Association Rule Mining

Prepared By: Kimia Aksir

What we will Learn..

- Association Rule Mining (ARM)
- Criterion of ARM and example
- ARM: Market Basket Analysis
- Different example scenarios of Market Basket Analysis
- Measures used in ARM:
  - Support
  - Confidence
  - Lift
- Work-through Example of ARM
- Advantages, Disadvantages and Applications of ARM

Association Rule Mining (ARM)

Association Rule Mining is about finding frequent patterns, correlations, association or, casual structure among the observations/datapoints from a transactional or relational database and/or, other data repositories.

Association Criterion in Association Rule Mining

If (Antecedent) then (Consequent)

Association Rule Mining (ARM) Example

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

For example,

If **Bread,** then **Milk**
Bread ⇒ Milk

If **Soda,** then **Chips**
Soda ⇒ Chips

If **Bread,** then **Jam**
Bread ⇒ Jam

ARM: Market Basket Analysis

Analyzing shopping basket of a Customer:

- Items customer(s) place in their shopping basket
- Association/correlation of the item(s) customers are buying together
- Frequency of the item from the item set a customer/customers are buying

Market Basket Analysis and it's ultimate purposes…

- Where should jam be placed in the shop to maximize its sale?
- Are fruits bought with milk or, typically it's banana that is brought with milk?
- Placing Eggs close to pasta is better than placing Eggs close to bread?
- A new Jam brand has been launched, which customers should we target to send the advertisement to (in store/online)?
- …..

Association Rule for Market Basket Analysis

Association Rule Mining is primarily used when you want to identify an association between different items in a set, then find frequent patterns from the transactional records.

Association Measures in ARM: Support

The frequency/occurrence percentage of an item/itemset is considered as “support”.

This measure is used to determine the popularity of the item and can be expressed as a percentage, a ratio or a fractional number.

Support = ***Number of times item or itemset is involved in the transaction / Total transaction***

Association Measures in ARM: Support Example

The support for the item "Bread" is,

Support (Bread) = (4/8) *100

= 50%

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Association Measures in ARM: Confidence

The likelihood of association or, how much the rule is valid is determined by Confidence. It can be expressed as a percentage, a ratio or a fractional number.

Assuming, we have a pattern of buying Y after X is bought.

$$Confidence(X \Rightarrow Y) = Support(X,Y) \,/\, Support(X)$$

Association Measures in ARM: Confidence Example

The Confidence for the item “Jam” is bought when “Bread” is bought,

Confidence (**Bread ⇒ Jam**) = 4/4 *100

= 100%

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Association Measures in ARM: Lift

Lift is the ratio of Confidence (Association) and Expected Confidence.

Expected Confidence is the support of Y for ***(X ⇒Y).*** So,

$$Lift(X \Rightarrow Y) = Confidence(X \Rightarrow Y) \,/\, Support(Y)$$

Hence, Lift can be considered as a measure of 'Interestingness' of a rule.

Association Measures in ARM: Lift Example

The Confidence for the item “Jam” is bought when “Bread” is bought,

Confidence (**Bread ⇒ Jam**) = 100

Support **(Jam)** = 62.5

So, Lift (**Bread ⇒ Jam)** = 100/62.5 = 1.6

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Interpretation of “Lift” in Association

IF (Lift > 1) for (*X* ⇒Y) :

The rule has strong positive impact. Means people buy the items “X” and “Y” together than buying “Y” alone.

IF (Lift < 1) for (*X* ⇒Y) :

The rule has a negative/inverse impact. Means the items X and Y are substitute of each other.

IF (Lift = ~1) for (*X* ⇒Y) :

The rule will not impact much as it’s going to happen anyways irrespective of any association.

Algorithms used in Market Basket Analysis

- **Apriori Algorithm**
- AIS
- SETM Algorithm
- FP Growth

Etc..

Apriori Algorithm

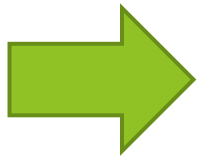
Using the three measures “Support”, “Confidence” and “Lift” we will now find the Associations from the Example set given here:

TID	Items
1	Bread, Peanuts, Milk, Fruit, Jam
2	Bread, Jam, Soda, Chips, Milk, Fruit
3	Steak, Jam, Soda, Chips, Bread
4	Jam, Soda, Peanuts, Milk, Fruit
5	Jam, Soda, Chips, Milk, Bread
6	Fruit, Soda, Chips, Milk
7	Fruit, Soda, Peanuts, Milk
8	Fruit, Peanuts, Cheese, Yogurt

Apriori Algorithm: Calculating Support First step

Creating the candidates(1-item set) and calculating Support for the items. Here, we assuming the support threshold to be 40%

Items	Support (%)
{Bread}	4/8*100 = 50
{Peanuts}	4/8*100 = 50
{Milk}	6/8*100 = 75
{Fruit}	6/8*100 = 75
{Jam}	5/8*100 = 62.5
{Soda}	6/8*100 = 75
{Chips}	4/8*100 = 50
{Cheese}	1/8*100 = 12.5
{Yogurt}	1/8*100 = 12.5



Frequent Items	Support (%)
{Bread}	4/8*100 = 50
{Peanuts}	4/8*100 = 50
{Milk}	6/8*100 = 75
{Fruit}	6/8*100 = 75
{Jam}	5/8*100 = 62.5
{Soda}	6/8*100 = 75
{Chips}	4/8*100 = 50

Apriori Algorithm: Calculating Support Second step

Creating the candidates (2-items set) and calculating Support for the items. Here, we assuming the support threshold to be 40%

Items	Support (%)
{Bread, Peanuts}	1/8*100 =12.5
{Bread, Milk}	3/8*100 =37.5
{Bread, Fruit}	2/8*100 =25
{Bread, Jam}	4/8*100 =50
{Bread, Soda}	2/8*100 =25
{Bread, Chips}	3/8*100 = 37.5
{Peanuts, Milk}	3/8*100 = 37.5
{Peanuts, Fruit}	4/8*100 =50
{Peanuts, Jam}	2/8*100 =25
{Peanuts, Soda}	2/8*100 =25
{Peanuts, Chips}	0/8*100 =0
{Milk, Fruit}	5/8*100 =62.5
{Milk, Jam}	4/8*100 =50
{Milk, Soda}	5/8*100 =62.5
{Milk, Chips}	3/8*100 =37.5
{Fruit, Jam}	3/8*100 =37.5
{Fruit, Soda}	4/8*100 =50
{Fruit, Chips}	2/8*100 =25
{Jam, Soda}	4/8*100 =50
{Jam, Chips}	3/8*100 =37.5
{Soda, Chips}	4/8*100 =50

Frequent Items	Support (%)
{Bread, Jam}	4/8*100 =50
{Peanuts, Fruit}	4/8*100 =50
{Milk, Fruit}	5/8*100 =62.5
{Milk, Jam}	4/8*100 =50
{Milk, Soda}	5/8*100 =62.5
{Fruit, Soda}	4/8*100 =50
{Jam, Soda}	4/8*100 =50
{Soda, Chips}	4/8*100 =50

Apriori Algorithm: Calculating Support Third step

Creating the candidates (3-items set) and calculating Support for the items. Here, we assuming the support threshold to be 40%

Items	Support (%)
{Bread, Jam, peanuts}	1/8*100 =12.5
{Bread, jam, milk}	3/8*100=37.5
{Bread, jam, chips}	3/8*100=37.5
{Bread, jam, soda}	3/8*100=37.5
{Bread, jam, fruit}	2/8*100=25
{peanuts, fruit, bread}	2/8*100=25
{peanuts, fruit, milk}	3/8*100=37.5
{peanuts, fruit, jam}	2/8*100=25
{peanuts, fruit, soda}	2/8*100=25
{Milk, Fruit, bread}	2/8*100=25
{Milk, Fruit, jam}	3/8*100=37.5
{Milk, Fruit, chips}	2/8*100=25
{Milk, Fruit, soda}	3/8*100=37.5
{Milk, jam, chips}	2/8*100=25
{Milk, jam, peanuts}	2/8*100=25
{Milk, jam, soda}	3/8*100=37.5
{Milk, Soda, bread}	2/8*100=25
{Milk, Soda, chips}	3/8*100=37.5
{Milk, Soda, peanuts}	2/8*100=25
{Fruit, Soda, bread}	1/8*100=12.5
{Fruit, Soda, jam}	2/8*100=25
{Fruit, Soda, chips}	2/8*100=25
{Jam, Soda, peanuts}	1/8*100=12.5
{Jam, Soda, fruit}	2/8*100=25
{Soda, Chips, bread}	3/8*100=37.5
{Soda, Chips, jam}	2/8*100=25

Here, the support none of the item sets is greater than the threshold (40%), so we cannot move forward with the 3-items set anymore.

This means, we have to find association for the frequent items from the 2-items set only.

Apriori Algorithm: Calculating Confidence

Rules	Support of both (X U Y) %	Support of X %	Confidence %
If Bread then Jam	50	50	50/50*100=100
If Jam then Bread	50	62.5	50/62.5*100=80
If Peanuts then Fruit	50	50	50/50*100=100
If Fruit then Peanuts	50	75	50/75*100=66.67
If Milk then Fruit	62.5	75	62.5/75*100=83.33
If Fruit then Milk	62.5	75	62.5/75*100=83.33
If Milk then Jam	50	75	50/75*100=66.67
If Jam then Milk	50	62.5	50/62.5*100=80
If Milk then Soda	62.5	75	62.5/75*100=83.33
If Soda then Milk	62.5	75	62.5/75*100=83.33
If Fruit then Soda	50	75	50/75*100=66.67
If Soda then Fruit	50	75	50/75*100=66.67
If Jam then Soda	50	62.5	50/62.5*100=80
If Soda then Jam	50	75	50/75*100=66.67
If Soda then Chips	50	75	50/75*100=66.67
If Chips then Soda	50	50	50/50*100=100

Considering the confidence less than 70% will be filtered out.

Apriori Algorithm: Calculating Lift

Rules	Support of Y %	Confidence %	Lift
If Bread then Jam	62.5	50/50*100=100	1.6
If Jam then Bread	50	50/62.5*100=80	1.6
If Peanuts then Fruit	75	50/50*100=100	1.33
If Milk then Fruit	75	62.5/75*100=83.33	1.11
If Fruit then Milk	75	62.5/75*100=83.33	1.11
If Jam then Milk	75	50/62.5*100=80	1.06
If Milk then Soda	75	62.5/75*100=83.33	1.11
If Soda then Milk	75	62.5/75*100=83.33	1.11
If Jam then Soda	75	50/62.5*100=80	1.06
If Chips then Soda	75	50/50*100=100	1.33

The rules with the highest lift will be considered as having a higher probability of correct associations between items from the transactions in the dataset.

To understand the sale of individual item “Lift” sometimes makes better sense, hence we use it.

Advantages of ARM (In terms of Apriori Algorithm)

- The execution is straight forward
- Memory usage is smaller in this algorithm than any other algorithms used for this ARM

Disadvantages of ARM (In terms of Apriori Algorithm)

- At a time allows to have a single Support Threshold and Confidence Threshold only
- This is sometimes considered as a slow process as it scans the database several times

Applications of Association Rule Mining

- Market basket Analysis
- Medical Diagnosis
- Planning profitable/useful services using the Census data held by Government
- Analyzing Protein sequence in cell

and many more….


Do you have any
QUESTIONS?