M.TECH/CSE/2ND SEM/CSEN 5231/2022

DATA PREPROCESSING AND ANALYSIS
(CSEN 5231)

Time Allotted : 3 hrs **Full Marks : 70**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

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

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**

(i) A statistical measure that indicates the extent to which changes in one factor are accompanied by changes in another
(a) Standard deviation (b) Correlation coefficient
(c) Quartile deviation (d) Range.

(ii) Which method shows hierarchical data in a nested format?
(a) Treemaps (b) Scatter Plots
(c) Area charts (d) Population pyramids.

(iii) Which of the following is an example of raw data?
(a) Original swath files generated from a sonar system
(b) Initial time-series file of temperature values
(c) A real-time GPS-encoded navigation file
(d) All of the mentioned.

(iv) K-Means clustering suffers from which deficiency?
(a) Hyper binding (b) Chaining
(c) Inversion (d) Collision.

(v) Which of the following hierarchical clustering is not monotonic?
(a) Group Average (b) Single Link
(c) Centroid (d) Complete Link.

(vi) Which of the following statement is incorrect about the hierarchal clustering?
(a) The hierarchical type of clustering is also known as the HCA
(b) In general the splits and the merges both are determined in a greedy manner
(c) The choice of an appropriate metric can influence the shape of the cluster
(d) All of the above.

(vii) Which one of the following can be defined as the data object which does not comply with the general behaviour of model or available data?
(a) Outliner Analysis (b) Evaluation Analysis
(c) Classification (d) Prediction.

(viii) Which of the following statistic for a test data set may be lie outside the actual data set?
(a) Median (b) Mean (c) Mode (d) Maxima.

(ix) Which of these data structures is most apt to store and arrange spatial data?
(a) Array (b) Linked List (c) Binary Tree (d) Quad Tree.

(x) In order to understand the classroom teaching-learning process, which of the following research tool is most appropriate?
(a) Rating scale (b) Observation schedule
(c) Questionnaire (d) Interview schedule.

Group- B

2. (a) Can a relational database scale horizontally? Explain with suitable examples. [(CO1)(Acquire/LOCQ)]
(b) Can XML documents be queried like relational databases, and if so, how fast and convenient is the process expected to be? [(CO1)(Acquire/LOCQ)]
(c) '*Unstructured data can come from Facebook, Twitter and Presentations*' - Justify the statement. [(CO1)(Acquire/LOCQ)]
4 + 4 + 4 = 12

3. (a) Illustrate some of the areas where EDA is deployed. [(CO2)(Remember/LOCQ)]
(b) When do you require to scale the database? What are the different Scaling methods? [(CO3)(Understand/IOCQ)]
(c) What is real time analysis? Explain in brief. [(CO4)(Analyse/HOCQ)]
4 + (2 + 3) + 3 = 12

Group - C

4. (a) Table - 1 of patient records:

Name	Age	Gender	Blood Group	Weight (kg)	Height (m)	Temp.(°F)	Diabetes
P. Das	35	Female	A+	50	1.52	98.7	0
M. Roy	52	Male	O	115	1.77	98.5	1
A. Sen	45	Male	B+	96	1.83	98.8	0
S. Giri	79	Female	A-	41	1.55	98.6	0
R. Paul	24	Male	O	79	1.82	98.7	0
D. Rai	43	Male	O	109	1.89	98.9	1
T. Mitra	68	Male	A+	73	1.76	99.0	0
N. Roy	77	Female	B-	104	1.71	98.3	1
L. Das	45	Female	A-	64	1.74	98.6	0
R. Pal	28	Female	A+	136	1.78	98.7	1

Assign the following variables from the table to one of the following categories: Constant, Dichotomous, Binary, Discrete and Continuous.
(i) Name, (ii) Age, (iii) Gender, (iv) Blood Group, (v) Weight, (vi) Height
[(CO4)(Analyze/IOCQ)]

(b) Assign the following variables from the table to one of the scales given as Nominal, Ordinal, Interval, Ratio.
(i) Name, (ii) Weight, (iii) Temperature, (iv) Diabetes. [(CO4)(Analyze/IOCQ)]

6 + 6 = 12

5. (a) Explain discrimination and aggregation with suitable examples. [(CO4)(Understand/LOCQ)]
(b) How are Data Quality problems solved? [(CO2)(Understand/IOCQ)]
(c) Describe with an example the central tendency of EDA. [(CO1)(Analyze/HOCQ)]

4 + 5 + 3 = 12

Group - D

6. A premium golf ball production line must produce all of its balls to 1.615 ounces in order to get the top rating (and therefore the top dollar). Samples are drawn hourly and checked. If the production line gets out of sync with a statistical significance of more than 1%, it must be shut down and repaired.
A sample of 18 balls has a mean of 1.611 ounces and a standard deviation of 0.065 ounces.
(i) State the Null and Alternate Hypotheses.
(ii) Draw the Test Diagram and assign the correct Reject and/or Fail to Reject regions.
(iii) What could be the Type-1 and Type-2 errors in this case? Determine the Degree of Freedom
(iv) Calculate the Test Statistic. State which statistic you chose and why.
(v) Conclude and summarize whether you would shut down giving proper argument.
[(CO3)(Apply/IOCQ)]

12

7. (a) What is null hypothesis? [(CO1)(Remember/LOCQ)]
(b) What does z-score represent? Illustrate an example to calculate z-score. [(CO3)(Understand/IOCQ)]
(c) Explain Chi-Square Test. [(CO1)(Analyze/HOCQ)]

4 + 5 + 3 = 12

Group - E

8. Explain the following concepts in your own words. You may use diagrams to illustrate your understanding. [(CO5)(Evaluate/HOCQ)]
(i) Spatial data and use of cartograms.
(ii) Usage of space and non-space filling methods, and node-link graphs.

(6 + 6) = 12

9. (a) Problems on designing effective visualization . [(CO1)(Remember/LOCQ)]

(b) Differentiate line phenomenon from point phenomenon and area phenomenon. [(CO2) (Understand/IOCQ)]

(c) Explain and analyze the time series graph of the given figure.

Fig-1

Year	Inflation Rate
1990	6.1
1991	3.1
1992	2.9
1993	2.7
1994	2.7
1995	2.5
1996	3.3
1997	1.7
1998	1.6

[(CO2)(Analyse/HOCQ)]

4 + 5 + 3 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	29	45	26

Course Outcome (CO):

After the completion of the course students will be able to

1. Acquire knowledge in a broad range of methods based on statistics and informatics for data preprocessing and analysis and tools for visualizing the main characteristics of data.
2. Understand the whole process line of gathering relevant data, preprocessing the data, performing exploratory analysis on the data and visualizing the implicit knowledge extracted from data.
3. Apply suitable methods for unveiling the underlying structure of the data, testing underlying assumptions in various fields.
4. Analyze the results of experiment with the help of various visualization tools and statistical tests.
5. Evaluate the performance of not only a computational method after obtaining different results by using different parameter values in order to choose the correct parameter value, but also, all similar methods in order to find out the best performing algorithm for a dataset.
6. Get familiar with relevant literatures, derive theoretical properties of the existing methods and come up with novel approach or pipeline for analyzing data across various fields by solving assignment problems

*LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question