Tribhuvan University
Institute of Science and Technology
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 answer in their own words as far as practicable.
Attempt All questions.

Group A [5 × 3 = 15]

1. Data science is often considered a blurry subject. What do you think are the reasons behind it?
2. Discuss the importance of data validation in data science. Name some common methods of data validation.
3. What do you mean by time series analysis? What are the frequent patterns observed in time series analysis?
4. Describe the common data enrichment techniques often used by data scientist.
5. Write short notes on:
   a. Stereotyping
   b. Data Lake

Group B [5 × 6 = 30]

6. Define and explain the TDSP lifecycle in data science.

OR

A leading retail chain in Nepal wants to use data science to enhance its customer experience and optimize inventory management. They have data from customer transactions, online browsing behavior, and social media interactions.

Briefly explain how data science can be applied in the retail industry to improve customer experience and optimize inventory management. Provide specific examples of data science techniques that could be used in this context.

7. Describe the common data quality issues with tabular data and their mitigation techniques with appropriate examples.
8. How does machine learning differs from traditional learning? Explain the various type of machine learning techniques.

9. Explain the generic process of real time data analytics in big data in context to Apache Kafka.

OR

Apply map-reduce paradigm to the following set of data:

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

10. You are analyzing a dataset containing information about customer orders for an e-commerce platform. However, upon initial inspection, you notice several data quality issues that may impact the reliability of your analysis.

Describe the common data quality issues that you may have identified in the dataset, providing specific examples for each issue. Explain the potential consequences of these issues on your analysis and propose strategies to address them effectively.

Tribhuvan University
Institute of Science and Technology
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 foranswering question numbers 6-10 with R scripts. R scripts must be knitted as PDF with the outputs/interpretation of question number 6-10 and it must be saved in a folder with the R script and the exam roll number for grading.

Attempt ALL Questions.

Group A [5 × 3 = 15]

1. Explainthe following concepts with R codes:
   a) Numeric variable and its type with example
   b) Categorical variable and its type with example
2. Explainthe following concept with focus on R software:
   a) Manipulating row and column of data frame in dplyr package with an example
   b) Extract, Transform and Loadin dplyr package with an example
3. Explain the following concepts with focus on R software:
   a) Boxplot with five number summaries with example
   b) Boxplot with outliers with example
4. Explain the following concept with focus on R software:
   a) Leverage in linear regression with example
   b) Multicollinearity in logistic regression with example
5. Explain the following conceptswith focus on R software:
   a) Biplot from principal component analysis
   b) Biplot from classical multidimensional scaling

Group B [5 × 6 = 30]

6. Load the "igraph" package in R studio and do the basic SNA as follows with R script and HTML output:
   a) Define "g1" as graph object with ("R", "S", "S", "T", "T", "R", "R", "T", "U", "S") as its elements
   b) Plot "g1" with node color as green, node size as 30, link color as red and link size as 5 and interpret it
   c) Get degree, closeness and betweenness of g1 and interpret them carefully
   d) Get hub and communities of this data and interpret them carefully

OR

Do the following in R Studio using "airquality" dataset with R script:
a) Replace missing values of "Ozone" variable with median of this variable as corrected Ozone
b) Get the histogram of the corrected Ozone variable using base R plot and interpret it carefully
c) Get the boxplot of Wind variable using based R plot and interpret it carefully
d) Get the Wind variable outliers using median and interquartile range and compare them with boxplot outlier values with justification

7. Do as follows in R Studio and do as follows with R script and HTML outputs:
   a) Open R and then go to Help and Manuals in PDF and open "An Introduction to R" file
   b) Import this pdf file in R using "pdftools" package
   c) Perform pre-processing and create 'corpus' afterwards
   d) Find the most frequent terms, create its bar diagram and interpret carefully

OR

Do the following in R Studio using "airquality" dataset with R script:
   a) Get the boxplot of Temp variable using ggplot2package and interpret it carefully
   b) Create class intervals of Temp variable using dplyr package and show it as frequency distribution
   c) Get pie chart of Temp variable class intervals using ggplot2 package and interpret it carefully
   d) Get scatter plot of corrected Temp and Wind variables using ggplot2 package and interpret it carefully

8. Do the following in R Studio using "airquality" dataset with R script:
   a) Perform Shapiro-Wilk test on "Wind" variable to check if it follows normal distribution or not
   b) Perform Bartlett test on "Wind" variable by "Month" variable to check if the variances of Wind are equal or not on Month variable categories
   c) Fit1-way ANOVA to compare "Wind" variable by "Month" variable and interpret the result carefully
   d) Fit the TukeyHSD post-hoc test with 95% confidence interval and interpret the result carefully

9. Do the followingsin R Studio using "USArrests" dataset with R script:
   a) Divide the mtcars data into train and test datasets with 70:30 random splits
   b) Fit a supervised linear regression model and KNN regression model on train data with "Urban population – UrbanPop" as dependent variable and all other variables as independent variable
   c) Predict the UrbanPop variable in the test datasetsusing these two models and interpret results carefully
   d) Compare the fit indices (R-square, MSE, RMSE)of the two predicted models and choose the best model

10. Use the first four variables "iris" data and do as follows in the R Studio with R Script:
   a) Fit a hierarchical clustering model using average linkage and get the dendogram for this model
   b) Get the best value of number of clusters to form (k) using the fitted model above
   c) Fit the k-means clustering with the best value of k identified above and interpret it carefully
   d) Compare k-means result with the last variable of this data usig confusion matrix and interpret the result carefully

Tribhuvan University
Institute of Science and Technology
2081
✡

Subject: **Data Structures and Algorithms** *Full Marks*: 45
Course No: **MDS 502** *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. Compare big-oh (O), omega (Ω), and theta (θ) notations. (3)
2. How do you implement push and pop operations of stack? (3)
3. Define hashing. What is linear probing? (1 + 2)
4. Explain postorder traversal with example. (3)
5. Explain adjacency matrix representation of a graph. How is it different from incidence matrix representation? (2 + 1)

Group B **(5 × 6 = 30)**

6. Define queue. How do you implement circular queue using array data structure? (1 + 5)

OR

What is priority queue? Explain tail recursion with suitable example. (1.5 + 4.5)

7. How linked list differs from array? How do you insert nodes in doubly linked list? (2 + 4)

OR

What is header node in linked list? How do you implement queue using linked list? (1.5 + 4.5)

8. Explain merge sort along with its time complexity. Trace the execution of merge sort algorithm with the array of numbers 30, 20, 15, 37, 45, 9, 23, 15, and 3. (2 + 4)
9. What is AVL tree? Construct AVL tree for the sequence 26, 65, 81, 11, 6, 15, 28, 8, and 7. (1 + 5)
10. Define minimum spanning tree. Explain Prim's algorithm to find minimum spanning tree with suitable example. (1 + 5)

Tribhuvan University
Institute of Science and Technology
2081
✡

Subject: **Data Base 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 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. What are the advantages of using DBMS?
2. How is cost-based optimization of SQL queries done?
3. How is no waiting algorithm of deadlock prevention different from cautious waiting?
4. Define a statistical database with an example.
5. What is three schema architecture?

Group B [5 × 6 = 30]

6. Describe 1NF, 2NF and 3NF with appropriate examples. [6]

OR

Describe the dependency preservation and lossless join property. Describe 4NF with an example. [3+3]

7. Describe how lost update and dirty read problems occur in concurrent execution of transactions? Illustrate with examples. [6]

OR

How is exclusive lock different from shared lock? Consider any three transactions T1, T2 and T2 are performing read and write operations over data items A, B, and C. Now create a schedule containing T1, T2, and T3 that suffers from deadlock. [2+4]

8. How database records are mapped to files using fixed length record and variable length record approaches? Justify with examples. [6]

9. Design an ER diagram that contains at least three entities. One of the entity must be weak entity. There should be many to many relationship between the two strong entities. Use your own assumptions for other requirements. Now convert the so constructed ER diagram into equivalent database table schema. [3+3]

10. Consider the following relations defining the hotel management system. [6]
Food(*Fid, Fname, PriceRate, Cookedby, Inspectedby*)
Cook(*Cid, Cname, Cspeciality*)
FoodInspector(*Fid, Fname, Experience*)
Reportdetail(*Fid, Cid, Report_date*)

Tribhuvan University
Institute of Science and Technology
2081
✡

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 [3×5 = 15]

1. Show that
   a) The line $x_2 = ax_1$ is a subspace $\mathbf{R}^2$.
   b) The line $x_2 = ax_1 + b$ is not a subspace $\mathbf{R}^2$ for $b \neq 0$.
2. Find the maximum value of $Q(x) = x_1^2 + x_2^2 - 10x_1x_2$ subject to the constraints $x^Tx = 1$. Find a unit vector x in $\mathbf{R}^2$ at which $Q(x)$ is maximized.
3. Find the singular values of the matrix $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix}$.
4. Find a basis for the solution space of the equation $x + y - z = 0$.
5. When a system of linear equations said to be consistent? Determine if the following system is consistent. Do not completely solve the system.

$x_1 + 3x_3 = 2$, $x_2 - 3x_4 = 3$, $-2x_2 + 3x_3 + 2x_4 = 1$, $3x_1 + 7x_4 = -5$.

Group B [6×5 = 30]

6. a) Let A be an $m \times n$ matrix, B an $n \times p$ matrix, and C a $p \times q$ matrix, so that $(AB)C$ and $A(BC)$ are defined. Prove that $(AB)C = A(BC)$. Determine the size of the matrix $A(BC)$.
   b) Let $x = x_1e_1 + x_2e_2$, where e_1 and e_2 are unit basis vectors of $\mathbf{R}^2$. When does the equality $x = 0$ hold? What does $x = 0$ mean?
7. Let V be a subspace of $\mathbf{R}^n$ and w a vector in R^n.
   a) If $\{v_1, v_2, \ldots, v_k\}$ is an orthogonal basis for V, then prove that
   $$\frac{w \cdot v_1}{v_1 \cdot v_1}v_1 + \frac{w \cdot v_2}{v_2 \cdot v_1}v_1 + \cdots + \frac{w \cdot v_k}{v_k \cdot v_k}v_k,$$
   is the projection of w onto V.
   b) Moreover, if $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis for a vector space V, then prove that
   $w = (w \cdot v_1)v_1 + (w \cdot v_2)v_2 + \ldots + (w \cdot v_k)v_k$,
   is the projection of w onto V.

OR

Give a geometric description of span(v) and span(u, v). Consider the vectors $u=\begin{pmatrix}1\\2\end{pmatrix}$ and $v=\begin{pmatrix}-1\\1\end{pmatrix}$.

a) Write the vector $w=\begin{pmatrix}3\\2\end{pmatrix}$ in terms of the vectors u and v.

b) Show that the vectors u and v span R^2.

8. 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 $A=\lambda_1 v_1 v_1^T+\lambda_2 v_2 v_2^T$.

b) Find an orthogonal matrix V which diagonalizes the matrix $\begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}$. Also check that $V^TAV=\Lambda$, where Λ is a diagonal matrix.

OR

a) Prove that if $A\in R^{3\times3}$ with a quadratic form in 3 variables, then there is a symmetric matrix $B\in R^{3\times3}$ such that
$\forall x\in R^{3\times1}\ x^TAx=x^TBx$.

b) Classify the quadratic form:$-x_1^2-2x_1x_2-x_2^2$. Then make achange of variable, $x=Py$, that transforms the quadratic forminto one with no cross-product term. Write the new quadratic form. Determine P.

9. Let A be an $m\times n$ matrix. Prove that

a) A^TA is a square matrix.

b) A^TA is symmetric and so it is orthogonally diagonalizable.

c) All the eigenvalues of A^TA are non-negative.

10. What is a reduced rowechelon form? Solve the following linear system by transferring the augmented matrix in reduced rowechelon form.

$6x-3y+z=31,\qquad 5x+y+12z=36,\qquad 8x+5y+z=11.$