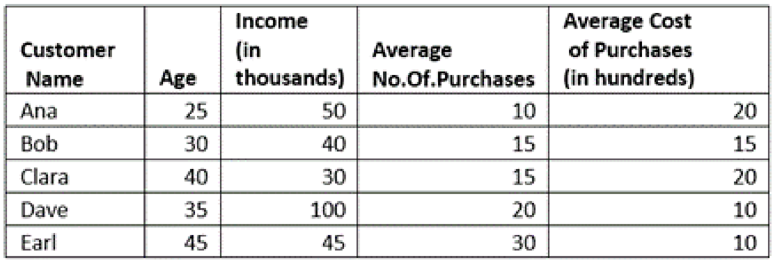
Course Number : PCAM ZC221
Course Title : Unsupervised Learning and Association Rule Mining
Type of Exam : Closed Book (Regular)
Weightage : 30 %
Date of Exam : 13/June/2020

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

Duration: 2 hours
Session : FN

Q1

Given below are certain transactions made by customers in a retail store.



Customer Name	Age	Income (in thousands)	Average No.Of.Purchases	Average Cost of Purchases (in hundreds)
Ana	25	50	10	20
Bob	30	40	15	15
Clara	40	30	15	20
Dave	35	100	20	10
Earl	45	45	30	10

To target customers who differ significantly from other customers, it is recommended to group the customers into a community for study. It's observed that a customer community is best portrayed by their purchasing capacity which is defined by given attributes ***Income*** and ***Average Cost of Purchases***.

[Note: *Use Manhattan distance only to calculate distance matrix. Use the given data as is i.e., No scaling or normalization is needed. Start the community building with customer '**Clara'**. Upload the screenshot of your written paper as an answer if required. Show step wise clustering.]*

A. A customer who is similar to at least 3 other customers is considered as a part of a community w.r.t to global data distribution. Apply DBSCAN with Epsilon = 20.

B. List which customers are classified as Border Point, Noise Point, and Core Point

6 + 1 = 7M

Q1

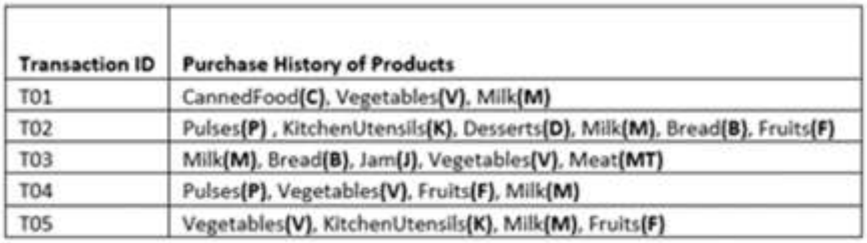
Distance matrix using manhattan (as specifed in QP) = 2m
MinPts = 4 (3+1-center point) = 1m
Correct ordered algorithm implementation Starting with "Clara"/"C" = 2m
Correct data structure updation at every steps as per exact algorithm: Core{},Border={},Noise={},CurrentClusterNeighborhood={} = 1m
Correct identification of below sets = 1m
Ans: Core={A,B}, border={C,E}, Noise={D}

	Ana	Bob	Clara	Dave	Earl
Ana	0	15	20	60	15
Bob	15	0	15	65	10
Clara	20	15	0	80	25
Dave	60	65	80	0	55
Earl	15	10	25	55	0

Iteration 1		**Iteration 2**		**Iteration 3**		**Iteration 4**		**Iteration 5**	
Clara (C)	Neighbors {A, B}	Randomly select any other customer **Ana (A)**	Neighbors {B,C,E}					The only unvisited customer is analysed **Dave (D)**	{}
Potential Noise point ={C}		Core Point = {A}	Border = {B, C, E}					**Noise point ={D}**	
			→	**Bob (B)**	Neigbors = {A,C,E}				
				Core Point = {A, B}	Border = {C,E} →	Note that A, C are already visited. **Earl (E)**	Neig hbors {A,B}		
						Border = {C,E}			

Q2

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



Transaction ID	Purchase History of Products
T01	CannedFood**(C)**, Vegetables**(V)**, Milk**(M)**
T02	Pulses**(P)** , KitchenUtensils**(K)**, Desserts**(D)**, Milk**(M)**, Bread**(B)**, Fruits**(F)**
T03	Milk**(M)**, Bread**(B)**, Jam**(J)**, Vegetables**(V)**, Meat**(MT)**
T04	Pulses**(P)**, Vegetables**(V)**, Fruits**(F)**, Milk**(M)**
T05	Vegetables**(V)**, KitchenUtensils**(K)**, Milk**(M)**, Fruits**(F)**

In the table given, items purchased by each customer (i.e. attribute "***Purchase History of Products"***) can be used to analyze which products were purchased together and hence are to be placed closer. Answer the following questions in this context to help better items placement.

A. Use apriori algorithm and find the frequent itemsets with 40% support. Show all the steps.
B. Find the interesting rule with more than 80% confidence containing the product "***KitchenUtensils***". Use your answer to part (A) to answer this.

6 + 2 = 8M

Q2

Conversion into binary transaction table & Usage of pruning principle= 1m
Support Thresohold >=2 identification = 0.5m
Level 1 = {B, F, K, M, P, V} = 1m
LEvel 2 = {BM, FK, FM, FP, FV, KM, MP, MV} (8 itemsets added) = 1m
LEvel 3 = {FKM, FMV, FPM} (3 itemsets added) & Usage of pruning principle = 1m
Level 4 = {} = Empty set NONE added & Usage of pruning principle = 1m
Final answer : Frequent Itemset = 0.5m
{ B, F, K, M, P, V , BM, FK, FM, FP, FV, KM, MP, MV , FKM, FMV, FPM }
Or
{Bread, Fruits, KitchenUtensils, Milk, Pulses, Vegetables , {Bread,Milk}, {Fruits,KitchenUtensils}, {Fruits,Milk}, {Fruits,Pulses}, {Fruits,Vegetables}, {KitchenUtensils,Milk}, {Milk,Pulses}, {Milk,Vegetables} , {Fruits,KitchenUtensils,Milk}, {Fruits,Milk,Vegetables}, {Fruits,Pulses,Milk}

Apriori Freq.set Generation:
Confidence Formula = 0.5m
Following Rules generation alone is expected where the first five needs to be identified as strong as per 80% confidence threshold:(1.5m)
K-->F, K-->M, K-->FM,MK-->F,FK-->M, MF-->K, F-->K, M-->K, F-->KM, M-->FK

Q3

In this question, you will propose versions of basic k-means discussed in the class [from PRML] to suit each of the following requirements. For each of the requirements, you have to suggest a suitable variation of basic k-means version, choice of initial cluster centers, and number of centers, convergence criteria along with your comment on convergence & performance tradeoffs.

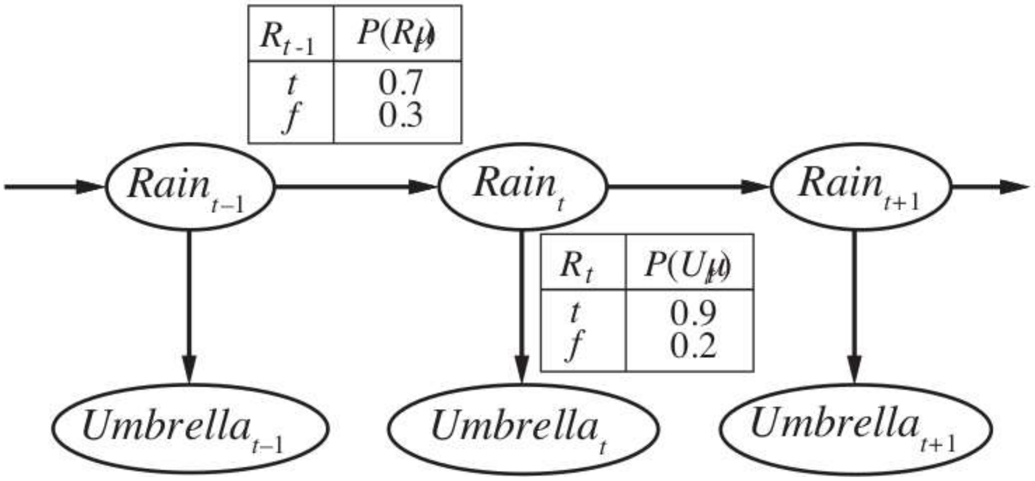
a. Given the performance of students over the past 20 years in the first year courses [say'**d**' courses], in BITS, find a suitable cutoff to regrade the students around k groups. It is expected that the clusters should give directions on students' academic inclinations which will help the university to propose specializations. Ignore academic aspects related to evaluation from consideration.

b. Assume that you have large data to be clustered around k clusters in a personal laptop. This has to be done in a few mins. Assume the data is in your disk and does not fit in primary memory entirely.

c. Assume that the data arrives in streams. Consider for example the click streams that google news page receives for example / or the orders that amazon continuously receives from across the globe. It is necessary to learn the evolution of customer behavior by monitoring how the clusters evolve over

3 x 2= 6 M

	time. Please note that in this setting, the data points cannot be seen/processed twice. Data must be processed as it arrives.	
Q3	a. Simple k means Cluster centers – Given (2) Convergence – when cost/error does not reduce significantly performance - O(tNK) b. Mini batch k means Cluster centers - random Convergence - User defined threshold [# of iterations, duration of execution] Performance: O(tNK) c. Streaming adaption of k means clustering [one pass k means, covered in contact session] Cluster centers - first k points / anything in this line. Convergence - Runs indefinitely. Performance - Constant time / record.	
Q4	a. How do you compare the objective functions of EM Algorithm and K-Means Algorithm? b. Provide a clustering scenario where EM performs better than K-Means algorithm. Explain.	**4 M**
Q4	a. Behavior of objective functions as the iterations progresses EM computes more parameters than k- means EM is probabilistic model while Kmeans is not. b. 1.Scenario requiring of soft clustering over hard clustering 2.Kmeans for circular shape and EM for elliptical shape clusters	
Q5	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:	**3+2= 5 M**

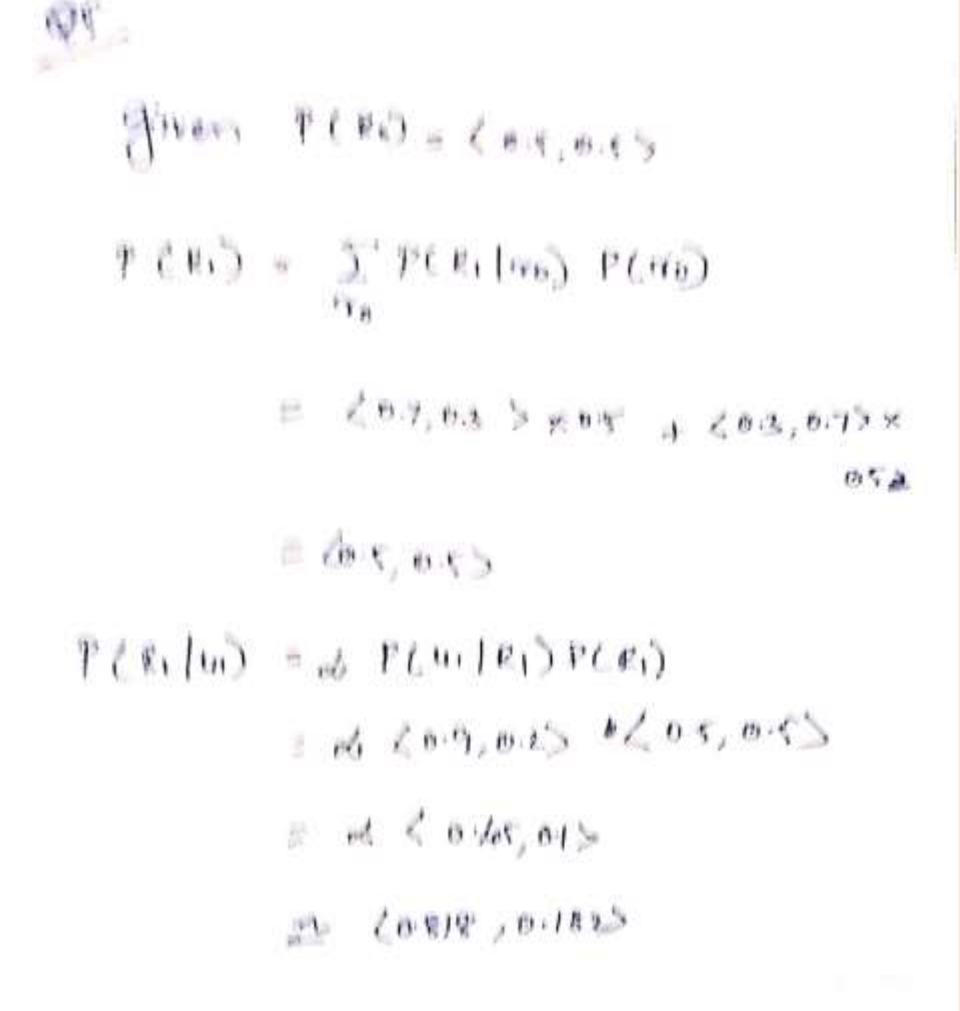


We assumed that $P(R_0) = [0.5, 0.5]$.

a. Compute $P(R_3 \mid U_1 = True)$

b. How accurate will be our prediction, $P(R_k \mid U_1 = True)$, when $k >> 1$? Explain.

Q5



Q5

Given $P(R_0) = \langle 0.5, 0.5 \rangle$

$P(R_1) = \sum_{r_0} P(R_1 \mid r_0)\, P(r_0)$

$= \langle 0.7, 0.3 \rangle \times 0.5 + \langle 0.3, 0.7 \rangle \times 0.5$

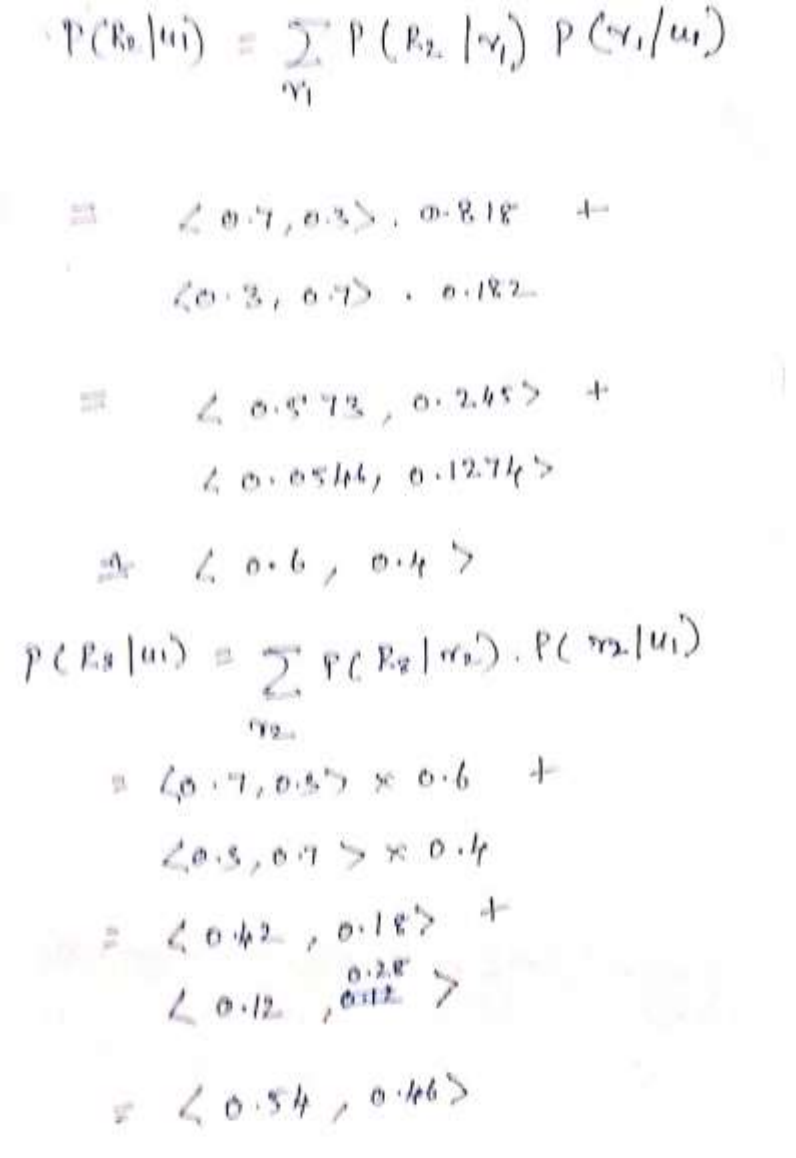
$= \langle 0.5, 0.5 \rangle$

$P(R_1 \mid u_1) = \alpha\, P(u_1 \mid R_1)\, P(R_1)$

$= \alpha \langle 0.9, 0.2 \rangle * \langle 0.5, 0.5 \rangle$

$= \alpha \langle 0.45, 0.1 \rangle$

$= \langle 0.818, 0.182 \rangle$



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

$$= \langle 0.7, 0.3\rangle . 0.818 +$$

$$\langle 0.3, 0.7\rangle . 0.182$$

$$= \langle 0.573, 0.245\rangle +$$

$$\langle 0.0546, 0.1274\rangle$$

$$= \langle 0.6, 0.4\rangle$$

$$P(R_3|u_1) = \sum_{r_2} P(R_3|r_2) . P(r_2|u_1)$$

$$= \langle 0.7, 0.3\rangle \times 0.6 +$$

$$\langle 0.3, 0.7\rangle \times 0.4$$

$$= \langle 0.42, 0.18\rangle +$$

$$\langle 0.12, \overset{0.28}{0.112}\rangle$$

$$= \langle 0.54, 0.46\rangle$$