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

Lecture-61
Bayesian Networks: Maximum Likelihood Learning

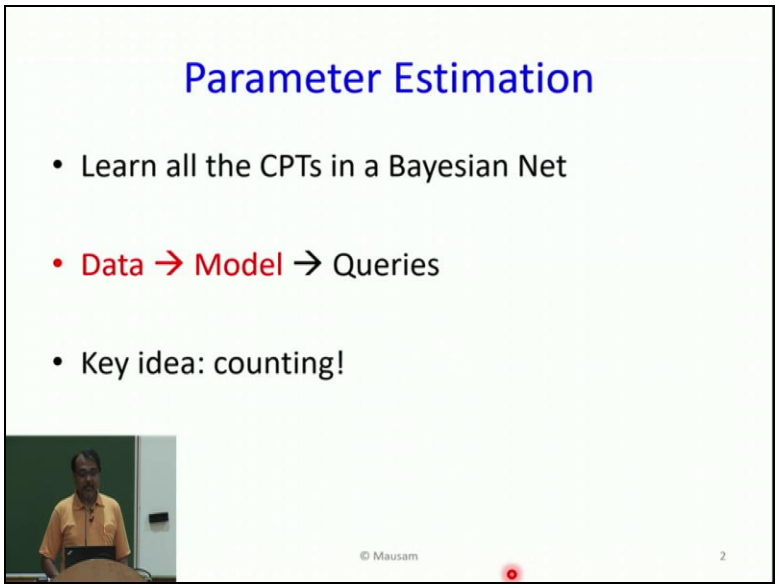
Up until now we have been talking about Bayesian networks as a knowledge representation language and we have talked about its syntax we have talked about its semantics we have talked about the probability distribution that a Bayesian network represents. We have talked about how to compute probability of queries like if I give you some evidence variables and I want to I want you to compute the probability of some query variables how do I do this?

And we have done 2 types of algorithms you have done an exact algorithm which is obviously and be hard and we have also done a an approximation algorithm which scales with the number of samples that you sort of want. So, the more time you give it the better the approximation to the optimal right. But we have not answered the question where do these probabilities come from right. We are assuming that this Bayesian network structure and the conditional probability tables are given to us and then we start working on our inference procedures.

But of course in real life nobody can give us the exact probability tables it is too much to expect from a human designer. A human designer might give you the structure the human designer might say that look your age should determine how often you would have an accident and your education level could determine how often you would have an accident. But it would not be able to say that you know that if the age is greater than 30.5 then it will be this probability and if it is greater than 32 then it will be that probability. A human designer cannot do that.

So but what it can what we can get is that we can get historical data of you know which kind of person had how many accidents and so on so forth. And that might and that should lead us to estimating the conditional probability tables which can then lead to downstream inference right. And this is what is called learning in Bayesian networks.

(Refer Slide Time: 02:24)

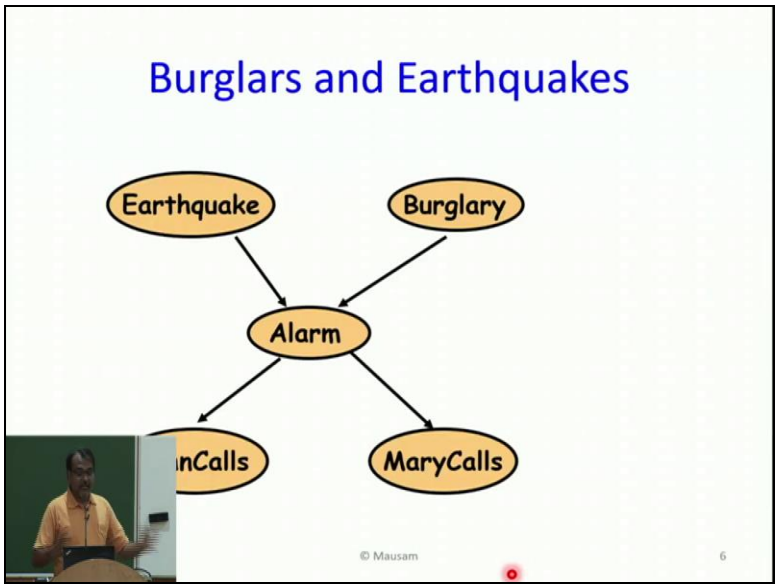


The idea being and there are different kinds of learning and for most part in our class we will talk about learning the parameters which means learning the conditional probability tables. Up until now we have been doing given the model give me the solution to a query. So, model to query this is what is inference but data to model and that is what is learning. I give you a historical data and on the basis of that you estimate the model that is called learning right.

And believe it or not the key idea that we will learn today we will learn 2 or 3 different ideas actually but the main idea the first main idea that we are going to learn today is very simple it is called counting. And the idea being that if I tossed a coin 100 times and 50 times I got heads in 50 times I got tails and I said what is the probability of heads? I given you some data and I am asking you to estimate a probability and how would you estimate the probability?

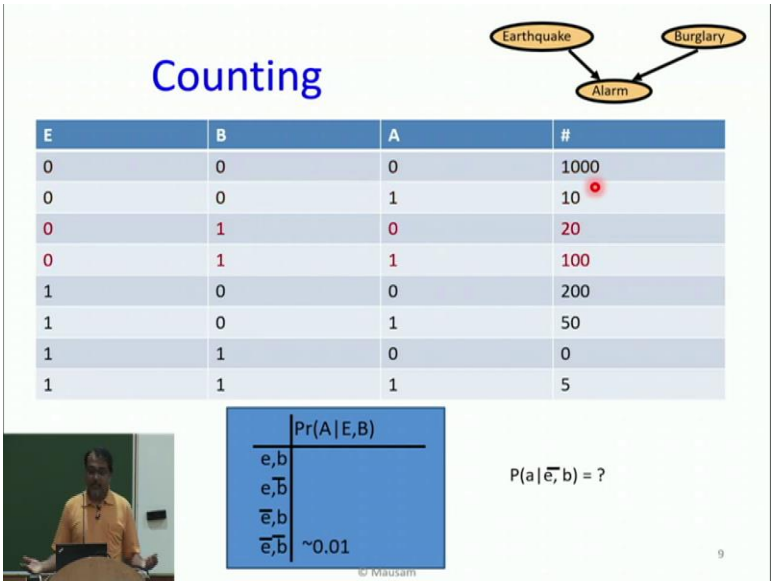
50 divided by 50 + 50 and that would be as estimation. So, all probability computation a lot of probability computation is nothing more than counting.

(Refer Slide Time: 03:41)



Of course we will discuss this in the context of our favorite example earthquakes and burglary and alarm and right calls. And let us just look at the part where I am trying to estimate probability that the alarm will go off.

(Refer Slide Time: 03:55)



Given that earthquake may have happened burglary may have had happened and allow me aback let's say that somebody some company goes around installing these alarms and some test houses and then gets all these samples. So, this is what my data says my data says that if there is no earthquake and no burglary 1000 times there was no alarm but 10 times there was some alarm right and so on and so forth right.

So this is how I have represented the whole data I am saying that this is the number of times a situation like this happened. And now I want to estimate the conditional probability table of alarm given earthquake and burglary right. And you can do this with me right this is going to be extremely simple. So, for example what is the probability of alarm going off assuming neither earthquake happens no burglary happens.

Somebody, what is in the numerator? 10 what is in the denominator? 1010, right. So, when you think about this for this particular question only the top 2 rows are important because we are saying given earthquake did not happen in burglary did not happen. So, we are only interested in these top 2 rows and in these top 2 rows you will see that 1000 times alarm did not go off 10 times around did go off so the probability estimated would be 10 over 1010 this looks simple enough right.

What can go wrong let us keep moving. So, if we have to estimate probability of alarm going off assuming there is burglary but no earthquake that will be 100 over 120 a high enough number that is pretty good. Alarm given earthquake happen but no burglary happened that is a smaller number 50 over 250 so far so good. Now let us think about probability of alarm given that there is both earthquake and burglary.

What is this probability? It is 5 over 5, if I give you this table with the probability 1 in it is that enough are we done or is there something we need to also secondarily take care of we should think about it. Is it a good idea to have probability 1 in the conditional probability table or 0 for that matter? Why not? Lots of people are nodding their head sideways, why not? What can go wrong? What happens if I have probability 1 or 0 in my condition probability?

There is at least one algorithm that we have studied that does not work which one which algorithm does not work? Markov chain Monte Carlo algorithms do not work Poorva says remember we discussed this at the time of Gibbs sampling we discussed this that you know if I have 1 and 0's then I had to create modes in my state space. I create 0 probability transitions and it is possible that there is a lot of probability mass here but if I start in one mode I will never be able to jump to a different mode that is one problem.

But there is another problem, what happens is that this look so can it never happen that there is earthquake and burglary but alarm did not go off? Is it at all a possibility? It is sort of possible right maybe there was the battery was not on. A battery had drained out. You know some crazy thing might have happened some technical glitch could have happened today's earthquake there is burglary but the alarm could not sense it or alarm was not active or something like that.

Now what has happened is that make by making this probability 1 we have said that that state is quote and quote impossible. Now if I have a discrete state space then 0 probability means, impossible. If I have a continuous state space and 0 probability actually does not mean impossible rather that that I will let you to figure this out right. We have done all done math and probability right but we are interested in the discrete spaces for now so we will not worry about it.

So when you have a 0 probability state in a discrete state space in a finite discrete state space then 0 probability actually means impossible. And what we are doing is that we are saying something is just disallowed it can just not happen. So, for example if I give you a scenario where earthquake has happened burglary has happened in alarm does not go up my Bayesian network has no idea that denominator becomes 0 it just does not know how to react.

Because in its model of the world such a state cannot exist and in reality things that are very there will get 0 probabilities because when you do data collection many of these the real states will not even occur in your data. So, even if you think about this for the top case we had 1010 samples of course because in most cases neither earthquake nor burglary. Then you simulated burglary, how did you simulate burglary? When somebody is trying to break it so that is alright.

So when you simulate burglary you might get more cases. Then how can you simulate earthquakes? Very difficult to simulate earth quack, so, you literally wait that there is an earthquake then I will get my probability estimate. But how would you simulate a situation then there is earthquake and there is burglary? Very rare situation to simulate as soon as there is a earthquake somebody calls this earthquake go burgle the house so that we can get the data.

Just not going to happen by that time the earthquake has passed away. So, you will not get that much data for that situation because you do not have enough data the conditional probabilities that you will learn from that very small amount or attack will be completely wrong basically. And furthermore you are calling it 1 that is just make going to make a model completely wrong. So, what do we do? Our objective function is we do not like 1's or 0's.

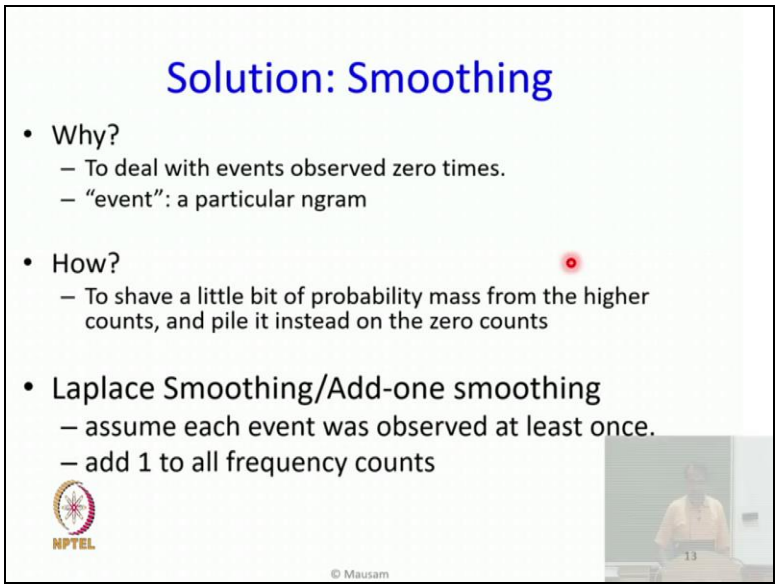
What can we do any guesses any ideas any suggestions. nNow these are heuristics at this point right so technically if you think about the theory of probability what we have computed is accurate but we will make it practical right. So, you will see over time and we know this is not a machine learning course so I am not going to delve into the details of this. But machine learning does not mean fitting the training data and that is very interesting.

This fits the training data perfectly. Machine learning does not mean optimizing such that you fit the training data. Machine learning means optimizing such that you sort of fit the training data but you generalize to unseen data really, really well. And if you do not get this it is ok there is a full-course waiting for you in the wings where you will learn about machine learning in detail. There is a big difference or small, depending upon how you want to think about it but there is a significant difference between optimization which is what is fitting to the data.

And generalization which is what is fitting to the data just enough but not completely such that you are robust enough to unseen data points to generalize to new data points. So, now at this point we have to employ some tricks of the trade so that we get away from 0's and 1's any suggestions? Yes what is your name? Jai, so Jai says some data points have just not been observed. Let us artificially observe them sometimes let us say once.

And this is an artificial construct before we start estimating we assume that every world has been seen once. For example and this by the way is called smoothing where is it called smoothing?

(Refer Slide Time: 12:52)

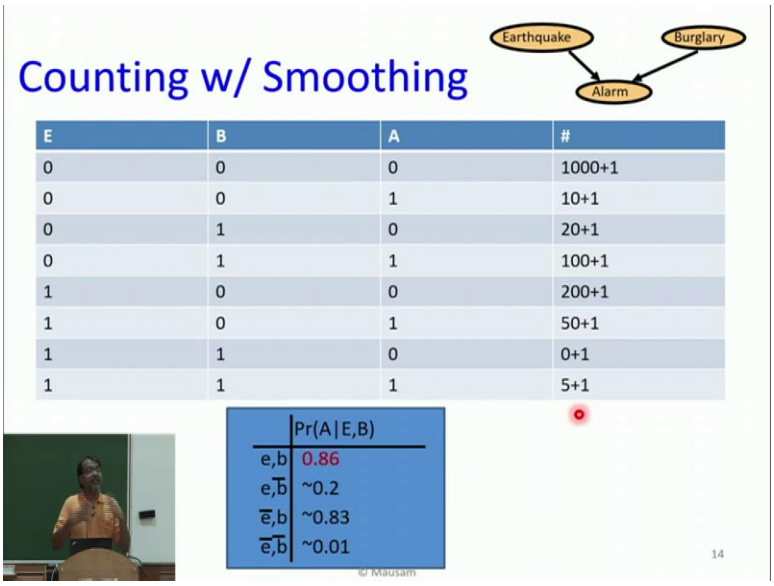


It is called smoothing because it tries to deal with events that have been observed 0 times by taking away some probability mass some events that have been observed and giving it to the events that have not been observed you smooth the probability distribution your probability distribution was too spiky you smooth it out such that everything has some mass. And what we do is we assume for example what Jai says is what is called the Laplace Edwin smoothing or Edwin smoothing or Laplace smoothing.

The earth right probably Laplace came up with this I am not sure. So, we assume that each event was observed at least once and then more right. And you can also do atom smoothing m could be greater than one m could be less than one they have different variants. What m to use depends upon the situation depends upon how many events have 0 probability etcetera. Believe it or not when we do NLP course we spent 2o lectures one full week on just smoothing.

It smoothing can get quite complicated in some situations but for our purposes in the AI class we will just say we will add one right. So, let us say we add one if we added one.

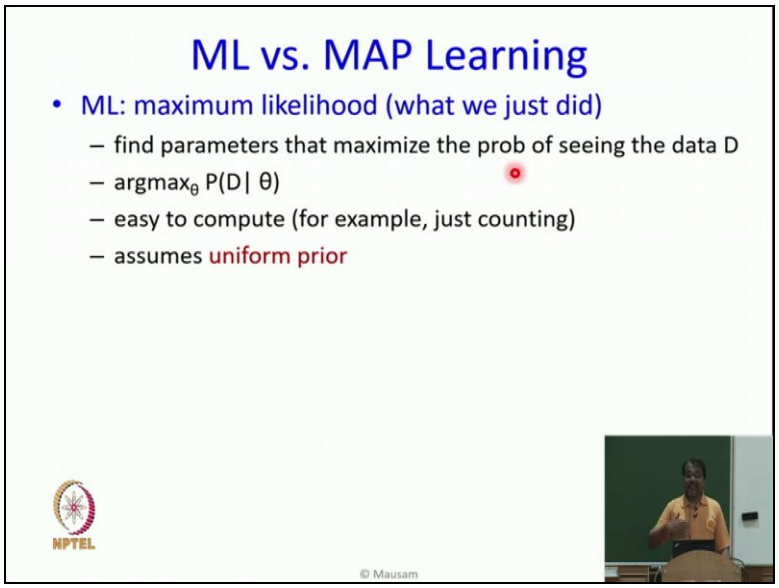
(Refer Slide Time: 14:05)



We got you know the first event which has happened 1000 times now happens 1001 times big deal. But the event that happens 0 times now happens once and the event which happened 5 times now happens 6 times. So, now when we are re-estimate the probabilities we do not get the one we get a high enough number. Now you can argue that this number is too low we may want to add less than one etcetera but that is, it does not matter. Notice that other numbers do not change by much.

If you go back the numbers were 0.2, 0.8, 3.01 those numbers sort of remain the same not much difference. However the value that we were estimating on a very small amount of data that makes some significant change. So this is about smoothing in learning and what we just did and we did not do the math is what is called the maximum likelihood learning. So, let us take a few minutes and think about what it means.

(Refer Slide Time: 15:09)



I gave you a data some data and I wanted you to estimate the parameters. Mathematically what kind of parameters do I want to estimate? I want to estimate those parameters that fit the data. What does it mean to fit the data? That means I am optimizing over the parameter space such that I maximize some objective function and this objective function should be you know fitness to the data. And what is that objective function that objective function is probability of the data given the parameters.

I want to find such parameters which maximize the likelihood of generating the; if that Bayesian network with those parameters was used to generate that specific data that I was given as training very interesting. It is saying that now I have parameters for a given set of parameters I can create a Bayesian network for a different given set of parameters I can create a different Bayesian network. I want to find that Bayesian network which maximizes the probability of the generating the data that I have seen.

Remember Bayesian networks are generative models it can use be used to generate data for every data point I can generate it using Bayesian network which basically means I can figure out what probability I would have generated this data point with if I was this Bayesian network was the true Bayesian network of the model. I can take all those probabilities and multiply them and that is the probability of generating this training data.

And I want to find those parameters where this probability is maximized this is called the maximum likelihood learning right. Finding the parameters that maximizes the likelihood of seeing the data B. Believe it or not if you write down the mathematical expression and if you take the derivative with respect to each parameter and set it to 0 you get nothing more than counting. So, what we did intuitively that we said get 10 over 1010 there is actually a mathematical basis for, what I might should be optimizing.

So that I get exactly that particular parameter I am not going to the proof of maximum likelihood but it you can do it as an exercise actually. You may not do it in two minutes it may take you half an hour of an hour. But it is not at all hard just write down the expression of the data that you have observed with respect to the parameters. Parameter is the conditional probability of a given transition and then takes the derivative with respect to that parameter and set it to 0 you will get exactly the numerator and denominator that we are interested.