M.TECH/CSE/1ST SEM/CSEN 5141/2021

DATA SCIENCE
(CSEN 5141)

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) How do we perform Bayesian classification when some features are missing?
- (a) We assume the missing values as the mean of all values.
- (b) We ignore the missing features.
- (c) We integrate the posteriors probabilities over the missing features.
- (d) Drop the features completely.

(ii) Which one of the following statements is False in respect of KNN Algorithm?
- (a) For a very large value of K, points from other classes may be included in the neighbourhood.
- (b) For a very small value of K, the algorithm is very sensitive to noise.
- (c) KNN is used only for classification problem statements.
- (d) KNN is a lazy learner.

(iii) Which one of the following statements is True?
- (a) Outliers should be identified and removed always from a dataset.
- (b) Outliers can never be present in the testing dataset.
- (c) Outlier is a data point that is significantly close to other data points.
- (d) The nature of our business problem determines how outliers are used.

(iv) What kind of distance metric is suitable for categorical variables to find the closest neighbours?
- (a) Euclidean distance
- (b) Manhattan distance
- (c) Minkowski distance
- (d) Hamming distance.

(v) Which one of the following statements is False about Correlation and Covariance?
- (a) A zero correlation does not necessarily imply independence between variables
- (b) Correlation and covariance values are the same
- (c) Covariance and correlation values are always of the same sign
- (d) Correlation is the standardized version of Covariance.

(vi) Manhattan distance can be used for:
(a) continuous variables only
(b) categorical variables only
(c) both categorical as well as continuous variables
(d) neither categorical nor continuous variables

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

(viii) Which of the following steps is performed by data scientist after acquiring the data?
(a) Data Replication (b) Data Integration
(c) Data Cleansing (d) None of the above.

(ix) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively. Now, you have added 2 in all values of X (i.e. new values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 and D3 respectively. How do the values of D1, D2 and D3 relate to C1, C2 and C3?
(a) D1= C1, D2 < C2, D3 > C3 (b) D1 = C1, D2 = C2, D3 = C3
(c) D1> C1, D2 < C2, D3 =C3 (d) D1 < C1, D2 < C2, D3 < C3

(x) Which of the following is a classification problem?
(a) Predicting the gender of a person by his/her handwriting style
(b) Predicting house price based on area
(c) Predicting the number of copies, a music album will be sold next month
(d) None of the above.

Group- B

2. (a) What is the significance of studying Data Science? [(CO2) (Understand /LOCQ)]
(b) *'If we want better performance we can buy better hardware, unfortunately we cannot buy a more maintainable or reliable system'* – from the perspective of data integration, justify the statement. [(CO1) (Understand /LOCQ)]
(c) *What are* Apache Sqoop and Apache Flume excel? [(CO2)(Learn and understand/IOCQ)]

2 + 4 + 6 = 12

3. (a) Why *Cleansing, integrating, and transforming data* are important segments of data science process? [(CO2) (Understand /IOCQ)]
(b) What do you gain by studying Different Distributions? [(CO2)(Understand/LOCQ)]
(c) Name 4 types of Distribution and explain Binomial Distribution. [(CO1)(Analyze/IOCQ)]

6 + 2 + 4 = 12

Group - C

4. (a) Imagine that a researcher wanted to know the average weight of 5th-grade boys in a high school. He randomly sampled 5 boys from that high school. Their weights were 54.5 kg, 45 kg, 46 kg, 39.5 kg and 63.5 kg. What is the standard error of the mean? [(CO4)(Demonstrate/LOCQ)]

   (b) (i) Consider a binomial experiment. With regard to the standard deviation, illustrate how and why the larger the deviation number, the more difficult it would be to predict an outcome.

   (ii) A stock had returns of 5%, -21%, 16%, 12%, and -4% over the past five years. What is the standard deviation of these returns? [(CO1)(Analyze/IOCQ)]

6 + (4 + 2) = 12

5. Consider a Multiple-Choice Exam that contains 10 multiple-choice questions with 4 possible choices for each question, only one of which is correct.
Suppose a student is to select the answer for every question randomly. Let X be the number of questions the student answers correctly. Then, X has a binomial distribution with parameters n = 10 and p = 0.25. (Convince yourself that all assumptions for a binomial distribution are reasonable in this setting.) [(CO4) (Demonstrate /LOCQ)]
   (i) What is the probability for the student to get no answer correct?
   (ii) What is the probability for the student to get two answers correct?
   (iii) What is the probability for the student to fail the test (i.e., to have less than 6 correct answers)?

(3 × 4) =12

Group - D

6. (a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student? [(CO3) (Understand and apply /LOCQ)]

   (b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student? [(CO4) (Suggest /LOCQ)]

   (c) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke. [(CO3) (Understand and apply /IOCQ)]

5 + 2 + 5 = 12

7. (a) Cluster the following eight points (with (x, y) representing locations) into three clusters:
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 centres are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as:

P(a, b) = |x2 – x1| + |y2 – y1|

Use K-Means Algorithm to find the three cluster centres after the second iteration. [(CO1) (Analyze/HOCQ)]

(b) The sales of a company (in Crore Rupees) for each year are shown in the table below:

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

(i) Find the least square regression line y = a x + b

(ii) Use the least squares regression line as a model to estimate the sales of the company in 2012. [(CO2) (Understand/LOCQ)]

7 + (3 + 2) = 12

Group - E

8. (a) Name the three functionalities of visualization? [(CO5)(Develop/LOCQ)]

(b) Hierarchical clustering is a powerful technique that allows us to build tree structures from data similarities. Explain with Python code or Give a Case Study how can Dendrogram help in Hierarchical clustering? [(CO1)(Remember/HOCQ)]

(c) Why and when do we use Graph database? [(CO2)(Analyse/IOCQ)]

3 + 7 + 2 = 12

9. (a) One of the formal definitions of visualization says "... is the process of extracting salient *features* from the *sets of data* and *displaying* the features in an *intuitive* and *expressive* way? Comment on all the italicized terms in the definition. [(CO4) (Remember /LOCQ)]

(b) What are Mackinlay's design criteria? [(CO3) (Understand/LOCQ)]

(c) Give examples for each of the following: [(CO4) (Analyse/IOCQ)]

(i) Structural visualization

(ii) Temporal visualization

(iii) Geospatial visualization

(iv) Multidimensional visualization

5 + 3 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	51.04%	34.38%	14.58%

Course Outcome (CO):

After the completion of the course students will be able to

CO1: Explain how data is collected, managed and stored for data science;

CO2: Understand the key concepts in data science, including their real-world applications and some of the popular techniques used by data scientists;

CO3: Build skills in data management;
CO4: Demonstrate proficiency with statistical analysis of data;
CO5: Develop ability to build and assess data-based models;
CO6: Apply data science concepts and methods to solve real-world problems;

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

Department & Section	Submission Link
CSE	https://classroom.google.com/c/NDA1NzY4MzgxMjI3/a/MjI3ODg5MzExODU2/details