Total No. of Questions : 12]

SEAT No. :

PA-673

[Total No. of Pages : 3

[5928]-119

M.E. (Computer Engineering) (Data Science)

BASICS OF DATA SCIENCE

(2017 Pattern) (Semester - I) (510302)

Time : 3 Hours] *[Max. Marks : 50*

Instructions to the candidates:

1) ***Answer Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8, Q.9 or Q.10, Q.11 or Q.12.***
2) ***Neat diagrams must be drawn wherever necessary.***
3) ***Figures to the right indicate full marks.***
4) ***Assume suitable data, if necessary.***
5) ***Use of logarithmic tables slide rule, mollier charts electronic pocket calculator and steam tables is allowed.***

Q1) a) What is data science? Explain first two steps involved in data science process. **[5]**

b) Differentiate between - **[4]**

i) Structured Data and Unstructured Data

ii) Big Data and Little Data

OR

Q2) a) What is the role of a Data Scientist in the industry? **[5]**

b) Explain following steps involved in data science process. **[4]**

i) Data preparation

ii) Data exploration

Q3) What is data distribution? Explain following representation of a distribution with exmaple. **[8]**

a) Probability Mass Function

b) Cumulative Distribution Function

OR

Q4) Explain Summarizing the Data in EDA process in detail with example. **[8]**

P.T.O.

Q5) Write K-nearest Neighbors algorithm. Suppose you have given the following data where weight and *y* are the height input variables and Class is the dependent variable. **[9]**

Weight	height	Class
51	167	underweight
62	182	normal
69	176	normal
64	173	normal
65	172	normal
56	174	underweight
58	169	normal
57	173	normal
55	170	normal

Apply KNN algorithm to predict the class of new data point weight = 57 kg and height = 170 cm using Euclidean distance. Assume $k = 3$.

OR

Q6) Write K means algorithm. Using k-means algorithm, cluster following data into two clusters. Show each step of clustering. **[9]**

X	Y
185	72
170	56
168	60
179	68
182	72
188	77

Q7) a) What is data visualization and explain for what it used for? **[4]**

b) Explain following bivariate data visualization techniques. **[4]**

i) Bar plot

ii) Scatter plot

OR

Q8) a) What is data visualization? Explain different types of visualization. **[6]**

b) Explain data encoding in detail. **[2]**

Q9) a) Below is a utility matrix representing ratings by users A, B, C, D, E and F for items a through *f*. Calculate the Jaccard distance between item *a* and *b*. **[4]**

User\Item	*a*	*b*	*c*	*d*	*e*	*f*
A	1	1	0	0	0	0
B	1	1	1	0	0	0
C	1	0	0	0	0	0
D	0	1	0	1	0	0
E	0	0	0	0	1	1
F	0	0	0	0	1	1

b) What is utility matrix in recommendation system? Explain key problems for recommender system to figure out values in utility matrix. **[4]**

OR

Q10) What is a Collaborative-Filtering based recommendation system? Explain the advantages and disadvantages of Collaborative-Filerting based recommendation system. **[8]**

Q11) What is social network? What are the essential characteristics of a social network. **[8]**

OR

Q12) Explain Girvan-Newman Algorithm with example. **[8]**