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) Data science is the process of diverse set of data through?
(a) Organizing data (b) Processing data
(c) Analyzing data (d) All of the above.

(ii) Point out the correct statement
(a) Raw data is original source of data
(b) Preprocessed data is original source of data
(c) Raw data is the data obtained after processing steps
(d) None of the mentioned.

(iii) Which of the following is performed by Data Scientist?
(a) Define the question (b) Create reproducible code
(c) Challenge results (d) All of the mentioned.

(iv) The value of the correlation coefficient lies in the range of
(a) $(-\infty, +\infty)$ (b) $[0, +\infty)$ (c) $[-1, +1]$ (d) $(-1, +1)$.

(v) The k-nearest neighbours (k-NN) algorithm is
(a) a supervised learning algorithm (b) an unsupervised learning algorithm
(c) Both (a) and (b) (d) None of the above.

(vi) Which of the following allows you to find the relationship you didn't know about?
(a) Inferential (b) Exploratory
(c) Causal (d) None of the mentioned.

(vii) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?
(a) K-NN (b) Random Forest
(c) Linear Regression (d) all of these.

(viii) Which of the following is the common goal of statistical modelling?
(a) Inference (b) Summarizing
(c) Sub-setting (d) None of the above.

(ix) Which of the following focuses on the discovery of (previously) unknown properties on the data?
(a) Data mining (b) Big Data
(c) Data wrangling (d) Machine Learning.

(x) Let $\mathbf{e}_1$ and $\mathbf{e}_2$ are the first and second principal component vectors, what statements are correct about them.
(a) $\mathbf{e}_1$ is parallel to $\mathbf{e}_2$ and variance along $\mathbf{e}_1$ is bigger than that along $\mathbf{e}_2$.
(b) $\mathbf{e}_1$ is orthogonal to $\mathbf{e}_2$ and variance along $\mathbf{e}_1$ is bigger than that along $\mathbf{e}_2$.
(c) $\mathbf{e}_1$ is parallel to $\mathbf{e}_2$ and variance along $\mathbf{e}_2$ is bigger than that along $\mathbf{e}_1$.
(d) $\mathbf{e}_1$ is orthogonal to $\mathbf{e}_2$ and variance along $\mathbf{e}_2$ is bigger than that along $\mathbf{e}_1$.

Group – B

2. (a) What are the differences between supervised and unsupervised learning?

(b) You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

(c) What are dimensionality reduction and its benefits?

(d) How should you maintain a deployed model?

(3 × 4) = 12

3. (a) Why data cleaning plays a vital role in analysis?

(b) Mention what are the various steps in an analytics project?

(c) List out some of the best practices for data cleaning?

(d) List out some common problems faced by data analyst?

(e) Mention the name of the framework developed by Apache for processing a large data set for an application in a distributed computing environment?

(f) Mention what are the missing patterns that are generally observed?

(2 × 6) = 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: 120 lbs, 99 lbs, 101 lbs, 87 lbs, 140 lbs. What is the standard error of the mean?

(b) Define and explain Central Limit Theorem.

6 + 6 = 12

5. (a) Consider a Multiple-Choice Examination that contains 10 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.
   (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)?

   (b) We want to predict the probability of heart attack in future based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case? Explain the method.

(2 × 3) + 6 = 12

Group – D

6. (a) One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric.
   (i) If object ***a*** is among the ***k-nearest*** **neighbors** of object ***b***, why is ***b*** not guaranteed to be among the ***k-nearest*** **neighbors** of ***a***?
   (ii) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

   (b) A study looks to model the relationship between the number of cookies left for Santa Claus and the number of presents received. Data collected over the last 5 Christmas is given in the table.

Year	**2015**	**2016**	**2017**	**2018**	**2019**
Cookies Left	2	4	2	6	8
Presents Received	1	1	4	5	5

   (i) Using the method of Least Squares, find the linear equation, describing the number of presents received as a function of the number of cookies left, that best fits the given data.
   (ii) Based on your linear model, if you want 7 presents this year, how many cookies should you leave Santa Claus?

(3 + 3) + 6 = 12

7. 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?
   (i) Given the above information is a randomly chosen college student more likely to be a graduate or undergraduate student?
   (ii) Repeat the same information assuming that the randomly chosen student is a smoker.

(iii) Suppose 30% of the graduate students live in dorm but only 10% of the undergraduate students live in 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.

(4 × 3) = 12

Group – E

8. (a) State how Mackinlay's Retinal Variables help us in finding out 'What We Do (and Don't) Know About Data Visualization'?

(b) Justify the assertion: "There's a story behind your numbers. Visualizing data brings them to life."

(c) Explain Informational Visualization?

(4 × 3) = 12

9. (a) What makes data visualisation good?

(b) Explain the difference between Bitmap and Pixmap?

(c) Give examples for any two of the following:
(i) Structural visualization, (ii) Temporal visualization, (iii) Geospatial visualization, (iv) Multidimensional visualization

3 + 3 + 6 = 12

Department & Section	Submission Link
CSE	https://classroom.google.com/c/MjIxODA3NzY3MzM4/a/MjkzOTI5NjQ5NzM0/details