9th semester IMCA Examination- 2019

9.2 Data Mining & Data Warehousing

Full Marks – 70

Time – 3 Hours

Answer All Questions

1	a)	What is data reduction? Explain the different strategies for data reduction.	[7]
	b)	What is data cleaning? Explain the basic methods of data cleaning.	[7]
		OR	
	c)	What is data mining? Discuss the architecture of a data mining system.	[7]
	d)	What is data normalization? What is its usefulness? Use the two methods below to normalize the following group of data: 150, 280, 560, 760, 980, 1200 i) min-max normalization by setting min=0 and max=1 ii) Z-score normalization	[7]
2	a)	Explain the basic differences between OLTP & OLAP.	[7]
	b)	Suppose that a data warehouse consists of the following four dimensions: (date, spectator, location and game) and two measures *count* and *charge*, where charge is the fare that a spectator pays when watching a game on a given date.	[7]

		Spectators may be students, adults or seniors with each category having its own charge rate. i) Draw a star schema diagram for the data warehouse. ii) Starting with the base cuboid [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by the student spectators at Mumbai in 2014. OR	
	c)	What is data warehouse? Explain the basic feature of data warehouse along with its architecture.	[7]
	d)	With reference to the Q. No- 2(a), i) Draw the snowflake schema diagram for the data warehouse. ii) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations should one perform in order to list the total number of students watched cricket in 2014.	[7]
3		A database has four transactions. Let min-support=60% and min-conf=80%. **TID** | **Items-bought** T1 | {I1, I2, I3, I4} T2 | {I5, I6, I7, I8, I9} T3 | {I10, I6, I9, I8} T4 | {I9, I2, I3} Write the steps of Apriori algorithm .Find all frequent item sets using Apriori Algorithm and determine the strong rules from any one frequent item set. OR	[14]

	a)	With reference to the database given in Q.No.-3, find all frequent item sets using FP-growth.	[10]
	b)	Explain the various kinds of Association rules.	[4]
4		Explain the working process of Naïve Bayesian classifier. Consider the following training data set.	[14]

Outlook	**Temp**	**Humidity**	**Wind**	**Play-Tennis**
sunny	hot	high	strong	no
sunny	hot	high	weak	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

Given a data tuple having the values “sunny”, “hot”, “normal”, ”weak” for the attributes outlook, temp, humidity, wind respectively, what would a naive Bayesian classification of

		the status of the tuple be? OR What is Decision tree? Explain the decision tree learning algorithm. Determine the root node using information gain in decision tree induction using the dataset given in No. 4.	[14]
5	a)	What is clustering? Discuss the different types of clustering methods.	[6]
	b)	Suppose the data mining task is to cluster the following eight points into three clusters: A1(3, 8), A2(4, 9), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9) The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 As the center of each cluster respectively. Write the K-means algorithm and use it to show only the three cluster centers after the first round execution.	[8]
	c)	OR Write short notes on interval-scaled variables and binary variables.	[6]
	d)	With reference to the clustering problem given in Q.No-5-(b), use k-medoid clustering algorithm to show the three clusters after the first round execution.	[8]