NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-81
Frequent Itemset Mining

Prof: Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

Okay, anyway so remember we actually spoke about frequent pattern mining very briefly at the very first lecture right, where I was introducing different machine learning tasks to you right. So this is a form of unsupervised learning and the statisticians call this as I told you that also, statisticians call this as bump hunting right. So you have a remarkably flat probability distribution okay.

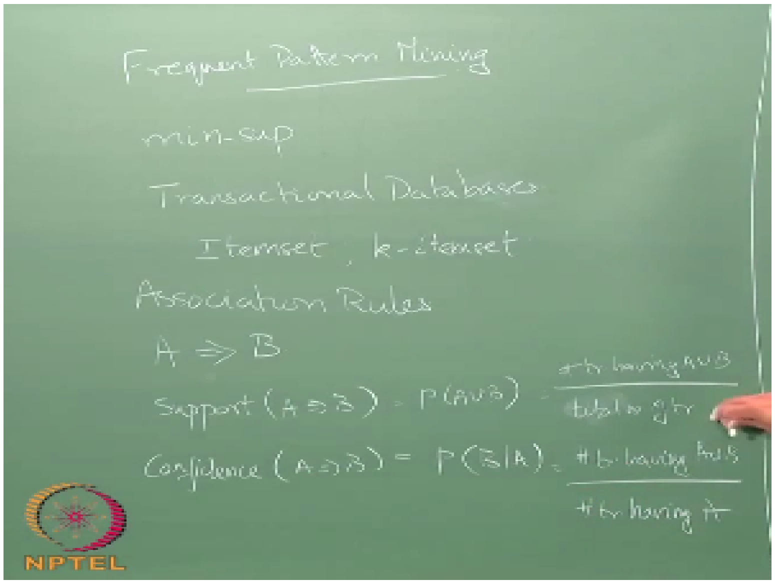
Well there are small bumps somewhere, so what does it mean slightly more frequent in the data than the rest right. So that is essentially what bump hunting is. So I have a very large pace right I have an exponentially large set of possible outcomes right, and I am going to have an extremely low probability of seeing any one outcome right. But there will be small bumps here and there which are places which are slightly more frequent than what I would see normally right.

So what in the modern context right, so something like Amazon's logs if you look at who bought what and Amazon right. So if you think about it Amazon has millions and millions of transactions right. So say suppose they have in a month let us say they have like 10 million transactions on Amazon that if somebody bought say 10,000 copies of the same item not somebody like some item was sold 10,000 times on Amazon right.

That is a very frequent occurrence, but think about what fraction of 10 million is 10,000 1%, 0.1%, 0.01%, 0.1% right. That is why I call this bump hunting it is remarkably flat right, so I have a huge inventory in my Amazon base, my medic Amazon catalog has a huge inventory right, and I am looking for frequent items there which is we like something sold 10,000 times is frequent for me.

Even though the overall transaction is 10 million right, so this is essentially what I am talking about when I talk about frequent pattern mining okay. So it is a take the frequent path here with a pinch of salt right, but all the examples and illustrations I will give you a frequent will be like 50% of the data or something that is because I cannot draw 10 million things on the board right. But in reality when you actually use these kinds of things the numbers would be the fractions will be very different just keep that in mind.

(Refer Slide Time: 03:07)



Frequent Pattern Mining

min-sup

Transactional Databases

Itemset , k-itemset

Association Rules

$A \Rightarrow B$

$\text{Support}(A \Rightarrow B) = P(A \cup B) = \frac{\#\text{tr having } A \cup B}{\text{total no of tr}}$

$\text{Confidence}(A \Rightarrow B) = P(B|A) = \frac{\#\text{tr having } A \cup B}{\#\text{tr having } A}$

So I am, so frequent patterns are those that are above a certain, suppose above a certain minimum support that I am looking for in a, so what is the support of a pattern in a database, that is the support count okay. So support of an item is essentially the fraction of times it occurs right so take the support count the number of times it occurs divided by the total number of items it gives you the support of the item right.

So I will call an item as B or all call a pattern as being frequent if it is above the minimum support. So minimum support is a parameter that I define sorry it is less than one, it will be less than 1, but it is a parameter that I define right. So occasionally what people do is, they actually translate the wind support also into account and essentially you take the fraction you multiplied by the total number of items it gives you a count.

So see sometimes easier to think of it as okay, I have a transaction of say, I have a total database of 10 possible transactions, I look for a mint support too right. So that is like 20%, but then you could think of it that way as well. So mint support is essentially the minimum support level at which I will consider something as frequent right. So classically right frequent pattern mining was applied to transactional databases. So where I have a collection of transactions right transaction essentially could be things like okay, these are the items you bought together right, or these are the items you checked out together from a bookstore or library okay, these are the items you borrowed from a library together, some kind of a transaction.

So something get went from A to B okay. So that is the usual classically where they applied this yet, and as you are mentioning market basket analysis is a place where they did this first right. So what why it is called market basket analysis, so you go to a shop , you go to a supermarket you buy something in a basket right, generally you bought a basket along with you start putting things from the shop into the basket, then you come and get it checked out right.

So everything that goes together into the basket okay, so we call at single transaction right you might go, you might buy some cereal, you might buy some milk, or whatever you want to buy right some vegetables everything you put together and you bring it to the bill right. So all those things that go together we call a transaction, so market basket analysis is essentially analyzing what goes into your basket right.

So this is the kind of things we will have right. And so these transactions will essentially be defined over a universe of items right, and each transaction will be thought of as a subset of these items right or as the data mining people call it will be refer to as an itemset okay. So what is an itemset, it is set of items that is it, okay. So instead of calling it a grammatically correct fashion a set of items okay for whatever reason they introduced a new known call itemset, it is a single word okay.

They introduced a word called itemset and then they started calling it frequent itemset mining and so on so forth. But so, we will use typically be using the word itemset and we will also use a term K itemset what do you think is K itemset? Set of size K, set of items of size K that is K itemet right. And so we also, so you also have something called association rules that we talked about in the context of frequent pattern mining.

So what is an association rule is our rules of the form A implies B what does it mean it means set okay, if you buy items in the itemset A right, then you are likely to buy the items in the itemset B so A and B upsets they are not individual items A and B are sets right. So if you buy things in the itemset A so for example if you buy the usual thing offer is if you buy milk and bread you are likely to buy eggs or something.

So you are basically going out shopping for breakfast items right. So you buy milk and bread and then you also buy eggs right. Now very famous or in famous example that people had what is it, yeah if you go out to buy beer you also buy diapers. So why you think that is causal effects, you are not kids you are supposed to know this, do not laugh at this. Yeah, some of things, yeah the masters and PhD students can laugh.

I am just kidding right, so yeah, so why do you think that was the case any theories come on I should be knowing, give me some theories people have been giving you horror stories were having kids, this one. And so there is another, then people did some analysis and they actually found all this is like this, the spike was happening on Sundays right. So in the US Sunday is football day, every Sunday they have football playing okay.

These guys are actually buying beer okay to drink during the football watching the football game. So that same time they also pick up diapers, because they do not want to be disturbed by the baby during the game or something right. So they probably slap a couple of diapers on the baby let us say okay, do not call me well the game is going on okay. So that is basically what was happening.

Of course it had anyway, so there is a larger point to this right. So it is not enough for you to association tool mining. Now some of the associations you discover just from the data might not immediately have any meaning to you. For example, you cannot say buy to whatever two crates of beer and get a diaper free. So you cannot use it for promoting safe and also look really weird if you start stacking diapers next to the beer, beer cans in the shop right,

So some rules are you I always use an example to illustrate the fact that hey statistics is all fine right, but you need something more than statistics in order to get any useful intelligence out of data. So you need to think of other ways of doing it, but we are not going to do that right great. So how do I know this A implies B is useful other than I mean ignore the discussion we had right. Normally how will I say whether it is useful or not.

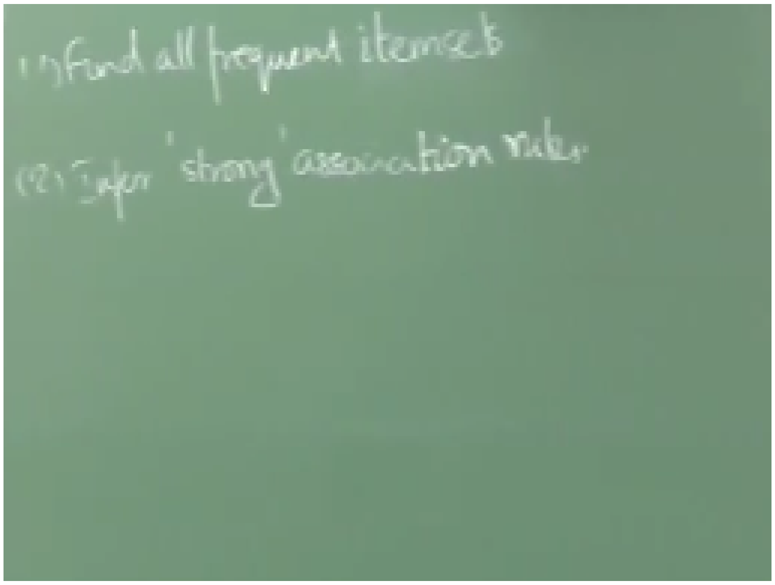
So I have different measures by which I can measure the usefulness of a room right. So two most popular ones or call support right, so what do you think is support how many times A and B have occurred together right. So this is essentially the probability of AUB, so AUB is a set union right. So how many times A and B have occurred together right. And the second thing that we look at this okay.

Yeah, yeah a set union, as A is a set, B is a set right, so when I do the union of that all the elements in A and B what is the probability that all the elements in A and B have occurred. This is how it is usually denoted right, so it is a set union. So this had been literals right, then I would have put A and B what is the probability of A and B, but since this is a set A and B okay. And confidence is essentially the probability of B given A.

I am saying A implies B that means then I have to figure out okay how many times B occur when A occurred right. So the probability of B implies if A implies B is confidence of A implies B is probability of B given A. So this is essentially the number of transactions having right. So the number of transactions having A U B divided by the total number of transactions this is number of transactions having A U B divided by the number of transactions having A okay.

So how do you find these association rules you first do frequent pattern mining find all patterns that are frequent right find all patterns that are frequent okay then you will find that okay A is frequent and some A U B is also frequent and then from that I can start inferring these kinds of association rules. Finding out but another frequently yeah you are right but it is still a hard problem so that is what the rest of the class is going to be a book how will you do this efficiently right.

(Refer Slide Time: 14:57)



(1) Find all frequent itemsets

(2) Infer 'strong' association rules.

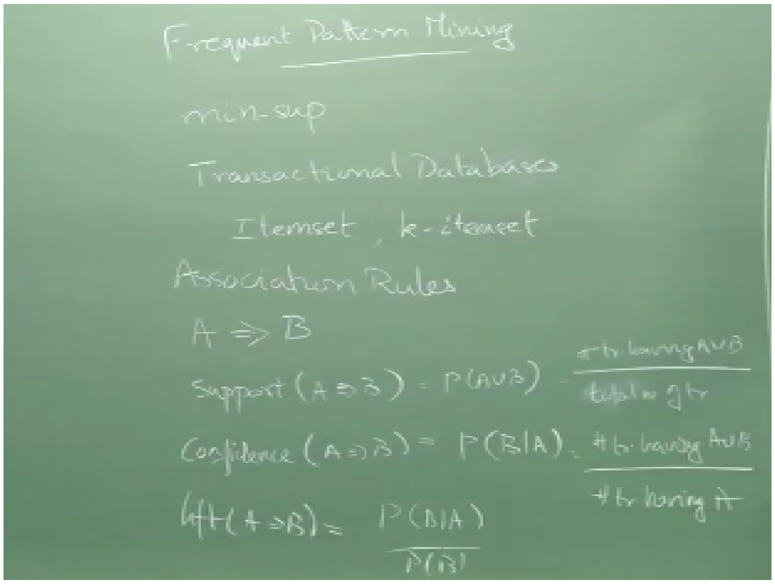
So usually how you do this is the first step is find all frequent patterns second step is infer strong association rules right, so all frequent patterns or itemsets right let us get into the habit of calling patterns as itemsets right so all frequent itemsets our itemsets which have min support right all strong association rules will be those association rules which have min support right and as well as a min confidence like minimum support under minimum confidence I want both, right for a strong association rule.

But there is a caveat right thus they are having strong support and strong confidence alone is not enough, right so you will not also have to see what is the probability of B in isolation I look at probability of B given a and I say it is 0 . 6 okay that looks like a good association rule but if I remove A I just look at what is the probability of B and if I say the probability of B is 0.75 right so what happens a implies a depression in B actually right I should not say that if A occurs then B will also occur.

A we occurs and the chances of B occurring will go down right, so this is something that again a classical example is when people are analyzing data from a store called blockbuster right, so blockbuster rents videos they also sell video games okay and they found out that if people rent videos from the shop I am sorry if people buy video games from the shop they also rent videos that rule had a confidence of 0. 6 right but then if you do not buy a video or if just anybody who comes to the shop whether they buy a video or not.

I am sorry by the game or not right that is a probability 0 .75 of them renting a video so essentially if you go to the shop to buy a game you are less likely to rent a video right, so you have to be careful about that right.

(Refer Slide Time: 17:52)



Frequent Pattern Mining

min-sup

Transactional Databases

Itemset, k-itemset

Association Rules

$A \Rightarrow B$

$\text{Support}(A \Rightarrow B) = P(A \cup B) = \frac{\#\text{ tr. having } A \cup B}{\text{total no. of tr}}$

$\text{Confidence}(A \Rightarrow B) = P(B|A) = \frac{\#\text{ tr. having } A \cup B}{\#\text{ tr. having } A}$

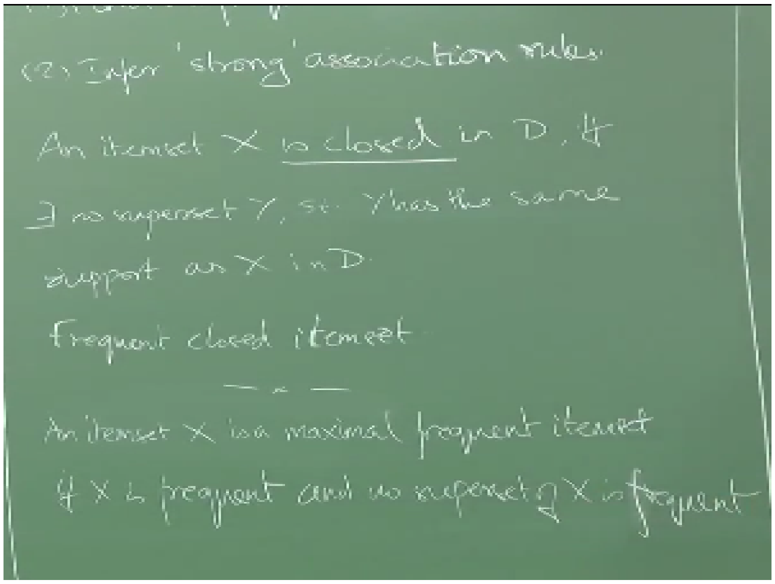
$\text{lift}(A \Rightarrow B) = \frac{P(B|A)}{P(B)}$

So there is another thing which people use which is essentially the ratio of these two things right, so if I am if this is greater than one then knowing A is useful that if it is less than one then knowing A is not useful this quantity is sometimes called lift okay, so lift also as another interpretation sometimes people take the difference between the two and that is also known as lift sometimes people take the ratio I think nowadays ratio as kind of become the standard way of defining and lift.

So and I should tell you that association rule mining is a very, very popular subfield of data mining and I given you three different ways of measuring usefulness of rules and there are about 100 right and I believe none then covers a good fraction of those hundred in his courses right, so if you want to know more about it go to dooms right so that there are lots of different ways of measuring this but these three are pretty common support and confidence are based the base right.

And then people build a lot of things on top of support and confidence okay, so that is basically it I am not going to talk about association rules any more right the interesting problem as you could have rightly surmised by now is finding all frequent itemsets right, so just a couple of other you know definitions here.

(Refer Slide Time: 19:57)



(2) Infer 'strong' association rules.

An itemset X is closed in D, if

∃ no superset Y, s.t. Y has the same support as X in D.

Frequent closed itemset

— x —

An itemset X is a maximal frequent itemset if X is frequent and no superset of X is frequent

An itemset X is closed in a particular data set D if there is no superset of X that has the same support as X, so it has to have a lesser support then X then you call that a closed itemset right what is a frequent closed itemset, so closed itemset whose frequency is higher than min of right, so if I give you the counts of all the closed itemsets in my data set and I am sorry all the frequent closed itemsets in my data set right you can recover the counts for all the frequent itemsets in the data set right.

If I give you the counts for all the frequent closed itemsets right you cannot recover the cones for all the items experience and a fairly straightforward right because what is what when would the itemset be part of the clothes itemset if a superset has a lesser count than it right but the superset could still be frequent let us say I have a frequency threshold of two right and some itemset has a count of five and add one more item to it has a count of only four right but still that is also frequent it will be part of the closed frequent itemset as long as that is a superset of that that has a smaller count.

Right if it does not what does it mean if I say that ABC is a closed itemset right and there are no two itemsets that are closed what does it mean the count of AB the count of AC and the count of BC is the same as the count of ABC to get that, so if I say that if we say that ABC is a frequent closed itemset right is a frequent close item say that means that that is under has a count of five that means that AB is also a frequent itemset right AC is also a frequent itemset and BC is also a frequent itemset.

And what are the counts of AB, AC and BC five right, so if I give you the frequency of all the frequent closed itemsets then I can recover the frequency of all the frequent itemsets that we do not yourselves we close the reason is called closed essentially this is sufficient for me to recover the entire later right, so typically which you are trying to come up with a new frequent new counting algorithm for itemsets right I have to make sure that you return the closed the complete set of closed frequent itemsets right.

A itemset is a maximal frequent itemset if X is frequent and nothing larger than X is right any no superset of X is frequent right, so it is a severe case of close right so closed what the closed condition is the superset should not have the same count right should have a lesser count than the subset right but it could also be frequent here I am saying not only should the superset have a lesser count but the count should be, so less that it is no longer frequent, so the set o maximal a frequent itemsets will be smaller than the right the set of all maximal frequent itemsets will be smaller than the set of all closed frequent itemsets right or frequent closed itemsets.

So that will be smaller and is the maximal set sufficient for you to recover all the all the frequent itemsets no not necessarily sure sorry we did not say that, subset could have a lesser frequencies so I mean there is been equal frequency, yeah but I cannot give you the frequency of those but I can say which are which are all frequent but I cannot give you the frequency of those subsets while in this case I can give you the frequency of the substance because they will be the same right so I can still recover something about the something about frequent itemsets but I will not be able to tell you what is the frequency.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in