M.TECH/CSE/1ST SEM/CSEN

DATA SCIENC

(CSEN 5141)

Time Allotted : 3 hrs

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and <u>any 5 (five)</u> from Group B to E, taking <u>at least one</u> 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) Point out the correct statement:
(a) The mean is a measure of central tendency of the data
(b) Empirical mean is related to "centering" the random variables
(c) The empirical standard deviation is a measure of spread
(d) all of the above.

(ii) In a fictitious state-wide database for road accidents, grouping of "accidents involving a truck" or "accidents at night" is an example of ________.
(a) classification (b) clustering
(c) decision table (d) decision tree.

(iii) Which of the following can act as possible termination conditions in K-means clustering algorithm?
(I) For a fixed number of iterations
(II) Assignment of observations to clusters does not change between iterations, except for cases with a bad local minimum
(III) Centroids do not change between successive iterations.
(a) I and II (b) I and III (c) II and III (d) I, II and III.

(iv) With respect to a typical 'Big Data Ecosystem', the technologies / tools that are **not** usually meant for 'Big Data Scientists' are ________________, ________________, and security.
(a) benchmarking, deployment (b) data integration, filesystem
(c) databases, scheduling (d) programming, databases.

(v) The two basic types of structured data are 'Numeric' and 'Categorical'; the former usually comes in two forms, 'Continuous' and ______________ while 'Binary' and ____________ are two useful cases for the latter.
(a) 'Discrete', 'Natural' (b) 'Discrete', 'Ordinal'
(c) 'Discontinuous', 'Primal' (d) 'Discontinuous', 'Secondary'.

Group – B

2. (a) Compare Data Science (DS) with:
(i) Machine Learning (ML) – mention any two points;
(ii) Databases (DB) – mention any two points;
(iii) Business Intelligence (BI) – mention any one point.

(iv) What should be the best choice for number of clusters based on the following results (as shown in Figure 1):

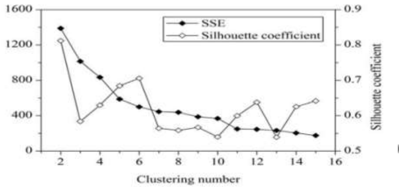


Figure 1

(a) 2 (b) 4 (c) 6 (d) 8

(vii) Which one of the following would be the most likely to produce selection bias in a survey?
(a) Using questions with biased wording
(b) Only receiving responses from half of the people in the sample
(c) Conducting interviews by telephone instead of in person
(d) Using a random sample of students at a university to estimate the proportion of people who think the legal drinking age should be lowered.

(viii) Variables X and Y (each with measured data) are said to be ____________ correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are ____________ correlated.
(a) negatively, low, negatively (b) negatively, high, positively
(c) positively, low, negatively (d) positively, high, negatively

(ix) Which of the following machine learning algorithm can be used for inputting missing values of type categorical and continuous?
(a) Linear regression (b) k-NN
(c) Logistic regression (d) all of the mentioned.

(x) Consider the following confusion matrix (as shown in table below). What is the value of recall?

	Predicted Yes	**Predicted No**
Actual Yes	100	5
Actual No	10	50

(a) 0.95 (b) 0.90 (c) 0.92 (d) 0.91.

(b) You are given a set of data about teenage boys and girls, aged between 16 and 19 years, residing in diverse localities within Kolkata, coming from various ethnic communities, studying in different schools under respective boards, and their respective overall percent marks in class-XII Board Examinations. You are expected to do some analysis on this data set to find out whether and how overall percent marks might be impacted by gender, ethnic community, residential locality, and school board. While carrying out the necessary 'data preparation' steps,

(i) mention any two typical issues you might face.

(ii) outline your possible approach(es) towards handling such issues with suitable examples.

(iii) highlight the relative advantages and disadvantages thereof. **(2 + 2 + 1) + (2 + 2 + 3) = 12**

3. (a) What roles does exploratory data analysis play in a data mining project?

(b) Does the experiment below suggest any bias? Justify your answer.

Experiment: Mall shoppers are asked to fill out and return a form rating their shopping experience at each of the 26 stores to identify the most popular stores in each of four categories. **(1 + 1 + 1) + (2 + 1) + (1 + 3) + (1 + 1) = 12**

4. (a) Define quantile and inter quartile range (IQR). How are 50th percentile and median related?

(b) Work out the IQR and the median for the given data set {9,2,7,6,3,5,1,3}. Explain in brief whether the two will necessarily be equal or not.

(c) What is 'expected value' (EV)? Work out one single EV of a seminar attendee for the following case:

A marketing company promotes a new cloud technology that offers two levels of service: one priced at USD 300 per month and another at USD $50 per month. It offers free seminars to generate leads, and it figures that 5% of the attendees will sign up for the USD 300 service, 15% for the USD 50 service, and 80% will not sign up for anything.

(d) What is 'random sampling'? Explain, in brief, the difference between 'sampling with replacement' and 'sampling without replacement'. **(1 + 1 + 1) + (2 + 1) + (1 + 3) + (1 + 1) = 12**

5. (a) A student counted the number of words in an essay she had written, recording the total every 10 lines (as shown in table below). How many words has she written if she writes 65 lines? Are you happy with the estimate? If not, what would you do to make yourself satisfied.

No. of Lines (x)	10	20	30	40	50	60	70	80
No. of Words (y)	75	186	210	291	868	441	519	789

(b) Consider the data given in table below. Show the contingency table and compute the distance between Jack and Mary using Jaccard's Coefficient.

Name	Fever	Cough	Medical Test 1	Medical Test 2	Medical Test 3	Medical Test 4
Jack	Yes	Negative	Positive	Negative	Negative	Negative
Mary	Yes	Negative	Positive	Negative	Positive	Negative
Jim	Yes	Positive	Negative	Negative	Negative	Negative

(5 + 1 + 2) + (2 + 2) = 12

(c) If the relationship is that one variable decreases when the other increases, but the amount is not consistent (as depicted in figure below), then will the values of Pearson and Spearman correlation coefficients be the same?

(d) If the relationship is that one variable decreases when the other increases, but the amount is not consistent (as depicted in figure below), then will the values of Pearson and Spearman correlation coefficients be the same?

(e) How does Pearson product moment correlation differ from Spearman rank order correlation?

Group – D

6. (a) K-means does not explicitly use a fitness function. Which fitness function does it implicitly optimize? In general, K-means is limited to find clusters having complex shapes. What could be done to enable K-means to find clusters having arbitrary shapes? **4 + 2 + 1 + 1 + 4 = 12**

(b) Which one of the two approaches of hierarchical clustering algorithms is robust to outliers? Justify your answer.

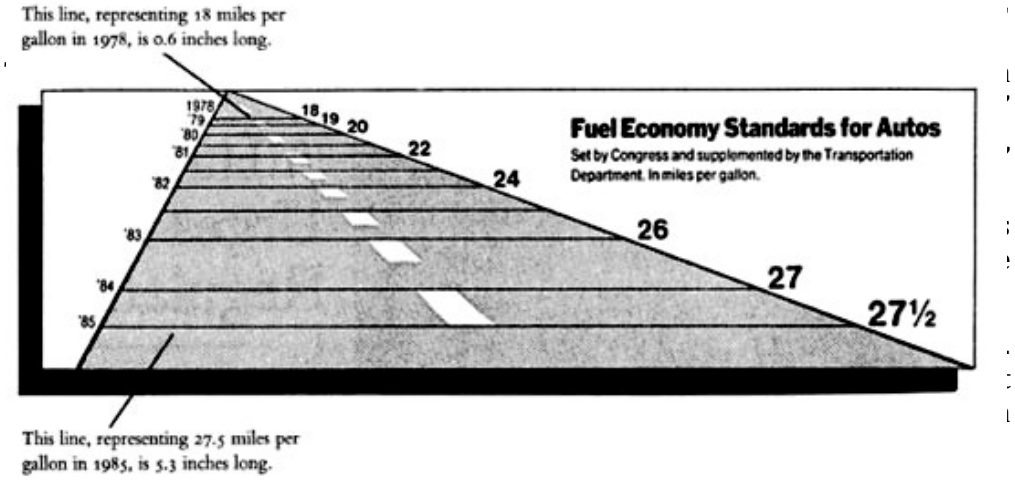
(c) Use the "Anscombe's Quartet" [Francis Anscombe, 1973] example to illustrate the importance of looking at a set of data graphically before starting to analyse it according to a particular type of relationship, and the inadequacy of basic statistical properties for describing real-life datasets.

5 + 3 + 4 = 12

9. (a) Explain expressiveness and effectiveness in the context of visual encoding.

(b) Give two examples of each of the following:
(i) Quantitative Mackinlay's retinal variables, (ii) Ordinal Mackinlay's retinal variables and (iii) Nominal Mackinlay's retinal variables.

(c) Calculate the lie factor for the following figure below:



(2 + 2) + (2 + 2 + 2) + 2 = 12

(d) Agglomerative hierarchical clustering algorithm merges the pair of clusters that are the closest to each other. Is this always a good approach? Justify your answer.

(d) By definition, Bayesian classification works with categorical variables. Explain, in brief, the two typical approaches for applying naïve Bayes technique to numerical variables.

2 + 6 + 2 + 2 = 12

Group – E

8. (a) One formal definition of 'visualization' says "... is the *process* of extracting salient *features* from *sets of data* and *displaying* the features

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