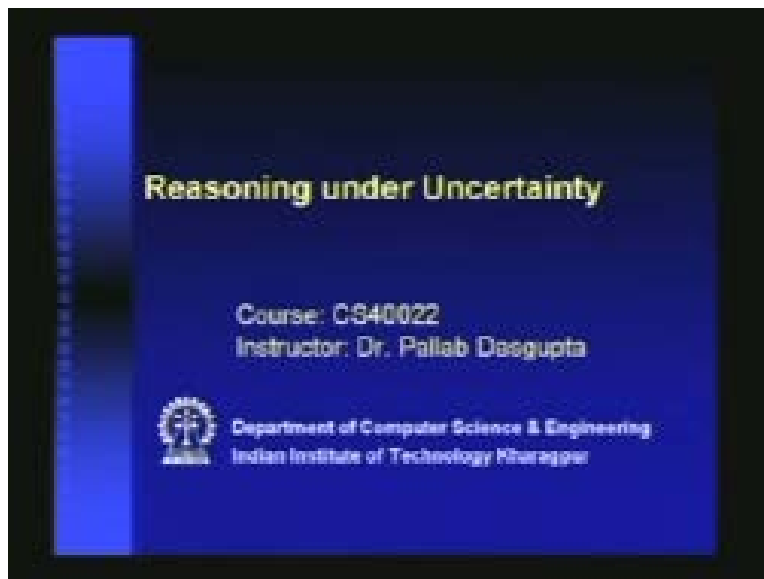


Artificial Intelligence
Prof. P. Dasgupta
Department of Computer Science & Engineering
Indian Institute Technology, Kharagpur

Lecture No - 21
Reasoning Under Uncertainty

We are going to start a new topic from this lecture and this topic is also a very important topic in AI. Namely, we will be studying reasoning under uncertainty.

(Refer Slide Time: 01:14)

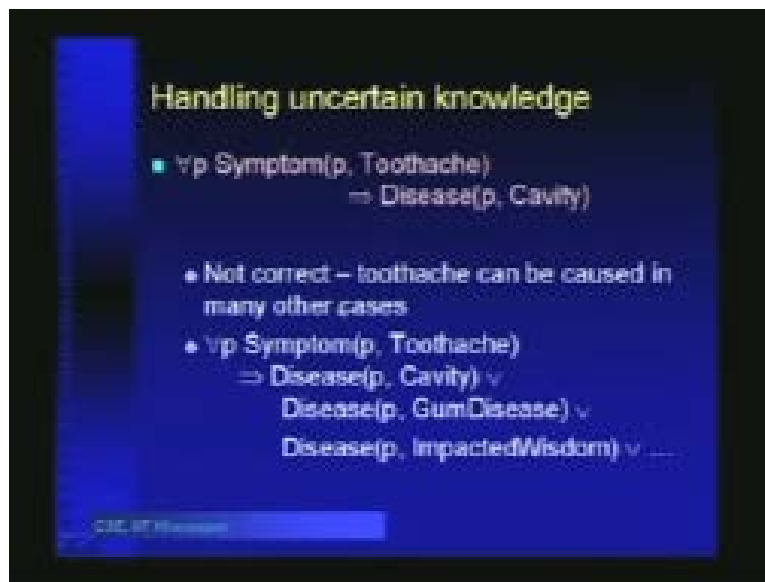


We have to rush up our school fund on probability; there will be quite a bit of probability coming up for you, so, let us get started. Why do we need reasoning under uncertainty? Let us see a few examples to get a clearer picture about why this is at all necessary. Firstly, the problem of handling uncertain knowledge. Suppose we are given the rule that

for all p symptom, p toothache implies disease p cavity. This says that whenever 1 has a toothache, then, that person has the disease which is a cavity in the tooth.

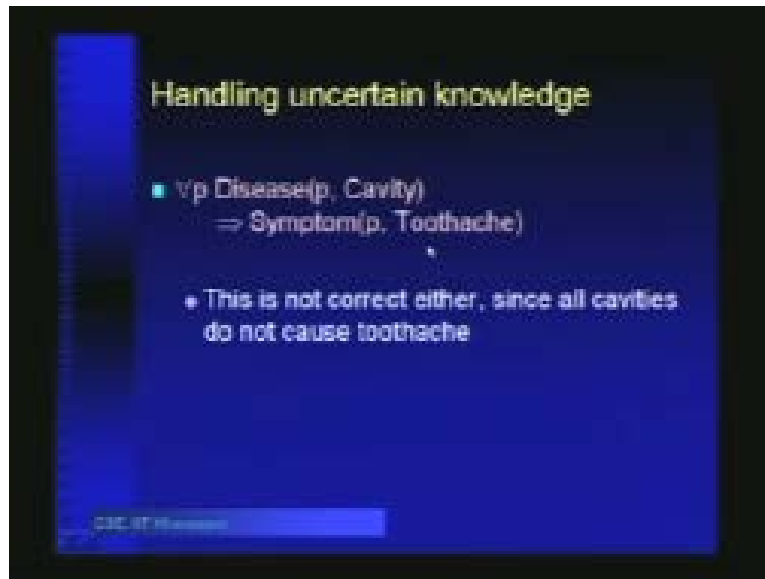
Now, this is not correct, because toothache can be caused in many other cases, so, we cannot say that whenever we have to take it is because of cavity. If we have to actually analyze the case, the problem of toothache, then, we have to comprehensively specify that what are the different causes for which this can happen. Cavity is one, gum disease is one, impacted wisdom is one, and so many other cases, for which we can have a toothache.

(Refer Slide Time: 02:10)



Now, enumerating all of this is a problem because all of this may not be known and you can miss out quite a few, like, for example, you can have toothache also, if somebody has hit you on the teeth, which you would probably not put in this root, right? But for diagnostic system, unless you have the complete set of causes, then, this rule will formally not be correct because we are using implication here. Let us try to model this in the other direction.

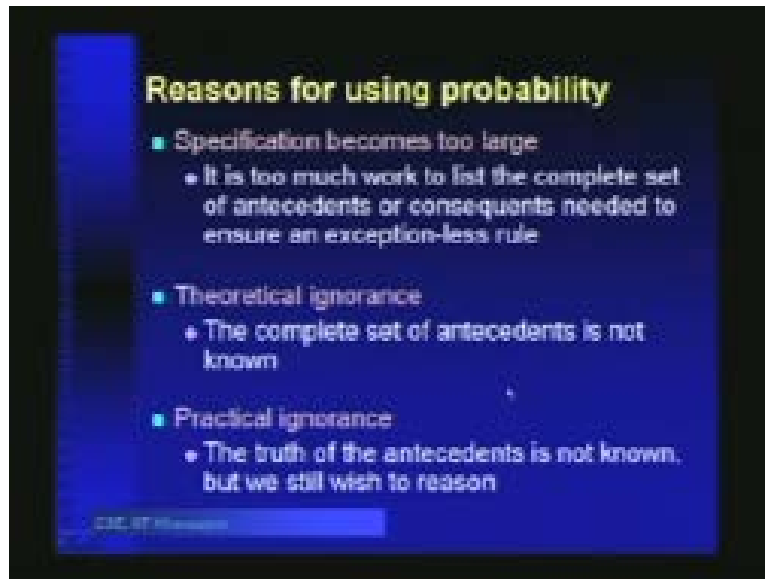
(Refer Slide Time: 04:02)



Suppose we want to say that for all p disease, p cavity implies symptom p toothache. We are saying that whenever there is cavity, then, the symptom is toothache. But this is not correct either, because all cavities do not cause toothache. So, even if someone has a cavity, that person may not have a toothache. This is a kind of scenario which is very difficult to represent unless we use some kind of statement which says okay, sometimes the cavity can cause toothache.

And if we have toothache, then, sometimes it is because of the cavity. Now, what are the reasons for using probability? 1 is that without using probability, the specification becomes too large. As we were seeing, we have to explicitly enumerate the complete set of antecedents or consequents for an exception-less rule, which in practice is very difficult to achieve.

(Refer Slide Time: 04:58)



It can also be because of theoretical ignorance. The complete set of antecedents is not known; we do not know what are the different kinds of diseases which can cause toothache. And the third 1 is practical ignorance, where the antecedents is known but there, truth is not known. So, if we do not have a mechanism of determining the truth of those antecedents, in practice, we will never be able to apply those rules. So, to apply a rule, the antecedents will have to be there in the knowledge base. Now, if 1 of the antecedents is something which we cannot experimentally measure, then, we will never be able to add that antecedent into the fact base.

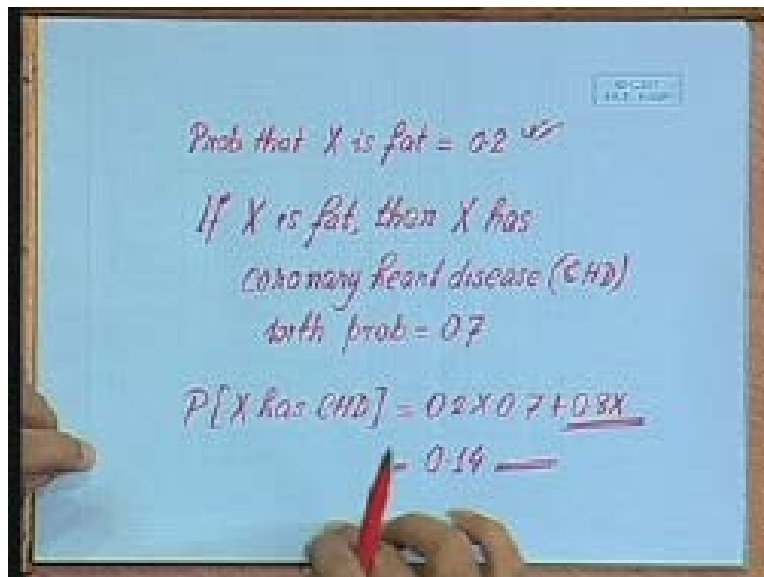
And therefore, we will never be able to apply that rule. In the cases where something is difficult to measure, we can leave that to chance, based on previous experience about percentage of cases where that actually happen. Before we come to the axioms of probability, let me explain a few subtle points about this. See, 1 thing is that when we are unable to model a set of rules exactly and completely, then, we resort to things like probability. People also have resorted, instead of using probability, to other kinds of logic, like fuzzy logic. Now, the difference between possibility, which is part of fuzzy

logic and probability, which is the classical bayesian analysis that we do- the difference is very subtle.

For example, suppose we are talking about how the obesity of a person is related to cardiac diseases. So, we are saying that if the person is fat, then, with certain probability, that person will have a cardiac problem. Now, when we are talking about a given person, then, we have to talk about what is the probability that he is fat. Now, the scenario is that we do not know that person; we have not seen that person, and based on some other information, we have some certainty that yes, this person is fat, with so much probability.

Suppose we want to reason about the probability of cardiac diseases among Indian people; now, we have statistics about what fraction of Indian people are fat, so therefore, we can have a certainty or a probability that a given x is fat, and then, with that, if we multiply it with the probability that if this person is fat, if x is fat, then, x will have cardiac disease with, say, 90% probability. Then, okay- let me write it down. Becoming complex, okay.

(Refer Slide Time: 11:12)



Suppose we know that probability- that this- we know from the population of x from Indians; if x is Indians, we know what percentage of Indians are fat. We know that if x is an Indian, what is the probability that is fat? Then, we are given that if x is fat, then, x has- let us call that CHD with probability of 0.7. Then, we ask that what is the probability that x has CHD? This we can say that it is 0.2 times 0.7, which is equal to 0.14. (Student speaking). Yes. (Student speaking). Plus- (Student speaking). 0.8 into- (Student speaking). Why should we have this? (Student speaking). Yes. Right, so, we also have to consider the case where x is not fat, but still has a heart disease.

So, that probability will also have to be added with this, so, we will see how we do that analysis once we go into the Bayesian analysis. But the point I am trying to make here is that this is a probability that we are given and we can apply that to determine the probability of some other event. But fuzzy logic deals with a slightly different philosophy. There, we can see the person x is given to us, but we are trying to say he is fat, but how fat? How fat is not a true/false value. You cannot say that this person is that. If the weight is above this, then he is fat. If the weight is below this, then he is not fat. Fat is actually a gradation, right? From thin to fat, we have a distribution.

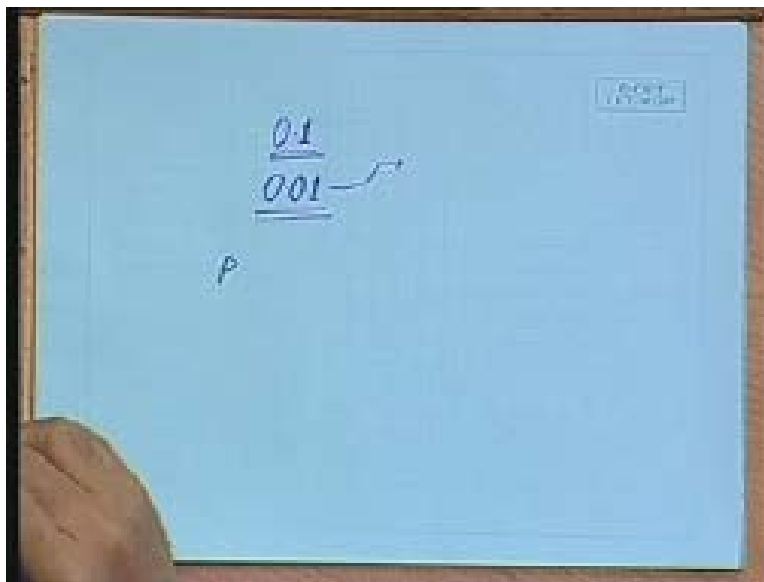
So, fuzzy logic tries to reason in that sense. It tries to look at the truth value not as Boolean, but as a value between 0 and 1. So, you say that x is very fat. Then, that means that truth of x being fat is 0.8 and x is moderately fat, which is 0.6, and so on. Now, you see, it is not a question of whether x is fat or not. It is a question of how fat is x and the rules, therefore, also has to be graded that way, right? So, that is a different kind of analysis which people do. So, I just want to introduce at the beginning, because later on, when we talk about these different kinds of reasoning, you have to be sure that which is what. 1 is probability and the other is the gradation of the truth.

Axioms of probability. Let me ask a simple puzzle; this is a puzzle which goes like this: that we are given that the probability that 1 person carries a bomb into the aircraft is 0.1. So, probability that someone carries a bomb into the aircraft is 0.1; then, what is the probability that 2 people carry a bomb into the aircraft? It is 0.01, right? Now, 1 professor

sees that okay, the probability of 2 people carrying a bomb into the aircraft is significantly lesser than the probability of 1 person carrying a bomb. So, what he does is, he carries a bomb with him; he carries a bomb with him into the aircraft. So, the question is, does that reduce the probability that another person will come up with the bomb? (Student speaking). Right. Why is that? What is the difference between these 2?

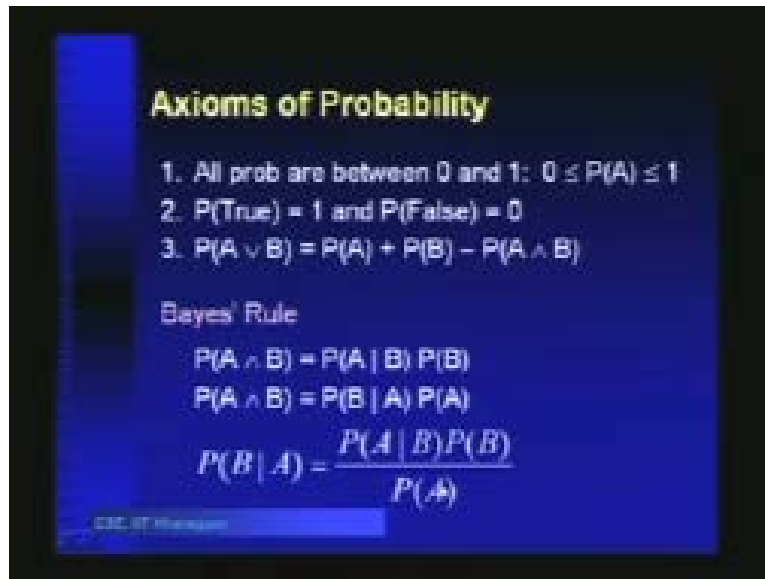
The difference is conditional probability.

(Refer Slide Time: 15:00)



The probability that 2 people carry a bomb into the aircraft is this, fine. But probability that another person carries a bomb into the aircraft, given that 1 person has carried it, is again 0.1, right? So, that is what we have as conditional probability.

(Refer Slide Time: 15:33)



Axioms of Probability

1. All prob are between 0 and 1: $0 \leq P(A) \leq 1$
2. $P(\text{True}) = 1$ and $P(\text{False}) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Bayes' Rule

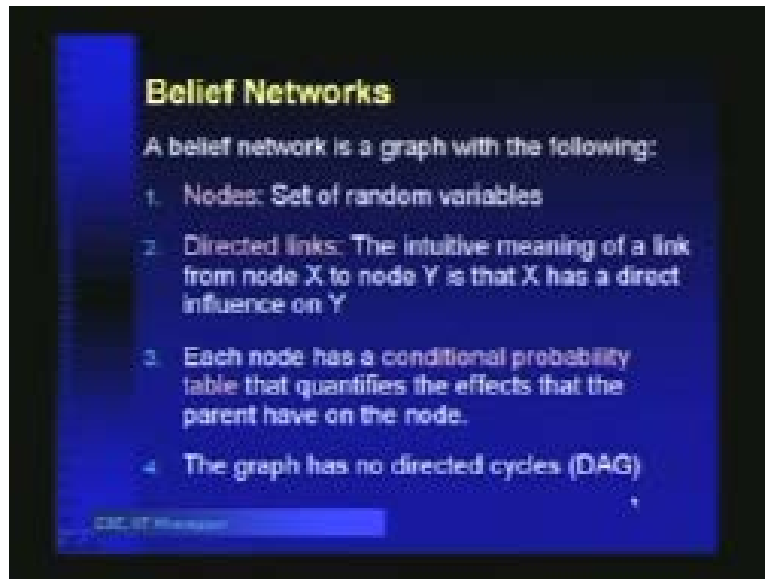
$$P(A \cap B) = P(A | B) P(B)$$
$$P(A \cap B) = P(B | A) P(A)$$
$$P(B | A) = \frac{P(A | B) P(B)}{P(A)}$$

© 2011 MIT

The important thing is to realize that which events are independent and which are conditional to each other. That will be the code thing that will found the basis of all our inferencing. These are school book axioms of probability $P(A \cup B)$ is $P(A)$ plus $P(B)$ minus $P(A \cap B)$. Then, we have Bayes rule. Everyone remember Bayes rule? Okay. So, let us move into the next- belief networks. Belief networks are networks with the following: we have a set of random variables as nodes, we have directed links.

The intuitive meaning of a link from node X to node Y is that X has a direct influence on Y. Cause-effect relationship between this. Now, each node has a conditional probability table that quantifies the effects that the parent have on the node and the graph has no directed cycle, because we assume that there is a cause-effect relationship between the events and there is no feedback. Networks with feedback has been also studied, but in this particular lecture, we are going to consider that.

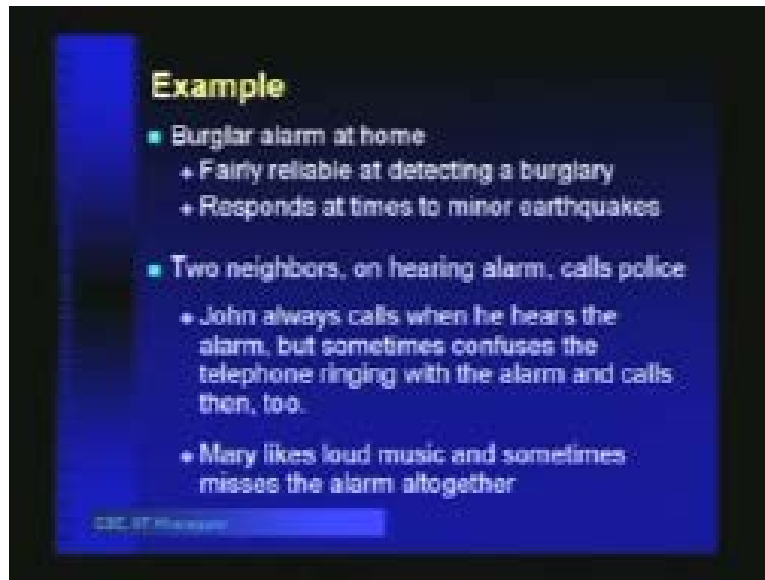
(Refer Slide Time: 16:29)



Let us start with an example. This is an example where we have a burglar alarm installed at home. The alarm is fairly reliable at detecting a burglary. With some probability, whenever there is a burglary, the alarm will go; it will ring, but it also responds at times to minor earthquakes. So, if there is a minor earthquake, then also, the burglar alarm can ring. Actually, this particular example is due to (unclear word) who was or who is in fact resident of LA, and that is why he is interested in earthquakes.

They have it quite frequently, right? 2 neighbors, on hearing the alarm, call the police. John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then too, right? And many likes loud music and sometimes misses the alarm altogether, because of the music, right?

(Refer Slide Time: 18:03)

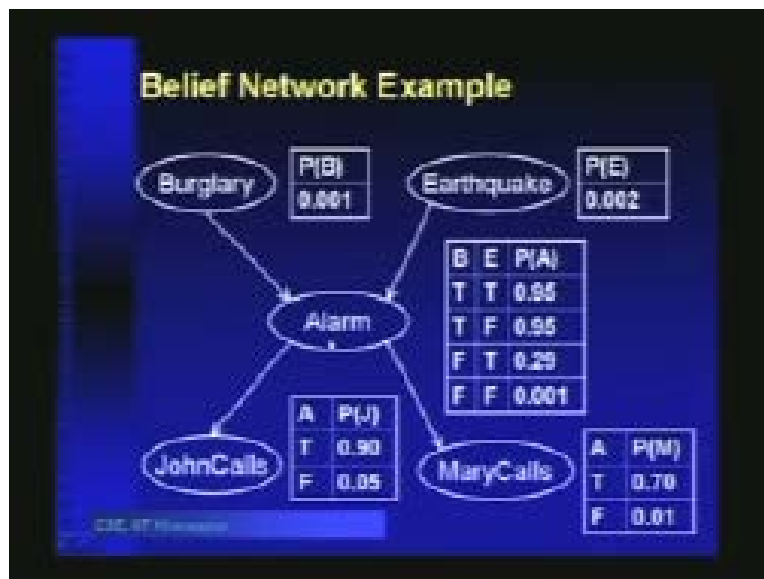


This is what we are given, so, we will model this as a belief network. This is an example of the belief network. Slides please. Let us see what we have here; these are the events, so, burglary earthquake alarm, John calls and Mary calls, and this ordering that we have done is our choice of doing this. You can construct for the same scenario, different belief networks, but the correct 1 will have certain desirable property, so, we will study this 1 first and then later on, we will see that if we model the same thing in a different topology, then, we will get a different kind of probability.

So, what this says is, it says that burglary is an independent event and the probability that a burglary occurs is 0.001. Probability that an earthquake occurs is 0.002. These figures are all from Los Angeles; the probability of earthquake is more than the probability of burglary, then, we have this alarm and this table tells us that given that it is a burglary and given that there is an earthquake, what is the probability that the alarm goes up? So, if both burglary and an earthquake has taken place, then, it goes off with 0.95 probability.

If there is a burglary and there is no earthquake, then also, with 0.95 probability, the alarm rings. If there is no burglary but there is an earthquake, then, the alarm goes off with probability 0.29. Sometimes, when there is no burglary but there is an earthquake, the alarm mistakenly goes off. If there is no burglary and no earthquake, then, the probability that the alarm goes off is very less- 0.001. If you sum up all these probabilities, then, you will find that you have one. (Student speaking). No, I think this should be 0.01. (Student speaking). Wait. Just a minute. If these cases were exhaustive, then, you would get one, right?

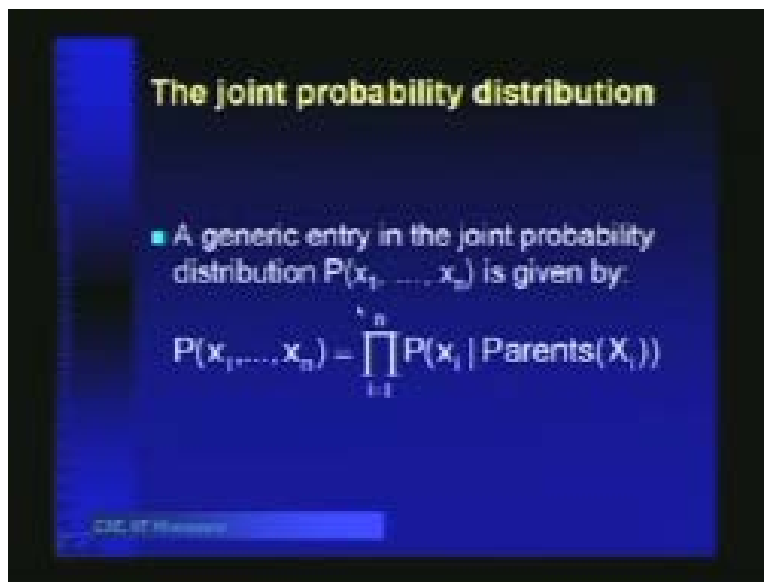
(Refer Slide Time: 18:59)



If these cases were exhaustive, then, you would get one, but these are not always independent. That is why you will get this thing- this overlap is there. Okay, I will come back to this and explain this later. Here, we have John calls and given the alarm takes place, the probability that John calls is 0.9, and if the alarm does not go off, then, the probability of John calls is 0.05, because this is when John mistakes the telephone ring to be the alarm. And when we look at Mary calls, then, when there is an alarm, then, Mary calls with 0.7. The remaining 0.3 is because of the music, and when there is no alarm, then also, Mary might call, but with a very small probability.

This is the belief network. Now, how do we use this network? First, we will see how we can use this network to answer different kinds of queries and we will also see how we can construct this network. Now, the joint probability distribution of a set of variables is given by this. Now, let us see what we mean by this. The probability that we have that X_1 to X_n take a given value, is given by the probability of each X_i , given that the parents have taken place. Let us see an example. Probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred and both Mary and John called- let me write this down here. So, I have probability of John calls and Mary calls and there is an alarm but no burglary and no earthquake.

(Refer Slide Time: 22:43)



What we do here is, we will start baking up this as follows. We will break this up, as probability of J given A and times probability of- okay, why J given A? Because if you recall, in the belief network that we had here- slides please- in the belief network that we had here, we had J only. Alarm was the predecessor of John calls. That is why, for the probability of John calls only, given the alarm is what is (unclear word).

(Refer Slide Time: 23:20)

The joint probability distribution

- Probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both Mary and John call:

$$\begin{aligned} P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J | A) P(M | A) P(A | \neg B \wedge \neg E) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &= 0.00062 \end{aligned}$$

We have here PJ given A and then PM- text please- given A, because Mary calls given alarm and then probability of A given not B and not E and then times probability of not B and probability of not E. It is the product of also this. So, what does this mean? This means that when I have a set of literals here, then, the probability of the conjunction of this in the belief network is given by the probability of each of those events, given the values of the predecessors of those in the belief network. Now, if some of these were missing, like, if A was not there, then also, we will have to bring this here and then listen about probability of A.

If we just wanted to know about probability of J, for example, then, we will have to first look at probability of J given A times probability of A. (Student speaking). Yes, and also for not A. Then, plus probability of J given not A times probability of not A. In this case, because we were given A, that is why these terms disappear. So, we will have this and then, again, this probability of A will be broken up similarly in terms of the probability of A, given what were the pre-predecessors of A, B and E.

(Refer Slide Time: 30:21)

$$\begin{aligned}
 &P(J|M,A,A-BA-E) \\
 &= P(J|A) * P(M|A) * P(A|-BA-E) \\
 &\quad * P(-B) * P(-E) \\
 &= 0.9 * 0.7 * 0.001 * (1-0.001) * (1-0.002) \\
 P(J) &= P(J|A) * P(A) + P(J|-A) * P(-A) \\
 P(A) &= P(A|B,E) * P(B,E) + \\
 &\quad P(A|-B,-E) * P(-B,-E) \dots
 \end{aligned}$$

So, B and E. Then, probability of A given not B and not E; probability of A given B and not E, and so on. We have to break it up like this and analyze them. So, the idea is that for any event, if you have to compute the probability, then, we have to compute it in terms of the predecessors of these events in the belief network, and those probability values are the ones that are given in the table. For example, when we have PJ given A, that probability is given in the belief network, so, if you look at the belief network, then, probability of J given A is 0.9, right? For this, we will take this down as 0.9 times probability of M, given AM. Given A here is 0.7, and then if you look at this probability of A given not B and not E not B and not E, is this 1.001?

Probability of not B is what? 1 minus 0.001, right? And probability of not E is 1 minus 0.002. Slides please. (Student speaking). Here- text- yes. (Student speaking). No, this is not product; this will also have to be done with PB and E and then, plus, with- right. Now, this makes sense, that when we have the joint probability distribution of a set of events, then, what we have here is the product of the P Xi given parents of Xi for each of them. (Student speaking). Yes, the belief network is attempt to model the cause-effect relationship between the events along with their probability value. There are 2 phases in

this reasoning; the first phase is to learn or the model the belief network. How do we do that?

One way could be that we know the events and we know the cause-effect relationship between the events. Suppose for a medical diagnostic system, the doctor tells us that if this happens, then, that will happen with so much probability, and so on. And we just write it down in the form of a belief network. That is 1 way of constructing the belief network. And then, when we actually do the diagnostic and we want to find out that if that person has fever, then, with what probability does he have leukemia? Then, we can find out those probabilities by analyzing the belief network. That is the use of the belief network.

But there has been also a significant amount of research on learning of belief networks from experimental data and 1 of the most interesting work that has been done in the last couple of years is on the following things: If you have a genome: a genome is the DNA sequence that we have, and there is a thing called DNA microarray, with which you can do experiments and find out that, for a given sample, which are the genes that are being expressed. Suppose we take a cancer patient and we analyze for a given protein injection, that what is the set of genes that are being expressed. We have some 20 samples of them, so we know that these genes have been expressed and similarly, we have for the healthy people also, a set of genes which are being expressed.

Then, what we try to do is, we try to find out cause-effect relationships between these genes, because expression of 1 gene produces some protein, which in turn, causes some other gene to express. For example, when we are born, we are born with only 1 cell and that cell multiplies and creates all the organs, etc. How does this happen? It happens because there is a genetic pathway through which it happens. So, depending on what we have in the cell sap, certain genes will express more at that time. There are genes in our body which are expressed only during the first couple of weeks of fertilization and thereafter, they are not expressed anywhere any time in the future. So, what happens is,

those genes will create some proteins which will cause other genes to express and slowly, this sequence of expressions will cause the entire organism to develop.

There is a lot of research to model this in terms of Bayes networks and to discover that what is this cause-effect relationship. What are the steps that we have to do there? We have to first decide which way the links will go; we have the set of events- the events are- this gene expresses, this gene expresses, etc., but we do not know which causes which. 1 thing is to learn the direction of those links and the other is to learn the probabilities of those links. Learning that from experimental data is a subject on its own, which we will not address into in detail.

But once we have the belief network, then, we can always reason and find out the probabilities of more complex events from the belief network, which is what we will study in more detail. The key feature of the belief network is conditional independence. Now, we have noted here- okay, let me go back to this slide- we have noted here, that the joint probability distribution in a belief network is given by probability of X_i , given the parents of X_i , and take their products and you will get the probability of X_1 to X_n . Now, how does that come from? Where does that that formula come from? It comes from here. So, I start with $P_{X_1 \text{ to } X_n}$, so, I can use Bayes rule to break it up like this, then, I can use Bayes rule on this part to again break it up like this.

This 1 will again break up into X_n minus 1, given the remaining ones and the probability of X_n minus to through X_1 , which recursively can get broken down and then, we will have this as the result. Now, what is this ordering that we have? This ordering of the variables is 1 topological order of the belief network. So, the larger indexed variable is towards the bottom right, so, when I am analyzing X_i , then, all this X_i X_1 to X_i minus 1 are all variables which are topologically having a lesser number than X_i , right?

(Refer Slide Time: 35:48)

Conditional independence

$$P(x_1, \dots, x_n)$$
$$= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \dots P(x_n|x_1, \dots, x_{n-1})$$
$$= \prod_{i=1}^n P(x_i | x_1, \dots, x_{i-1})$$

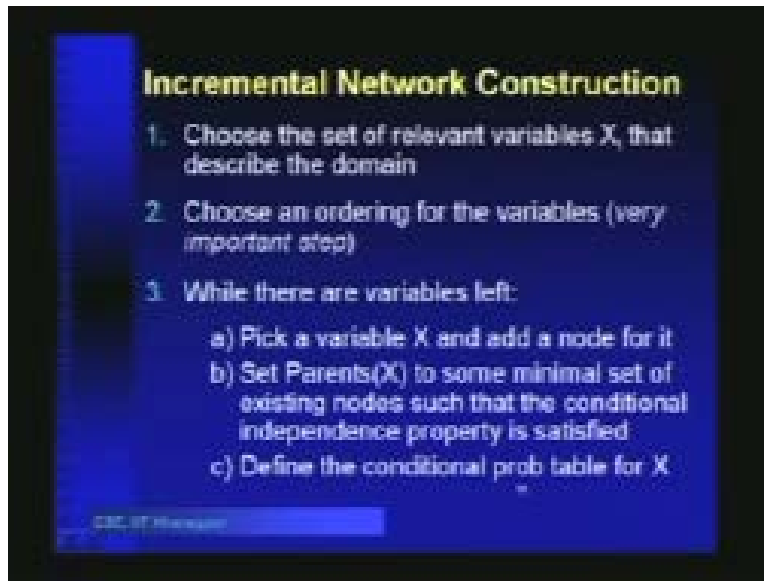
- The belief network represents conditional independence:
$$P(X_i | X_1, \dots, X_n) = P(X_i | \text{Parents}(X_i))$$

These set of variables is the super set of all those variables which can influence the value of X_i , but all these are not going to influence X_i , not directly. I can express the cause. The cause of X_i expressing in terms of the parents of X_i in the belief network. So, the belief network tells me that okay, it is these parents of X_i whose values are instrumental in determining the value of X_i . You could have other factors influencing it transitively-parents, parent and any ancestors, like that. But if I knew the values of the immediate parents, then, I can get the probability of X_i by looking up the probability table that is there in the belief network.

Of these, which are not parents of X_i can be dropped, because this term is going to remain the same and that is what we mean by conditional independence. It means that this term can be simplified to just the parents of X_i $P(X_i | \text{Parents}(X_i))$. That is the conditional independence and that makes our analysis much simpler than having to do with all the variables together. How do we construct the network? I will just outline this today and we will discuss it in more details in the later classes. Choose the set of relevant variables X_i that describe the domain: that is selecting the set of event.

Choose an ordering for the variables; this is very important- the ordering has to be such that we have the cause-effect relationship in the proper direction, but what happens if you choose an incorrect ordering? We will discuss that later. You will still be able to construct a belief network, but there will be certain problems. We will come to that later, then, while there are variables left, pick a variable X and add a node for it. Set parents of X to some minimal set of existing nodes, such that the conditional independence property satisfied.

(Refer Slide Time: 41:05)



So, a minimal set which, such that I can write probability of X , given all its predecessors, is equal to the probability of X , given its parents, so, that is what we mean by conditional independence. And then, define the conditional probability table for X . This can be given or it can be extracted from the experimental data. Now, see, this step 2 is the most important step. If you are able to do this properly from your existing knowledge, then, learning the conditional probability table is much easier, because then, you can actually just check out what is the probability of X , given the parents of X , by just analyzing experimental data.

But if you do not know this ordering, this ordering is incorrect, then also, you will get conditional probability tables, but they will not be- what will happen is that this set of parents of X is going to be very large, and the larger the parent set of X , the larger is the size of your conditional probability table. If you have 2 parents of X , then, you have 4 entries in the conditional probability table. If you have ten parents of X , then, you have 2 to the power of 10 entries in the conditional probability table. So, if you have the ordering appropriately done, then, you will find that the number of parents of X for every X will be limited, otherwise it will become large. We will see in the next lecture, that how we are able to deduce this to some extent.