**NPTEL**

**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine learning**

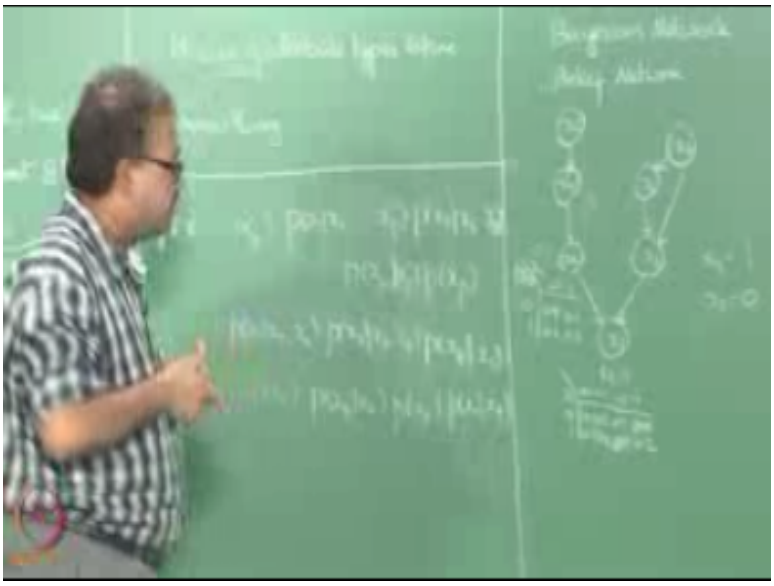
**Lecture-64**
**Bayesian Networks**

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

So one of the things that you could do along the lines of this independence assumptions is try to be more nuanced about your independence, so what do I mean by that? So just do not make this assumption that everything is independent of one another given the class right so think of something like this I want to look at this joint distribution I want to look at this joint distribution $X_1$ to $X_P$.

So I am going to say things like okay, I am going to write something work I am not stopping so what I am saying here is given X to write $x_1$ is independent of everything else if you know what the value $X_2$ is $X_1$ is independent of everything else. Before that let me let me motivate it this way this makes it easier for you to grasp people agree right.

(Refer Slide Time: 01:30)



So I can write probability of $x_1$ to $X_P$ as a probability of all of these conditionals yes okay great. So now I am saying hey I am saying I am trying to find out what is the probability of $x_1$ given $x_2$ $x_3$ $x_4$ all the way up till $X_P$ likewise I'm trying to find the probability of $x_2$ given $x_3$ I mean either some arbitrary ordering of choice and $X_1$ to $X_P$ right it could be any other ordering now probability of $X_2$ given $x_3$ 2 $X_P$ and so on so forth.

Now I am going to tell you that okay this is when you can always add a conditioning g if you want right this makes my life easier if you do not put everything conditioned on g right if you want you can do that as well okay. So what I am going to say is this is really not going to happen you know how likely is it that each of my variable like $X_1$ is going to depend on the value taken by every other variable in my system especially if I am going to have 30 and 40 variables how likely is that $x_1$ is going to depend directly on all the other 30 40variables in the system right is not going to happen right.

So what will happen is? Let us say that this is equivalent to say it is something like probability of $x_1$ even $X_2$ m, $X_3$probability of X 3 given $X_6$, $x_7$the probability of blah, blah, blah right all the way up to get another example oh well just okay right so maybe my system is like this so does it means of so $x_1$is dependent only on $x_2$ and $x_3$ given $x_2$ and $x_3$ $x_1$ is independent of all the other variables in the system likewise given x6 and $x_7$ $x_3$ is independent of all the other variables in the system right given $X_4$ and X will give an $X_5$ $X_4$ is independent of everything else.

And x5 is independent of everything else just by itself right it is independent of everything else and x6 depends only on $x_7$ and $x_7$ is independent of everything else and it is just one way of writing it right, whenever I say $x_6$ is dependent on $X_7$ I can always flip it around and say $X_7$ is dependent on $X_6$ I am not talking about causal directions here I am not saying that $X_7$ cause of $x_6$ right.

And they saying that the probability distribution can be factored in the form of x 6 given xml into x 7 other I can also do it as probability of x 7given x 6 into probability of x 6 I could I could try and do the factorization the other way okay, it just want you to keep in mind there is nothing sacrosanct about the way I have done the factorization right just a convenient way of representing these things okay.

Does it make sense? Like I said if you are worried about the classification scenario you can add a conditioning on G everywhere right that is typically what we do but this kind of way of factoring things is more powerful than just using it for classification and you can use it for learning about any probability distribution okay does not have to necessarily be about classification you can use it for representing any probability distribution okay.

So one way of I mean this looks a little hard to track right, so one way of specifying these kinds of conditional independence relations is to use a graph, so what will I do in this case I will have a graph that has seven nodes. So one node corresponding to each feature right what are the features here more generally the features here are random variable say $X_1$ is a random variable that will take the value in whatever range $x_1$ can take and so on so forth these are all random variables.

So I am going to have there something that is missing here, so I have connected the graph note so I have x 1okay, so $x_1$ depends on $X_2$ and $X_3$ so I will put arrows $x_3$ depends on $x_6$ and $x_7$ and $x_2$ depends on $X_4$ $X_4$ depends on $x_5$ andx6 depends on $X_7$ so this graph structure right gives me the dependency or independence conditional independence relation I wrote in that expression that right makes sense.

So if you remember when I was talking to you about the interpretation of conditional independence that I said if you do not know what the class is then the onion cricket might become dependent but if you know what the class is the only in cricket or independent right I mean occurrence of the words right I was telling you about that at likewise right if I know what

$x_2$ is right $x_4$ and $x_1$or independent right is it clear if I know what $X_2$ is then $x_4$ and $x_1$ are independent.

But if you do not know what $x_2$ is then $x_4$ and $x_1$ become dependent, so what do I mean by that if I know something about $X_4$ then I can tell you something about $X_1$. So if there is a little confusing we will try to make this concrete let us say this can take values 0 and 1 and this can take values0 and 1 not that it is confined to binary things right Boolean things what makes it easy for me to write.

So let us say that the probability of something like they said so X to basically copies x4 with a high probability right and likewise so that is $x_2$ $x_3$ so I will have to write a table like this for $x_1$ right and yeah. So basically it says that when $x_2$is zero okay the probability of $x_1$ being zero is slow and probability of $x_1$ being one is high right and likewise and $x_2$ is one the probability of $x_1$ being zero is high the probability of $x_1$ being zero is slow that is essentially what I am saying that okay, so now if I know what $X_2$is right. Let us say that I tell you that$x_2 = 0$.

Now the fact that $x_4 = 1$ and if I say $x_2$0 then you know that the probability of $x_1$ being one will be high right regardless of the value of $x_3$ because that is the way I have written this thing down but if I know value of $x_3$ also then I will know that okay whether it is 0.9 or 0.8 right. Now if I tell you that $x_4$ is 1 it does not matter because the only way $x_4$ will give me any information about x1 is through $x_2$ but I know $x_2$ is already 0 but suppose I do not know that $x_2$ 0 I but I tell you that $x_4$ is right immediately what do you know that the probability of x to being one is higher right therefore the probability of $x_1$being zero is higher right if I had not told you the value of $x_2$.

But if you are told you the value of $x_2$ I say that okay $x_4$ is 1 right but there is a small chance at $X_2$ can be 0 right so X to actually become 0 in which case the conclusions you can draw about x1completely changes is a very dramatic example but this is not always be so dramatic but the point I am making is because of the way I have drawn these arrows right if $x_2$ is not known knowing $x_4$ will tell me something about $X_1$ if $x_2$ is known the knowing $x_4$ will not tell me anything more about x.

So everything that I can know about $x_1$ by knowing $x_4$ I already extracted by knowing $x_2$ is it clear so this is this whole idea of conditional independence and why this kind of graphical

representation helps us right. So knowing $x_4 x_3$ right not $x_2$ I know only $x_3$ but not $x_2$ still does not disconnect me from $x_4$ right because the paths are very different so $x_2$ $x_4$ can still leaks influence $x_1$ if I do not know $x_2$ but no $x_3$ is it clear okay I come to that right.

So this is the initial setup right, so what these kinds of graphical models do or rather this kind of these are called a Bayesian network right sometimes call sometimes called a Bayesian belief network and then sometimes called a Bayesian belief network okay, so we will find all the terminology in literature you will find a Bayesian network belief networks and belief networks and so in the Bayesian network is a dag it has to be a cyclic graph and because if it has cycles in it you are basically missed up right because the semantics of the thing right so we will actually talk about a graph representation which does not have any arrows even right which actually is undirected graph.

So there are undirected graph you can start talking about cycles we will come about come to that in the next class but when you are talking about a directed graph representation right it has to have no cycles because it has cycles then x1depends on $X_2$ $X_2$ depends on $X_3$ and $X_3$ will in turn depend on $X_1$ therefore what will happen this thing will get completely messed up right. So you cannot write out a factorization like that right.

So one way of thinking about it is this thing is going to give you a set of conditional probability distributions right, so each node is going to have a set of conditional probability distribution so $x_1$ is going to have a distribution which gives you $X_1$ given $X_2$ $X_3$ likewise $x_2$ will have a distribution associated with it which will give you probability of $X_2$ given $X_4$. So if I take the product of all these conditional probability distribution I will recover the Joint Distribution of those variables okay.

So that is basically the semantics associated with it right take the product of all this conditional probability distribution I should recover the Joint Distribution of all the variables so if you are going to have cycles then that property will no longer be satisfied we do not want cycles in this case right. And what is this here? So it is a dag where each node is a random variable okay and each edge represents a conditional dependence great.

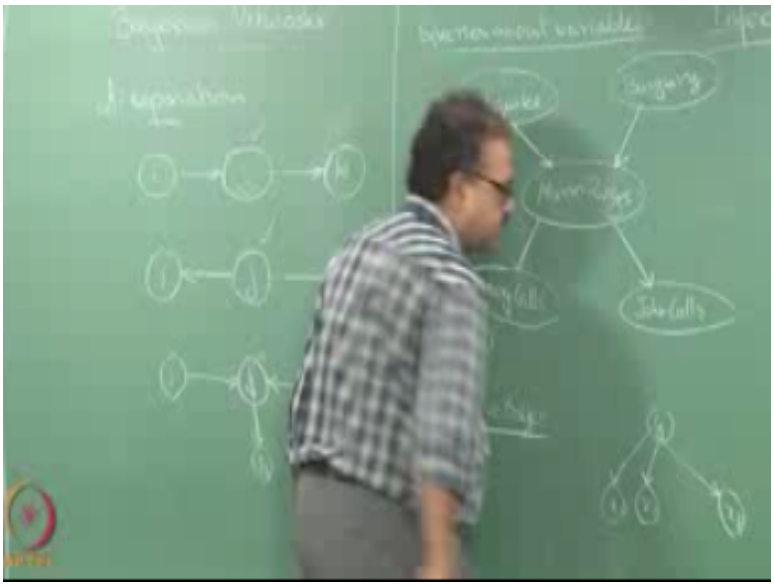
So because of the nature of the graph we are drawing right so the graph encodes a lot of separation rules so what we win by separation rule it tells me that $x_1$is independent of $X_4$ given X

$_2$ right. So I would say that $X_4$ and $x_1$ are separated by $x_2$ right, so likewise can you say something about $x_6$ and $x_1$separated by $x_3$ what about $x_7$ and $x_1x_3$ right what about $x_6$ the next six separate $x_7$ and $x_1$ $x_6$ will not separate $X_L$ what about x 6 will it separate $x_7$ and $x_3$ know right.

So you can see where we are going with this right, so I have shown you one separation rule but if here is a directed path from $X_3$ X 1 right directed path from $X_j$ - $x_I$ any node along the path will separate $X_J$ and $X_I$ provided that is the only path, if there are multiple directed paths okay it has to appear on every one of those directed paths then it will separate it or you have to have a set of nodes these two nods together will separate $X_J$ and $X_I$.

So you will have to select one representative from each of those directed paths then it will separated because we will have to consider directed edges here this is not called separation it is called directed separation or D separation come on else obvious registry make them so D separation.

(Refer Slide Time: 19:02)




So there are three D separation rules very simple d separation rules okay, JD separates I NK that is a rule we already saw okay, so what do you think about this hmm if I know J INK are independent right they are separated if we do not know J they are connected, so knowing j okay

separate itself likewise knowing J will it separate INK here yes again think about this right suppose I did not know J right but I know I let us say something like this right.

So I know that x1 this has a value of one I know x1 has a value of one right and I know that $X_1$ will be 1right only when or whether I will know that I is one with a high probability if J is 0 right, so if I know that I is one then I know that the probability of J being 0 is higher as soon as you know the probability of J being 0 is higher than I will know something about k right because there is a direct influence from J to K.

But if a new j I know that J is 0, I do not care what I is I can be anything but knowing I will not tell me anything more about k then I get by knowing J right. So in this case knowing J separates INK in this case knowing J separates INK is there anything else that we need to worry about any other combination sorry sir IJK all the way yeah anything else is substantially different convergent say.

So what do you think? Knowing actually something else it does, not knowing j separates INK knowing J connects INK think about it knowing J connects I NK because let us go back here right. So I know that $X_I$ is 0 I know that $X_I$ is 0 I know that $x_2$ is one now I know that $x_2$ is also 0 right I know that x is 0 and I know that $x_2$is 0 then what about x3 both are 0 then the probability of $X_3$ being one goes slightly higher right.

Because this is the case where both zeros occur with a higher probability right so we have this situation here, so if I know if I know $x_2$ is 0 I do not know anything about $X_1$ I cannot say anything about $x_3$ does not matter at x1 and $x_3$ are independent and $x_2$ and $x_3$ are independent if I do not know $x_1$ right if I know $x_1$ then $x_2$ and$x_3$ become connected let that make sense is right. So that is these are the 3d separation rules in both these cases knowing it separates in this case not knowing J separates in fact is slightly stronger also I can look at any descendant of $J_1$ I knowing any of the descendants of $J_1$ also will end up connecting INK not only should I not know J I should not know the value of any of the descendants of Jay also because as soon as you know the value of $J_1$ I can make an inference about of J and now that will help connect INK right not only surely not no J none of the descendants as well okay.

So these are the 3d separation rules so it is great to see the D separation rules do not talk about actually talk about the values of the probabilities right it is just a representational thing so I can

plug in whatever values I want all I am saying is just from the structure of the network I can tell you something about the separation probable properties right. So the actual probability values could come in later the values I use there was only for illustration purposes did not necessarily be that right this is the structure of the network itself tells you that what are the separation properties okay.

Could any questions on this? This is clear I can give you a very large graph right and ask you okay or A and B separated if I know C D and E okay, what should you do then? Sorry you have to find out all the pus directed and undirected between a and B because I mean these are all undirected things this is only one that goes along the direction directed path right this is actually an undirected thing right.

So I have to look for all such paths between a and B and figure out now do CD and DI given you CD and DS variables that are known right, see figure out whether knowing those variables disconnects the path and all the other variables are unknown right. So you have to apply for the third rule for the unknown variables as well right then if all the paths get disconnected between a and B then you say that a and B are D separated by C D and D so this is this is a kind of analysis that you can do to make sure that you understand your system properly right.

So one of the original motivations for proposing these kinds of belief networks it is kind of dag representation for the variables was to study causality was to figure out causal relationships and so on so forth so these kinds of networks are also sometimes called as causal networks and typically when you talk about causal networks you do not associate the conditional probabilities with things so just you are talking about a causes B kind of relations right you are not really worried about the probabilities in that in that kind of a setting.

So the same representation can be used for representing causality also let a causes B right so that kind of relationships can be represented using the same representation but in general when you are using this as a Bayesian network you do not imply any causality is something which you have to keep in mind when you are using it in practice right. So you are not implying any kind of causality right when you are using this direction this does not mean that I actually believe that x 2 causes x1 right.

When you are using it as a causal Network model yes okay when you are putting in an arrow here that means you have thought about it and you believe from the physics of the system or

whatever it is that X2 actually causes x1 and it turns out that when you are trying to do this learn this graphical structure by just observing the system looking at the data and trying to infer this kind of graphical structure between the variables the most compact structure that you can derive will turn out to be the one that corresponds to the actual causal nature of the system right.

If you do it in a incorrect fashion and you will end up adding a lot more spurious dependencies between variables that if you are doing it in the correct causal ordering then you will end up having a much more compact graph then you would if you are doing it at wily nilly right. So what is the use of doing all of these things I have talked about this major modeling business what is the use of doing all of this?

So essentially we are interested in answering queries about interested in answering queries about variables right, so no class no lecture on graphical models or Bayesian networks is complete without you looking at the earth quake network at least once okay the very, very popular network for historical reasons. So you have a burglar alarm in your house okay and alarm rings okay the alarm could ring because of two things right it could ring because there is actually a is actually a burglary it could also ring because as okay.

So this network was originally made up by today a pearl okay is one of the early pioneers in the study of causal networks and belief networks and so on so forth so Jerry Apple lived in la lived in California it was not thinking of wild animals you think you have something else remember what I call this network okay. So probably the most two common occurrences in California earthquake and burglaries fire alarm know it is a burglar alarm I am not interested in the fire alarm I am interested in burglar alarms okay.

And it turns out that pearl had two very nice neighbors right who would call him at his office and tell him hey your burglar alarm rang okay with some probability, so this is basically so if the alarm rings Mary or John will call Paul in his office and tell them that hey your alarm rang so you can think about the causal directions here right so alarm will be cause d by the earthquake or burglary right and then Mary will call and John will call if the alarm rings came both of the might call or none of the might call because they are all probabilistic things right.

So now I can ask questions like this Mary called me and said that there is a she thought she heard the alarm okay so what is the probability that the alarm rang both Mary and John called me and

said both of them thought they heard an alarm what is the probability that the alarm rang I know there was an earthquake in my place but both Mary and John did not call me what is the probability that alarm rang?

No, but if that is going to happen see that that is going to happen I should have had an arrow like this okay since they do not have the arrow I am going to assume that it is not likely that is what I am saying right if the earthquake is actually going to directly influence John and Mary's behavior whether they are the question of their mortality or not or other things right, so I would I would need to put an arrow directly between earthquake and Mary just assume that Paul and Mary live in different earth quake zones right there is this one small fault line which will only shake Judea Paul's house and go away go on guy this illustrative example do not take in too, too, too, too, too, much too hard.

So the point here is I can ask all kinds of queries on variables on this right I can ask even other things like that hey Mary called what is the probability that a burglary actually happened right Mary and John call what is the probability that a burglary actually happened? Things change or not change or not eating this change no you know everything answer that question yeah it is change because if both Mary and John call then the my belief that alarm rang goes up right if it believes the long line goes up then whatever belief I had about burglary happening will also automatically go.

Now we know why these are called belief networks right so when I say Mary call then my belief on whether the alarm actually rang or not changes right base I have some belief on alarm ringing so I think okay if nobody calls the law must not run if Mary call cell basically I will flip this backwards right here this will actually be only probability of Mary given alarm right. So this is this is the probability I will have here but now given that many of us one what is the probability of a being one and given that Mary and John or one what is the probability of a being one and given the probability of a being one what is the probability of burglary happening right.

So all of these things I can do all kinds of reasoning a word about the system based on just the this whole model that I am learning right, so I can ask all kinds of questions I can ask questions about joint distributions so what so many call okay what is the probability that the alarm rang and that John will call it kind of redundant probability but still you can try to ask us questions like this right or I know that the alarm rang somehow I know that the alarm rang what's the

probability that media and John will call me so I can ask all kinds of questions I can also conditional questions I cannot join probability questions they can ask conditional probability questions right.

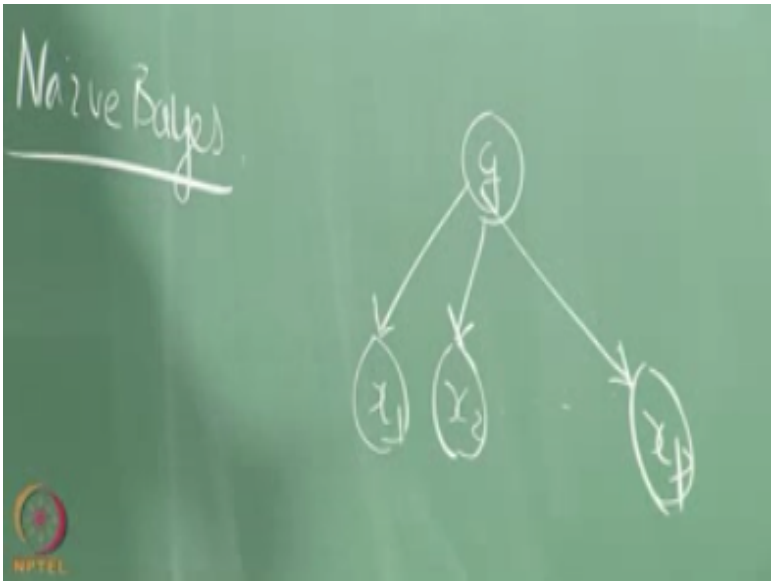
And if you think of this as classification problem I can ask questions about okay I know these five variables what is the probability that this is class one you got around our neighbors problem right I said what if you observe some variables whose values I never see before how will you estimate the probability I can still do that I can just assume that the variable is unobserved and I can estimate the probabilities that will give me actually give me valid answers that given that I know ABC what is the probability that class is one I might have another d e f g h i j k which I have not observed that is okay.

So I cannot call all these kinds of questions that given partial data now I can ask questions about classification right, our given class labels I can ask questions about class conditional densities write the given that it is document on cricket right how often will I do not know let me not pick on any cricketers forecast given that it is a document about football how often will the word Ronaldo and goal occur together in my document right yeah if you leave it to my son he'll see you say the probability should be 0 but I know this is almost religious right the camps anyway right.

So that is the whole idea right so these kinds of questions these kinds of queries and I asked about these variables so we call this problem as the problem of inference on the graphical model is the problem of inference on the graphical model essentially is to figure out all these conditional probabilities or the marginal probabilities that we are interested in right. So we looked at Nave base right.

So can you think of drawing the Nave base assumption as a graphical model every node is independent every node is independent is it so that will be like this is that my graphical model know what is the Nave's base assumption given that class they are independent, so where should the class been? The top or the bottom and of course I can ride wherever question is the direction of the arrows if you let me draw it at the top.

(Refer Slide Time: 38:36)




So people tell me how the arrow should go down right we surprised how many times people actually draw the arrow up the reasoning is the variable values are the one that causes the class to happen right. So if $x_1$ is this $x_2$ is this $x_3$ is that then they should all be influencing the class variable that for the error should go up well that is a fairly valid argument it is just that it does not capture the name basis option that is a different kind of assumption that you are modeling there right.

So each variable is somehow affecting the class this is essentially the opposite of naïve base okay instead curve is essentially a complete model right, so since all the variables are influencing the class that if in fact if you think about it like given that class what happens in that case all the variables get connected it will be this case right. So if all the arrows were going up it will be this case so if J was no Nathan all the variables get connected right in this case its opposite of Nave base at given the class all the variables are dependent on one another that is the assumption if you draw the arrows upwards the arrow should go down and up and down its relative rate error should go away from the class node.

**IIT Madras production**

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

www.nptel.ac.in