[216] SEAT No._____

No. of Printed Pages: 03

SARDAR PATEL UNIVERSITY
M.Sc. (Applied Statastics) (III Semester) External Examination, 2018
COURSE No.: PS03CAST21
(Knowledge Discovery and Data Mining)
(Total Marks 70)

Date: 26/10/2018, Friday **Time:02:00 PM – 05:00 PM**

Q.1 Write the correct answer (each sub question carries **one marks**). **[8]**

(1) _____ is the first stage in genetic algorithm.

a. Evaluation of each string. c. Selection of string.

b. Creation of population of string. d. Genetic manipulate

(2) A _____ is necessary condition for KDDs effective implement.

a. Data set b. database c. Data warehouse d. Data

(3) An abstraction concept for building composite object from their component object is called :

a. Specialization b. Normalization c. Generalization d. Aggregation

(4) Strategic value of data mining is _____

a. cost-sensitive b. work-sensitive c. Time-sensitive d. Technical-sensitive

(5) _____Maps data into predefined groups.

a. Regression b. Time series analysis c. Prediction d. Classification

(6) The supervised learning algorithm, that uses a number of weak classifiers to form a strong classifier is

a. Learning Vector Quantization c. Singular Value Decomposition

b. Support Vector Machines d. Boosting

(7) Which of these is also called Fuzzy K-Means?

a. K-Means Clustering c. Probabilistic Clustering

b. Hidden Markov Models d. Hierarchical Clustering

(8) If we are sampling from a Normal population, when can we say that the results of Logistic Regression Analysis will be somewhat close to that of LDA?

a. If the sample size is large enough

b. Number of categories is high

c. If estimates are obtained using Maximum Likelihood Method

d. Number of regressors/covariates is high

Q.2 Answer any **SEVEN** the following questions (each sub-question carries **TWO marks**). **[14]**

(a) What is the classification of association rules based on various criteria?

(b) Explain number of methods to determine the number of factors.

(c) Define RDBMS and data attributes.

(d) Discuss Data Compression Algorithms in Unsupervised Learning

(e) Differentiate between OLAP and OLTP.

(f) Describe ANN computing and discuss its suitability to data mining.

(g) When do we use Hamming Distance? Cite at least one classification algorithm where it might be used.

(h) What are the challenges associated with the implementation of Unsupervised Learning?

(i) What are Autoencoders?

Q.3 (a) '80% of KDD is about preparing data and 20% of mining' Justify this statement with the help of knowledge discovery process. [6]

(b) Define Data mining. What are the data mining problems and issues? [6]

==OR==

(b) What is data warehouse? Explain characteristic of data warehouse. [6]

Q.4 (a) Your company is considering whether it should tender for two contracts (MS1 and MS2) on offer from a government department for the supply of certain components. The company has three options: Tender for MS1 only; or Tender for MS2 only; or Tender for both MS1 and MS2. If tenders are to be submitted the company will incur additional costs. These costs will have to be entirely recouped from the contract price. The risk, of course, is that if a tender is unsuccessful the company will have made a loss. [6]

The cost of tendering for contract MS1 only is £50,000. The component supply cost if the tender is successful would be £18,000. The cost of tendering for contract MS2 only is £14,000. The component supply cost if the tender is successful would be £12,000. The cost of tendering for both contracts MS1 and contract MS2 is £55,000. The component supply cost if the tender is successful would be £24,000. For each contract, possible tender prices have been determined. In addition, subjective assessments have been made of the probability of getting the contract with a particular tender price as shown below. Note here that the company can only submit one tender and cannot, for example, submit two tenders (at different prices) for the same contract.

Option	Possible tender prices			Acceptance Probability (Respective price)		
MS1	130000	115000		0.2		0.85
MS2	70000	65000	60000	0.15	0.80	0.95
MS1 AND MS2	190000		140000	0.05		0.65

In the event that the company tenders for both MS1 and MS2 it will either win either contracts (at the price shown above) or no contract at all. What do you suggest the company should do and why?

(b) In a County, 51% of the adults are males. It learned that the selected survey subject was smoking a cigar. 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars (based on data from the Substance Abuse and Mental Health Services Administration). Use this information to find the probability that the selected subject is a male. [6]

==OR==

(b) An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Airline Manufacturing Company makes 70% of the ELTs, the Bryant Company makes 20% of them, and the Chartair Company makes the other 10%. The ELTs made by Altigauge have a 8% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Chartair ELTs have a 12% rate of defects. [6]

1. If an ELT is randomly selected from the general population of all ELTs, find the probability that it was made by the Airline Manufacturing Company.
2. If total production ELT Device 10000 units then found the number of defective quantity produce by The Bryant ELTs.

Q.5 (a) Distinguish the Method of K-Means and K-Nearest Neighbors.

(b) What is Hierarchical method of clustering? Also describe agglomerative and divisive hierarchical clustering? [6]

==OR==

(b) Decompose the following Matrix Using SVD: [6]

$$\begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

Q.6 (a) What is the need of dimensionality reduction? Describe two methods for dimensionality reduction. [6]

(b) Discuss E- commerce and usefulness of data mining in e-commerce. [6]

==OR==

(b) The discovery of the association rule can be formulated in terms of relational algebraic operations Support and Confident. Explain with example. [6]

— X —