Birla Institute of Technology & Science, Pilani

Work Integrated Learning Programmes Division
Comprehensive Exam

Course Number	:	PCAM ZC221
Course Title	:	Unsupervised Learning and Association Rule Mining
Type of Exam	:	Closed Book (Make-UP)
Weightage	:	30 %
Date of Exam	:	20/June/2020

No. of Pages: 2
No. of Questions: 5

Duration: 2 hours
Session : FN

Two branches B1, B2 of a retail store XYZ were asked to submit their sales data for yearly audit.
The precomputed similarity (proximity) matrix for all the records is given below. The records are identified with branch #- quarter #. For ex, "B1-Q1" means the data corresponds to the data for quarter-1 submitted by Branch-1. Questions 1 & 2 is based on this.

	B1-Q1	B1-Q2	B1-Q3	B2-Q1	B2-Q2	B2-Q3
B1-Q1	0	25	30	35	55	50
B1-Q2	25	0	15	10	30	35
B1-Q3	30	15	0	25	25	20
B2-Q1	35	10	25	0	20	45
B2-Q2	55	30	25	20	0	25
B2-Q3	50	35	20	45	25	0

Q1 It's been found from a reliable source that two of the report submissions are forged in the profit and sales details. Detect the forged records in given audit input data and remove them using appropriate density-based outlier detection algorithm. Consider 2-NN as the parameter. **7M**

*[**Note**: Upload the screenshot of your written paper as an answer if required. Show all step wise solution]*

Q1

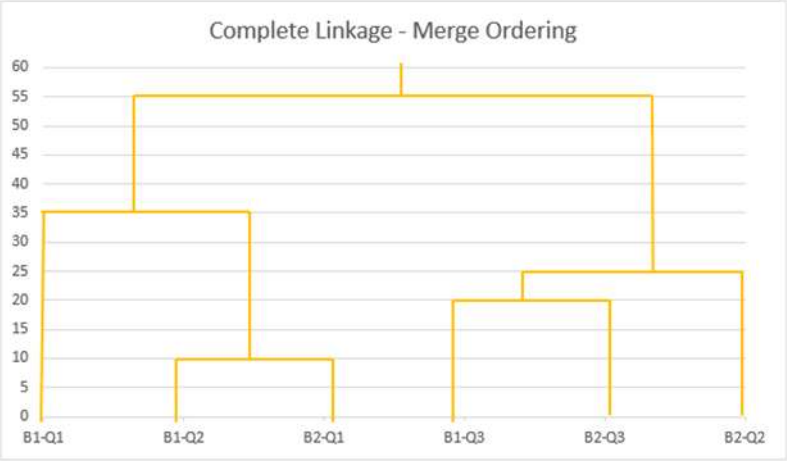
1m - Identification of Local Outlier Factor with Kth neighbor = 2
1m - Local reachability density & Local Outlier Factor
1m - Correct other numerical values like K-NN listing per point and Dist(K-NN,p)= {30,15,20,20,25,25} values
1.5m - lrd(p) calculation = 1.5m
1.5m - lof(p) calculation = {For B1-Q1 : 1.375, For B2-Q2:1.12 }
1m - Correct top 2 outlier detection = {B1-Q1 , B2-Q2}

K=2	Kth NN	dist $_{MinPts}$ (p)	N $_{MinPts}$ (p)	lrd $_{MinPts}$ (p) **(1m)**	Local Outlier Factor **(1m)**	isOutlier as per problem statement?
B1-Q1	B1-Q3	30	B1-Q2, B1-Q3	2/ 55	**1.375**	Yes
B1-Q2	B1-Q3	15	B2-Q1, B1-Q3	2/ 40	1	
B1-Q3	B2-Q3	20	B1-Q2, B2-Q3	2/ 40	0.944	
B2-Q1	B2-Q2	20	B1-Q2, B2-Q2	2/ 40	0.944	
B2-Q2	B1-Q3, B2-Q3	25	B2-Q1, B1-Q3 or B2-Q1, B2-Q3	2/ 45	**1.125**	Yes
B2-Q3	B2-Q2	25	B1-Q3, B2-Q2	2/ 45	1.062	

Q2 Apply complete linkage algorithm. Draw the dendrogram obtained alone with correct merge order and distance values mentioned. **4M**

*[**Note**: Upload the screenshot of your written paper as an answer if required. No need to show all step wise matrix recomputation in solution]*

Q2



Q3

In retail business, better product placement strategies help increase sales. Below are some of the transactions performed by customers in a retail branch.

Order ID	Items Sold
Transaction 1	Tissue(**T**), Oats(**O**), Corn(**C**), Milk(**M**)
Transaction 2	Medicine(**MD**), Milk(**M**), Corn(**C**), Wine(**W**), Bread(**B**)
Transaction 3	Milk(**M**), Wine(**W**), Oats(**O**), Egg(**E**)
Transaction 4	Egg(**E**), Salad(**S**), Nuts(**N**), Bread(**B**), Milk(**M**), Oats(**O**)
Transaction 5	Oats(**O**), Bread(**B**), Egg(**E**), Corn(**C**), Milk(**M**), Salad(**S**)

*[**Note**: Show stepwise solutions. Upload the screenshot of your written paper as an answer if required in correct order]*

A. Use apriori algorithm and find the frequent itemsets with minimum support count of 3. Show all the steps

B. List the closed frequent item sets and maximal frequent itemsets. Differentiate the significance of both with appropriate justification. Use your results from part (A) to answer this.

(5+3 = 8marks)

Q3

Part - A:

1m - From the attribute extract the items and construct the transaction table or find the support of all items

1m - Level 1 = {B, C, E, M, O}

1m - Level 2 = {BM, CM, EM, EO, MO} (5 itemsets)

1m - Level 3 = {EMO} (Only one frequent 3-itemset)

1m - Application of Apriori Pruning Principle & Final answer : Frequent Itemset

{ B, C, E, M, O, BM, CM, EM, EO, MO, EMO }

Or

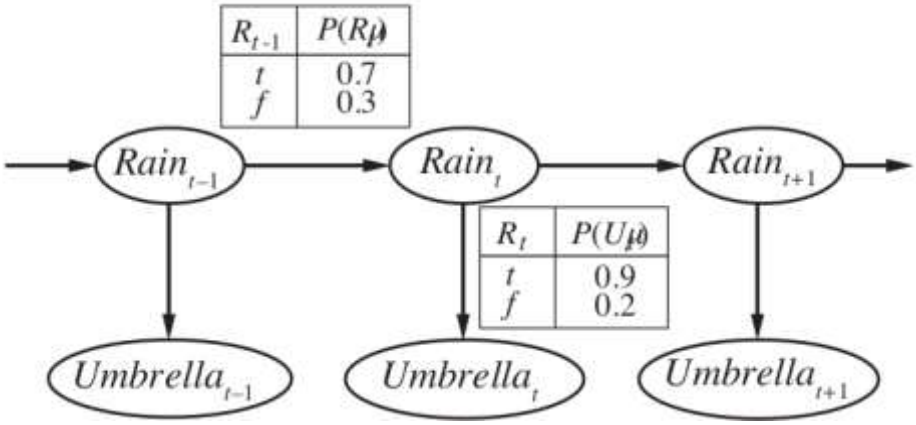
{ {Bread}, {Corn}, {Egg}, {Milk}, {Oats}, {Bread,Milk}, {Corn,Milk}, {Egg,Milk}, {Egg,Oats}, {Milk,Oats}, {Egg,Milk,Oats} }

Part-B: Answer key (for your reference)

1m - Correct Closed frequent itemsets = {M , BM, CM, MO, EMO}

1m - Correct Maximal frequent itemsets = {BM, CM, EMO}

1m - Significance = Maximal frequent: compact form from which all frequent itemset generation, Closed frequent itemsets: additional support counts info

Q4	Recollect the case we have used in our HMM classes for predicting the weather by a guard posted in an underground installation. We have used the following model of the scenario:  We assumed that P(R0) = [0.5, 0.5]. Compute P(R2 | U1 = False, U2 = False, U3 = True,). Show all steps.	**5 M**

Q4

$P(R_2 | u_1 u_2) = \alpha f_{1:2} \times b_{3:3}$

① Find $f_{1:2}$

ie. $P(R_2 | u_1, u_2)$

day 0 $P(R_0) = \langle 0.5, 0.5 \rangle$

day 1 u_1 = False

$$P(R_1) = \sum_{r_0} P(R_1 | r_0) P(r_0)$$

$= \langle 0.5, 0.5 \rangle$ [5.3 Page 572]

$P(R_1 | u_1) = \alpha P(u_1 | R_1) P(R_1) =$

$\alpha \langle 0.1, 0.8 \rangle \langle 0.5, 0.5 \rangle$

$= \alpha \langle 0.05, 0.4 \rangle$

$= \langle 0.111, 0.889 \rangle$

day 2 u_2 = False

$$P(R_2 | u_1) = \sum_{r_1} P(R_2 | r_1) \cdot P(r_1 | u_1)$$

$= \langle 0.7, 0.3 \rangle \times 0.111 +$

$\langle 0.3, 0.7 \rangle \times 0.889$

$= \langle 0.3466, 0.6556 \rangle$

$P(R_2 | u_2) = \alpha P(u_2 | R_2) \; P(R_2 | u_1)$

$= \alpha \langle 0.1, 0.8 \rangle \langle 0.3444, 0.6556 \rangle$

$= \alpha \langle 0.034444, 0.52448 \rangle$

$= \langle 0.06161, 0.9390 \rangle$

$b_{3:3} = P(u_3 | R_2)$

Lets set $P(u_{4:3} | R_3) = P(. | x_t) \cdot 1$

$O_3 = \text{True}$

$P(u_3 | R_2) = \sum_{r_2}$

$P(u_3 | R_2) = \sum_{r_3} P(u_3 | r_3) \cdot P(. | r_3) \cdot P(r_3 | R_2)$

$= (0.9 \times 0.1 \times \langle 0.7, 0.3 \rangle) + (0.2 \times 1 \times \langle 0.3, 0.7 \rangle)$

$= \langle 0.69, 0.41 \rangle$

$P(R_2 | u_1, u_2, u_3) = \alpha \langle 0.06161, 0.9390 \rangle \cdot \langle 0.69, 0.41 \rangle$

$= \alpha \langle 0.75161, 1.349 \rangle$

Q5	Given the samples $X_0 = \{4, 4\}$, $X_1 = \{1, 4\}$, $X_2 = \{4, 1\}$, $X_3 = \{2, 2\}$, $X_4 = \{5, 3\}$, and $X_5 = \{3, 5\}$. Suppose that the samples are randomly clustered into two clusters $C_1 = \{X_0, X_1, X_3\}$ and $C_2 = \{X_2, X_4, X_5\}$. A. Apply one iteration of the K-means, and find a new distribution of samples in clusters. What are the new centroids? How can you prove that the new distribution of samples is better than the initial one? B. Apply one more iteration and explain the new distribution of samples is better than the previous one.	**(3+3=6Marks)**
Q5	A. 1. Compute Initial centroids $C_1 = \{X_0, X_1, X_3\}$ and $C_2 = \{X_2, X_4, X_5\}$ & J0 X_C1 = (X1 + X2 + X3)/3 Y_C1 = (Y1 + Y2 + Y3)/3 X_C2 = (X5 + X4 + X6)/3 Y_C2 = (Y4 + Y5 + Y6)/3 Calculate the Euclidian distance, J and assign the points to the clusters 2. Repeat Step 1 and get the new centroids. 3. J has reduced from step J0 to J1. J1 < J0 B. It is to be solved in the same as Part A. 1.Apply one iteration and get New centroids & J2 2. Comment that J has reduced from step J1 to J2.	