

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

**Lecture-63**  
**Bayesian Networks: Bayesian Learning**

What this learning have one significant disadvantage? The disadvantages and this will take you some time to recognise the beauty of the disadvantages that they are looking for one Bayesian network that best fits the data. Suppose there are three Bayesian networks. This has probability 0.3 this has probability 0.3 this has probability 0.4, let us say this is first second and third. First and third has probability 0.3 each and the second highest probability 0.4.

Which Bayesian network will be picked? Bayesian network number 2 I will find one set of parameters that are the best and that Bayesian network will become my model and alternative way to handle this would be do not pick a Bayesian network. Maintain all this 3 Bayesian network in as possible hypothesis and then if you are given a query like tell me probability of A given B whatever, anything like this. I do it based on all 3 major networks. And then I give you the best value.


For example, probability A given B is false was 0 by Bayesian Network 1 and then it was 0 by Bayesian network 3 but it was the 1 by Bayesian network 2 then if I only pick Bayesian network 2 then probability of A given B will be 1. But when I kept all 3 in my system and then try to estimate A given B it turn out to be 0.4. This kind of approach where I do not find the best set of parameters but maintain the full probability distribution over the set of parameters. This is called Bayesian learning.

**(Refer Slide Time: 02:33)**

### Example



Suppose there are five kinds of bags of candies:

- 10% are  $h_1$ : 100% cherry candies
- 20% are  $h_2$ : 75% cherry candies + 25% lime candies
- 40% are  $h_3$ : 50% cherry candies + 50% lime candies
- 20% are  $h_4$ : 25% cherry candies + 75% lime candies
- 10% are  $h_5$ : 100% lime candies



Then we observe candies drawn from some bag: ●●●●●●●●●●

What kind of bag is it? What flavour will the next candy be?

This was the rage in 2000 for theoreticians that AI theoreticians the probability theoreticians would absolutely love Bayesian learning because actually this is the best we can do given any kind of data. And to understand this let us look at this particular example. Let us say that there are five kinds of bags think of them as hypothesis. One bag has 100% Cherry the red candies some bag 100% lime the green candies. Other bags are 75-25, 50-50 and 25-75 split. Let us say I have only 10% of the first bag 10% of fifth bag.

20% of the second and fourth bag and 40% of the third bag, now I give you a bag and you start to draw out candies and you say oh green oh another green oh another green. I give you 10 greens the question that you ask is what kind of bag is it? And what flavour will be the next candy? Ok, this is my question. Now notice that what kind of bag is it? It is a learning question. I give you five hypotheses, 5 models of the world.

And I said I give you some data from that distribution the based on the data you figure out which model of the world you are in that is the learning problem. Given the modern to figure out which flavour the candy would be that is the inference for the next.

**(Refer Slide Time: 04:17)**

### Full Bayesian learning

View learning as Bayesian updating of a probability distribution over the hypothesis space

$H$  is the hypothesis variable, values  $h_1, h_2, \dots$ , prior  $P(H)$

$j$ th observation  $d_j$  gives the outcome of random variable  $D_j$   
training data  $\mathbf{d} = d_1, \dots, d_N$

Given the data so far, each hypothesis has a posterior probability:

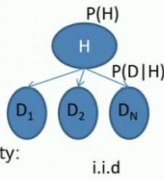
$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$


where  $P(\mathbf{d}|h_i)$  is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$P(X|\mathbf{d}) = \sum_i P(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i P(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!





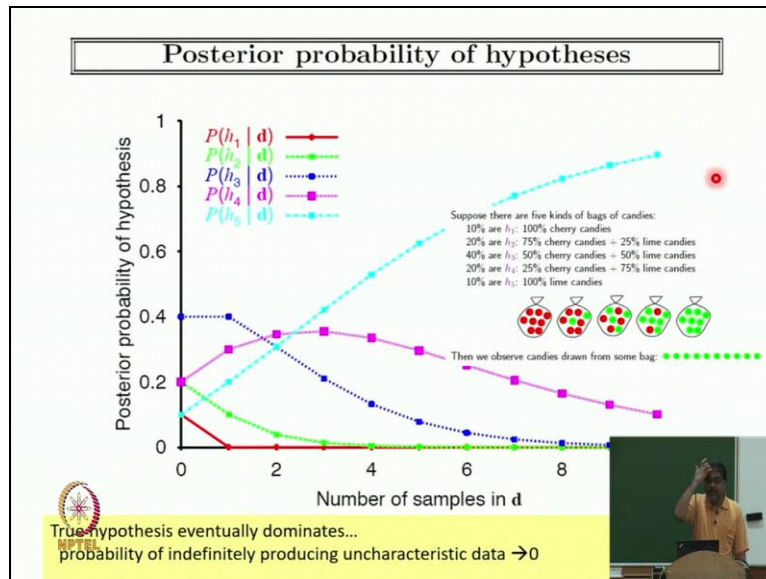
This is called full Bayesian learning view learning as a Bayesian update of the probability distribution over the hypothesis space. I do not output one set of parameters as the best output of full probability distribution of the parameter space. So, all given set a parameter is a hypothesis. Think of the hypothesis is a random variable and that random variable is generating the parameters. And if I give you the full set up a reminder that becomes 1 model 1 Bayesian network on which you can do infinite.

So you can actually model a mate-Bayesian network for a Bayesian network learning problem where I have a random variable which is the set of hypothesis. And this d1 is the probability of generating the data given the specific hypothesis. So, H is hypothesis variable its value is h1, h2, hn are sort of determining which Bayesian network was a-priori more probable. In this case the probability value should be 10%, 20%, 40%, 20% and 10%.

The Jth observation  $D_j$  will be the outcome of the random variable  $D_j$  in the training data  $D$ . So,  $D_j$  for everything. And given the data so far each is hypothesis will have a posterior probability that would be probability of hypothesis given the data you can write down as probability of data given hypothesis in a-posteriori estimation. Now if I take a maximum amount of all hypotheses that is MAP. But if I do not take a maximum? And when I asked the query, like probability of  $X$  given  $D$  or probability that the next Candy would be green given that I have seen 10 candies.

I do a summation over this hidden variable  $H$ , which is the summation of all possible hypotheses if I do this that is called Bayesian learning full Bayesian learning that is we do not pick one best get we maintain the full hypothesis distribution in my space.

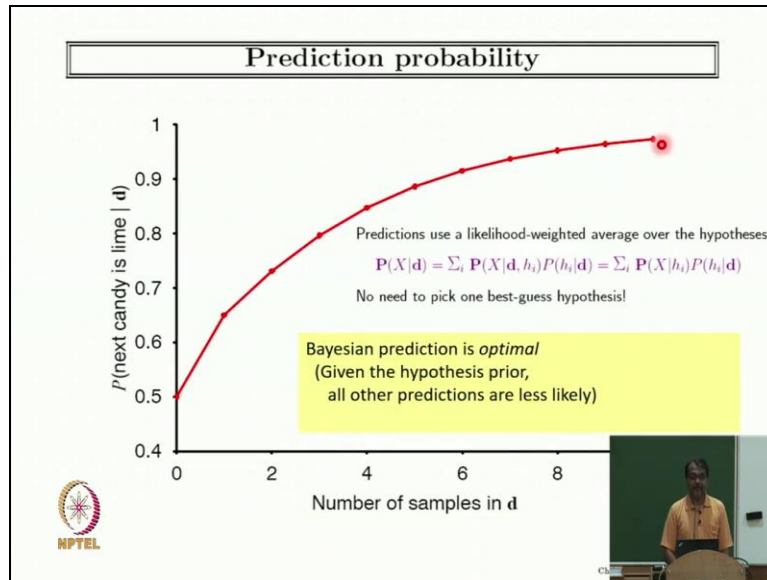
(Refer Slide Time: 06:29)



Compute this will not be very complicated for you can try it out. So initially if I have seen nothing the estimates 0.4, 0.2, 0.2 and 0.1, 0.1 as soon as I see 1 green candy which hypothesis is just not possible anymore?  $H_1$  is not anymore and notice that its probability goes down to 0. The probability  $h_2$  start to go down probability of  $h_3$ ,  $h_4$ ,  $h_5$  also start to vary but over time as have only seen the green candies you notice that all other Probability go down to 0 and only the blue one which is 5 that remains and becomes highest.

Now given this probability distribution of the hypothesis, if I ask what is the probability of the next candy I can do sum over this probabilities where all the 3 would be very close to 0. This would be 0.1 and this is very close to 0.9 and that probability will in turn out to be very close to 1.

(Refer Slide Time: 07:49)



You can prove that if I give you some data then Bayesian learning is optimal. This is the best you can do you cannot do anything better than this. However is it going to be practical? If I give you a structure and I wanted to estimate probabilities. Then you will have to maintain probability distribution over the probability space. If I have  $N$  probabilities to estimate then you will be representing a probability distribution over  $N$  continuous variables. That will be very difficult to estimate.

So in practice full Bayesian learning in a Bayesian network parameter estimation setting is not tractable and nobody does that but theoretician love to do it because that is the optimal think and love to find specific cases where you can make this tractable. This is called Bayesian learning. So what have you learnt so far in the class? You talked about maximum likelihood estimation. We talked about maximum a-posteriori estimation that allows us to give some inside about which parameter is more likely less.

But they still compute 1 set of parameters, is my model. Then I maintain the full model space and that is called Bayesian learning. Bayesian learning is not to be considered as Bayesian inference and Bayesian learning means it is very specific.