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) is a component of Hadoop file system.
(a) Namenode (b) Datanode
(c) Both (a) and (b) (d) None

(ii) Hierarchical agglomerative based clustering is a
(a) Bottom up Approach (b) Top down Approach
(c) Both (a) and (b) (d) None of the above

(iii) ROCK clustering means
(a) Robust hierarchical clustering
(b) Robust hierarchical clustering with links
(c) Runtime hierarchical clustering with links
(d) Robust clustering with kmedoids

(iv) Naive Bayes and Decision Tree are
(a) Supervised Learning (b) Unsupervised learning
(c) both (d) None

(v) Job Tracker is a master process and its slave process is
(a) Data Node (b) Task Traker
(c) Reducer (d) Mapper

(vi) 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

(vii) A density based clustering algorithm is
(a) PAM (b) STIRR
(b) ROCK (d) DBSCAN

(viii) Sensitivity of a classification model is also known as
(a) True positive (b) True negative
(c) True positive rate (d) True negative rate

(ix) Which database is horizontally scalable
(a) HBase (b) RDBMS
(c) MongoDB (d) None

(x) The algorithms based on partitioning paradigm
(a) K-means (b) STIRR
(c) Both (d) None of the above

Group – B

2. (a) Group the following data points using k-means clustering technique, where k=3 and each data point represented in the form of (x_coordinate, y_coordinate). Consider 1, 3, and 4 as the initial cluster centers.

Datapoints	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	15.0	17.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

(b) Discuss the limitations of K Means clustering algorithm.

9 + 3 = 12

3. (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 3), considering the threshold =1/3.

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

(b) Explain the process of Naïve Bayes classification technique.

(4 + 3) + 5 = 12

Group – C

4. (a) Using Fuzzy C means method cluster the group of 2D data objects, having level of fuzziness 1.28, C =2.
Data Objects are ([0.1,0.3], [0.3,0.3], [0.7,0.2], [0.9,0.4], [0.5,0.5], [0.2,0.3], [0.8,0.6], [0.3,0.6])
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).
Only update the membership matrix with respect to 2 iteration, i.e., twice.

(b) Explain how noise points are handled in DBSCAN clustering algorithms with example.

8 + 4 = 12

5. (a) Construct a Decision Tree model based on ID3 algorithm using the weekend spending data, as given in the following Table. Consider Weather type,Humidity, and Money Expended as the features to classify, using binary classification.

Week End	Weather type	Humidity	Money Expended	Decision
Week1	Hot	Extreme	7000	Stay Indoor
Week2	Chilled	Low	10000	Outdoor visit
Week3	Wet	Low	2500	Outdoor visit
Week4	Wet	Extreme	2000	Stay Indoor
Week5	Hot	Low	4000	Outdoor visit
Week6	Chilled	Extreme	3500	Outdoor visit
Week7	Hot	Low	5000	Outdoor visit
Week8	Chilled	Low	1500	Outdoor visit
Week9	Chilled	Extreme	3400	Outdoor visit
Week10	Wet	Extreme	1560	Stay Indoor

(b) Explain how ID3 has been modified in C4.5 classification technique.

8 + 4 = 12

Group – D

6. (a) Explain how parallelism is encountered in Map Reduce paradigm, with the help of an example.

(b) Describe with the help of a diagram the architecture of Hadoop Distributed File System.

6 + 6 = 12

7. (a) Suppose you have a word file with the following text.
"The world is going through a huge crisis. God save the world. The world is beautiful" The file size is 110 MB. Explain how the file gets broaken into input splits and explain the overall steps in mapper and reducer.

(b) Explain the functions of namenode, datanodes, job tracker and tasktracker

7 + 5 = 12

Group – E

8. (a) Explain with an example how HBase is horizontally scalable

(b) Explain the architecture of HBASE distributed database.

5 + 7 = 12

9. Write short notes on (i) MongoDB, (ii) HBase vs. RDBMS.

6 + 6 = 12

Department & Section	Submission Link
IT	https://classroom.google.com/c/MzAwMzg1ODQ4MDI2/a/MzY0NTQwNTc3NjMz/details