Printed Page:- Subject Code:- ACSBS0301

Roll. No:

□□□□□□□□□□□□□□□

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech.

SEM: III - THEORY EXAMINATION (2021 - 2022)

Subject: Computational Statistics

Time: 03:00 Hours Max. Marks: 100

General Instructions:

1. All questions are compulsory. It comprises of three Sections A, B and C.

- Section A - Question No- 1 is objective type question carrying 1 mark each & Question No- 2 is very short type questions carrying 2 marks each.
- Section B - Question No- 3 is Long answer type - I questions carrying 6 marks each.
- Section C - Question No- 4 to 8 are Long answer type - II questions carrying 10 marks each.
- No sheet should be left blank. Any written material after a Blank sheet will not be evaluated/checked.

SECTION A 20

1. Attempt all parts:-

1-a. If X follows N_p (μ, Σ) , then Squared Mahalanobis distance follows to which distribution? (CO1) 1

1. Normal Distribution
2. Multivariate Normal distribution
3. Chi-square distribution
4. None of these

1-b. The two basic groups of multivariate techniques are: (CO1) 1

1. dependence methods and interdependence methods
2. primary methods and secondary methods
3. simple methods and complex methods
4. None of these

1-c. Which of the following is not the objective of discriminant analysis? (CO2) 1

1. To find a linear combination of predictor variables that discriminate best between categories of dependent variable
2. To find statistical significance of discriminant function
3. To evaluate the accuracy of classification
4. All of the above statements are true

1-d. Multiple regression analysis is used when (CO2) 1

1. there is not enough data to carry out simple linear regression analysis.
2. the dependent variable depends on more than one independent variable.
3. one or more of the assumptions of simple linear regression are not correct.
4. the relationship between the dependent variable and the independent variables cannot be described by a linear function.

1-e. Which of the following techniques would perform better for reducing dimensions of a data set? (CO3) 1

1. Removing columns which have too many missing values
2. Removing columns which have high variance in data

3. Removing columns with dissimilar data trends
4. None of the above

1-f. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA? (CO3) 1

(a) PCA is an unsupervised method
(b) It searches for the directions that data have the largest variance
(c) Maximum number of principal components <= number of features
(d) All principal components are orthogonal to each other

1. a) and b)
2. a) and c)
3. a) ,b) ,c)
4. All of the above

1-g. Which of the following is not the part of the exploratory factor analysis process? (CO4) 1

1. Extracting factors
2. Determining the number of factors before the analysis
3. Rotating the factors
4. Refining and interpreting the factors

1-h. Which of the following can be used to determine how many factors to extract from a factor analysis: (CO4) 1

1. Scree plots
2. Eigen values and percentage of variance explained by each factor
3. Factor loadings
4. All of the above

1-i. What is the minimum number of variables/ features required to perform clustering? (CO5) 1

1. 0
2. 1
3. 2
4. 3

1-j. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering? (CO5) 1

(a) Single-link
(b) Complete-link
(c) Average-link

1. a) and b)
2. a) and c)
3. b) and c)
4. a), b) and c)

2. Attempt all parts:-

2-a. If X_1 and X_2 are independent with 2
find the value of c , such that X_2 and ($-X_1 + X_2 - X_3$) are independent. (CO1)

$$\mu = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & c \\ 1 & c & 2 \end{bmatrix}$$

2-b. Define Multicollinearity. (CO2) 2

2-c. What does a PCA do? (CO3) 2

2-d. What are the assumptions of factor analysis? (CO4) 2

2-e. What is a good cluster? (CO5) 2

SECTION B 30

3. Answer any <u>five</u> of the following:-

3-a. Write down the properties of Multivariate normal distribution? (CO1) 6

3-b. If X distributed as N3 (μ, Σ), where (CO1) 6

$$\mu = \begin{bmatrix} 5 \\ 3 \\ 7 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{bmatrix}$$

(a) Find the probability $P(X_1 > 6)$

(b) Find the $P(4X_1 - 3X_2 + 5X_3) < 80$

3-c. Define the term Wilks' Lambda and What is the difference between Metric and Non-metric Variables. (CO2) 6

3-d. What are the objectives of Linear Discriminant Analysis (LDA) and explain LDA technique with the model. (CO2) 6

3-e. What is the importance of using PCA before the clustering? (CO3) 6

3-f. Differentiate between EFA and CFA. (CO4) 6

3-g. Cluster the following eight points (with (x, y) representing locations) into three clusters using K means/K medoid (CO5) 6

A_1(2, 10), A_2(2, 5), A_3(8, 4), A_4(5, 8), A_5(7, 5), A_6(6, 4), A_7(1, 2), A_8(4, 9).

SECTION C 50

4. Answer any <u>one</u> of the following:-

4-a. Define Multivariate analysis and How does multivariate analysis differ from univariate and bivariate analysis? What is the process of conducting multivariate analysis? (CO1) 10

4-b. What is Squared Mahalanobis distance in multivariate normal distribution? (CO1) 10

If X distributed as $N_4(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 1 & -2 & 3 & 5 \\ -2 & 9 & -3 & 6 \\ 3 & -3 & 2 & 3 \\ 5 & 6 & 3 & 9 \end{bmatrix}$$

(a) Find the distribution of $Z = 4Y_1 - 2Y_2 + Y_3 - 3Y_4$.

(b) Find the distribution of $Z_1 = Y_1 + Y_2 + Y_3 + Y4$ and $Z_2 = -2Y_1 + 3Y_2 + Y_3 - 2Y_4$

5. Answer any <u>one</u> of the following:-

5-a. A farmer applies three types of fertilizers on 4 separate plots. the figure on yield per acre are tabulated below: (CO2) 10

Fertilizers	Yield			
	A	B	C	D
Nitrogen	6	4	8	6
Potash	7	6	6	9
Phosphates	8	5	10	9

Perform the analysis of variance and comment. Here, given that $F_{(2,6)}$ = 5.14, F $_{(3,6)}$= 4.76 at 5% level of significance.

5-b. What are the assumptions of discriminant analysis and the applications of discriminant analysis. (CO2) 10

6. Answer any <u>one</u> of the following:-

6-a. What is the importance of using PCA before the clustering? Choose the most complete answer. 10

6-b. Explain all the steps of a PCA algorithm. (CO3) 10

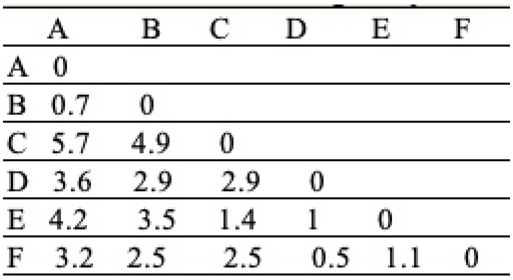
7. Answer any <u>one</u> of the following:-

7-a. Explain the curse of dimensionality. Also, elaborate the factor analysis model. (CO4) 10

7-b. Explain the following terms: (CO4) 10
a) Observed variable
b) Latent variables.
c) Communality
d) Factor Loading
e) Score Matrix

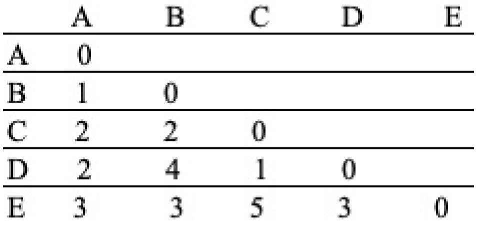
8. Answer any <u>one</u> of the following:-

8-a. Perform DBSCAN on the given problem with €=2and minimum point=3. (CO5) 10



	A	B	C	D	E	F
A	0					
B	0.7	0				
C	5.7	4.9	0			
D	3.6	2.9	2.9	0		
E	4.2	3.5	1.4	1	0	
F	3.2	2.5	2.5	0.5	1.1	0

8-b. Find the clusters and draw the dendrogram using Hierarchical Clustering (Divisive approach). (CO5) 10



	A	B	C	D	E
A	0				
B	1	0			
C	2	2	0		
D	2	4	1	0	
E	3	3	5	3	0