B.TECH/IT/6TH SEM/INFO 3202/2022

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 any 5 (five) from Group B to E, taking at least one 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) Hierarchical agglomerative based clustering is a
(a) bottom up approach (b) top down approach
(c) both (a) and (b) (d) none of the above.

(ii) Consider a dataset to be clustered using soft clustering into 3 clusters. If the membership value of first data point to first cluster and second cluster is .41 and .21 respectively, then what will be the membership value of the point to the third cluster?
(a) 1 (b) 0.36 (c) 0.38 (d) 0.

(iii) The number of mappers required in a hadoop system when number of input splits of a dataset are 4, are
(a) 4 (b) 1 (c) any number (d) none.

(iv) Gain Ratio has an advantage over 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
(d) numerical with large number of duplicate values.

(v) _____ is an example of document database
(a) HBase (b) MongoDB (c) both (a) and (b) (d) None

(vi) Consider two different execution instance of K means clustering with the same dataset, where, k=2. The sum squared error is calculated as 3.23 and 3.12 respectively. Which statement is/are true?
Statement1: Intra cluster similarity between clusters of Execution instance 1 is more than instance 2
Statement2: Intra cluster similarity of Execution instance 1 is less than instance 2
(a) Both (b) Statement 1 (c) Statement 2 (d) None is true.

(vii) A _____ point in DBSCAN algorithm can become a part of cluster later.
(a) core (b) border (c) noise (d) dissimilar

(viii) C4.5 algorithm falls under the category of
(a) Unsupervised learning algorithm
(b) Reinforced learning algorithm
(c) Supervised learning algorithm
(d) Prone to errors in classification problems with many classes.

(ix) Which of the following statement/s is/are true with respect to HBase?
Statement1: There can be several regions in HBase.
Statement2: A region can have many HFiles.
Statement3: An HFile can be associated with multiple column families.
(a) Statements 1 and 2 (b) Statements 1 and 3
(c) Statements 1, 2, and 3 (d) None.

(x) Cassandra and HBase are NO-SQL databases of which category?
(a) Wide-column oriented (b) Document based
(c) Graph based (d) Key Value.

Group- B

2. (a) Consider the transactional database below. Using the concept of **ROCK** clustering, find out the neighbors of each object and also find the link between (object 1 and 4), considering the threshold =1/2. [(CO1)(Apply/IOCQ)]

Transaction Id	Items Bought
T1	A,C
T2	D,F,G,R
T3	A,C,D,R
T4	A,R,C,F
T5	A, F

(b) Apply k means algorithm to 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 A2, A6, and A7 as the initial cluster centroids.
Data Points: A1(2.3,2.4); A2(4.4, 5.2); A3(1.8,1.4); A4(18.5,18.6); A5(17.4, 15.3); A6(16.4,14.4);A7(100.2,112.2); A8(117.5,149.6). [(CO1)(Apply/IOCQ)]
6 + 6 = 12

3. (a) Explain the **DBSCAN** clustering algorithm with example, and also discuss the advantage of DBSCAN over kmeans algorithm. [(CO5)(Analyze/IOCQ)]

(b) Using Fuzzy C means method, cluster the group of 2D data objects, having level of fuzziness 1.25, C =2. Only update the membership matrix with respect to 2 iterations.
Data Objects are ([0.2,0.4], [0.2,0.2], [0.8,0.3], [0.9,0.5], [0.6,0.7], [1.6,1.3], [1.8,1.6], [1.3,1.6]).
The initial membership value for the data points belonging to cluster 1 are (0.67, 0.51, 0.88, 0.30, 0.33, 0.44, 0.50, 0.18). [(CO1)(Evaluate/HOCQ)]
(3 + 2) + 7 = 12

Group - C

4. (a) State Bayes' theorem/rule. What is maximum likelihood? What is the basis of assigning Prior probability? [(CO2)(Understand/LOCQ)]

   (b) Following is your flu data set:

Chills	runny nose	head ache	fever	Flu
Y	N	Mild	Y	N
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

   Construct the posterior probability P(Flu = Y|X), where,
   X = {Chills, Runny nose, headache, fever}. [(CO2)(Analyze/IOCQ)]

   (c) Develop your predictive model and compute the accuracy of your model. [(CO5)(Evaluate/HOCQ)]

(1 + 1 + 2) + 4 + 4 = 12

5. (a) Define sensitivity and specificity and accuracy with the help of confusion matrix. [(CO2) (Understand/LOCQ)]

   (b) "ID3 algorithm is supervised machine learning algorithm" — Justify the statement with appropriate example. [(CO2)(Analyze/IOCQ)]

   (c) "Information entropy is defined as the expectation of logarithm of probability with a negative sign" — Justify the truth of this statement with proper example. [(CO2)(Analyze/IOCQ)]

3 + 4 + 5 = 12

Group - D

6. (a) Consider a 630 MB data file. A task is to be executed on the data file. In order to efficiently execute the task, explain with the help of diagram how in Hadoop paradigm the data as well as task gets executed in parallel. [(CO3)(Analyse/IOCQ)]

   (b) Given a set of data points with set of initial centroids and the value of k. Can you convert the serially executed k means algorithm to a Map Reduce approach, by designing map and reduce algorithm such that total execution time gets reduced (Hints: Each mapper accepts input splits, all centroids and k as input)? [(CO3,CO6)(Create/HOCQ)]

6 + 6 = 12

7. (a) Explain with an example how HBase is horizontally scalable with the help of its architecture. [(CO4)(Analyse/IOCQ)]

   (b) Consider a relation R in RDBMS consisting of 3000 records, with attributes Customer-id, Customer-name, Customer-email, Customer-salary, Customer-department, Customer-Commission. If the above relation is converted to a columnar database HBase, model the relation R following the data modelling

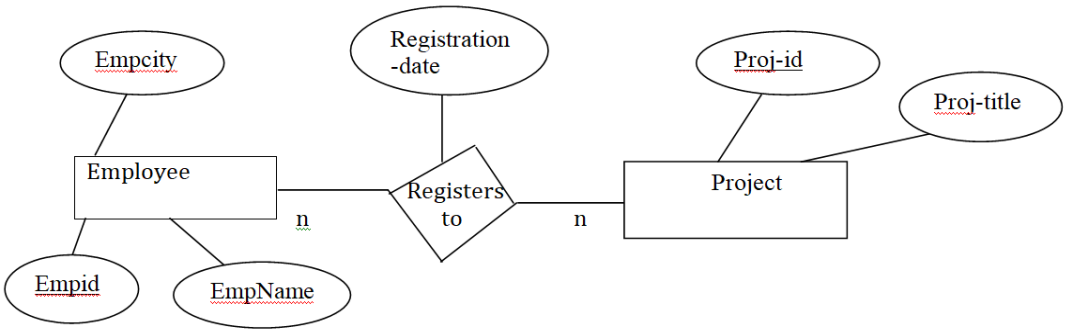
principle of HBase. If a region holds 1200 records, how many regions can be there? How many Hfiles each regions will have? [(CO4)(Create/HOCQ)]

5 + (5 + 1 + 1) = 12

Group - E

8. (a) With example explain the data model of Mongo-DB. [(CO4) (Understand/LOCQ)]

(b) A company's employees register themselves in different projects. An employee can register to multiple project, and a project can have multiple employees registered to it. From the ER model provided below, design the relational model and next convert the relational model to document based model.



[(CO4,CO6)(Create/HOCQ)]

5 + 7 = 12

9. Write short notes *on any three:* **(4 × 3) = 12**

(i) Graph Databases
(ii) HDFS Architecture
(iii) KNN Algorithm
(iv) Expectation Maximization Algorithm. [(CO1,CO4,CO3,CO5))(Understand/LOCQ)]

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	25	42.71	32.29

Course Outcome (CO):

After the completion of the course students will be able to

1. Compare among the different clustering algorithms and apply a suitable algorithm to cluster real life data sets.
2. Apply appropriate classification algorithm to classify an unknown dataset.
3. List the components of Hadoop and Hadoop Eco-System.
4. Access and Process Data on Distributed File System.
5. Analyze the performance of the algorithms.
6. Develop Big Data Solutions using Hadoop EcoSystem.

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