MDS501-2078 ✡

MDS I Year, I Semester

Tribhuvan University
Institute of Science and Technology
2078

✡

Master Level / I Year /First Semester/ Science
Data Science (MDS 501)
(Fundamentals of Data Science)

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 and explain in short three main limitations of Data Science.
2. List three main differences between TDSP lifecycle and OSEMN framework.
3. List the differences between supervised and unsupervised machine learning methods including examples.
4. Define and briefly describe neural networks. List a few example use-cases where neural networks can be used.
5. What are decision trees? What does impurity of a node mean in context of decision tree?

Group B

(5×6=30)

6. Explain CRISP-DM in detail with a diagram and an example.

OR

What are the benefits of using OSEMN framework over CRISP-DM? In what scenarios will you use OSEMN instead of CRISP-DM?

7. Define with an example the K-means along with its limitations.
8. List major differences between linear and logistics regressions with examples.
9. Explain map-reduce programming paradigm. Include details on how it applies to Hadoop ecosystem.
10. Describe in detail how the quality of data can be assessed during data munging.

OR

List and describe how you would address various kinds of issues during data cleanup while doing data munging.

Tribhuvan University
Institute of Science and Technology
2078

✡

Master Level / I Year /First Semester/ Science
Data Science (MDS 502)
(Data Structure and Algorithms)

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 asymptotic notation. Explain theta (θ) notation with example. (1 + 2)
2. Explain circular queue. What are the benefits of using circular queue over linear queue? (1 + 2)
3. Explain tail recursion with suitable example. (3)
4. How do you implement queue using linked list? (3)
5. Explain binary search algorithm with example. (3)

Group B (5×6=30)

6. Explain algorithm to evaluate the value of postfix expression using stack with suitable example. (6)

OR

Compare linear queue with circular queue. How do you implement enqueue and dequeue operations in queue? Explain. (2 + 4)

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

OR

Explain shell sort. Trace the shell sort algorithm with the array of numbers 34, 13, 16, 13, 45, 71, 21, 15, 8, and 1. (2 + 4)

8. Starting with empty binary search tree, show the effect of successively adding the following numbers as keys: 20, 23, 10, 21, 30, 15, 5, 22 and 40. Also, traverse this tree in preorder and postorder. (4 + 2)
9. Explain quadratic probing. Suppose, the set of keys is {7, 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)
10. Define spanning tree and minimum spanning tree. Explain Kruskal's algorithm to find minimum spanning tree with suitable example. (2 + 4)

Tribhuvan University

Institute of Science and Technology

Final Examination 2078

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 in R Notebook**. R scripts must be knitted as HTML or PDF with the outputs/interpretation of question number 6-10 and it must be saved in'a folder with the R notebook and knitted HTML/PDF file with your exam roll number for grading.*

Attempt ALL Questions.

Group A **[5 × 3 = 15]**

1. Describe the following concepts with focus on R software:
   a) Loops
   b) Function
   c) Pipe
2. Explain following concepts with examples focusing on R software:
   a) Big data
   b) Data wrangling
   c) Tidy data
3. Explain the following concept with examples focusing on R software:
   a) Measures of central tendency
   b) Measures of dispersion
   c) Measures of relative position
4. Explain the following concepts with examples focusing on R software:
   a) Correlation
   b) Parametric tests
   c) Non-parametric tests
5. Compare following model with focus on R software:
   a) Naïve Bayes and Support Vector Machine
   b) Decision Tree and Random Forest
   c) Feed-forward and feed-backward neural network

Group B [5 × 6 = 30]

6. Do the following in R Studio with R script so that it can be knitted as PDF:
   a) Prepare a column vector of miles per gallon (mpg) variable with random range between 10 to 50 of 500 values, **do not forget to use your exam roll number as random seed to replicate the result**
   a) Plot histogram of this "mpg" variable and interpret it carefully
   b) Refine the histogram by filling the bars with "blue" color and changing number of bins to 8
   c) Add a vertical abline at the arithmetic mean of the mpg variable
   d) Plot Q-Q plot of mpg variable, add normal Q-Q line of red color on it and interpret it carefully
   e) Plot density plot of mpg variable without the border, fill it with yellow color and interpret it

OR

Use the "ggplot2" package and do as follow in R studio:
   a) Define first layer of the ggplot object with diamond data, carat as x-axis and price as y-axis
   b) Add layer with geometric aesthetic as "point", statistics and position as "identity"
   c) Add layers with scale of y and x variables as continuous
   d) Add layer with coordinate system as Cartesian
   e) Add layer with appropriate title and interpret the resulting graph carefully

7. Do the following in R Studio with R script so that it can be knitted as PDF:
   a) Prepare a data with 100 random observations and two variables: miles per gallon (mpg) with random range between 10 to 50 and transmission gears (gear) as random binary variable (3=3 gear, 4=four gear and 5=five gears), **do not forget to use your class roll number as random seed to replicate the result**
   b) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if it follows normal distribution or not
   c) Perform goodness-of-fit test on miles per gallon (mpg) variable to check if the variances of mpg are equal or not on gears variable categories
   d) Perform the best 1-way analysis of variance test based on goodness-of-fit results with justification.
   e) Can you use this test for this data? Interpret the result carefully, if applicable.

8. Do the followings in R Studio using R script so that it can be knitted as PDF:
   b) Prepare a data with 200 random observations and four variables: miles per gallon (mpg) with random range between 10 to 50; transmission (am) as random binary variable (0=automatic, 1=Manual), weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   c) Divide this data into train and test datasets with 70:30 random splits **with your exam roll number as random seed for replication**
   d) Fit a supervised linear regression model for the train data
   e) Explain the model fit and BLUE coefficients for the fitted model
   f) Predict the mpg variable in the test data, get fit indices and interpret them carefully

9. Do the following in R Studio with R script so that it can be knitted as PDF:
   a) Prepare a data with four random variables and 300 observations: miles per gallon (mpg) with random range between 10 to 50; transmission (am) as random binary variable (0=automatic, 1=Manual), weight (wt) with random range of 1 to 10 and horse power (hp) with random range of 125 and 400, **do not forget to use your exam roll number as random seed to replicate the result**
   b) Divide this data into train and test datasets with 80:20 random splits **with your exam roll number as random seed for replication**
   c) Fit a supervised logistic regression model on train data with transmission (am) as dependent variable and miles per gallon (mpg), horse power (hp) and weight (wt) as independent variable
   d) Predict the transmission variable in the test data and interpret the predicted result carefully
   e) Get the confusion matrix, sensitivity, specificity of the predicted model and interpret them carefully

10. Do as follows using "mtcars" dataset in R studio with R script so that it can be knitted as PDF:
   a) Check the head and the structure of the dataset
   b) Create a "cars scale" using the Principal Component Analysis (PCA) model based on nine numerical variables with centering and scaling of the variables
   c) Based on the PCA summary result, how many components must be extracted? Why?
   d) Get the bi-plot of the fitted model and interpret it carefully
   e) Improve the fitted model with VARIMAX process and interpret the results carefully

OR

Do as follows using "USArrests" dataset in R studio with R script so that it can be knitted as PDF:

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

***

MDS504-2078

Tribhuvan University

Institute of Science and Technology

2078

Master Level / I Year /First Semester/ Science

Data Science (MDS 504)

(Mathematics for Data Science)

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

Attempt All Questions. Write your answer in detail as far as possible.

Group A

(3×5=15)

1. Show that
   a. The line $x_2 = ax_1$ is a subspace $\mathbb{R}^2$.
   b. The set of points that is the union of two lines through the origin is not a subspace.
2. Let

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},\ v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}.$$

   Show that $B = \{v_1, v_2\}$ is an orthonormal basis for $\mathbb{R}^2$. Find a vector $x \in \mathbb{R}^2$ with respect to the basis B.
3. Without calculation, find one eigenvalue and two linearly independent eigenvectors of

$$A = \begin{pmatrix}4 & 4 & -4\\4 & 4 & -4\\4 & 4 & -4\end{pmatrix}.$$

   Justify your answer.
4. Let $Q(x) = 3x_1^2 + 9x_2^2 + 8x_1x_2$. Find (a) the maximum value of $Q(x)$ subject to the constraint $x^Tx = 1$, (b) a unit vector u where this maximum is attained, and (c) the maximum of $Q(x)$ subject to the constraints $x^Tx = 1$ and $x^Tu = 0$.
5. Describe and compare the solution sets of $x_1 + 9x_2 - 4x_3 = 0$ and $x_1 + 9x_2 - 4x_3 = 2$

Group B

(6×5=30)

6. 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 $\begin{pmatrix}3\\2\end{pmatrix}$ in terms of the vectors u and v.
   b. Show that the vectors u and v span $\mathbb{R}^2$.

7. Let $u_1 = (2\ 0\ 0)^T, u_2 = (0\ 1\ 1)^T$ and $u_3 = (0\ 1\ -1)^T$. Find the orthonormal set associated with the set $S = \{u_1, u_2, u_3\}$.

   Prove that an orthogonal set of nonzero vectors in a vector space is linearly independent.

8. consider the matrix : $A = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$. What can you say about the action of A on an arbitary vector? What are examples of eigenvalues/eigenvectors of this matrix? What does this discussion for this example illustrate?

   OR

   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 $A = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T$. (1)

9. Prove that if B is a symmetric bilinear function on $\mathbb{R}^n$, then it is of the form $B = B_A(v, w) = v^T A w$, for some unique symmetric matrix A.

   OR

   Express the quadratic form $Q(x) = x_1 x_2 - x_1 x_3 + x_2 x_3$ as a sum of squares.

10. Find the SVD of $A = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$. If A an invertible $n \times n$ matrix, what is the relationship between the singular values of A and A^{-1}?

Tribhuvan University
Institute of Science and Technology
2078

✡

Master Level / I Year /First Semester/ Science
Data Science (MDS 505)
(Data Base Management Systems)

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. Differentiate between DDL, DML, and DCL.
2. What is Embedded SQL? Provide an example.
3. Explain Armstrong's axioms of functional dependencies.
4. Highlight on the desirable properties of transactions.
5. Briefly explain spatial database.

Group B **(5×6=30)**

6. Draw ER diagram for online voting system. Make necessary assumptions.
7. Explain B-tree Index files.
8. Perform Normalization on the following relation.

ID	CODE	SALARY	PROJECT MANAGER	PLACE
1	2	10000	P1	KTM
2	1	20000	P4	BKT
3	2	30000	P3	BKT
1	4	10000	P2	PTN

9. Explain the concept of recoverable, cascade less, and strict schedule.

OR

Which of the following schedule is (conflict) serializable? For each serializable schedule, determine the equivalent serial schedules.

i) r1(x):w1(x):r2(y):w2(y)
ii) r1(y):r2(y):r3(y):w2(y):w1(y):w3(y)

10. Explain different types of RAID.

OR

Differentiate between classification and clustering. Explain any one clustering algorithm.