of age and body mass index variable separately and interpret the results carefully
e) Create histogram of age and body mass index variable separately and interpret the results carefully

7. Do the following in R studio and with R script to knit HTML output:
a) Define an object "rating" with 9, 2, 5, 8, 6, 1, 3, 2, 8, 4, 6, 8, 7, 1, 2, 6, 10, 5, 6, 9, 6, 2, 4, 7 values.
b) Replicate the given table obtained from SPSS software for the rating object in R.

rating

Valid	Frequency	Percent	Valid Percent	Cumulative Percent
1	2	8.3	8.3	8.3
2	4	16.7	16.7	25.0
3	1	4.2	4.2	29.2
4	2	8.3	8.3	37.5
5	2	8.3	8.3	45.8
6	5	20.8	20.8	66.7
7	2	8.3	8.3	75.0
8	3	12.5	12.5	87.5
9	2	8.3	8.3	95.8
10	1	4.2	4.2	100.0
Total	24	100.0	100.0	

8. Do the following in R Studio with R script:
a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with random 250 cases of each variable. Your exam roll number must be used to set the random seed.
b) Create scatterplot of age and body mass index variable and interpret it carefully.
c) Which correlation coefficient must be used based on the interpretation of the scatterplot? Why?
d) Compute the best correlation coefficient identified from the scatterplot and interpret it carefully.
e) Test whether this correlation coefficient is statistically valid or not and justify its value.

OR

Do the following in R Studio with R script:
a) Create a dataset with following variables: age (18-99 years), sex (male/female), educational levels (No education/Primary/Secondary/Beyond secondary), socio-economic status (Low, Middle, High) and body mass index (14 – 38) with random 250 cases of each variable. Your exam roll number must be used to set the random seed.
b) Check if body mass index variable follows normal distribution using suggestive plot and confirmative tests and interpret the results carefully.
c) Check if body mass index variables have equal variance for sex variable using suggestive plot and confirmatory test and interpret the results carefully.
d) Which independent sample t-test must be used to compare body mass index by sex? Why?
e) Perform the independent sample t-test identified above and interpret it carefully.

9. Do the following in R Studio using "mtcars" dataset with R script:
   a) Divide the mtcars data into train and test datasets with 70:30 random splits.
   b) Fit a supervised logistic regression model and naïve bayes classification models on train data with transmission (am) as dependent variable and miles per gallon, displacement (disp), horse power (hp) and weight (wt) as independent variable.
   c) Predict the transmission (am) variable in the test data for both the models and interpret the result carefully
   d) Get the confusion matrix, sensitivity, specificity of both the models using predicted transmission variable on test data and interpret them carefully.
   e) Which supervised classification model is the best for doing prediction? Why?

10. Do as follows using given dataset of 10 US cities in R studio with R script:

City	Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington D.C
Atlanta	0	587	1212	701	1936	604	748	2139	2182	543
Chicago	587	0	920	940	1745	1188	713	1858	1737	597
Denver	1212	920	0	879	831	1726	1631	949	1021	1494
Houston	701	940	879	0	1374	968	1420	1645	1891	1220
Los Angeles	1936	1745	831	1374	0	2339	2451	347	959	2300
Miami	604	1188	1726	968	2339	0	1092	2594	2734	923
New York	748	713	1631	1420	2451	1092	0	2571	2408	205
San Francisco	2139	1858	949	1645	347	2594	2571	0	678	2442
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	2329
Washington D.C	543	597	1494	1220	2300	923	205	2442	2329	0

   a) Get this data in R and compute dissimilarity distance as city.dissimilarity object.
   b) Fit a classical multidimensional model using the city.dissimilarity object.
   c) Get the summary of the model and interpret it carefully.
   d) Get the bi-plot of the model and interpret it carefully.
   e) Compare this model with the first two components from principal component analysis model in this data.

OR

Use the first four variablesof "iris" data file into R Studio and do as follows with R script:

   a) Fit a k-means clustering model in the data with k=2 and k=3.
   b) Plot the clusters formed with k=3 in the **single graph** and interpret them carefully.
   c) Add cluster centers for the plot of clusters formed with k=3 above and interpret it carefully.
   d) Compare the k=3 cluster variable with Species variable of iris data using confusion matrix and interpre the result carefully.