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 (2022 - 2023)

Subject: Computational Statistics

Time: 3 Hours Max. Marks: 100

General Instructions:

IMP: Verify that you have received the question paper with the correct course, code, branch etc.

1. This Question paper comprises of three Sections -A, B, & C. It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.

2. Maximum marks for each question are indicated on right -hand side of each question.

3. Illustrate your answers with neat sketches wherever necessary.

4. Assume suitable data if necessary.

5. Preferably, write the answers in sequential order.

6. 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 What is the best description of a point estimate? (CO1) 1

(a) any value from the sample used to estimate a parameter

(b) a sample statistic used to estimate a parameter

(c) the margin of error used to estimate a parameter

(d) All of the above

1 All of the following are examples of dependence methods of analysis EXCEPT? (CO1) 1

(a) multiple regression analysis

(b) multiple discriminant analysis

(c) multivariate analysis of variance

(d) cluster analysis

1-c. In multiple discriminant analysis, the dependent variable must be ________ ,while in Multiple regression analysis,the dependent variable must be ________ . ? (CO2) 1

(a) nominal, nominal

(b) nominal , metric

(c) metric, metric

(d) none of these

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

(a) there is not enough data to carry out simple linear regression analysis.

(b) the dependent variable depends on more than one independent variable.

(c) one or more of the assumptions of simple linear regression are not correct.

(d) the relationship between the dependent variable and the independent variables cannot be described by a linear function.

1-e. Imagine, you have 1000 input features and 1 target feature in a machine learning problem. You have to select 100 most important features based on the relationship between input features and the target features. Do you think, this is an example of dimensionality reduction? (CO3) 1

(a) Yes

(b) No

(c) None of the above

(d) Cant determined

1-f. For the projected data you just obtained projections ($(-\sqrt{2})$, (0), $(\sqrt{2})$). Now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, What is the reconstruction error? (CO3) 1

(a) 0

(b) 0.1

(c) 0.3

(d) 0.4

1 What will a factor loading in an orthogonal solution represent? (CO4) 1

(a) Correlation

(b) Partial correlation

(c) Multiple correlation

(d) Eigenvalue

1 Which of the following is not a typical model fit index used in SEM? (CO4) 1

(a) Root mean squared error of approximation (RMSEA)

(b) Adjusted R-square

(c) Comparative fit index (CFI)

(d) Tucker-Lewis index (TLI)

1 What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithm for the same dataset? (CO5) 1

(a) Proximity function used

(b) of data points used

(c) of variables used

(d) All of the above

1 Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: 1

C1: {(2,2), (4,4), (6,6)}

C2: {(0,4), (4,0)}

C3: {(5,5), (9,9)}

What will be the cluster centroids if you want to proceed for second iteration? (CO5)

(a) C1: (4,4), C2: (2,2), C3: (7,7)

(b) C1: (6,6), C2: (4,4), C3: (9,9)

(c) C1: (2,2), C2: (0,0), C3: (5,5)

(d) None of these

2. Attempt all parts:-

2.a. Given a normal distribution with $\mu=150$ and $\sigma=10$, find the following probabilities. P(150<x) ? (CO1) 2

2.b. What is the model used for one way classification of ANOVA? (CO2) 2

2.c. What are the properties of Principal Components in PCA ? (CO3) 2

2.d. What does rotation do in factor analysis ? (CO4) 2

2.e. Why do you prefer Euclidean distance over Manhattan distance in the K means Algorithm ? (CO5) 2

SECTION B 30

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

3 The mean weight of 500 male students at a certain college is 151 lbs and the standard deviation is 15 lbs. Assuming that the weights are normally distributed. Find how many students weight i) between 120 lbs and 155lbs. ii) more than 155 lbs. ? (CO1) 6

3 If X distributed as $N_3(\mu,\Sigma)$, where $\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$ 6

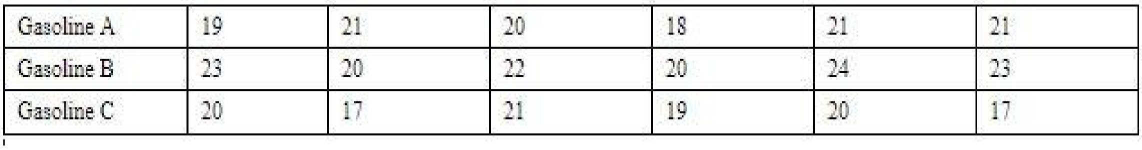
$\Sigma = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{bmatrix}$

Check whether i)(X_1 , X_3) and X_2 are independent or not.

ii) X_1 and ($X_1 + 3X_2 - 2X_3$) are independent or not.? (CO1)

3 Define canonical correlation and coefficient of determination.? (CO2) 6

3 In order to compare the mileage yields of 3 kinds of gasoline several tests were run and the following results were obtained (each figure represents the no. Of miles obtained with a gallon of the respective gasoline) 6



Gasoline A	19	21	20	18	21	21
Gasoline B	23	20	22	20	24	23
Gasoline C	20	17	21	19	20	17

Calculate F and assuming that the necessary assumptions can be met, test at a level of significance of 0.05 ? (CO2)

3.e. What are the main advantages and disadvantages of PCA transformation ? (CO3) 6

3.f. Explain factor structure. Discuss the benefits of Factor analysis ? (CO4) 6

3.g. Cluster the following eight points (with (x, y) representing locations) into three clusters using K means A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9) .Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2) ? (CO5) 6

SECTION C 50

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

4 A) If X distributed as $N_3(\mu,\Sigma)$, where 10

$\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$

$\Sigma = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{bmatrix}$

Check whether

i) X_1 and ($X_1 + 3X_2 - 2X_3$) are independent or not.

ii) ($X_1 + X_2$) and ($X_3 - X_1$) are independent or not.

B) If X and Y are not independent , then what is the cov(y/x) for conditional distribution. ?

(CO1)

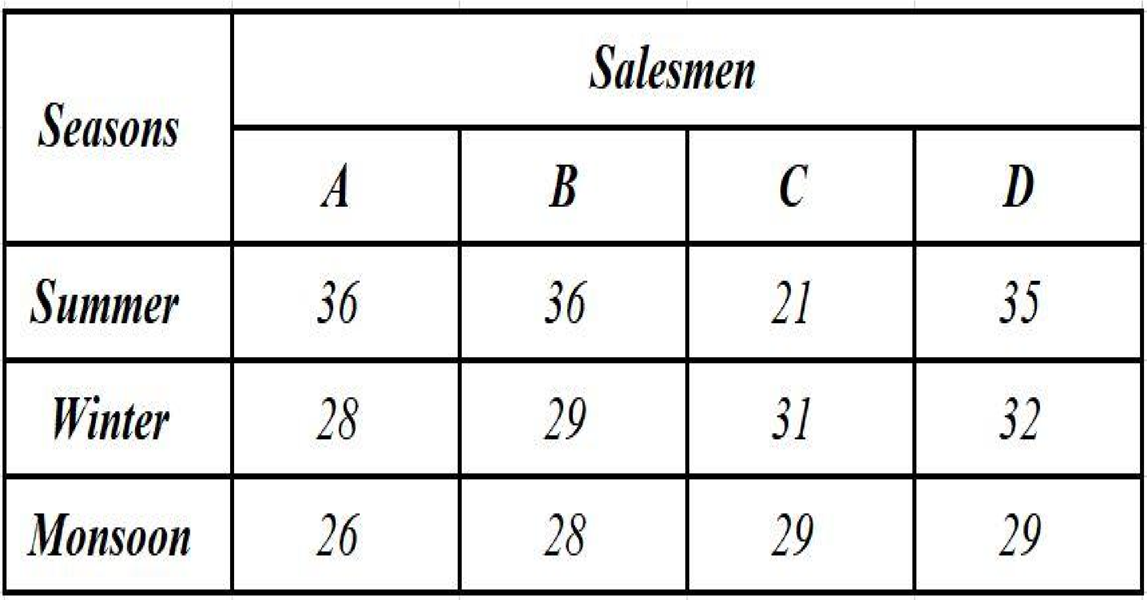
4 A company ships 5000 cell phones. They are expected to last an average of 10,000 hours before needing repair; with a standard deviation of 500 hours. Assume the survival time of the phones are normally distributed. If a phone is randomly selected to be tracked for repairs find the expected number that needs repair, 10

a) after 11,000 hours

b) before 9500 hours ? (CO1)

5. Answer any one of the following:-

5 A company appoints four salesmen A, B, C and D and observes their sales in three seasons: summer, winter and monsoon. The figures (in lakhs) are given in the following table: 10



Seasons	Salesmen			
	A	B	C	D
Summer	36	36	21	35
Winter	28	29	31	32
Monsoon	26	28	29	29

Carry out an analysis of variance. ? (CO2)

5 What is the difference between linear regression and logistic regression. Also give the practical utility of logistic regression. ? (CO2) 10

6. Answer any one of the following:-

6 Given data = { 2, 3, 4, 5, 6, 7 ; 1, 5, 3, 6, 7, 8 }. 10

Compute the principal component using PCA Algorithm. ? (CO3)

6 Describe the feature selection , feature extraction .how they are related to dimensionality reduction ? (CO3) 10

7. Answer any one of the following:-

7 Differentiate between Exploratory Factor analysis and Confirmatory Factor analysis ? (CO4) 10

7 Define 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 Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: 10
1)C1:{(2,2),(4,4),(6,6)}
2)C2:{(0,4),(6,6)} 3)
C3:{(5,5) , (9,9)}
What will be the cluster centroids if you want to proceed for second iteration. ? (CO5)

8 Discuss the following : 1)Correlations and distances 2) Clustering Profiling 3) Interpreting clusters ? (CO5) 10