Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

Subject: **Statistical Computing with R**
Course No: **MDS 503**
Level: **MDS /I Year /I Semester**

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

*Candidates are required to write answers with examples for answering question numbers **1-5**in the answer sheet and use laptop for answering question numbers **6-10 with R scripts**. R scripts must be knitted as PDFwith the outputs/interpretation of question number 6-10 and it must be saved in a folder with the PDF file/s alongside the name/class roll number for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Describe data visualization with focus on:
   a) Concept of grammar of graphics with Wilkinson's approach
   b) Layers in grammar of graphics with ggplot package's approach
   c) Statistical transformations in grammar of graphics
2. Describe followings for checking fit of the multiple linear regression model:
   a) Outliers
   b) Cook's distance
   c) Leverage
3. Describe supervised learning **classification** regression model with focus on:
   a) Model fit indices
   b) Confusion matrix with an example
   c) Prediction accuracy with ROC curve
4. Describe following with example on it use:
   a) Poisson regression
   b) Zero-inflated Poisson regression
   c) Negative binomial regression
5. Describe supervised linear regression model with focus on:
   a) Cross-validation
   b) K-fold cross-validation
   c) Repeated k-fold cross-validation

Group B **[5 × 6 = 30]**

6. Do the following in R Studio using ggplot2 package with R script to knit PDF output:
   a) Create a dataset with following variables: age (10-99 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 200 cases of each variable. Your roll number must be used to set the random seed.
   b) Create scatter plot of age and body mass index variables using ggplot2 package and interpret the result carefully.
   c) Create classes of body mass index variable as: <18, 18-24, 25-30, 30+ and show it as pie chart using ggplot2 package and interpret it carefully
   d) Create histogram of age variable with bin size of 15 using the ggplot2 package and interpret it carefully
7. Do the following in R Studio using "airquality" data set of R with R script to knit PDF output:

a) Perform goodness-of-fit test on Temp variable to check if it follows normal distribution or not
b) Perform goodness-of-fit test on Temp variable by Month variable to check if the variances of mpg are equal or not on am variable categories
c) Discuss which independent sample test must be used to compare "Temp" variable by "Month" variable categories based on the results obtained above
d) Perform the best independent sample statistical test for this data now and interpret the results carefully

8. Do the following in R Studio using "Arrests" dataset of **car** package with R script to knit PDF output:
a) Divide the **Arrests** data into train and test datasets with 80:20 random splits
b) Fit a supervised logistic regression and naïve Bayes classification models on train data with "released" as dependent variable and colour, age, sex, employed and citizen as independent variable
c) Predict the released variable in the test datasets of these models and interpret the result carefully
d) Compare and decide which classification model is better for this data

9 & Do as follows using in-built "iris" dataset with R script to knit PDF output:
a) Create a "flower scale" of first four variables of iris dataset using the Principal Component Analysis
b) Compute the eigenvalues and interpret the PCA result carefully using Kaiser's criteria
c) Show the Scree plot and decide on the number of components to retain with careful interpretation
d) Revise the flower scale with 3 components using VARIMAX rotation and interpret the result carefully

OR

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 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

10. Use the first four variables of the "iris" data and do as follows with R Script to knit PDF output:
a) Fit a hierarchical clustering model using single linkage and get the dendogram for this model
b) Fit a hierarchical clustering model using complete linkage and get the dendogram for this model
c) Fit a hierarchical clustering model using average linkage and get the dendogram for this model
d) Find the best hierarchical clustering model model for this data and locate the number of clusters for it

OR

Use the first four variables of "iris" data file into R Studio and do as follows with R script to knit PDF output:

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 and interpret it carefully
d) Compare the k=3 clusters with Species variable using confusion matrix and interpret the result carefully

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

Subject: **Data Structure 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. Why do you study efficiency of algorithms? What is big-oh (O) notation? (1.5 + 1.5)
2. What is priority queue? Why do we need this queue? (1.5 + 1.5)
3. Compare recursion with iteration. Write a recursive function to find greatest common divisor. (1.5 + 1.5)
4. What are the benefits of using linked list? Compare singly linked list with doubly linked list. (1.5 + 1.5)
5. Compare linear search with binary search? What are their time complexities? (2 + 1)

Group B [5×6 = 30]

6. Explain algorithm to convert an infix expression to postfix using stack. Use this algorithm to convert A * B + C / D – E into postfix. (3 + 3)

OR

List some applications of stack. Explain algorithm to evaluate a postfix expression using stack with suitable example. (1 + 5)

7. Explain merge sort. Trace the execution ofmerge sort algorithm with the array of numbers34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)

OR

Explain shell sort. Trace the execution of shell sort algorithm with the array of numbers 34, 23, 17, 31, 45, 7, 21, 15, 8, and 1. (2 + 4)

8. Explain linear probing. Suppose, the set of keys is {17, 12, 14, 10, 49, 58, 9, 50}, m = 10, and $h(x) = x \bmod 10$. Show the effect of successively inserting these keys using quadratic probing. (2 + 4)
9. What is AVL tree. Construct AVL tree for the sequence 21, 26, 30, 9, 4, 14, 28, and 18. (1 + 5)
10. What is graph traversal? Explain breadth first search (BFS) algorithm for traversing graphs with example. (1 + 5)

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

Subject: **Database Management Systems** | *Full Marks*: 45
Course No: **MDS 505** | *Pass Marks*: 22.5
Level: **MDS /I Year/I Semester** | *Time*: 2 hrs

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

Group A **[5 × 3 = 15]**

1. State non-subversion rule in Codd's Rule.
2. Why group by and having clauses are used in SQL? Support your answer with suitable examples.
3. How division operation is done in relational algebra? Illustrate with example.
4. Define multivalued dependency.
5. Write SQL statement to create a table and insert value into it. Use your own assumption.

Group B **[5 × 6 = 30]**

6. Differentiate discretionary access control from mandatory access control mechanism. [6]

OR

How database is created using XML? Create a XML file to store information about School. Use your own assumptions if required. [2+4]

7. Describe different architectures in parallel databases. [6]

OR

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]

8. What is query optimization? How materlization is different from pipelining? Prepare the query tree and optimized query tree for following SQL statement; [1+2+3]
SELECT FNAME, ADDRESS FROM EMPLOYEE INNER JOIN DEPARTMENT ON DNO = DNUMBER WHERE DEPARTMENT.DNAME='Research';

9. Describe the rules for converting ER diagram into relational table. Illustrate with suitable example reflecting each cases. [3+3]

10. How can you determine whether a schedule is conflict serializable or not? Given following schedule, determine whether it is conflict serializable or not and justify your answer. [1+2+3]

T_1	T_2
read_item(X); $X := X - N$;	
	read_item(X); $X := X + M$;
write_item(X); read_item(Y);	
	write_item(X);
$Y := Y + N$; write_item(Y);	

Schedule C

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

Subject: **Fundamentals 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 answers in their own words as far as practicable.
Attempt ALL questions.

Group A [5 × 3 = 15]

1. Explain in brief the concept of learning through data and experience.
2. Define entropy. How is it used with decision tree?
3. Explain why data ethics is crucial in data science.
4. What is deep learning? How is it similar or different from neural network.
5. Write short notes on:
   a) Bias blind spot
   b) Predictive Analytics

Group B [5 × 6 = 30]

6. Explain the working mechanism of KNN algorithm for classification and regression.

OR

Explain the forward propagation and backward propagation of neural networks.

7. What do you mean by Random Forests? Why is random forest commonly used for feature selection despite being a machine learning model? Is random forest prone to overfitting? Why or why not?
8. What do you mean by support vectors? Explain with appropriate example of your own how supports vectors are useful in machine learning ?
9. 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.

10. A healthcare organization developed a machine learning model to predict patients' risk of developing certain medical conditions based on their electronic health records (EHR). The model was trained using historical patient data, including diagnoses, treatments, and outcomes. However, it was later discovered tha the model exhibited bias against patients from lower socioeconomic backgrounds. The training data disproportionately represented patients from wealthier neighborhoods who had better access to healthcare services and resources. As a result, the model erroneously associated higher socioeconomic status with lower health risks, leading to underestimating the risk of certain conditions for patients from disadvantagec backgrounds.

***

Tribhuvan University
Institute of Sciences and Technology

SCHOOL OF MATHEMATICAL SCIENCES

Second Assessment 2081

Subject: **Mathematics for Data Science**
Course No: **MDS 504**
Level: **MDS /I Year /I Semester**

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

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

Group A [5×3 = 15]

1. If $\lambda = 1, 5$ eigenvalues of the matrix $\begin{pmatrix} 7 & 4 \\ -3 & -1 \end{pmatrix}$, find a basis for the eigenspace corresponding to each eigenvalue.
2. Find the maximum value of $9x_1^2 + 4x_2^2 + 4x_3^2$ subject to the constraints $x^Tx = 1$ and $x^Tu_1 = 0$, where $u_1 = (1, 0, 0)$. Find x where it is attained. Here, u_1 is a unit eigen vector corresponding to the greatest eigenvalue $\lambda = 9$ of the matrix of the quadratic form.
3. Find the singular values of the matrix $\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 1 \end{pmatrix}$.
4. Consider the quadratic form $Q(x) = 2x_1^2 + 4x_1x_2 - 4x_3x_1 - x_2^2 + 8x_3x_2 - x_3^2$. Decide whether this quadratic form is positive, negative or indefinite.
5. Determine if the following homogeneous system has a nontrivial solution. Then describe the solution set.
   $3x_1 + 5x_2 - 4x_3 = 0, \quad -3x_1 - 2x_2 + 4x_3 = 0, \quad 6x_1 + x_2 - 8x_3 = 0.$

Group B [5×6 = 30]

6. Consider the matrix: $A = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$.
   a) What can we say about the action of A on an arbitrary vector ?
   b) What are examples of eigenvalues and eigenvectors of this matrix?
   c) What does the discussion for this example illustrate?

   OR

   a) Let v_1, v_2 be the eigenvectors associated with the eigenvalues λ_1, λ_2 of a 2×2 symmetric matrix A respectively. Prove that if $A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ and $V = (v_1 v_2)$, then $A = VAV^T$.
   b) Find all 2×2 matrices A which admit the normalized eigenvectors $v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ with the corresponding eigenvalues λ_1 and λ_2.
7. a) Let A be an $n \times n$ matrix. Prove that if A has n linearly independent eigenvectors, then A is diagonalizable
   b) Show that the matrix $A = \begin{pmatrix} 2 & -1 \\ 1 & 4 \end{pmatrix}$ is not diagonalizable.

8. a) Prove that if A is a symmetric $n \times n$ matrix and $B_A(v,w) = v^T Aw$, then $B_A(v,w)$ is linear in the first variable v.

b) Write the quadratic form $10x_1^2 - 8x_1x_2 + 4x_2^2$ as $x^T Ax$. Transform it into a quadratic form without the cross product term using eigenvalues and eigenvectors.

9. Find an SVD of the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 2 \end{pmatrix}$.

OR

a) Prove that if A is an $m \times n$ matrix, then all the eigenvalues of $A^T A$ are non-negative.

b) Find the eigenvalues and eigenvectors of $A^T A$ where $A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ -2 & 1 \end{pmatrix}$.

10. What is reduced row echelon form? Illustrate with an example of an augmented matrix of order 4×5. Solve the following linear system by placing the augmented matrix in reduced row echelon form.

$2x + y - z = 2, 4x + 3y + 2z = -3, \ 6x - 5y + 3z = -14.$

***