GANPAT UNIVERSITY

M. TECH SEMESTER - II – Information Technology

3IT203: Data Mining &Data Warehousing

Regular Examination April June -2015

Time: 3Hour] **[Total Marks: 60**

SECTION-I

Q-1 **A** Explain KDD process with diagram. [5]

B Explain various attribute subset selection methods. [5]

OR

Q-1 **A** Apply OLAP operations (slice and dice, pivot, drill down, roll up) on given dataset. [5]

Location = "Valsad"					Location = "Surat"				
	Item					Item			
Time	Computer	Monitor	CPU	Keyboard	Time	Computer	Monitor	CPU	Keyboard
Q1	201	132	123	233	Q1	238	189	143	239
Q2	520	124	435	142	Q2	508	193	434	542
Q3	234	322	532	144	Q3	238	392	532	544
Q4	433	433	534	233	Q4	439	493	544	223
Location = "Mehsana"					**Location = "Patan"**				
	Item					**Item**			
Time	Computer	Monitor	CPU	Keyboard	Time	Computer	Monitor	CPU	Keyboard
Q1	224	135	163	273	Q1	270	137	173	237
Q2	250	143	474	172	Q2	505	125	437	147
Q3	215	336	562	174	Q3	237	327	572	147
Q4	243	233	574	273	Q4	437	437	537	273

B What is Data warehousing? How are organizations using the information from data warehouses? [5]

Q-2 **A** Differentiate OLAP and OLTP. [5]

B Given a set of four-dimensional samples with missing values: [5]

X1 = { 0, 1, 1, 2 } X2 = { 2, 1, - , 1 } X3 = { 1, -, -, 0 } X4 = { -, 2, 1, - }

If the domains for all attributes are [0, 1, 2], what will be the number of "artificial" samples if missing values are interpreted as "don't care values" and they are replaced with all possible values for a given domain?

OR

Q-2 **A** What is fact constellations schema? Describe with suitable example. [5]

B What is Normalization? Which are three methods of it. Suppose the minimum and maximum values for the attribute income are $12,000 and $98,000, respectively. Map income to the range [0,1] using any one method of it. [5]

Q-3 **A** Define data mining as an influence of multiple disciplines. [5]

B What is Lattice of cuboid? Draw a diagram of lattice of cuboid of starting from 0-D to 4-D for various dimensions. [5]

SECTION-II

Q-4 **A** How Market Basket Analysis relates to data mining. Explain it with suitable example. Also discuss support and confidence. [5]

B What is Association Rule Mining? Find out the frequent item sets and strong association rules for the above example using Apriori Algorithm. [5]

Transaction Id	Items
1	1,3,4,6
2	2,3,5,7
3	1,2,3,5,8
4	2,5,9,10
5	1,4

OR

Q-4 **A** [5]

	A1	A2	A3	A4	A5	A6
O1	True	True	True	False	False	True
O2	False	True	True	False	True	False

For the above given objects having asymmetric attributes, where True is more significant than False; find a)Dissimilarity measure and b) Jaccard coefficient.

B Discuss the methods to improve Apriori algorithm: [5]

Q-5 **A** Use the k-means algorithm to cluster the following 8 objects into three clusters.X1=(2,5), X2=(2,10), X3=(8,4), X4=(5,8), X5=(7,5), X6=(6,4), X7=(4,9), X8=(1,2).Take initial clusters as X2, X4 and X8 and distance measure as Euclidean distance.1) Find final three clusters and their centroids formed after 3 iterations. [5]

B Explain Agglomerative & Divisive hierarchical clustering approaches. [5]

OR

Q-5 **A** Predict Class Label using Naïve Bayesian Classification algorithm for the given tuple & class label from weather.arff data values given below: X=(outlook=rainy, temperature=<=75, Windy=TRUE) where Class label to predict is play. [5]

outlook	temperature	windy	play
sunny	85	FALSE	No
Sunny	80	TRUE	No
Overcast	83	FALSE	Yes
Rainy	70	FALSE	Yes
Rainy	68	FALSE	Yes
Rainy	65	TRUE	No
Overcast	64	TRUE	Yes
Sunny	72	FALSE	No
Sunny	69	FALSE	Yes
rainy	75	FALSE	Yes
Sunny	75	TRUE	Yes
Overcast	72	TRUE	Yes
Overcast	81	FALSE	Yes
rainy	71	TRUE	No|

B Discuss one the method which overcomes the issues of Agglomerative clustering. [5]

Q-6 **A** Explain DBSCAN algorithm. Also explain various terms used in DBSCAN algorithm with example. [5]

B Differentiate supervised learning and unsupervised learning. [5]

2 OF 2

===END OF PAPER===