Unit - 4

Business Intelligence

- Linear Regression
- Apriori Algorithm Qs
- Clustering Theory
- Data mining
- Data Mining Outliers
- K Mediods
- K - Means Clustering

* Apriori Algorithm

Q1. Suppose there are 4000 customer transactions. Calculate the Support, Confidence and lift for two products, Biscuits & Chocolate. Out of the 4000 transactions, 400 contain biscuits, 600 contain chocalate, and these 600 transactions include 200 with both biscuits and chocolate.

$$\text{Support (Biscuits)} = \frac{\text{Transactions w/ biscuits}}{\text{Total transactions}}$$

$$= \frac{400}{4000} = 10\%$$

$$\text{Confidence} = \frac{\text{Transactions w/ biscuits \& chocolate}}{\text{Transactions w/ biscuits}}$$

$$= \frac{200}{400} = 50\%$$

$$\text{lift} = \frac{\text{confidence}}{\text{support}} \quad \frac{50}{10} = 5$$

(Q) Apply the Apriori Algorithm on the following product set.

Transaction Id	Rice	Pulso	Oil	Milk	Apple
t1	1	1	1	0	0
t2	0	1	1	1	0
t3	0	0	0	1	1
t4	1	1	0	1	0
t5	1	1	1	0	1
t6	1	1	1	1	1

Let min-sup ~~50%~~ ~~60%~~ >50%.

Ans

Product	Frequency
R	4
P	5
O	4
M	4
A	3

⇒ prune the apples

⇓

Product	Frequency
R	4
P	5
O	4
M	4

apply selfjoin

⇒

Product	Frequency
RP	4
RO	3
RM	2
PO	4
PM	3
OM	2

prune RO, PM, RM & OM

Product	Frequency
RP	4
PD	4
RD	3
PM	3

≥ 3

$\Rightarrow$

RPD is the customer's set of product

Example Apply the Apriori algorithm on

TID	items
100	1 3 4 6
200	2 3 5
300	1 2 3 5
400	1 5 6

min-sup = 0.5

Ans

TID	1	2	3	4	5	6
100	1	0	1	1	0	1
200	0	1	1	0	1	0
300	1	1	1	0	1	0
400	1	0	0	0	1	1

$0.5 \times \text{num}(TID) = 2$

Itr 1

Item	Freq
1	3
2	2
3	3
4	1
5	3
6	2

prune $\longrightarrow$

Item	Freq
1	3
2	2
3	3
5	3
6	2

join ⇒

Item	Freq
{1,2}	1
{1,3}	2
{1,5}	2
{1,6}	2
{2,3}	2
{2,5}	2
{2,6}	0
{3,5}	2
{3,6}	1
{5,6}	1

prune ⟶

Item	Freq
{1,3}	2
{1,5}	2
{1,6}	2
{2,3}	2
{2,5}	2
{3,5}	2

join

↓

Item	Freq
{1,3,5}	
{1,3,6}	X
{1,5,6}	X
{2,3,5}	

prune by ⇒ checking subsets

Item	Freq
{1,3,5}	1
{2,3,5}	2

prune by ⇒ minsup

Item	Freq
{2,3,5}	2

⇓

cannot join anymore

{2,3,5} is the Frequent Item set

Example: Use the apriori algorithm to find frequent itemsets.

Let min. sup = 2

T.ID	Items
T_1	1, 2, 5
T_2	2, 4
T_3	2, 3
T_4	1, 2, 4
T_5	1, 3
T_6	2, 3
T_7	1, 3
T_8	1, 2, 3, 5
T_9	1, 2, 3

T.ID	1	2	3	4	5
T_1	1	1	0	0	1
T_2	0	1	0	1	0
T_3	0	1	1	0	0
T_4	1	1	0	1	0
T_5	1	0	1	0	0
T_6	0	1	1	0	0
T_7	1	0	1	0	0
T_8	1	1	1	0	1
T_9	1	1	1	0	0

Item	Frequency
I_1	6
I_2	7
I_3	6
I_4	2
I_5	2

min sup = 2

⇓ no pruning needed

join

Item	Frequency
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

prune ⇒

Item	Frequency
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

join

Item	
$\{I_1, I_2, I_3\}$	
$\{I_1, I_2, I_5\}$	
$\{I_1, I_3, I_5\}$	X
$\{I_2, I_3, I_4\}$	X
$\{I_2, I_3, I_5\}$	X
$\{I_2, I_4, I_5\}$	X

⇒

Item	Frequency
I_1, I_2, I_3	2
I_1, I_2, I_5	2

⇓ join

$\{I_1\ I_2\ I_3\ I_5\}$ X

The frequent itemsets are $\{I_1, I_2, I_3\}$
$\{I_1, I_2, I_5\}$

* Clustering?

Cluster = a collection of data objects which are similar/dissimilar

→ cluster analysis involves - given a set of data points, partition them into a set of groups, which are as similar as possible

→ a type of unsupervised learning.

→ Good clustering gives:

(i) High intra-class similarity: cohesive within clusters

(ii) Low Inter-class similarity: distinctive between clusters

* Distance Measures

→ used to measure how the similarity of 2 elements is calculated & it influences the shape of the clusters.

These include:

(i) Euclidean Distance

(ii) Manhattan distance

(iii) Max Norm $d(x,y) = \max |x_i - y_i|$

(iv) Inner product space

(v) Hamming Distance

* Applications of Cluster Analysis

(i) intermediate step for other data mining tasks

(ii) data summarization, compression and reduction

(iii) dynamic trend prediction

(iv) multimedia data analysis

* Clustering Methods

A. Partioning Approach

→ construct various k partitions & then evaluate them by some criterion

eg. k-means, k-medoids, CLARANS

B. Hierachical Approach

→ create a hierarchical decomposition of the set of data using some criterion

→ classified using agglomerative (bottom-up) or divisive (top-down)

eg. Diana, Agnes, BIRCH, CAMELO

C. Density Based methods

→ based on connectivity and density functions

→ used to filter out noise & outliers

eg. DBSCAN, OPTICS

D. Grid-based Methods

→ based on multiple level granularity structure

→ quantize object space into finite no. of cells that form a grid structure

eg. STING, CLIQUE

* K - Means Clustering Algorithm

(1) Cluster the following 8 points into 3 clusters

A1 (2,10) A4 (5,8) A7 (1,2)

A2 (2,5) A5 (7,5) A8 (4,9)

A3 (8,4) A6 (6,4)

Consider the cluster centers to be:

A1 (2,10)

A4 (5,8)

A7 (1,2)

and the distance metric is defined as: $P(a,b) = |x_2 - x_1| + |y_2 - y_1|$

	Distance from C1 A1(2,10)	C2 A4(5,8)	C3 A7(1,2)	Cluster
A1 (2,10)	0+0 = 0	3+2 5	1+8 9	C1
A2 (2,5)	0+5 = 5	3+3 = 6	1+3 = 4	C3
A3 (8,4)	6+6 = 12	3+4 7	7+2 9	C2
A4 (5,8)	3+2 = 5	0+0 0	4+6 = 10	C2
A5 (7,5)	5+5 = 10	2+3 = 5	6+3 = 9	C2
A6 (6,4)	4+6 16	1+4 5	5+2 7	C2
A7 (1,2)	1+8 9	4+6 10	0+0 0	C3
A8 (4,9)	2+1 3	1+1 2	3+7 10	C2

C1

A1(2,10)

C2

A3(8,4)

A4(5,8)

A5(7,5)

A6(6,4)

A8(4,9)

C3

A2(2,5)

A7(1,2)

Find new cluster centers

C1 = (2,10) C2 = (6,6) C3 (1.5, 3.5)

Distance from

Point	C1 (2,10)	C2 (6,6)	C3 (1.5, 3.5)	belongs to Cluster
A1(2,10)	0+0 = 0	4+4 = 8	0.5 + 6.5 = 7	C1
A2(2,5)	0+5 = 5	4+1 = 5	0.5 + 1.5 = 2	C3
A3(8,4)	6+6 = 12	2+2 = 4	6.5 + 0.5 = 7	C2
A4(5,8)	3+2 = 5	1+2 = 3	3.5 + 4.5 = 8	C2
A5(7,5)	5+5 = 10	1+1 = 2	5.5 + 1.5 = 7	C2
A6(6,4)	4+6 = 10	0+2 = 2	4.5 0.5 = 5	C2
A7(1,2)	1+8 = 9	5+4 = 9	0.5 + 1.5 = 2	C3
A8(4,9)	2+1 = 3	2+3 = 5	2.5 + 5.5 = 8	C1

The clusters are:

C_1	C_2	C_3
$A_1(2,10)$	$A_3(8,4)$	$A_9(2,5)$
$A_8(4,9)$	$A_4(5,8)$	$A_7(1,2)$
	$A_5(7,5)$	
	$A_6(6,4)$	

Repeat this process.

*K-Medoids Clustering → also called PAM

Partitioning Around Medoid

A medoid is a point in the cluster whose dissimilarities with all the other points in the cluster are minimum

The cost in the K-Medoids algorithm is given as:

$$C = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

Algorithm:

1. Initialize k random points out of the n data points as medoids
2. Associate each data point to the closest medoid by using any common distance metric methods
3. While the cost decrease: For each medoid m, for each data point o, which is not a medoid

   (i) swap m and o, associate each data point to the closest medoid and recompute the cost

   (ii) If the total cost is more than that in the previous step, undo the swap

* Apply K-medoids clustering on the following dataset

X	Y
8	7
3	7
4	9
9	6
8	5
5	8
7	3
8	4
7	5
4	5

Let (8,5) and (4,5) be the initial medoids

Point	Distance from (4,5)	Distance from (8,5)	Cluster
(8,7)	4+2 = 6	0+2 = 2	C2
(3,7)	1+2 = 3	5+2 = 7	C1
(4,9)	0+4 = 4	4+4 = 8	C1
(9,6)	5+1 = 6	1+1 = 2	C2
(8,5)	4+0 = 4	0	C2
(5,8)	1+3 = 4	3+3 = 6	C1
(7,3)	3+2 = 5	1+2 = 3	C2
(8,4)	4+1 = 5	0+1 = 1	C2
(7,5)	3+0 = 3	1+0 = 1	C2
(4,3)	0	-	C1

C1 (4,5)

(3,7) 1+2 = 3

(4,8) 0+3 = 3

(4,6) 1+1 = 2

C1 cost = 8

C2 (8,5)

(8,7) = 0+2

(6,2) = 2+3

(4,0) = 4+5

(5,3) =

(5,1) =

(3,1) =

C2 cost

$C_1 = (4,5)$

(3,7) = 1+2 = 3

(4,9) = 0+4 = 4

(5,8) = 1+3 = 4

cost C_1 = 11

$C_2 = (8,5)$

(9,6) = 1+1 = 2

~~(8,8) = 0+0 = 0~~

(7,3) = 1+2 = 3

(8,4) = 0+1 = 1

(7,5) = 1+0 = 1

(8,7) = 0+2 = 2

= 9

Total cost = 20

choose some random point find distance for (4,5) and 0 ← random point

compute cost again.

* Outliers

→ An outlier is a point that lies abnormally far away from the other values in a dataset.

→ A data point is an outlier if it is 1.5 the interquartile range greater than the third quartile or 1.5 the interquartile range less than the first quartile.

Eg1 25th percentile (Q1) income = 15,000

75th percentile (Q3) income = 120,000

Find the range of outliers

IQR = 105,000

lower boundary = $Q_1 - 1.5 \times IQR$

= 15,000 - 1.5 * 105,000

= -142,500

Upper boundary = Q3 + 1.5 IQR

= 120,000 + 1.5 * 105,000

= 277,500

eg2 25th percentile (Q1) for how long individuals can hold their breath = 15 seconds

75th percentile (Q3) = 75 seconds

IQR = 60 sec

Lower boundary $= Q_1 - 1.5\,IQR$

$= 15 - 1.5(60)$

$= -75$

Upper boundary $= Q_3 + 1.5\,IQR$

$= 75 + 1.5(60)$

$= 165$

eg3 - 25th percentile ticket sale = $2 million

75th percentile ticket sales = $15 million

IQR = $13 million

Lower boundary $= Q_1 - 1.5\,IQR$

$= 2 - 1.5(13)$

$= -17.5$

Upper boundary $= Q_3 + 1.5\,IQR$

$= 15 + 19.5$

$= 34.5$