DMA assignment 2

Data Mining And Analytics (SRM Institute of Science and Technology)

18CSE355T – Data Mining and Analytics

Assignment-2

M. Taruni
RA2111003010888
Section: M1

1) Consider the following contingency table summarizing the transactions wrt game and video purchases.

	Game	Non-Game	Total Game
Video	4000	3500	7500
Non video	2000	500	2500
Total	6000	4000	10000

Perform correlation analysis using lift and find how game and video are correlated ?

Sol) The lift between the occurence of A and B can be measured by computing

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{P(B/A)}{P(B)} = \frac{conf(A \Rightarrow B)}{sup(B)}$$

From the table, we can see that the probability of purchasing a computer game is $P(\{game\}) = 0.60$, the probability of purchasing a video is $P(\{video\}) = 0.75$ and the probability of purchasing both is $P(\{game, video\}) = 0.40$.

By the equation of lift,

$$\text{lift} = \frac{P(\{game, video\})}{P(\{game\}) \times P(\{video\})} = \frac{0.40}{0.60 \times 0.75}$$

$$\text{lift} = (0.89) < 1$$

→ Because this value is less than 1, there is a negetive correlation between the occurance of {game} and {video}.

→ The numerator is the likelihood of a customer purchasing both, while the denominator is what the likelihood would have been if the two purchases were completely independent.

→ such a negetive correlation cannot be identified by a support-confidence framework.

---

2) cluster the following eight points into three clusters
A1(2,10), A2(2,5), A3(8,4), A4(5,8), A5(7,5), A6(6,4), A7(1,2), A8(4,9).

Initial cluster centers are A1(2,10), A4(5,8) & A7(1,2)

Distance function - $P(a,b) = |x_2 - x_1| + |y_2 - y_1|$

Use k-means Algorithm to find the three cluster centers after the second iteration.

Also solve the problem using Euclidean distance measure.

Given points	Distance from center (2,10) of cluster 1	Distance from center (5,8) of cluster 2	Distance from center (1,2) of cluster 3	Point belongs to cluster
A1(2,10)	0	5	9	C_1
A2(2,5)	5	6	4	C_3
A3(8,4)	12	7	9	C_2
A4(5,8)	5	0	10	C_2
A5(7,5)	10	5	9	C_2
A6(6,4)	10	5	7	C_2
A7(1,2)	9	10	0	C_3
A8(4,9)	3	2	10	C_2

Distance btw A1(2,10) & C_1(2,10)

$P(A_1, C_1)$

$= |x_2 - x_1| + |y_2 - y_1|$

$= |2-2| + |10-10|$

$= 0$

Distance btw A1(2,10) & C_2(5,8)

$P(A_1, C_2)$

$= |x_2 - x_1| + |y_2 - y_1|$

$= |5-2| + |8-10|$

$= 5$

Distance btw A1(2,10) & C_3(1,2)

$P(A_1, C_3)$

$= |x_2 - x_1| + |y_2 - y_1|$

$= |1-2| + |2-10|$

$= 9$

Now, calculate the centroid of clusters.

For cluster-1,

we have only one point A1(2,10) in cluster-01

So cluster centroid remains same.

For cluster-2

$$\text{centroid} = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6,6)$$

For cluster-3,

centroid = $\left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5)$

This is the completion of first iteration.

Now, we calculate the distance of each point from each of the center of the three clusters. using the given distance function.

Given points	(2,10)	(6,6)	(1.5,3.5)	Point belongs to cluster
$A_1(2,10)$	0	8	7	C_1
$A_2(2,5)$	5	5	2	C_3
$A_3(8,4)$	12	4	7	C_2
$A_4(5,8)$	5	3	8	C_2
$A_5(7,5)$	10	2	7	C_2
$A_6(6,4)$	10	2	5	C_2
$A_7(1,2)$	9	9	2	C_3
$A_8(4,9)$	3	5	8	C_1

centroid of C_1 centroid C_2 centroid of C_3

$= \left(\frac{2+4}{2}, \frac{10+9}{2}\right)$ $= \left(\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right)$ $= \left(\frac{2+1}{2}, \frac{5+2}{2}\right)$

$= (3, 9.5)$ $= (6.5, 5.25)$ $= (1.5, 3.5)$

∴ After second iteration, The center of the three clusters are

$C_1(3, 9.5)$, $C_2(6.5, 5.25)$, $C_3(1.5, 3.5)$

Using Euclidean distance measure,

$$P(a,b) = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

Given Points	Distance to (2,10)	Distance to (5,8)	Distance to (1,2)	cluster	New cluster
$A_1(2,10)$	0	3.61	8.06	1	1
$A_2(2,5)$	5.00	4.24	3.16	3	3
$A_3(8,4)$	8.49	5	7.28	2	2
$A_4(5,8)$	3.61	0	7.21	2	2
$A_5(7,5)$	7.07	3.61	6.71	2	2
$A_6(6,4)$	7.21	4.12	5.39	2	2
$A_7(1,2)$	8.06	7.21	0	3	3
$A_8(4,9)$	2.24	1.41	7.62	2	1

Initial centroids

C_1 : (2,10)

C_2 : (5,8)

C_3 : (1,2)

New centroids

C_1 : (2,10)

C_2 : (6,6)

C_3 : (1.5,3.5)

Given Points	Distance to (2,10)	Distance to (6,6)	Distance to (1.5,3.5)	cluster	New cluster
$A_1(2,10)$	0	5.66	6.52	1	1
$A_2(2,5)$	5	4.12	1.58	3	3
$A_3(8,4)$	8.49	2.83	6.52	2	2
$A_4(5,8)$	3.61	2.24	5.7	2	2
$A_5(7,5)$	7.07	1.41	5.7	2	2
$A_6(6,4)$	7.21	2	4.53	2	2
$A_7(1,2)$	8.06	6.4	1.58	3	3
$A_8(4,9)$	2.24	3.61	6.04	2	1

Current centroids

C_1 : (2,10)
C_2 : (6,6)
C_3 : (1.5,3.5)

New centroids

C_1 : (3, 9.5)
C_2 : (6.5, 5.25)
C_3 : (1.5, 3.5)

Given points	Distance to (3,9.5)	Distance to (6.5,5.25)	Distance to (1.5,3.5)	cluster	New cluster
A_1 (2,10)	1.12	6.54	6.52	1	1
A_2 (2,5)	4.61	4.51	1.58	3	3
A_3 (8,4)	7.43	1.95	6.52	2	2
A_4 (5,8)	2.50	3.13	5.70	(2)	(1)
A_5 (7,5)	6.02	0.56	5.70	2	2
A_6 (6,4)	6.26	1.35	4.53	2	2
A_7 (1,2)	7.76	6.39	1.58	3	3
A_8 (4,9)	1.12	4.51	6.04	1	1

current centroids

C_1 : (3,9.5)
C_2 : (6.5, 5.25)
C_3 : (1.5, 3.5)

New centroids

C_1 : (3.67, 9)
C_2 : (7, 4.33)
C_3 : (1.5, 3.5)

Given points	Distance to (3.67,9)	Distance to (7,4.33)	Distance to (1.5,3.5)	cluster	New cluster
A_1 (2,10)	1.94	7.56	6.52	1	1
A_2 (2,5)	4.33	5.04	1.58	3	3
A_3 (8,4)	6.62	1.05	6.52	2	2
A_4 (5,8)	1.67	4.18	5.70	1	1
A_5 (7,5)	5.21	0.67	5.70	2	2
A_6 (6,4)	5.52	1.05	4.53	2	2
A_7 (1,2)	7.49	6.44	1.58	3	3
A_8 (4,9)	0.33	5.55	6.04	1	1

$A_1(2,10)$, $A_4(5,8)$, $A_8(4,9)$ belongs to cluster 1.
$A_2(2,5)$, $A_7(1,2)$ belongs to cluster 3.
$A_3(8,4)$, $A_5(7,5)$, $A_6(6,4)$ belongs to cluster 2.

3) Describe each of the following clustering algorithms in terms of the following criteria:

i) Shapes of clusters that can be determined
ii) input parameters that must be specified and
iii) limitations.

a) k-medoids:

i) k-medoids can identify clusters of various shapes, including non-linear shapes.

ii) The number of clusters k, need to be specified. It also requires an initial set of medoids.

iii) k-medoids can be computationally expensive for large datasets. It is also sensitive to noise and outliers, and the choice of initial methods can impact the final clustering.

b) CLARA:

i) CLARA is designed for handling arbitary shaped clusters and large datasets.

ii) The number of clusters, k, as the sample size need to be provided as inputs.

iii) CLARA may not be suitable for high-dimensional data and can be computationally expensive, especially for large datasets. It also requires a fixed sample size, which can impact its accuracy.

c) BIRCH:

i) BIRCH can efficiently handle clusters of various shapes, particularly in large datasets.

ii) The branching factor, the threshold, and the number of clusters must be specified as inputs.

iii) BIRCH may not perform well with very large dimensions and can be sensitive to the choice of parameters. It can also struggle with clusters of varying densities.

d) CHAMELEON:

i) CHAMELEON is effective in identifying clusters with various shapes, densities and scales.

ii) Parameters such ~~that~~ the number of clusters and the distance function need to be provided.

iii) CHAMELEON's performance may degrade when dealing with high-dimensional data. It also requires parameter tuning, and the clustering results can be sensitive to the choice of parameters.

e) DBSCAN:

i) DBSCAN can identify clusters of arbitary shapes including those of non-linear and complex structures.

ii) The neighbourhood distance and the minimum number of points need to be provided as inputs.

iii) DBSCAN may struggle with datasets that have varying densities, and it may not perform well in high-dimensional spaces. Also, it can struggle with different density levels within the dataset.

4) For the one dimensional dataset {5, 8, 13, 28, 35}, perform hierarchial clustering and plot the dendogram to visualize it (use single and complete linkage).

i) single

	$P_1(5)$	$P_2(8)$	$P_3(13)$	$P_4(28$	$P_5(35)$
(5) P_1	0				
(8) P_2	(3) → min	0			
(13) P_3	8	5	0		
(28) P_4	23	20	15	0	
(35) P_5	30	27	22	7	0

ii)

	(P_1, P_2)	P_3	P_4	P_5
(P_1, P_2)	0			
P_3	(5) → min	0		
P_4	20	15	0	
P_5	27	22	0	

$d(P_3, (P_1, P_2)) = \min(d(P_3, P_1), d(P_3, P_2)) = \min(8, 5) = 5$

$d(P_4, (P_1, P_2)) = \min(d(P_4, P_1), d(P_4, P_2)) = \min(23, 20) = 20$

$d(P_5, (P_1, P_2)) = \min(d(P_5, P_1), d(P_5, P_2)) = \min(30, 27) = 27$

iii)

	(P_1, P_2, P_3)	P_4	P_5
(P_1, P_2, P_3)	0		
P_4	(15) → min	0	
P_5	22	(7)	0

iv) (P_1, P_2, P_3) (P_4, P_5)

(P_1, P_2, P_3) — 0

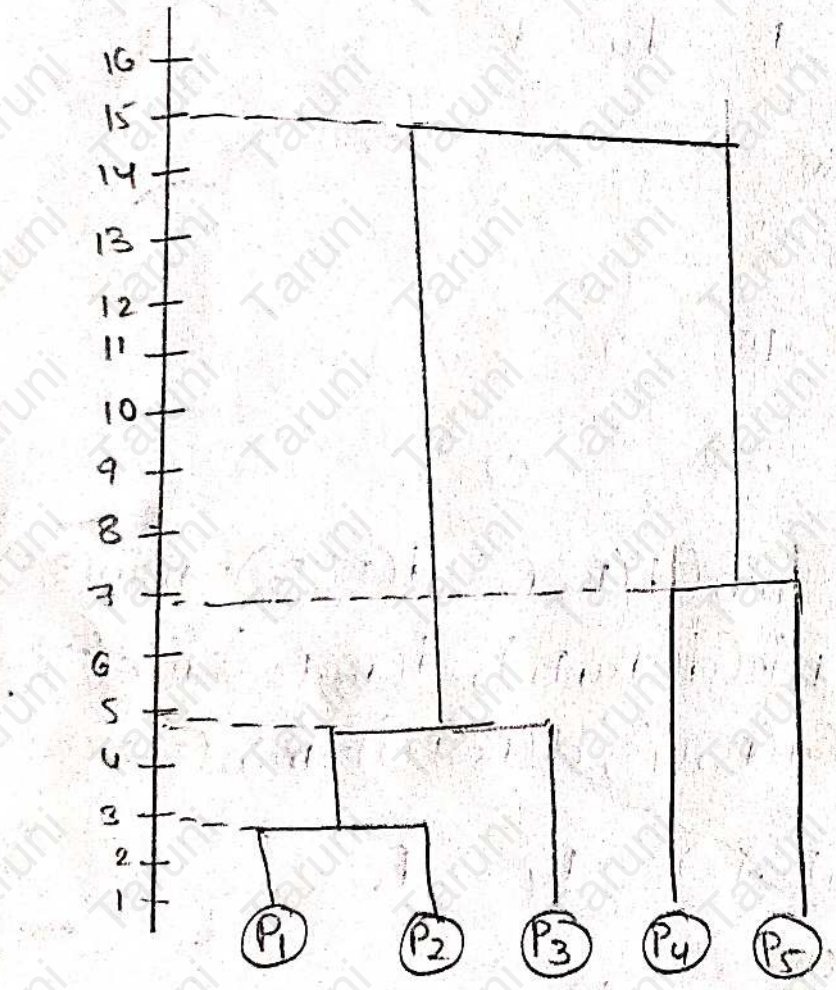
(P_4, P_5) — 15 — 0

$$d((P_4,P_5),(P_1,P_2,P_3)) = \min(d(P_4,P_1), d(P_4,P_2), d(P_4,P_3), d(P_5,P_1), d(P_5,P_2), d(P_5,P_3))$$

$$= \min(23, 20, 15, 30, 27, 22)$$

$$= 15$$

dendogram



complete linkage

i)

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	(3) → min	0			
P_3	8	5	0		
P_4	23	20	15	0	
P_5	30	27	22	7	0

ii)

	(P_1,P_2)	P_3	P_4	P_5
(P_1,P_2)	0			
P_3	8	0		
P_4	23	15	0	
P_5	30	22	(7) → min	0

$d(P_3,(P_1,P_2)) = \max(d(P_3,P_1), d(P_3,P_2)) = \max(8,5) = 8$

$d(P_4,(P_1,P_2)) = \max(d(P_4,P_1), d(P_4,P_2)) = \max(23,20) = 23$

$d(P_5,(P_1,P_2)) = \max(d(P_5,P_1), d(P_5,P_2)) = \max(30,27) = 30$

iii)

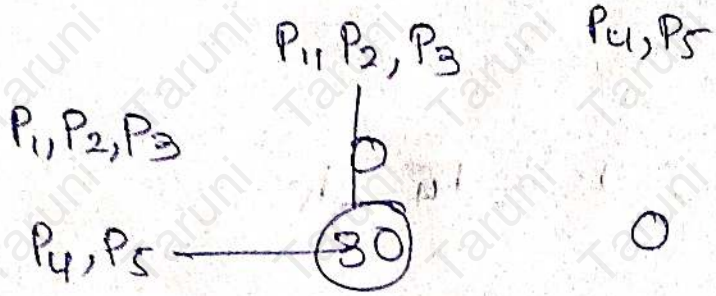
	(P_1,P_2)	P_3	(P_4,P_5)
(P_1,P_2)	0		
P_3	(8) → min	0	
(P_4,P_5)	30	22	0

$d((P_4,P_5),(P_1,P_2)) = \max(d(P_4,P_1), d(P_4,P_2), d(P_5,P_1), d(P_5,P_2))$

$= \max(23, 20, 30, 27) = 30$

$d(P_3,(P_1,P_2)) = \max(d(P_3,P_1), d(P_3,P_2)) = \max(8,5) = 8$

iv)

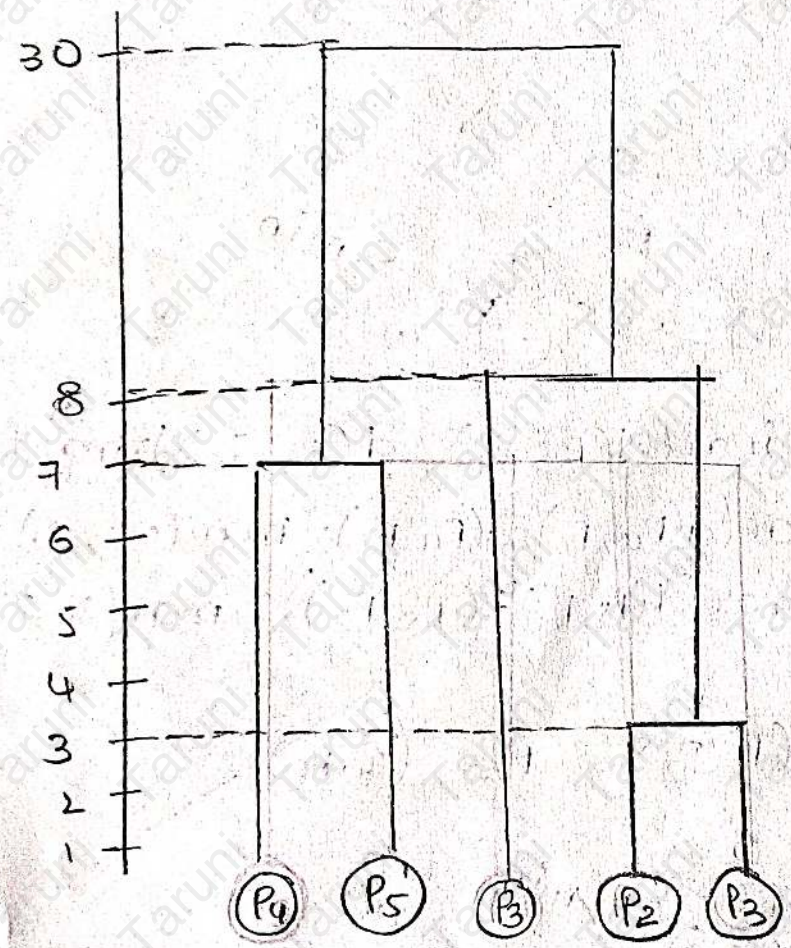


$$d((P_4,P_5),(P_1,P_2,P_3)) = \max(d(P_4,P_1), d(P_4,P_2), d(P_4,P_3), d(P_5,P_1), d(P_5,P_2), d(P_5,P_3))$$

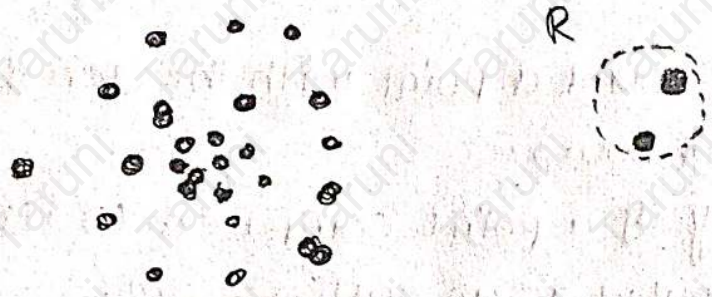
$$= \max(23, 20, 15, 30, 27, 22)$$

$$= 30$$

Dendogram



5) Explain the proximity-based outlier detection methods.



→ Proximity based methods assume that an object is an outlier if the nearest neighbors of the object are far away in feature space, that is, the proximity of the object to its neighbours significantly deviates from the proximity of most of the other objects to their neighbours in the same dataset.

→ In the above figure if we model the proximity of an object using its nearest neighbours, then the objects in Region R are substantially different from other objects in the data set.

→ There are two types of proximity based outlier detection

i) Distance based outlier detection

A representative method of proximity-based outlier detection uses the concept of the distance based outliers.

Formally, let r ($r \geq 0$) be a distance threshold and π ($0 < \pi \leq 1$) be a fraction threshold. An object, O, is a $DB(r, \pi)$-outlier if

$$\frac{||\{O' | dist(O, O') \leq r\}||}{||D||} \leq \pi$$

2) Density-based outlier detection

General idea

→ Compare the density around a point with the density around its local neighbours.

→ The relative density of a point compared to its neighbours is computed as an outlier score.

→ Approaches generally differ in how to estimate density.

Basic assumption

→ The density around a normal data object is similar to the density around its neighbours.

→ The density around an outlier is considerably different to the density around its neighbors.

→ local reachable density of o':

$$lrd_k(o) = \frac{||N_k(o)||}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$$

→ LOF of an object o is the average of the ratio of local reachability of o and those of o's k-nearest neighbours.

$$LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{||N_k(o)||} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$$

→ $$LOF_k(p) = \frac{\sum_{o \in KNN(p)} \frac{lrd_k(o)}{lrd_k(p)}}{Card(KNN(p))}$$

6) Write in detail about any one Data Mining application

→ Financial Data Analysis is one of the most important application of data mining.

→ Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality.

→ Design and construction of data warehouses for multidimensional data analysis and data mining.

- view the debt and revenue changes by month, by region, by sector and by other factors.
- Access statistical information such as, max, min, total, average, trend etc

→ Loan payment prediction / consumer credit Policy analysis

- Feature selection and attribute relevance ranking
- Loan payment performance
- consumer credit rating.

→ classification and clustering of customers for targeted marketing.

→ multidimensional segmentation by nearest neighbor, classification, decision trees etc to identify customer groups or associate a new customer to an appropriate customer group.

→ Detection of money laundering and other financial crimes.

→ Integration of multiple DBs (e.g., bank transactions, federal / state crime history DBs)