DATA ANALYTICS
(INFO 3202)

Time Allotted : 3 hrs **Full Marks : 70**

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) Which of the following statement is/are true with respect to clustering data objects?
St. 1 - An object belong to only one cluster in hard clustering technique
St. 2 - An object always maps to multiple cluster in soft clustering
St. 3 - Hierarchical agglomerative approach is a bottom up approach
(a) 1 & 3 (b) All
(c) 1 & 2 (d) 2 & 3.

(ii) __________ and __________are slave services in Hadoop ecosystem.
(a) Job Tracker and Task Tracker (b) Data node and Name Node
(c) Name Node and Job Tracker (d) Data Node and Task Tracker.

(iii) A 200 mb data file is fragmented into _______ input splits in a 64 mb block size of HDFS and has _______ mappers associated.
(a) 4, 4 (b) 3, 4
(c) 4, 3 (d) 3, 3.

(iv) Gain Ratio has an advantage over Information Gain of an attribute when the
(a) attribute is categorical with 3 possible values
(b) attribute has multiple values, and each is unique
(c) attribute is numerical with large number of duplicate values
(d) none.

(v) The algorithm for categorical attribute clustering is
(a) K-means (b) DBSCAN
(c) Both (a) and (b) (d) ROCK.

(vi) The order of the membership matrix in Fuzzy C Means is M * N, where M is number of objects, and N is
(a) number of data points (b) number of clusters
(c) number of samples (d) number of means.

(vii) With respect to clustering technique which of the following statement is/are true?
St. 1- K means algorithm is poor in handling outliers
St. 2- DB-SCAN doesnot require the value of number of clusters, as initial parameter
St. 3- In Fuzzy C means the initial membership matrix has equal membership values
(a) 1 & 3 (b) All (c) 1 & 2 (d) 2 & 3.

(viii) Bayes rule provides posterior probability of the system under experiment. The posterior probability is obtained as
(a) likelihood probability only
(b) product of likelihood probability and prior probability
(c) prior probability only
(d) likelihood probability prior probability.

(ix) Your dataset is a mixture of two gaussian distributions. Identification of two independent Gaussian distributions can be obtained by
(a) Principal component analysis
(b) Maximum likelihood method
(c) Expectation-Maximization algorithm
(d) Linear regression.

(x) The Fuzzy C Means algorithm is a _______ clustering technique.
(a) hard (b) soft (c) density based (d) semi hard

Group - B

2. (a) Group the following data points using k-means clustering technique, where k=2 and each data point represented in the form of (x_coordinate, y_coordinate). Consider P1, P5 as the initial cluster centroids. Iterate till centroid does not become same or iterate till 3 (whichever comes early).
Data Points: P1(2,5); P2(4, 5); P3(5,5); P4(5,4); P5(7, 5); P6(6,4); P7(100,112); P8(10,9); P9(2,3); P10(7,7). *[(CO1,CO3,CO6)(Apply/IOCQ)]*

(b) State the objective function of K means algorithm. *[(CO1)(Remember/LOCQ)]*

8 + 4 = 12

3. (a) Discuss the core, border, and noise points with respect to DBSCAN algorithm. *[(CO1)(Remember/LOCQ)]*

(b) Using DBSCAN clustering algorithm group the following 10 spatial data objects, with number of minimum neighbors required = 4 and ε = 4 .3. All steps should be shown. Each object has 2 dimensions. The objects are (4,5), (8,3), (4,4), (6,6), (5,7), (26,28), (27,26), (18,19), (23,19), (6,5). *[(CO1,CO6)(Create/HOCQ)]*

3 + 9 =12

Group - C

4. (a) Given the following dataset apply Naive Bayes algorithm, to construct P(Flu = Y|X; X={Chills, Runny nose, headache, fever}).

Chills	Runny nose	Head ache	Fever	Flu
Y	N	MILD	Y	N

Chills	Runny nose	Head ache	Fever	Flu
Y	Y	NO	N	Y
Y	N	STRONG	Y	Y
N	Y	MILD	Y	Y
N	N	NO	N	N
N	Y	STRONG	Y	Y
N	Y	STRONG	N	N
Y	Y	MILD	Y	Y

[(CO2,CO3,CO6)(Evaluate/HOCQ)]

(b) Explain in brief the naive bayes algorithm. *[(CO2)(Remember/LOCQ)]*

8 + 4 = 12

5. (a) Construct the decision tree model from the following training dataset, using gain ratio indices.

Name	**Hair**	**Height**	**Weight**	**Lotion**	**Result**
Sarah	brown	short	light	no	sunburned
ana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned

[(CO2,CO6)(Apply/IOCQ)]

(b) "ID3 has been modified in C4.5 classification technique" — Justify the statement.

[(CO3)(Evaluate/HOCQ)]

9 + 3 = 12

Group - D

6. (a) A Map Reduce paradigm is used to compute the number of unique words present in the following text:

"There is a Workshop in HIT. The workshop is on Big Data Analytics. Heritage is in Kolkata."

Explain with the help of a diagram how the parallel execution takes place in the mappers and reducers. *[(CO4,CO6)(Analyse/IOCQ)]*

(b) What are the four properties of Big Data? *[(CO4)(Remember/LOCQ)]*

8 + 4 = 12

7. (a) Suppose a 250 MB file is required to be stored in HDFS. Explain how data parallelism is encountered in hadoop ecosystem while storing the file.

[(CO4)(Analyse/IOCQ)]

(b) State the hadoop master and slave services and the relationship among them.

[(CO4)(Understand/LOCQ)]

7 + 5 = 12

Group - E

8. (a) Explain the architecture of HBase NO-SQL database. *[(CO5)(Understand/LOCQ)]*

(b) With the help of a diagram explain the data model of MongoDB.

[(CO5)(Analyse/IOCQ)]

6 + 6 = 12

9. Write short notes on the following (any two). **(6 × 2) = 12**
   (i) Expectation Maximization Algorithm
   (ii) Fuzzy C Means Clustering Algorithm
   (iii) K nearest neighbor classification algorithm. *[(CO5/CO2/CO1)(Analyse/IOCQ)]*

Cognition Level	*LOCQ*	*IOCQ*	*HOCQ*
Percentage distribution	*27.08*	*52.08*	*20.83*

Course Outcome (CO):

1. Apply the different clustering algorithms to cluster real life datasets.
2. Apply appropriate classification algorithm to classify an unknown dataset.
3. Analyze the performance of the Clustering or Classification Algorithms.
4. Identify the need of Big Data Paradigms, and will be able to Store and Process Data on Hadoop Distributed File System.
5. Identify the need of No-SQL Databases and be able to Convert Relational Model to different No-SQL Data Models.
6. Create Appropriate Classifiers or Clustering Models for Analyses of Big Data using Hadoop Eco System.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*