Total No. of Questions: 6 *Total No. of Printed Pages:3*

Enrollment No.....

Faculty of Engineering

End Sem Examination May-2023

CS3ED06 Data Science

Programme: B.Tech. Branch/Specialisation: CSE / All

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d. Assume suitable data if necessary. Notations and symbols have their usual meaning.

Q.1 i. Data science is the process of diverse set of data through- **1**

(a) Organizing data (b) Processing data

(c) Analysing data (d) All of these

ii. The modern conception of data science as an independent discipline is sometimes attributed to- **1**

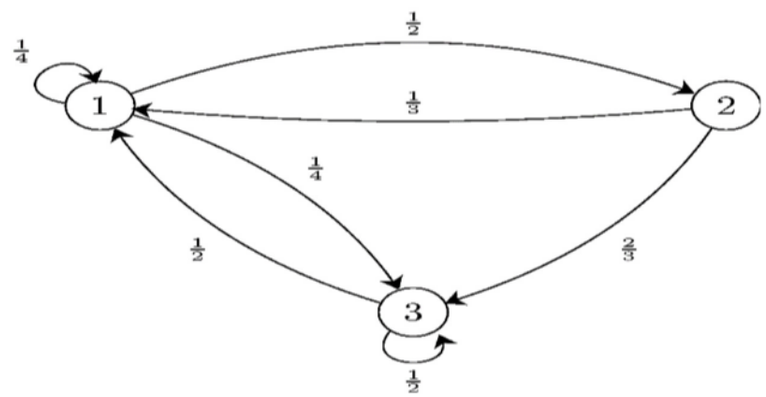
(a) William S. (b) John McCarthy

(c) Arthur Samuel (d) Satoshi Nakamoto

iii. Consider the Markov chain shown in Figure. Assume $X_0=1$, and let R be the first time that the chain returns to state 1, i.e., **1**

$$R=\min\{n\geq 1 : Xn=1\}.$$

Find $E[R|X_0=1]$.



(a) 8/3 (b) 7/3 (c) 4/3 (d) 5/3

iv. If A and B are two events such that P(AUB) = 5/6, P(A∩B) = 1/3, P(B) = 1/2, then the events A and B are- **1**

(a) Dependent (b) Independent

(c) Mutually exclusive (d) None of these

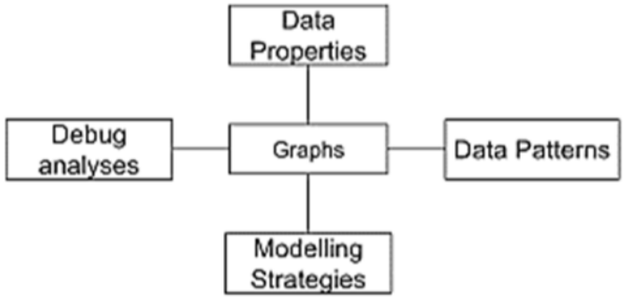
P.T.O.

v. Which of the following is not part of the data science process? **1**

(a) Communication Building (b) Operationalize

(c) Model planning (d) Discovery

vi. Which of the following graphs has properties in the below figure? **1**



(a) Exploratory (b) Inferential

(c) Causal (d) None of these

vii. What is true about Data Visualization? **1**

(a) Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.

(b) Data Visualization helps users in analyzing a large amount of data in a simpler way.

(c) Data Visualization makes complex data more accessible, understandable, and usable.

(d) All of these

viii. Data can be visualized using? **1**

(a) Graphs (b) Charts

(c) Maps (d) All of these

ix. Which of the following is not correct sub-packages of SciPy? **1**

(a) scipy.cluster (b) scipy.source

(c) scipy.interpolate (d) scipy.signal

x. Matplotlib is _____ plotting library. **1**

(a) 1D (b) 2D (c) 3D (d) 4D

Q.2 Attempt any two:

i. Explain the basic framework and architecture of data science? **5**

ii. What are the differences between data science, machine learning, and artificial intelligence? **5**

iii. How and why data science play important role in the today's business world? **5**

Q.3 Attempt any two:

i. You toss a fair coin three times: **5**

(a) What is the probability of three heads, HHH?

(b) What is the probability that you observe exactly one heads?

(c) Given that you have observed at least one heads, what is the probability that you observe at least two heads?

ii. What is normal distribution? Explain properties of normal distribution. **5**

iii. Define statistical inference. A bag contains about 2 green balls, 3 blue balls and 5 black balls. One of them is taken out. Find the probability that it is black. **5**

Q.4 Attempt any two:

i. What is exploratory data analysis in data science? Explain with example. **5**

ii. What is the philosophy of EDA in data science? **5**

iii. What are the four primary types of EDA? **5**

Q.5 Attempt any two:

i. How can we visualize more than three dimensions of data in a single chart? Explain with example. **5**

ii. What is a scatter plot? What types of data work best in scatter plots? **5**

iii. Explain the different types of visualizations you can use on data. **5**

Q.6 Attempt any two:

i. Explain the challenges and scope of data science project management. **5**

ii. What is linear graph? Use matplotlib to create linear graph to visualizations for following data set: **5**

x = [10, 20, 30, 40]

y = [20, 25, 35, 55]

iii. Write short notes on following: **5**

(a) NoSQL (b) Pylearn (c) SciPy

******

Marking Scheme
CS3ED06 Data Science

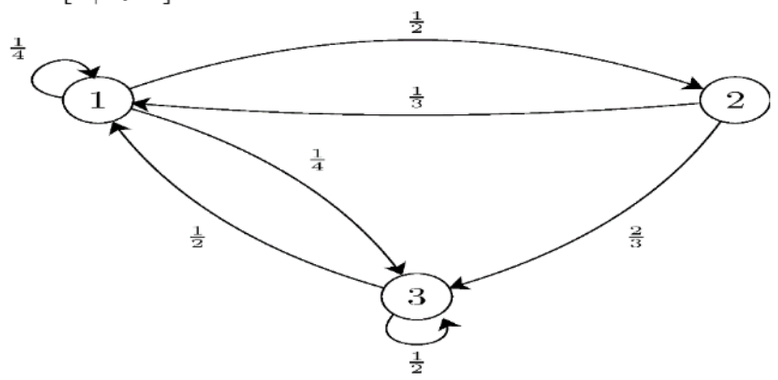
Q.1 **i)** Data science is the process of diverse set of data through? **1**
Answer: d. All of the above

ii) The modern conception of data science as an independent discipline is sometimes attributed to? **1**
Answer: a William S.

iii) Consider the Markov chain shown in Figure. Assume $X_0=1$, and let R be the first time that the chain returns to state 1, i.e., **1**

$$R=\min\{n\geq 1 : Xn=1\}.$$

Find $E[R|X_0=1]$.

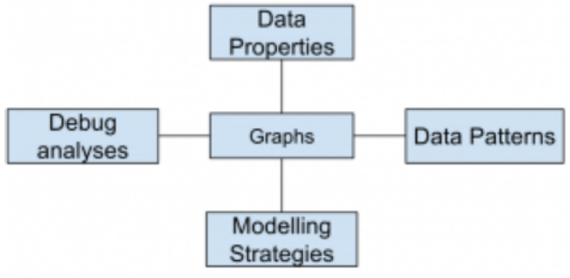


Answer: a 8/3

iv) If A and B are two events such that P(AUB) = 5/6, P(A∩B) = 1/3, P(B) = ½, then the events A and B are **1**
Answer: b Independent

v) Which of the following is not part of the data science process? **1**
Answer: a Communication Building

vi) Which of the following graphs has properties in the below figure? **1**



Answer: a Exploratory

vii) What is true about Data Visualization? **1**
Answer: d All of the above

viii) Data can be visualized using? **1**
Answer: d All of the above

ix) Which of the following is not correct sub-packages of SciPy? **1**
Answer: b scipy.source

x) Matplotlib is _____ plotting library **1**
Answer: b 2D

Q.2 Attempt any two:

i. Explain the basic framework and architecture of data science? **5**
Answer:
Basic framework 2.5 marks
Architecture of data science? 2.5 marks Diagram

ii. What are the differences between data science, machine learning, and artificial intelligence? **5**
Answer: Minimum five difference 1 marks for each

iii. How and why data science play important role in the today's business world? **5**
Answer: Minimum five role 1 marks for each

Q.3 Attempt any two:

i. You toss a fair coin three times: **5**
a. What is the probability of three heads, HHH?
b. What is the probability that you observe exactly one heads?
c. Given that you have observed at least one heads, what is the probability that you observe at least two heads?

Answer:

a. $P(HHH) = P(H) \cdot P(H) \cdot P(H) = 0.5^3 = \frac{1}{8}$.

b. To find the probability of exactly one heads, we can write

$$\begin{aligned} P(\text{One heads}) &= P(HTT \cup THT \cup TTH) \\ &= P(HTT) + P(THT) + P(TTH) \\ &= \tfrac{1}{8} + \tfrac{1}{8} + \tfrac{1}{8} \\ &= \tfrac{3}{8}. \end{aligned}$$

c. Given that you have observed *at least* one heads, what is the probability that you observe at least two heads? Let A_1 be the event that you observe at least one heads, and A_2 be the event that you observe at least two heads. Then

$$A_1 = S - \{TTT\}, \text{ and } P(A_1) = \frac{7}{8};$$

$$A_2 = \{HHT, HTH, THH, HHH\}, \text{ and } P(A_2) = \frac{4}{8}.$$

Thus, we can write

$$\begin{aligned} P(A_2|A_1) &= \tfrac{P(A_2 \cap A_1)}{P(A_1)} \\ &= \tfrac{P(A_2)}{P(A_1)} \\ &= \tfrac{4}{8} \cdot \tfrac{8}{7} = \tfrac{4}{7}. \end{aligned}$$

ii. What is normal distribution? Explain properties of normal distribution? **5**

What is normal distribution? 2 marks
Explain properties of normal distribution? 2 marks
Diagram 1 mark

iii. Define statistical inference? A bag contains about 2 green balls, 3 blue balls and 5 black balls. One of them is taken out. Find the probability that it is black. **5**

Define statistical inference? 2 Marks
Solution. 3 marks

Correct Answer:
The total number of bags are 10.
Probability of getting black balls= 5/10=½=0.5

Q.4 Attempt any two:

i. What is exploratory data analysis in data science? Explain with example? **5**

What is exploratory data analysis in data science? 3 marks
Explain with example? 2 marks

ii. What is the philosophy of EDA in data science? **5**

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

1. maximize insight into a data set;
2. uncover underlying structure;
3. extract important variables;
4. detect outliers and anomalies;
5. test underlying assumptions;
6. develop parsimonious models; and
7. determine optimal factor settings.

iii. What are the four primary types of EDA? **5**

1. Univariate Non-graphical 1 marks
2. Multivariate Non-graphical 1 marks
3. Univariate graphical 1 marks
4. Multivariate graphical 2 marks

Q.5 Attempt any two:

i. How can we visualize more than three dimensions of data in a single chart? Explain with example? **5**

To visualize data beyond three dimensions, we need to use visual cues such as

color, 1.5 marks
size, 1.5 marks
shape 2 marks.

Color is used to depict both continuous and categorical data.
Marker Size is used to represent continuous data. Can be used for categorical data as well. However, since size differences are difficult to detect, it is not considered the most appropriate choice for categorical data.
Shapes are used to represent different classes.

ii. What is a scatter plot? What types of data work best in scatter plots? **5**

What is a scatter plot? 2 marks
What types of data work best in scatter plots 2 marks
Diagram 1 mark

iii. Explain the different types of visualizations you can use on data? **5**

Minimum 2 types 2.5 marks for each

Q.6 Attempt any two:

i. Explain the challenges and scope of data science project management? **5**

[2] [3]

Explain the challenges 2.5 marks
and scope of data science project management? 2.5 marks

ii. What is linear graph? Use matplotlib to create linear graph to visualizations for following data set: **5**
x = [10, 20, 30, 40]
y = [20, 25, 35, 55]
What is linear graph? 2 marks
Use matplotlib to create linear graph 3 marks

iii. Write short notes on following: **5**
1. NoSQL 1.5 marks
2. Pylearn 1.5 marks
3. SciPy 2 marks

******