ARM:-

Data Warehouse and Data Mining – Video Lecture Series (For B.Tech, MCA, M.Tech)

Association Rule Mining: (Rules)

"ARM", Also Called as Market Basket Analysis (MBA) and Affinity Analysis.

↳ Set of items in a transaction is Called Market Basket.

↳ Mostly used in RETAIL.

↳ if 'A' then 'B' { $A \Rightarrow B$ } → Consequent.
↳ Product ↳ antecedent

= **Support:** (S). Percentage (%) of transactions (T) -that contains both 'A' and 'B'.

$("A" \Rightarrow "B") = P(A \cap B)$ } measures frequency of association.

= **Confidence:** (C). In a transaction set 'T' if 'C' is -the % of -times 'B' is present in all -the transactions containing 'A'. (Strength).

Conditional Probability. { $C = P(B/A) = \frac{P(A \cap B)}{P(A)}$ } Strength of association

Parameters:-

(i) Finding all items that appears frequently in transaction. } min. Support Count.

(ii) Finding Strong associations among frequent items } Confidence.

Data Warehouse and Data Mining – Video Lecture Series (For B.Tech, MCA, M.Tech)

Problems in ARM:-

- i) Levels of frequency of appearance determination.
- ii) Finding strong associations among frequent items.

Functions of ARM:-

- i) Finding set of items - that has significant impact on business.
- ii) Collating info^n from numerous tr^x.
- iii) Generating rules from Counts in tr^x.

Strengths of ARM:-

i) Easy interpretation.
ii) Easy-to start
iii) Flexible data formats
iv) Simplicity.

(1,2,3,4)

(1,2), (1,3), (1,2,5) ...

Weakness:-

i) Exponential Growth in computations
ii) bumping
iii) Rule Selection
iv) Rare items } frequent items

Data Warehouse and Data Mining – Video Lecture Series (For B.Tech, MCA, M.Tech)

Apriori Algorithm: Idea is to generate Candidate itemsets of a given size and then scan dataset to check if their Counts are really large. The Process is iterative.

(i) All Singleton itemsets are Candidates in the first pass. Any items with less than specified Support Value is eliminated.

(ii) Two member Candidate itemsets. {Numericals}

(iii) Three " " . . . .

eliminated.

(iv) Frequent itemsets Constitutes Set of frequent itemsets.

(v) Generate Association Rules which have Confidence Values greater than or equal to Specified min. Confidence.

Eg:-

Tid	Items
1	2,3
2	1,3,5
3	1,2,4
4	2,3

min Support = 2

Items	Support
1	2
2	3
3	3
4	1
5	1

eliminated.

Itemsets	Support
{1,2}	1
{1,3}	1
{2,3}	2

{2,3} → 2

Easy Engineering Classes – Free YouTube Lectures

For Engineering Students of GGSIPU, UPTU and Other Universities, Colleges of India

Data Warehousing and Data Mining- Lecture Series [Mumbai Univ, GTU, UPTU, GGSIPU, Pune Univ & others]

Ques.) For the following Given Transaction Data-set, Generate Rules using Apriori Algorithm. Consider the Values as SUPPORT = 50% and CONFIDENCE = 75%.

Transaction ID.	Items Purchased
1	Bread, cheese, Egg, Juice
2	Bread, cheese, Juice
3	Bread, Milk, Yogurt
4	Bread, Juice, Milk
5	cheese, Juice, Milk

$$\text{Support(Bread)} = \frac{n_{\text{Bread}}}{n}$$

Frequent Item set

	Items	Frequency	Support
1.	Bread	4	4/5 = 80%
2.	Cheese	3	3/5 = 60%
	Egg	1	1/5 = 20%
3.	Juice	4	4/5 = 80%
4.	Milk	3	3/5 = 60%
	Yogurt	1	1/5 = 20%

Remove these ∵ there Support is less than 50%

(APRIORI ALGORITHM SOLVED PROBLEM-1) (2)

Data Warehouse and Data Mining [Mumbai Univ, Pune Univ, GTU, Lecture Series [UPTU, GGSIPU, DU, PTU and other Universities]

Make 2-Items Candidate set and Wrote their Frequency.

Item Pairs	Frequency	Support
(Bread, Cheese)	2	2/5 = 40%
(Bread, Juice)	3	3/5 = 60%
(Bread, Milk)	2	2/5 = 40%
(Cheese, Juice)	3	3/5 = 60%
(Cheese, Milk)	1	1/5 = 20%
(Juice, Milk)	2	2/5 = 40%

There Support ≥ 50 %

For Rules → (Bread, Juice) (1)

→ (Cheese, Juice) (2)

(1) (Bread, Juice)

(Bread → Juice) (Juice → Bread)

$$\text{Confidence}(A \to B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

1.) (Bread → juice) = $\frac{S(B \cup J)}{S(B)} = \frac{3.5}{5.4} = \frac{3}{4} = 75\%$

2.) (Juice → Bread) = $\frac{3.5}{5.4} = 75\%$

(2) → (Cheese → juice) = $\frac{3.5}{5.3} = 100\%$

(juice → Cheese) = $\frac{3.5}{5.4} = 75\%$

All the Rules are Good.

Data Warehouse and Data Mining [Mumbai Univ, GTU, UPTU]
Lecture Series [GGSIPU, Pune Univ, PTU and other University]

(C2)

Itemset Pairs	Frequency	Support	
(I_1, I_2)	4	4/9 = 44.4%	
(I_1, I_3)	4	4/9 = 44.4%	
(I_1, I_4)	1	1/9 = 11.1%	®
(I_1, I_5)	2	2/9 = 22.2%	
(I_2, I_3)	4	4/9 = 44.4%	
(I_2, I_4)	2	2/9 = 22.2%	
(I_2, I_5)	2	2/9 = 22.2%	
(I_3, I_4)	0	0/9 = 0%	®
(I_3, I_5)	1	1/9 = 11.1%	®
(I_4, I_5)	0	0/9 = 0%	®

(C3)

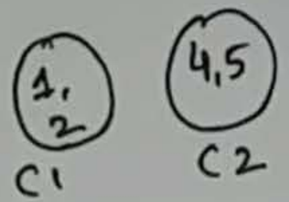
(Itemset)	(Frequency)	(Support)
(I_1, I_2, I_3)	2	2/9 = 22.2%
(I_1, I_2, I_5)	2	2/9 = 22.2%

Confidence.

- $(I_1, I_2) \rightarrow (I_5) = 2/4 = 50\% \rightarrow$ X
- $(I_1, I_5) \rightarrow (I_2) = 2/2 = 100\%$ ✓
- $(I_2, I_5) \rightarrow (I_1) = 2/2 = 100\%$ ✓
- $(I_1) \rightarrow (I_2, I_5) = 2/6 = 33\% \rightarrow$ X
- $(I_2) \rightarrow (I_1, I_5) = 2/7 = 29\% \rightarrow$ X
- $I_5 \rightarrow (I_2, I_1) = 2/2 = 100\%$ ✓

Major Data Mining Techniques:-

(i) **Cluster Detection:** clustering means forming groups.



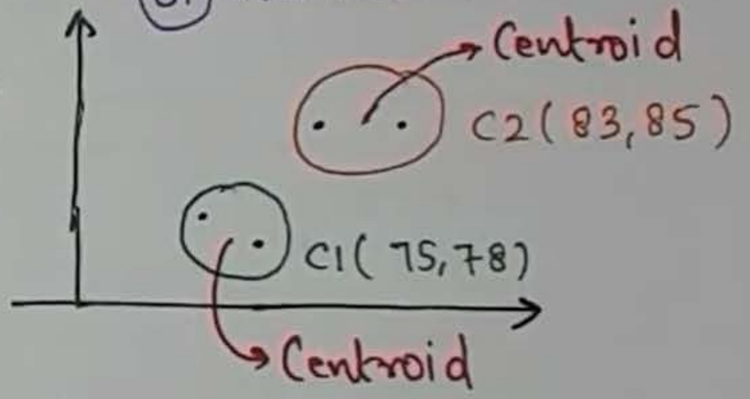
S.No.	Subject Codes	Marks
1	01	85
2	02	78
3	03	75
4	04	83

→ Earliest data mining techniques.

→ Unsupervised Learning → we don't know the class Labels. → no. of Labels.

↳ Algo searches for groups or clusters of data elements that are similar to one another.

↳ K-means Clustering } Example:-



Distance Measures:-

↳ (i) Euclidian distance

$$\text{Euclidian}(A,B) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$ } Imp.

Advantages:- (i) not affected by addⁿ of new object.

(ii) No need not worry about Signs.

(iii) Computation Process is Very Simple.

Applⁿ:-

(i) Medical:- Establish taxanomy of disease, Cure and symptoms.

(ii) WWW:- Social N/w Communities.

(iii) Seismology:- Epicentre of earthquake

(iv) Business/Marketing:- Targeted Customer finding.

[CLUSTERING: K-MEANS SOLVED PROBLEM-1]

Easy Engineering Classes – Free YouTube Lectures

For Engineering Students of GGSIPU, UPTU and Other Universities, Colleges of India

Data Warehousing and Data Mining - lecture Series [Mumbai Univ, GTU, UPTU, GGSIPU, Pune Univ and others]

Ques) Divide the given Sample Data in two (2) clusters using K-Means Algorithm [Euclidean Distance].

$C_1 \rightarrow \{1,4,5,6,7,8,9,10,11,12\}$
$C_2 \rightarrow \{2,3\}$ ANS..

	Height (H)	Weight (W)
1	185	72 ✓
2	170	56 ✓
3	168	60
4	179	68 ✓
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

$$\sqrt{(X_H - H_1)^2 + (X_W - W_1)^2}$$

O.V

X_H: observed value; H_1: centroid value; W_1: centroid value

i) Initialize two clusters.

	H	W	Centroid
C1	185	72	(185, 72)
C2	170	56	(170, 56)

$C_1 \left(\frac{185+179}{2}, \frac{72+68}{2}\right) = [182, 70]$ (C1)

E.D of Row 3:
- C1: $\sqrt{(168-185)^2 + (60-72)^2} = \sqrt{289+144}$ [(20.80)]
- C2: $\sqrt{(168-170)^2 + (60-56)^2} = \sqrt{4+16}$ [(4.48)]

$C_2 \left(\frac{170+168}{2}, \frac{60+56}{2}\right)$

$C_2 [169, 58]$

E.D of R4:
- C1: $\sqrt{(179-185)^2 + (68-72)^2}$ [6.32]
- C2: $\sqrt{(179-169)^2 + (68-58)^2}$ = [14.14]