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) (ONLINE)

Subject: Computational Statistics

Time: 02:00 Hours Max. Marks: 100

General Instructions:

1. *All questions are compulsory. It comprises of two Sections A and B.*

- *Section A - Question No- 1 has 35 objective type questions carrying 2 marks each.*
- *Section B - Question No- 2 has 12 subjective type questions carrying 3 marks each. You have to attempt any 10 out of 12 question.*
- *No sheet should be left blank. Any written material after a Blank sheet will not be evaluated/checked.*

SECTION A 35 x 2 = 70

1. Attempt ALL parts:-

1.1.a The shape of the normal curve depends on its 1

(a) Mean deviation

(b) Quartile deviation

(c) Standard deviation

(d) correlation

1.1.b A estimator which provides all the information provided by a sample with respect to the parameter is called 1

(a) unbiased

(b) consistency

(c) efficiency

(d) sufficiency

1.1.c Multiple regression analysis is used when 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) the relationship between the dependent variable and the independent variables cannot be described by a linear function.

(d) None of these

1.1.d T_n be an estimator of θ. If $T_n(P) \rightarrow \theta$, then 1

(a) T_n is a sufficient estimator of θ

(b) T_n is an unbiased estimator of θ

(c) T_n is a consistent estimator of θ

(d) T_n is an efficient estimator of θ

1.1.e Let $X_1, \ldots, X_{60}$ be a random sample of size 60 from a four-variate normal distribution having mean μ and covariance Σ, then The distribution of $\bar{X}$ is 1

(a) $\bar{X}$ is distributed as $N_4(\mu, \frac{1}{60}\Sigma)$

(b) $\bar{X}$ is distributed as $N_4(\mu, \Sigma)$

(c) $\bar{X}$ is distributed as $N_4(2\mu, \Sigma)$

(d) None of these

1.1.f Which type of analysis involves three or more variables? 1

(a) univariate statistical analysis

(b) Bivariate statistical analysis

(c) Multivariate statistical analysis

(d) All of the above

1.1.g The two basic groups of multivariate techniques are: (CO1) 1

(a) dependence methods and interdependence methods

(b) primary methods and secondary methods

(c) simple methods and complex methods

(d) None of these

1.2.a The error deviations within the SSE statistic measure distances: 1

(a) within groups

(b) between groups

(c) both (a) and (b)

(d) none of the above

1.2.b Analysis of variance is a statistical method of comparing the _____ of several populations. 1

(a) Standard deviations

(b) variances

(c) Means

(d) Proportions

1.2.c In the discriminant analysis, the dependent variable is 1

(a) Interval scale

(b) Ratio scale

(c) Nominal scale

(d) All are true

1.2.d Before, the results of discriminant function are interpreted, on must examine the 1

(a) The problem of multicollinearity among predictor variables

(b) Significance of predictor variable

(c) The significance of discriminant function

(d) None of the above

1.2.e Linear Discriminant Analysis is 1

(a) Unsupervised Learning

(b) supervised Learning

(c) Semi supervised Learning

(d) None of the above

1.2.f The coefficient of determination is 1

(a) r

(b) r^2

(c) z

(d) None of these

1.2.g The goal of discriminant analysis is 1

(a) to develop a model to predict new dependent values

(b) the develop a rule for predicting to what group a new observation is most likely to belong

(c) to develop a rule for predicting how independent variable values predict dependent values.

(d) none of these

1.3.a 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. 1

Do you think, this is an example of dimensionality reduction?

(a) Yes

(b) No

1.3.b It is not necessary to have a target variable for applying dimensionality reduction algorithms. 1

(a) TRUE

(b) FALSE

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

(a) Removing columns which have too many missing values

(b) Removing columns which have high variance in data

(c) Removing columns with dissimilar data trends

(d) None of the above

1.3.d 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

(a) a) and b)

(b) a) and c)

(c) a) ,b) ,c)

(d) All of the above

1.3.e What will happen when eigenvalues are roughly equal? 1

(a) PCA will perform outstandingly

(b) PCA will perform badly

(c) Can’t Say

(d) None of these

1.3.f PCA works better if there is? 1

1.A linear structure in the data
2.If the data lies on a curved surface and not on a flat surface
3. If variables are scaled in the same unit

(a) 1 and 2

(b) 1 and 3

(c) 2 and 3

(d) 1,2 &3

1.3.g PCA is technique for ____ 1

(a) feature extraction

(b) variance normalisation

(c) data augmentation

(d) dimentionality reduction

1.4.a Which of the following is not a typical structural equation model? 1

(a) Confirmatory factor analysis
(b) Latent path analysis
(c) Latent mean analysis
(d) Exploratory factor analysis

1.4.b Which of the following statements is wrong? 1

(a) Confirmatory factor analysis (CFA) is a type of SEM
(b) In CFA measurement error of indicators is removed during the estimation
(c) Both in EFA and CFA we specify the pattern of indicator-factor loadings
(d) CFA belongs to the common factor model family

1.4.c Which of the following criteria cannot be used to determine the number of factors in an EFA? 1

(a) Asking a group of researchers before the analysis
(b) Eigenvalue rule
(c) Scree test
(d) Parallel analysis

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

(a) Extracting factors
(b) Determining the number of factors before the analysis
(c) Rotating the factors
(d) Refining and interpreting the factors

1.4.e To determine which variables relate to which factors, a researcher would use: 1

(a) Factor loadings
(b) Communalities
(c) Eigen values
(d) Beta coefficients

1.4.f If a researcher wants to determine the amount of variance in the original variables that is associated with a factor, s/he would use: 1

(a) Factor loadings
(b) Communalities
(c) Eigen values
(d) Beta coefficients

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

(a) Scree plots
(b) Eigen values and percentage of variance explained by each factor
(c) Factor loadings
(d) All of the above

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

(a) 0
(b) 1
(c) 2
(d) 3

1.5.b For two runs of K-Mean clustering is it expected to get same clustering results? 1

(a) Yes
(b) No

1.5.c Which of the following algorithm is most sensitive to outliers? 1

(a) K-means clustering algorithm
(b) K-medians clustering algorithm
(c) K-modes clustering algorithm
(d) K-medoids clustering algorithm

1.5.d In which of the following cases will K-Means clustering fail to give good results? 1

Data points with outliers
Data points with different densities
Data points with round shapes
Data points with non-convex shapes

(a) 1 and 2
(b) 2 and 3
(c) 1, 2 and 4
(d) 1, 2, 3 and 4

1.5.e 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

(a) a) and b)
(b) a) and c)
(c) b) and c)
(d) a), b) and c)

1.5.f Which of the following is finally produced by Hierarchical Clustering? 1

(a) final estimate of cluster centroids
(b) tree showing how close things are to each other
(c) assignment of each point to clusters
(d) all of the mentioned

1.5.g Which of the following clustering requires merging approach? 1

(a) Partitional
(b) Hierarchical
(c) Naive Bayes
(d) None of the mentioned

SECTION B 10 X 3 = 30

2. Answer any TEN of the following:-

2.1.a What is the Moment generating function for Multivariate normal distribution? (CO1) 2

2.1.b Let X follows $N_3(\mu, \sum)$ with (CO1) 2

$$\sum = \begin{bmatrix} 1 & c & 0 \\ c & 1 & c \\ 0 & c & 1 \end{bmatrix}$$

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

2.2.a What is Wilks' Lambda? (CO1) 2

2.2.b what is Logistic Regression? (CO1) 2

2.2.c 2

what is the difference between simple and multiple discriminant analyses? (Co2)

2.3.a	Comment whether PCA can be used to reduce the dimensionality of the non-linear dataset.	2
2.3.b	What will happen when eigenvalues are roughly equal?	2
2.3.c	What are the properties of Principal Components in PCA?	2
2.4.a	How do Researchers Decide Whether to Use PCA or EFA?	2
2.4.b	What are the assumptions of factor analysis? (CO4)	2
2.5.a	What do you mean by Cluster Analysis?	2
2.5.b	What do you understand by Direct Density Reachable in DBSCAN?	2