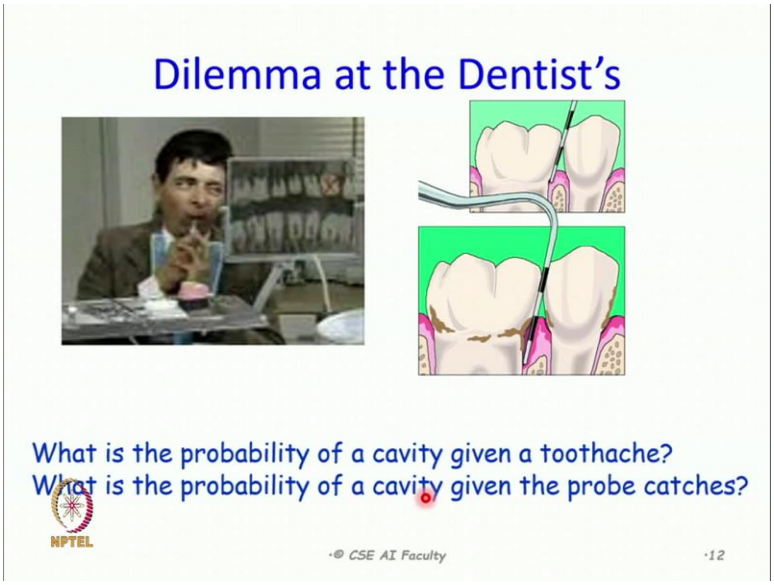
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology - Delhi

Lecture-52
Conditional Independence and Bayes Rule - Part-3

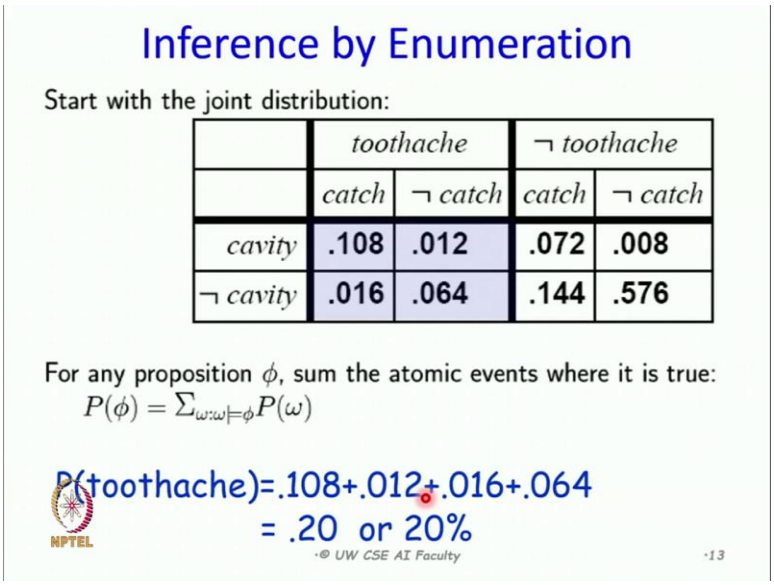
So, at this point we have quickly covered everything that you already know about probability distributions. So let us start with Mr. Bean.

(Refer Slide Time: 00:25)



The first thing I am going to show you is the joint probability can answer every question is obvious but you will take a minute to show this. Our dilemma is that Mr. Bean is going to the dentist office then he feels that I has tooth ache and he might have cavity. The dentist may probe something in his teeth and then put it to test. And we say the probe catches if the particular probe has found some microorganisms.

(Refer Slide Time: 00:55)

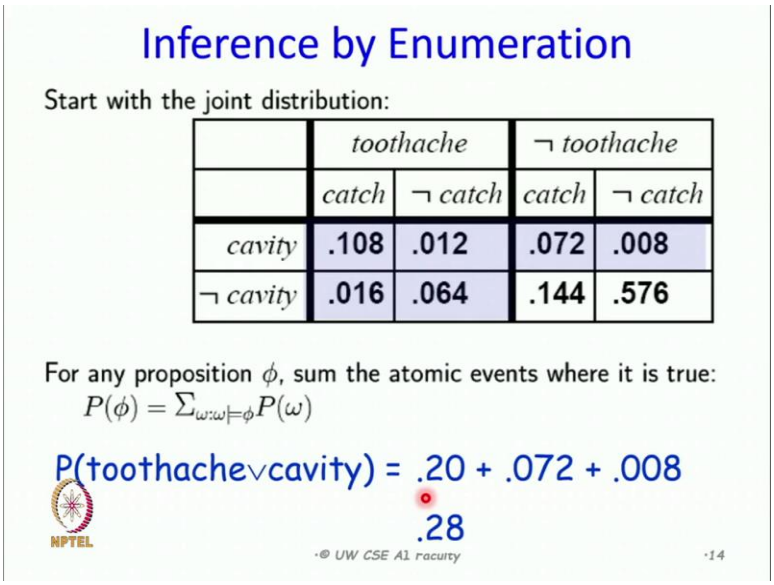


And let us say I am able to write down the full joint distribution, right. There are 3 variables to catch and cavity. They all have Boolean variable, 28 possible atomic events and 8 possible States. For every state, I have a value. This is my joint probability value. What is the probability that tooth ache is true, catch is true, cavity is true. That is 0.108. What is the probability that tooth ache is not true, catches not true and is no cavity is 1.14 and so on.

And now suppose I ask the question, what is the probability that Mr. Bean has to tooth ache? Right. Can you answer this question from this table? What would be the calculation if I have to ask, what is the probability that Mr. Bean has tooth ache, what are you going to do? You are going to sum. And how many values are you want sum? 4 Values, which 4 value is the left, right? Left side. So you are going to sum all these four values because you have no tooth ache, can happen with catch cavity, not cavity not catch cavity and catch cavity.

It is exhaustive, need to be added for you to get probability of Bean. So, and then, you can ask the question that what is the probability that Mr. Bean has tooth ache or cavity?

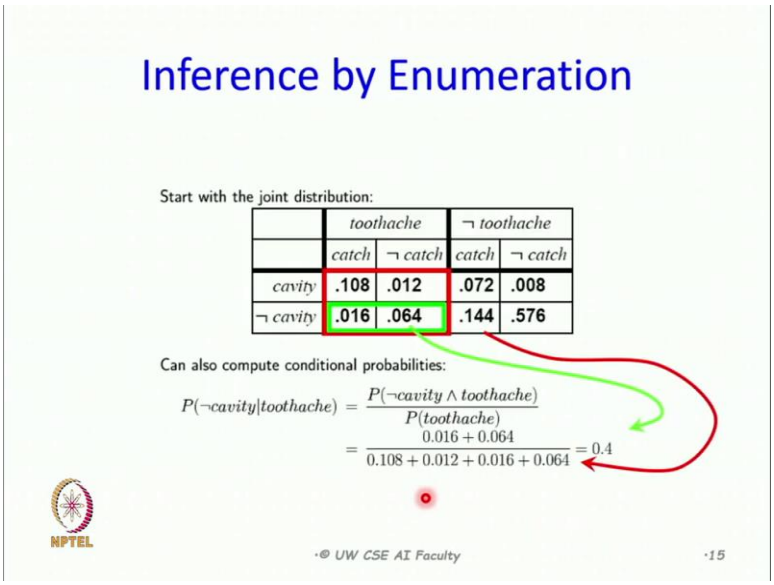
(Refer Slide Time: 02:21)



And again, you can compute this by this table how many values you are going to sum? 6 values, you are going to sum 6 values. Not only are you going to sum the the column which represents tooth cavity and you are also going to sum the cavity and sum of the common cavity and so you are going to add twice and you will get 0.28. And last but not the least suppose I ask the question, what is the probability that Mr. Bean does not have cavity given that he has tooth ache?

I tell you that he has tooth ache I am asking that if he has tooth ache, what is the probability that he does not have cavity? What are you going to do? Is it also going to be the sum, it is going to be a, it is going to be a division. And in the numerator I am going to add how many numbers two numbers in the denominator and how many numbers?

(Refer Slide Time: 03:26)

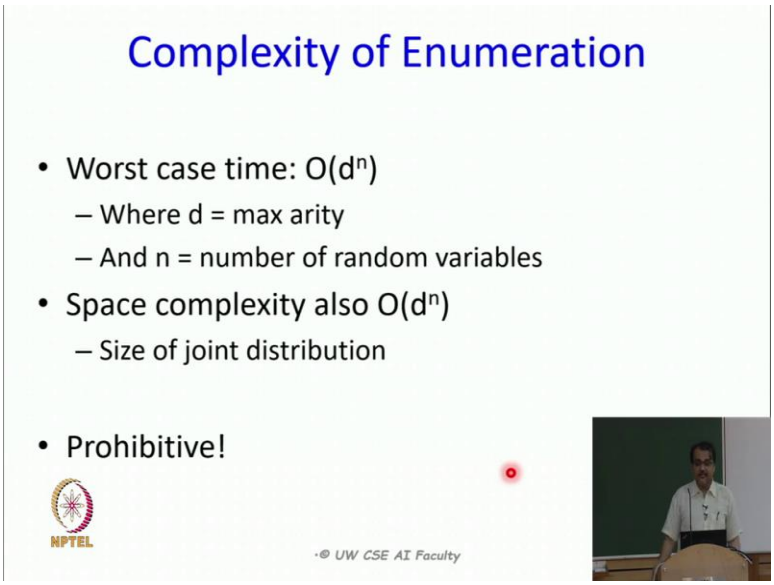


So, you can see that if I want to ask the question, what is the probability that Mr. Bean does not have cavity, given that he has tooth ache, then, basically I am interested in if I go by the computation of conditional probability. I am interested in probability of not cavity and tooth ache might be a probability of tooth ache not cavity and toothache is green box because the two numbers a cavity is false, tooth ache is true.

And the probability of tooth ache is the red box that is the set tooth ache is true and so you compute the condition. For What I have shown to you is, if I give you the full joint distribution, Complete all the numbers have been filled, then, I can add the appropriate numbers, divide the appropriate numbers, to answer any kind of query. I can answers probable conditional probability question; I can answer unconditional probability questions and answer all.

But unfortunately, what is the problem here? What is the problem that using the joint distribution for answering queries? Exactly, how many such values do I have to maintain in my system. And how many values we have to add and subtract.

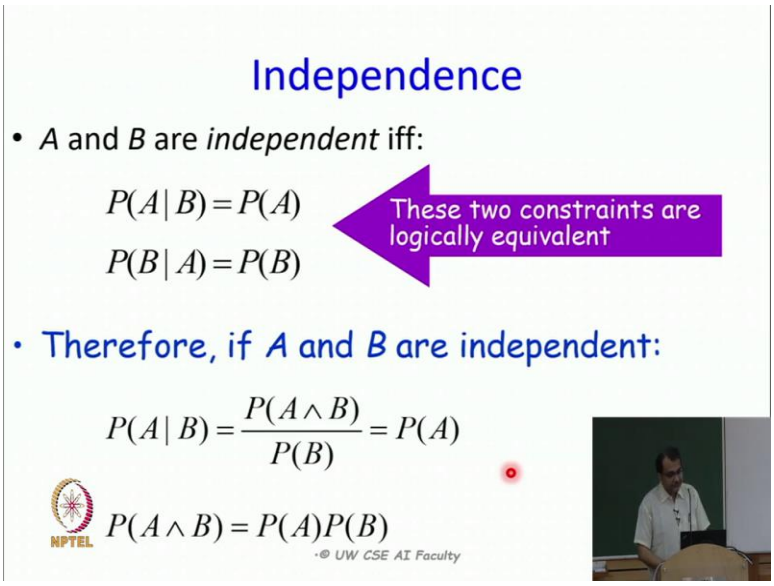
(Refer Slide Time: 04:45)



Worst case time complexity for answering a query is order to the power n where d is the domain size of each random variable and another number of random variables because that is the number of atomic number given, that is the number of states. And for each state I have to maintain a value and then how to add, subtract and even if I might be able to do that, but even space complexity result. Basically, we are just not going to be able to enumerate the full joint distribution, right?

So, with this very fast analysis, we have seen that full joint distribution is amazing, extremely helpful, we can do everything and is no way we can mention. And that is going to slowly lead us to where we want to go. But now I am going to introduce the most important concept, first of the two most important concepts, but this concept you already know it is the concept of Independence.

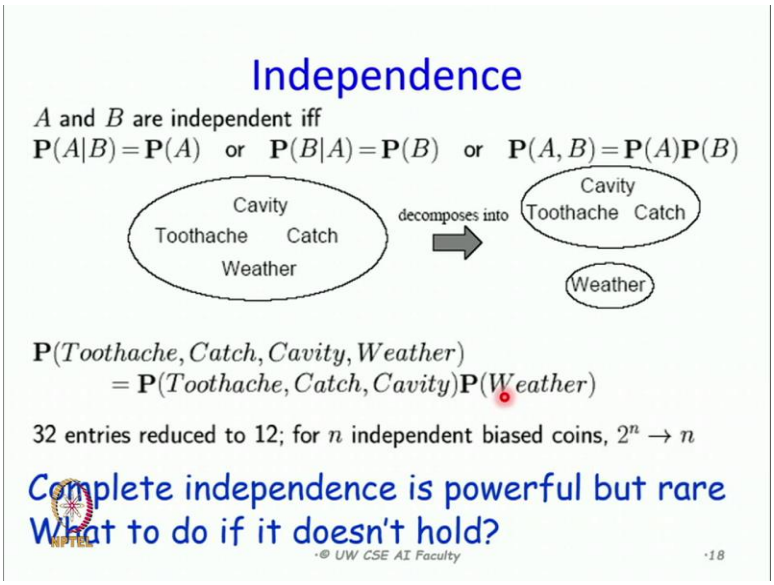
(Refer Slide Time: 05:46)



We are going to use this beautiful idea of independence to actually make our life simpler in reducing the values that we have, to store for a given distribution. So, first of all, when are two variables called independent? In, intuitively it says that B is called independent with A if by knowing B, I do not change my probability distribution of A. I do not get any new information and that does not change my belief about A.

And mathematically that means that by doing B, That is A given B, I do not change my distribution at least the probability of that area. And of course, probability of A intersection B is equal to probability of B given A, so these are equivalent statements. If A does not added any summation about B then B does not add any summation about A. And therefore, another definition of Independence is probability of A and B is equal to probability of A times probability of B comes from this very simple derivation. So now, why is independence important?

(Refer Slide Time: 07:11)



And this slide illustrates, suppose A and B are Independent. Let us say in one set I have cavity, toothache and catch in the other set I have the weather. Let us say sort of assume that weather does not affect, does not add evidence to any of the 3 whether you have cavity, whether you have catch or whether you have tooth ache. So now, let us think about the joint probability distribution. In the joint probability distribution of toothache, catch, cavity and weather, how many original values will I have to maintain? Ok.

So these are simple questions but I no need everyone to start this answer in simple questions because I am going to explain a very important concepts. So I want everybody's concentration. How many values for weather? 4; how many values for cavity, 2; how many values for catch; how many values for tooth ache; how many total values in my joint probability distribution 32. Do I need all 32 to maintain the distribution?

How many do I need? Because the sum to 1 I need. 31 ok this is where we are. We actually have 31 values in my joint probability distribution. Now let us suppose I tell you, that weather is independent of all these three other things. What does that mean, Right? What does that mean that weather is independent of all the other things? It means that suppose I use the product rule and tooth ache catch cavity weather.

It is equal to when I use the product rule I can say whether times tooth ache, catch cavity given weather. And because they are independent, toothache, catch, cavity given weather is going to be toothache, catch cavity. Weather does not matter and this is called factorization. I can factorise the original joint distribution into two factors. One factor is weather and one factor is toothache catch cavity. Make sense.

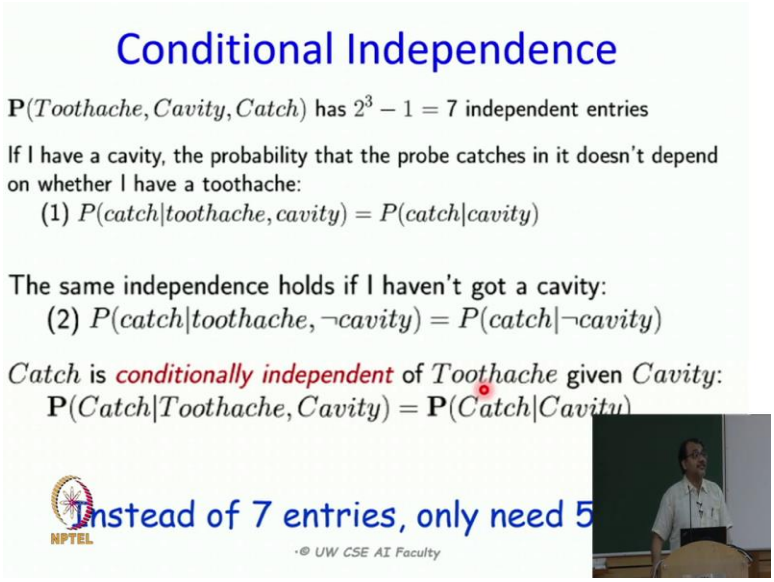
So now, how many parameters have I reduce this two? How many parameters do I need to maintain tooth ache, catch, cavity, joint distribution? 7 because 8 – 1. How many parameters do I need to maintain weather? 3, 4 -1. How many parameters do I need if I know this factorization to maintain the full joint distribution? So basically what have we done? We have said that by the probability of Independence something that requires 31 parameters now only request 10 parameters.

And this factorization could only happen because we had independent, independence of random variables. So, therefore, it is a very, very powerful tool. It might be sometimes able to make an exponential distribution linear, Amazing. Unfortunately, you can alternative ask the question, Why are you modeling weather if you are interested in tooth ache? Like why even model it, just get rid of it. You do not need it. It is an independent variable if a different set of random variable which are related to weather and different variables which are related to dentist.

If it is a dentist question, use those parameters, if weather you can use those parameters. Why do you maintain them in the same system and that is the reasonable question. There is no reason. So, why is complete Independence is incredibly powerful? It is useless. Features does not happen. You do not model it. You do not want to completely independent variables in a system, never. So then we are back to square one, what do we do? And so, what I am going to say is the most important Concept for today's class and this is where we want everybody in purpose.

In the idea of conditional Independence have you heard the idea of conditional Independence before? How many of you are aware of it? 1 person at least you know something. You must be feeling like is this an AI class or an eleventh class and finally we are back to AI. Ok.

(Refer Slide Time: 12:05)



And in order to explain this, we will again use example of toothache, catch, cavity and I am going to ask the question what can we say about probability of catch given toothache and cavity. Think about it. I know that a person has to think. I know that a person has cavity and now I am asking the question are they going to catch the microbes. Now intuitively, what does whether you can catch or not depend on? It depends on whether there are microbes or not in your system and that depends on whether you have cavity or not.

In fact cavity is the cause. And then, we have two symptoms of, two ways to find it. One is whether you have tooth ache, one is weather. Suppose, I tell you the value of cavity, suppose, I that tell you the value of weather I tell you the person has cavity. Then do you get any additional information? By also knowing that he has tooth ache in predicting the probability of catch. You understand the question?

This is important question; it is the way you ask the question. If you do not ask the question like this, you will get confused. By additional knowing that the person has tooth ache, does my Belief in catch changes if I already knew that they have cavity. It does not because actually catch does not depend on toothache. If I did not give you the value of cavity, that is also another way of thinking.

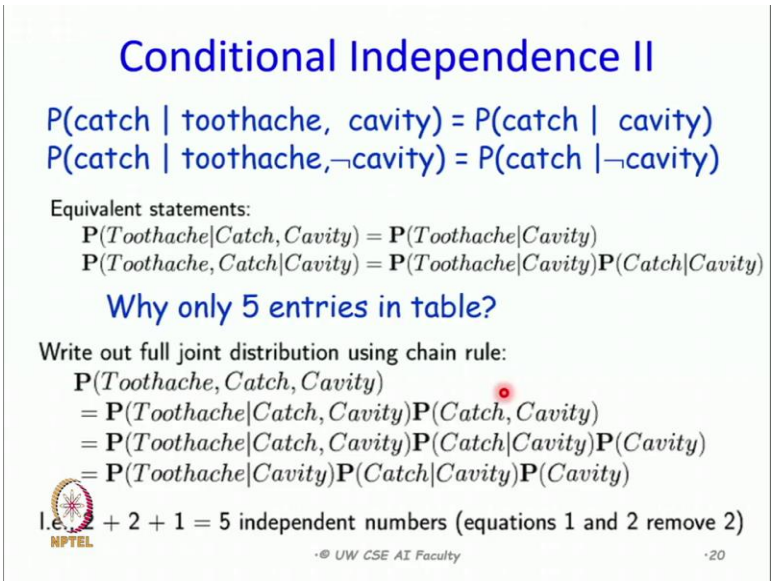
If I did not give me the value of cavity, I only told you that the person has toothache, would you say that increases the probability that the microbe is going to be caught? Important question we have to start thinking about it like this. So, suppose I did not give me the value of cavity. I only said oh, the person has come in, he has tooth ache, do you think that there is a higher probability that you are going to have a catch?

Compare to another patient who does not have tooth ache? Yes, why because he has tooth ache, he might have cavity because he might have cavity my tooth may catch. In other words, you should say not like this. It is wrong to state because he has tooth ache that increases the probability that they have cavity because the probability of cavity has increased therefore, the probability of catch has increased.

And therefore, they are dependent. But if you already give me the value of cavity by additionally knowing that they have tooth ache that does not change the probability as I already know the cause and why to bother about others. So, mathematically we can say the probability of catch given tooth ache, cavity is equal to probability of catch given tooth ache. And similarly, we can also say probability of catch given tooth ache not cavity is equal to probability of catch given not cavity, correct.

When this happens, we mathematically say that, catch is conditionally independent of toothache given cavity. Ok. Once I know cavity, catch and toothache are independent of each other, right A given B is equal to A thing that is a given B, C is equal to a given C which basically means A and B are independent given C. Now I am going to show you is that when this happens the full join distribution does not even need 7 values. It only needs 5. That is the most important part actually in some ways right.

(Refer Slide Time: 16:34)

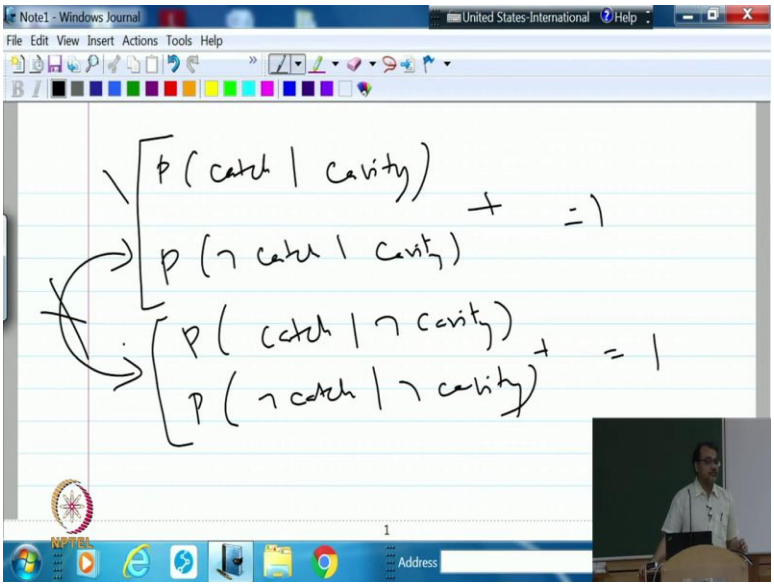


We know that catch given tooth ache, cavity, Catch given cavity, Catch given tooth ache cavity is catch given not cavity, Why do we need only 5 entries in a table, let us think about it? What is the joint probability distribution? It is probability of toothache, catch, cavity. Now, let us use product rule. We can use it in any way. Any subset way and let us say, we use it in this particular fashion. We say this would be equal to probability of toothache given catch, cavity times probability of catch, cavity.

Now by conditional Independence, what do we know? Probability of toothache given cash, cavity is equal to probability of tooth ache given catch. Moreover, we can also divide catch, cavity into two steps catch given cavity and probability of cavities. Overall, we get probability of cavity Times probability of catch given cavity time probability of toothache given cavity so far, so good. Now, let us think about how many parameters do I need if I give you this factorization? How many parameters do I need to represent probability of cavity? Ok Somebody from here.

How many parameters do I need to the represent just probability of cavity? What is your name? Chinnah says 1 because I can have probability of cavity or I can have probability of not cavity, but both of them sum to 1. So we can say 1 parameter. Is everybody with me? Now, how many parameters, it is an important question. How many parameters to represent the probability of catch given cavity? Somewhere from the pass back? How many parameters do I need to represent probability of catch given cavity?

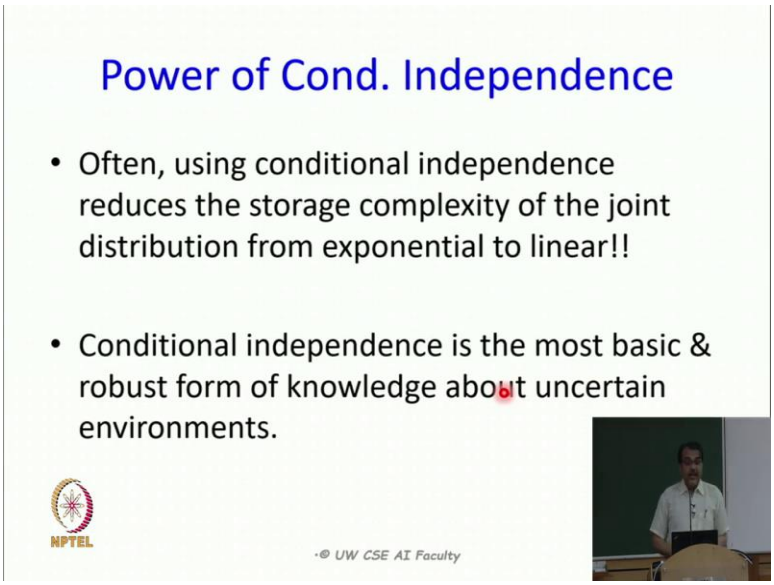
(Refer Slide Time: 18:57)



Now, what are the possibilities? One of the possibilities catch given cavity, one possibility is not catch given cavity, one possibilities catch given not cavity, One possibility not catch given not cavity. Now what sums to 1, what all things sums to 1? First these two sum to 1. If I know cavity either it is going to catch or not catch, both of possibilities sum to 1. And similarly, these two possibilities sum to 1. Is there any relation between the first set and the second set? No direct relation.

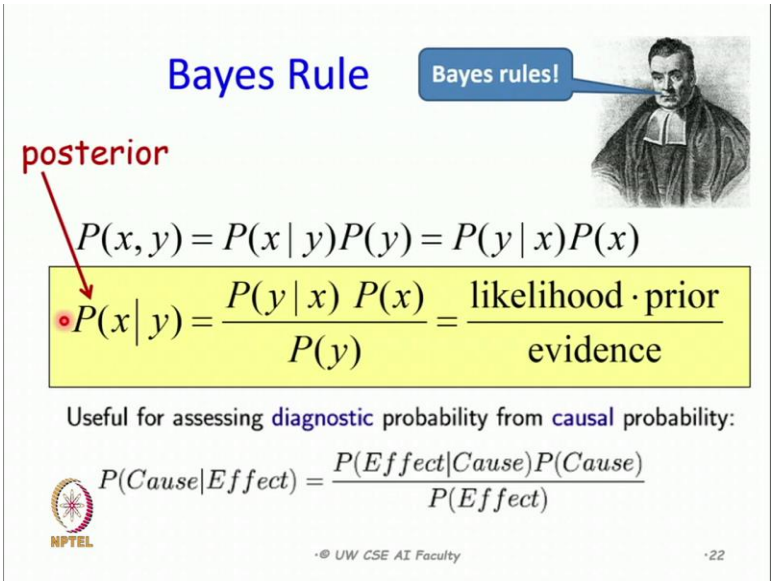
So, therefore to represent this first set, I need how many parameters? 1. to represent the second parameter, 1 therefore, totally I have how many parameters? So for probability of catch given cavity I need 2 parameters. So this kind of reasoning has to become given nature to you after the next 3,4 classes. You should try to start thinking about it in these terms and similarly to the cavity request parameters. Therefore what I have done is I have used the power of conditional independence to go from 7 parameters originally to 5 parameters in 5 independent numbers. And we can use the power of factorization to do all your computation.

(Refer Slide Time: 20:35)



Conditional dependency is extremely powerful. It often is the reason how we are able to reduce the storage complexity of a joint distribution, sometimes even some exponential to Linear. Conditional independence is the most basic but also most robust form of knowledge about uncertain environment ok. I will take 2 or 3 more minutes. I will quickly revise Bayes theorem for you. Everybody knows about it. Most important is this thing.

(Refer Slide Time: 21:06)



So, basically Bayes Theorem was developed by Mr. Bayes, right and he said that if I get additional evidence, then how do I compute the probabilities, what is equation and basically he he was able to say the property of x and y can be written down as probability of y given x, x

probability of x. This is why this all comes from the various ways you can factorise just probability of x, y.

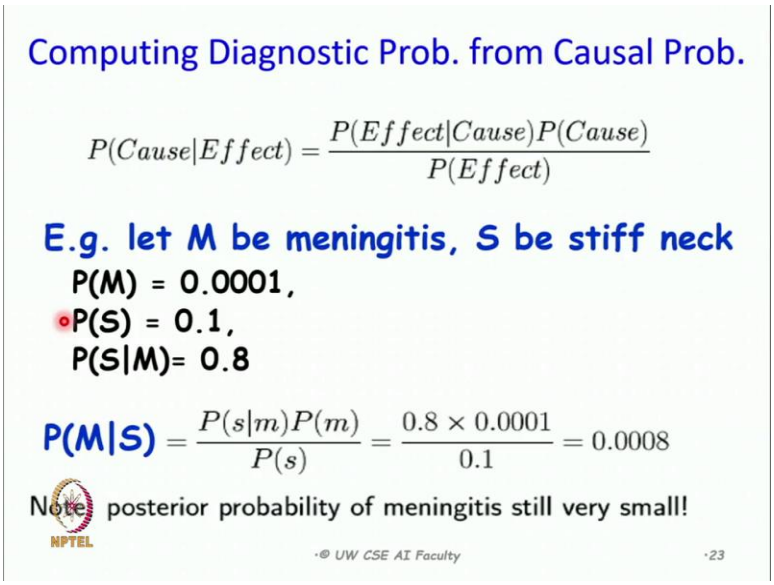
Now this has become the fundamental way we develop our probabilistic models. All this is based on Bayes Theorem. And there is a very famous person who got a Nobel Prize, not Nobel Prize, Nobel prize of computer science, which is the Turing award for developing all this theory about probabilistic models in Computer Science. And last time I asked you to go check it out. So what is the answer? Oh, you actually checked it out. Alright Happy about it.

I know you all want easy exams. Now, when does Bayes rule become important, Bayes rule becomes important because it is easy sometimes to compute the probability distribution in One Direction conditional but difficult in the other direction. So we are saying that x given y can be represented in terms of y given x. And typically, when we have to compute cause given effect, we find it easier to compute effect given cause, Ok because you know, when this happens then this happens, it is easy to compute probability of this.

When outcome happens, what is the reason this must have happened for the outcome to happen, this is harder reasoning for us right in practice? And so, we essentially say that cause given effect is written down as effect given cause and probability of cost divided by probability of Effect. But s effect is not important as I told you in the next slide. But this x given y is called the posterior distribution P of X is called the prior distribution for X and Y given X has a specific term, it is called the likelihood function.

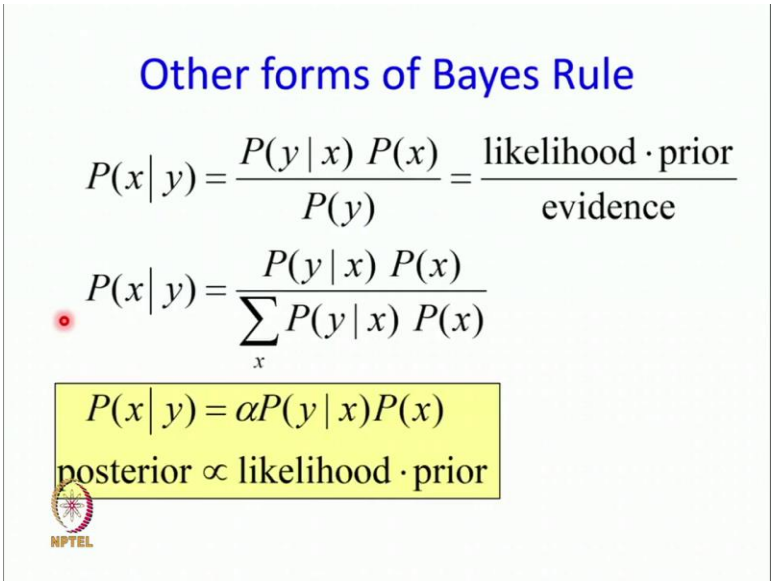
These are just terms; we just have to memorize them. So, posterior is equal to likelihood times prior divided by the evidence. And his is how we often write it down, right?

(Refer Slide Time: 23:27)



And we will skip the example you can check you all know how to use Bayes theorem to compute any value. So, you can double check this example, that's ok.

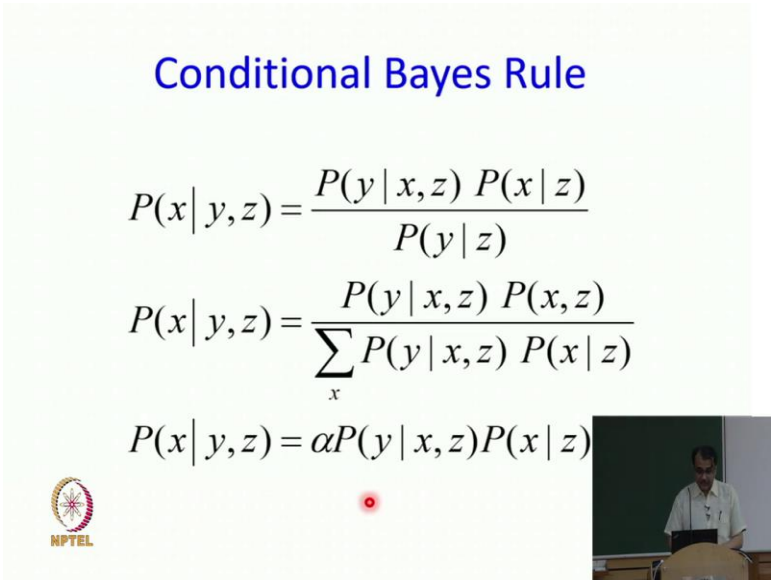
(Refer Slide Time: 23:35)



But the main part I wanted to show you is that notice what is P of y, while in the denominator? First of all, it is all possible causes x of P of x given y, x that can be written down. But moreover we do not care. Why do we not care about the denominator because this is going to be the same for every value of x. So impact is what we do is that we simply compute we can we can say that P of x given y, is proportional to likelihood times prior.

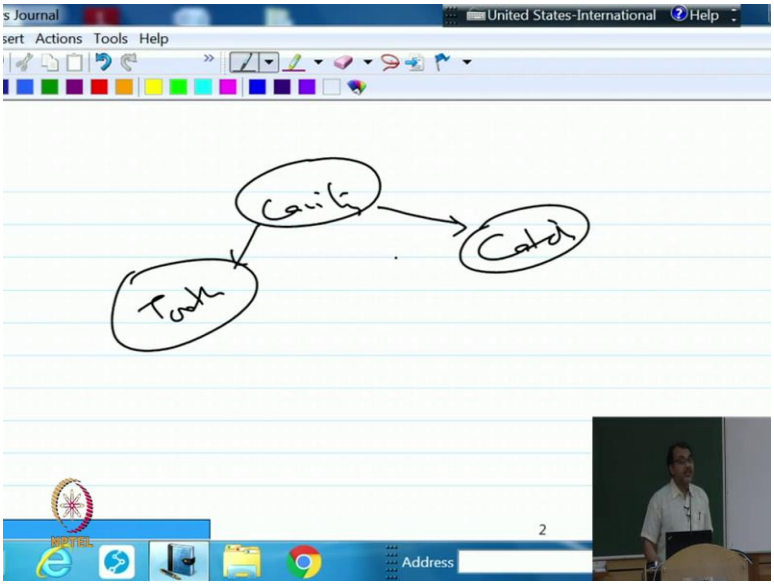
And so in practice when we have to actually compute x given y, we take all possible values of X and compute likelihood times prior for them and then normalise it to sum to 1. That is how we easily compute the Bayes probability of values.

(Refer Slide Time: 24:23)



And we can also have conditional Bayes rule, which basically says that additionally give you z all these equations for x given y, z is equal to y given x, z and x given z divided by y given z we can put given z in everything. That is all the Condition Bayes Theorem.

(Refer Slide Time: 24:55)

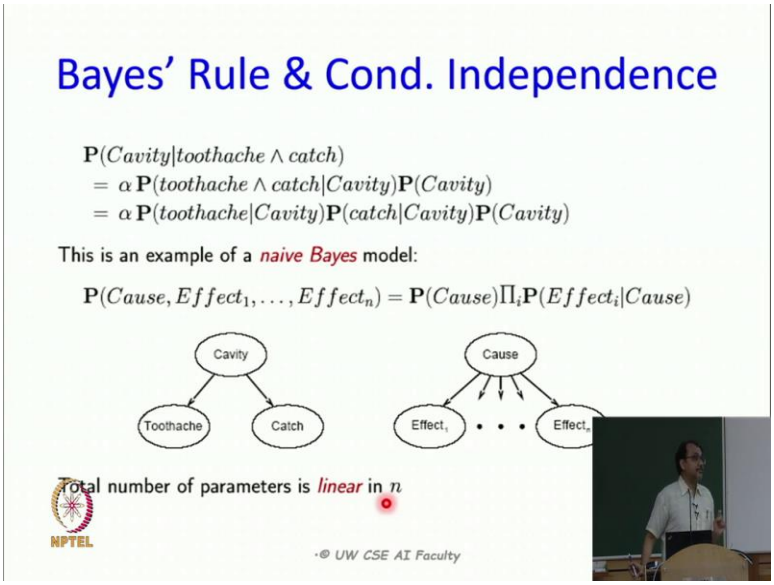


And last but not the least if you have to show me what depends on what let us say. I make it in a graphically I say there is cavity, toothache and catch what depends on what? Suppose there is

there is you know cavity, there is toothache and there is catch. If I have to draw an edge from what influences what? What edges will you draw? Cavity to toothache and cavity to catch, you will make this kind of a diagram, a graph to represent this model.

And what were going to do starting next class is we are going to give semantics to these kinds of graph. This is V-graph. This is the Bayesian network for this particular problem. You do not know what Bayesian network is but intuitively it says what directly influences, what let me just add an arrow ok and this particular way where there is one cause and many independent effect.

(Refer Slide Time: 26:04)



This model is called the naive Bayes model. And you will learn about Naive Bayes model in machine learning course. We will talk about general bayesian network, but naive Bayes is one specific model that you learn in machine learning. But basically, it is saying that if I know the cause all the effects are conditionally independent of each Other because sometimes it was a naïve omission model. So it is called Naïve Bayes model.

And you can check the total number of parameters here is linear in and why because you will have one parameter for cause, two parameters each for effect given cause because if 2 is allowed to N +1, ok. I will stop here. We are basically covered all the basics of probability. And so the starting next class, we will start talking about Bayesian networks the stuff that you have forgot

during our have three to four lectures, just talking about the knowledge representation in Bayesian networks.