

Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 04
Association Rules

Welcome to the 4th lecture. Now we start our algorithms for mining different type of patterns. So, the first pattern we will consider is something known as association rules. So, let me explain what this patterns mean.

(Refer Slide Time: 00:40)

Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

$\{Diaper\} \rightarrow \{Beer\}$,
 $\{Milk, Bread\} \rightarrow \{Eggs, Coke\}$,
 $\{Beer, Bread\} \rightarrow \{Milk\}$,

Implication means co-occurrence, not causality!

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, the origin of this kind of pattern was one of the earliest use of data mining in the retail shop. Say for example, you are going you are gone to a super market a mall, and you have bought some items. So, say for example, I note down say the bill after a person has bought something in his basket, you carry a basket, and you buy something and I am noted them down for example, you can see a table, where this rows are the different transaction. So, TID 1 is the transaction I'd 1, the say the one customers transactions.

Next row is the next customer transactions and so on. And along with that what is noted is that, what are the items purchased by that customer. So, you can see that in this table customer 1 has bought bread and milk, customer 2 has bought bread and diaper and beer and eggs customer 3 has bought milk and diaper and beer and coke and so on. So, these types of transactions are called market basket transactions. So, they act sort of consist of

2 part 1 is the ID of the transaction one particular customer, the second is the list of items purchased by the customer. So, note down this term. So, these things this milk, bread, diaper, eggs, coke, they are different items available in some store available in some store. So, a market basket transaction would consist of a set of items corresponding to a particular transaction. So, I can think of an item set.

So, in the first row it is a 2 elements say bread and milk and second 3rd 4th fifth row are 4 elements sets 4 item sets. You can imagine that if you sort of look at all the purchases in a supermarket, there will be millions of this kind of transaction entries, there will be millions of entries. So, maybe everyday thousand of person come, and do this kind of transactions. So, if you see over a period of say 1 or 2 years, there will be an enormous amount of data a very large this kind of transaction market basket database where everything would be noted down. And then what this data mining in fact, this this particular technology was developed by a group in IBM, where there a software which develop for this kind of analysis in the late 80s and early 90s where they had try to explore beyond the normal database analysis transaction analysis, they looked at this kind of database table.

And they find out that if you look at lot of such transactions, you can observe certain patterns in the form of rules we call them association rules. What are this rules? Maybe we can find out a pattern that people who buy diaper also buy beer. People who buy milk and bread, also buy eggs and coke say. So, each of these is an example of what is called a association rule. There are 2 parts of each rule, the left hand side is a item set; that means, a collection of items same diaper or milk and bread and so on. The right-hand side is also a collection of items it is a eggs and coke and so on. So now the question is can we find some association between these 2 items say, can we say that whenever people buy milk and bread people also buy egg and coke.

So, kind of this left-hand item set and the right-hand item set, they would be always found together in a particular market basket or a transactions or a transaction. So, you can see in the table to the left that for example, milk and bread whenever people bought in the 4th and fifth transaction, they also bought may be coke. In this case now in this case may be diaper both of the types they bought. Similarly, whenever people bought diaper in the second and 3rd and 4th and 5th they also bought beer, most of the time 3 out of 4 times not always, but most of the time; that means, there is a cooccurrences

between the left-hand item set and the right-hand item set. What is a coherence means? So, if you look at lot of market basket the trolleys.

Where people buy a item and then check out, if you look at lot of this trolleys that customer has bought if you have found out the left-hand side, you might have also found the right-hand side is also there they are bought together they are bought together left-hand side and right-hand side that together purchased. Now this kind of patterns had some commercial significance. For example, I could have say given some discount of opportunity, that if you buy beer and bread you can buy milk at a discounted rate.

Or I can place the racks where beer and bread kept together with the racks where milk can milk is kept. So, people will buy them. So, this kind of association rules have sort of significance. Note that I am explaining this with respect to a market basket, but this kind of coherence happen in many other instances. For example, um if I buy a train ticket I always buy following up by a taxi ticket or something. Or if in some in some stock some set of stocks go up another set of stock also go up.



So, this kind of associations are present in many type of data. So, what we will see now is that given this set of transaction millions of this kind of records of this people what people have purchased. Can I find out that what are this kind of valid association rules that one can extract from this market basket transactions. What are the valid rules that were, note that everything is data driven you cannot you cannot hypotheses a rule you can you can say that in this data the that is the fundamental principle of data mining that; in this data this kind of pattern is present. Similarly, in this kind of transaction this kind of association is observed or present.

(Refer Slide Time: 09:09)

Definition: Frequent Itemset

- **Itemset**
 - A collection of one or more items
 - Example: {Milk, Bread, Diaper}
 - k-itemset
 - An itemset that contains k items
- **Support count (σ)**
 - Frequency of occurrence of an itemset
 - E.g. $\sigma(\{\text{Milk, Bread, Diaper}\}) = 2$
- **Support**
 - Fraction of transactions that contain an itemset
 - E.g. $s(\{\text{Milk, Bread, Diaper}\}) = 2/5$
- **Frequent Itemset**
 - An itemset whose support is greater than or equal to a *minsup* threshold

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

IIT KHARAGPURNPTEL ONLINE
CERTIFICATION COURSES

So, let me now tell you a set of steps to find this. The first steps towards discovering this association rules is to detect what is called a frequent item set.

That means item set I already defined it is a collection of item say bread butter diaper and eggs this is a collection of items, and frequent items sets means I want to find out which set of items are popular. They are purchased by a lot of persons. They are popular for example, bread and milk is a popular product popular item set. Whereas, may be cricket bat and ball is not. So, popular item not. So, frequent item most of them will do not buy cricket bat and ball. May be bread and milk is more commonly purchased. So, here is the definition a collection of one or more items is we called a item set. Particularly I would call it a k item set, if it is a set of k items. So, milk bread diaper is a 3-item set.

So now for each of this item set, I kind of measure it is popularity. It is frequency, how common it is how commonly it is purchased. That count how many times it is purchased is called the support count of that item set denoted by sigma within bracket item set, it is called a support count; that means, out of total all the transactions; that you are observing how many times in how many transactions this 2 items set are present. So, you can see that if you consider this 5 transactions, your support count of this item set milk bread butter diaper there item set is actually 2, you can see it appears in milk bread diaper appears in in may be item number 5 item number 4 item 4 and 5 contains this 3 item sets.

In other words, this 3-item set is a subset of this 4 and 5 that is 2. (Refer Time: 12:02) out of 2 out of 5. What we do is that we kind of normalize this support count by the total number of transactions.

So, the support of an item set a denoted by s is the fraction of transactions in which this appear. So, in this example milk bread diaper the support is 2 upon 5. You might express this as a percentage also, may be 40 percent is the support in this case 2 of 5. So, next what we do is that, we set a threshold of support call it mean support, mean sup. May be 40 percent, may be 50 percent, may be 10 percent, and we say that any item set whose support value is greater than this mean sup, we denote it we call it frequent item set. So, may be your mean sup if it is say 30 percent, 30 percent then milk bread diaper would be considered as a frequent item set because it appears in 40 percent of the time which is more than the 30 percent.

So, to summarize a frequent item set is nothing but a set of items which appear more than mean sup fraction of the total transactions. Means sup has to be set by the user. If your set of transactions are very large may be in millions a typical value of mean sup is may be one percent one percent.

(Refer Slide Time: 14:37)

• Association Rule

- An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
- Example:
 $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

• Rule Evaluation Metrics

- Support (s)
 - ◆ Fraction of transactions that contain both X and Y
- Confidence (c)
 - ◆ Measures how often items in Y appear in transactions that contain X

Example:
 $\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us define what is a association rule. So, as I have already mentioned it is it has a left-hand item set and a right-hand item set x and y in this case; here is an example, x is milks milk and diaper y is beer. And we say that in order that this particular association

rule x associated with y is valid, then 2 conditions has to be satisfied. One is that both x and y has to be frequent item set; that means, say if we consider the rule milk and diaper associated with beer, the set milk and diaper has to appear in more than mean sup transactions, fraction of transactions.

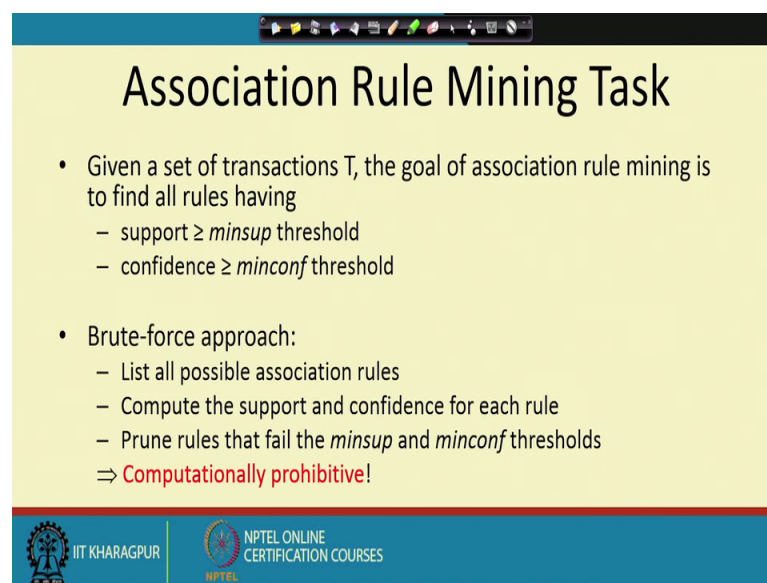
The item set beer also has to appear in more than mean sup fraction transaction. Both has to appear say if mean sup is 30 percent more than 30 percent. So, the first condition 2 condition to be satisfied first condition is called the support condition. It says that both x and y has to have more than mean sup support. In other words, both x and y has to be frequent item sets that is the definition of frequent item sets. There is a second condition that this should satisfy. It is called the confidence condition. What it says that in addition to x and y appear in lot of transaction being frequent, in most of the transactions that x appear y should also appear. So, suppose if I say a condition threshold of 90 percent, it should say that in 90 percent of this transactions were x appears y should also appear.

So, if I define with a explain with this example, suppose I have this rule; milk and diaper to be a this association rule, the support factor of the left-hand side is 2 of 5. So, total 5 transactions in 2 out of them milk and diaper and beer appear 4 and transaction 4 and 5 they appear. So, 0.4 is the support factor. What is the, note that one thing I have make a small note is that the condition for a support I stated that both x and y should appear in lot of transaction. So, how do I relate that to the definition of frequent items are? It says that x and y taken together in this case milk diaper and beer together this 3 items set 2 item set plus 1 item set 3 item set, the support of that is greater than the threshold.

So, support of milk diaper left hand side plus, beer right hand side, this 3-item set is 0.4×2 by 5 the second factor is out of how many times. The left-hand side appear what fraction of them that right hand side also appear. So, it is the ratio of the support count of all 3 together left right milk diaper and beer divided by only left, the support of only left. So, support of milk diaper beer divided by the support of only milk and diaper. So, let us see I have already calculated that milk diaper beer appear 2 times out of 5 and milk and diaper appears in transaction 3 transaction 4 and transaction 5. So, 3 times, 3 is the denominator. So, the confidence factor of this rule is $2 \text{ by } 3.67$. So, if I look at this particular rule milk and diaper associated with beer, let me tell you the steps first I support calculate the support of that rule.

How do I calculate the support of that rule? I calculate the support of that rule as I take both the left-hand side count how many times it appear out of total transaction. Then I calculate the confidence it is the ratio of how many times all 3 appear to how many times only the left-hand side appear, only x appear times x y both appear divided by the times only x appear. And support is the time x and y both appear. So, this is the support as of this rule and c of this rule. Now what I will do is that I will call a rule a valid association rule, if both the support and confidence that I have previously calculated a greater than some user supplied threshold. In this case mean sup and mean conf.

(Refer Slide Time: 20:58)



The slide is titled "Association Rule Mining Task" and contains the following content:

- Given a set of transactions T , the goal of association rule mining is to find all rules having
 - support \geq *minsup* threshold
 - confidence \geq *minconf* threshold
- Brute-force approach:
 - List all possible association rules
 - Compute the support and confidence for each rule
 - Prune rules that fail the *minsup* and *minconf* thresholds

⇒ **Computationally prohibitive!**

The slide footer includes the IIT KHARAGPUR logo and the NPTEL ONLINE CERTIFICATION COURSES logo.

If both are greater than this I call it a valid association rule.

So, suppose my mean sup threshold is 30 percent 0.3 and mean confidence threshold is 0.5, then this rule would qualify as an association rule. It qualifies as an association rule. So, what is the meaning of this? What is the meaning of this? Physical meaning; that means, for one particular rule to be a valid association rule, it not only has to be popular it not only has to be frequent, it also has to be valid; that means, it has also have high confidence; that means, most of the time left hand side is bought right hand side is also bought. So, even though it may be true that most of the time cricket bat is bought cricket ball is also bought, it is not a valid association rule, because the total number of times cricket bat and ball are purchased is much less is below the support threshold.

Whereas, may be milk and may be rice they may be individually bought a large number of time they may support satisfied a support threshold, but they are together. Not bought this is not satisfy the confidence threshold. So, that is also not an association rule. So, it has to be popular and it has to be confidence valid. As you can understand it is only then when the rule will have some meaningful pattern some meaningful commercial use. So now, let us say that how do we discover this rules. I just give you this table; I just give you a big transaction table like this, millions of such transactions. From them how do I extract this rules discover this rules. In other words how do I find the pair of item sets x and y , and check whether they satisfy the support and threshold a confidence criteria. The simplest way to do that as you can think of is that you take all possible combination of items.

So, if there are say 100 items in your that are sold in the shop in the store, you take all possible values of x and y milk bread butter, milk diaper coke all possible combination you take. And try to form all possible rules will call them as candidate rules, all possible rules. And for each of this rule you run through the database. You run through the database run through this table, and check whether the left and right-hand side, how many times they appear this is the counts how many time the left and right-hand side appear. And after finding how many times they appear you find out they support value from this definition, you find out their confidence value if the support and confidence values are above the threshold values, then they are rules then they are rules. The well this approach is definitely going to give you correct rules, but the problem is it is computationally prohibitive.

Why because suppose there are 100 items, how many possible x you can make, you can make 100 choose 2 to the power 100 x is similarly 2 to the power 100 y is or 2 to the power 100 into all possible 2 to the power 100 plus 1 different combination of x and y is you can take. So, x and y combinations you can take, and that is a large number 2 to the power 100 very, very large number and each of this possible rules, we have to go through the entire database which itself is very big and count how many times each of this rule appear. So, that would take a enormous time, and not just 100 a store would have thousands of items. So, it is prohibited you cannot do it that way. So, what do I do? I have to some find some more intelligent way of doing it.

So, what we will do is that we will explain you an algorithm called apriori algorithm, which will need not have so many computation need not try out all possible x and y's. So, in summary, what I do I have defined what is a association rule, I have defined support and confidence, I have defined what is a frequent items set. And I have defined if the support and confidence criteria satisfied I call that as a rule. And we have see there is seen that finding out all possible combinations brought force is not possible computationally. So, next in my next lecture, I will explain an intelligent algorithm which will reduce this computational complexity, and do it in a much faster way.

Thank you for now.