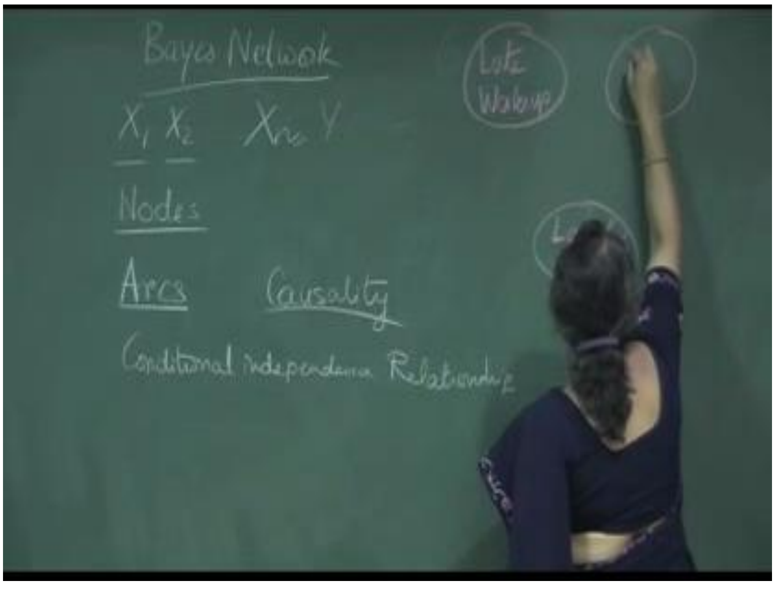
Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module – 4
Lecture - 18
Bayesian Network

Good morning. Today, we will start with the 4th module of our lecture on Probability and Bayesian learning. We have looked at the basics of probability and we have talked about Bayesian approach to learning. We have talked about bayes map classification, bayes optimal classifier and we have also looked at naive bayes. Now, as we have seen the bayes optimal classifier is not practical to apply it, because one has to apply all the classifiers towards particular problem.

Also, if we look at the entire join distribution of the probabilities involving all the variables the problem becomes intractable because the number of join distributions is too large to learn or to represent. We have seen that naïve bayes makes very restrictive assumptions, but it is very simple, very fast, easy to use learning algorithm and we have also considered that it works well under certain situations, but what we are going to study today is somewhere in the middle which is called the bayes network approach.

(Refer Slide Time: 01:39)

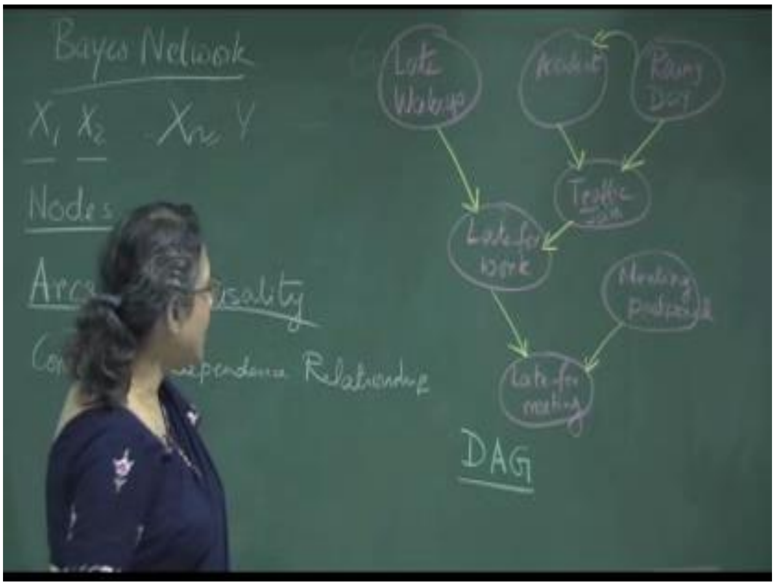


Bayes or belief network is a type of graphical model. In fact, it is a type of directed graphical model. There also other types of graphical model which represents certain independence relations or conditional independence relationships between the different variables in the domain. This helps us into make tractable inference in many cases where all the variables are not completely connected or completely dependent with each other.

This is a vast topic and we will just introduce what is Bayesian network and we will talk little bit about that, but for more details you have to study on your own or take a more advanced course. Now, what the bayes network represents is that we have variables suppose X 1, X 2, X n are the variables of interest in the domain; Y also can be included in as one of the variable. We have the different variables in the domain and we want to look at the types of dependences or other independences between them and, we can think these variables are represented by nodes. So, we use a graphical notation to represent the relations especially the independence relations between the variables. These variables are the nodes and we have arcs.

So, these arcs represent relations between the variables. The lack of arcs denotes the lack of relation between the variables in certain ways and we also have nodes, we have arcs and they represent conditional dependence, conditional independence relationships though it is not essential we can represent in Bayesian networks causality. So, causality can be represented. So, among the different variables in the domain some of these variables may be causes of the other, some may be effects of the other. If we represent this causality we get a compact Bayesian network structure. We can still have Bayesian network without representing causality, but the representation of causality makes this structure of the Bayesian network more efficient. So, what we will do is that we look at an example of Bayesian network.

(Refer Slide Time: 05:11)



So, let us say that this is a node corresponding to the variable ‘late wakeup’, this is a node corresponding to ‘late for work’, this is node corresponding to ‘accident on the highway’, this is a node corresponding to ‘rainy day’, this is a node corresponding to ‘traffic jam’, this is a node corresponding to ‘meeting postponed’ and this is a node corresponding to ‘late for meeting’. Suppose, these are the variables which are of our interest and we want to see, what is the relation between the variables?

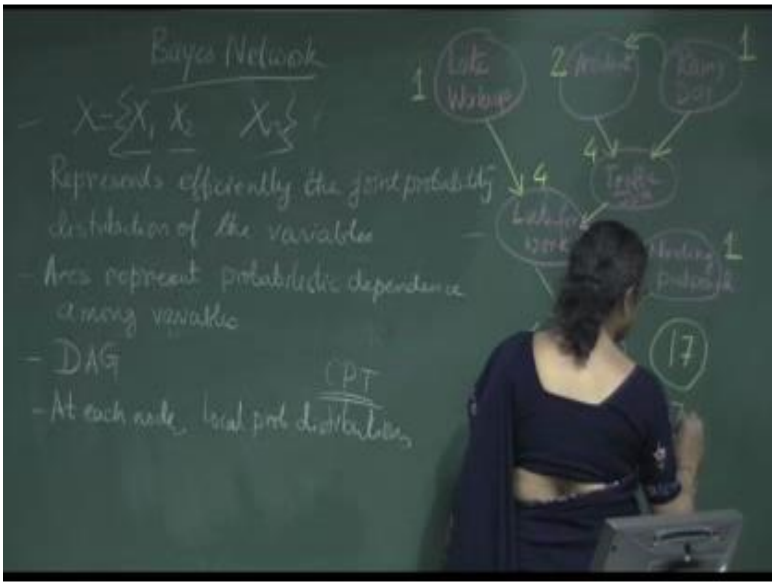
Let us say, waking up late influences whether you are late for work. If there is accident it is more likely there is a traffic jam. If it is a rainy day it is likely to be a traffic jam and maybe we can say if it is a rainy day there is higher likelihood for an accident. If you are caught in a traffic jam it affects whether you are late for work. If you are late for work affects whether you are late for a meeting and if the meeting is postponed that also influences whether you are late for the meeting. Now, these arcs that I have drawn denote causal relationships between these variables and from this structure we get a graph or more specifically, this is a directed acyclic graph, which represents the nodes and the relationships between them and from this network we can read different conditional independence relationships.

For example, we can say that late for work is influenced by whether you have woken up late, but if you are late work that directly influences whether you are late for meeting and therefore, we can say that if you wake up late that effect whether you are late for the meeting. However, if you know that whether the person is late for work or not; suppose you are not late for work in spite of waking up late then whether you are woken up late or not does not influence whether you late for the meeting, if you know whether you are late for work or not. So, waking up late and being late for meetings are normally not independent, but they are conditionally independent given that you know whether you are late for work.

Again, let us look at accident, whether an accident happened and whether you woke up late right. So, if you think of whether accident happened, whether you woke up late. These 2 variables are independent, if you have woken up late that is independent of whether an accident happened, but if you know that you are late for work, then these 2 variables do not remain independent if the value of the node, for example, you know that you late for work then, suppose you did not wake up late then accident could be reason why you are late for work. So, these 2 variables waking up late and accident are not independent if you are given the value of this.

So, similarly late for work and meeting postponed are independent, but they are not independent if these values are known. So, this particular graphical representation encodes certain conditional independence relationships. There are three specific conditions of these operations I will not cover them in detail we will just look at some examples in this class. Let us look at formally what a Bayesian method represents. So, a Bayesian network is a graphical representation which represents efficiently.

(Refer Slide Time: 11:01)



Bayes Network

- $X = \{X_1, X_2 \; X_n\}$

Represents efficiently the joint probability distribution of the variables

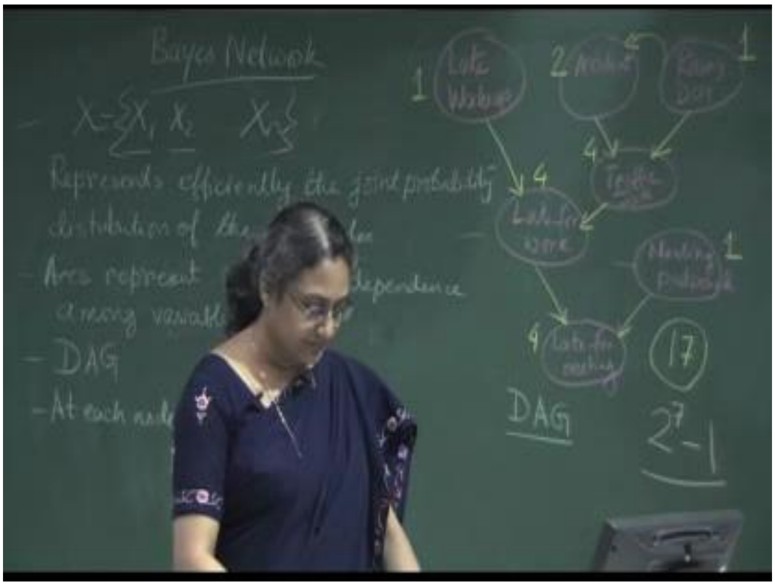
- Arcs represent probabilistic dependence among variables
- DAG
- At each node, local prob distribution CPT

So, this is an efficient representation of the joint probability distribution of the variables. Any probability or conditional probability of interest can be computed if you know the full joint probability distribution, but as we have discussed representing the entire joint probability distribution is intractable. If we have bayes network representation we can more compactly and efficiently represent the joint probability distribution, especially if this graph does not have too many edges. If this graph has less number of edges that means, these variables many of them are conditionally independent of others given some evidences and we can represent the joint distribution more efficient.

So, a Bayesian network is a set of nodes. Let us say it is a set of nodes X equal to X 1, X 2, X n these are the nodes or variable. Then arcs represent probabilistic dependence or independence among variables. In fact, absence of arc denotes independence or conditional independence the network structure is a directed acyclic graph or DAG and at every node we keep the local probability table, at each node we keep local probability distributions which is also called the conditional probability table. So, CPT or conditional probability table is associated with each node.

Now, let us try to look at you know what are the values what are the probability distributions that we need to fully specify this bayes network. For simplicity, let us assume all this nodes corresponds to variables which are Boolean. So, we have 1, 2, 3, 4, 5, 6, 7, we have 7 variables. So, they are all Boolean. There are 2 to the power 7 possible combinations and for each one of them we can find the probability of it. Now, if we follow this model at the conditional probability table we will see; we have to keep probability distribution of a node given the value of its parent. Now, if you look at this node this node does not have any parent. So, what we need to keep here is you know just we need 1 value what is a probability of late wake up. So, we require 1 value.

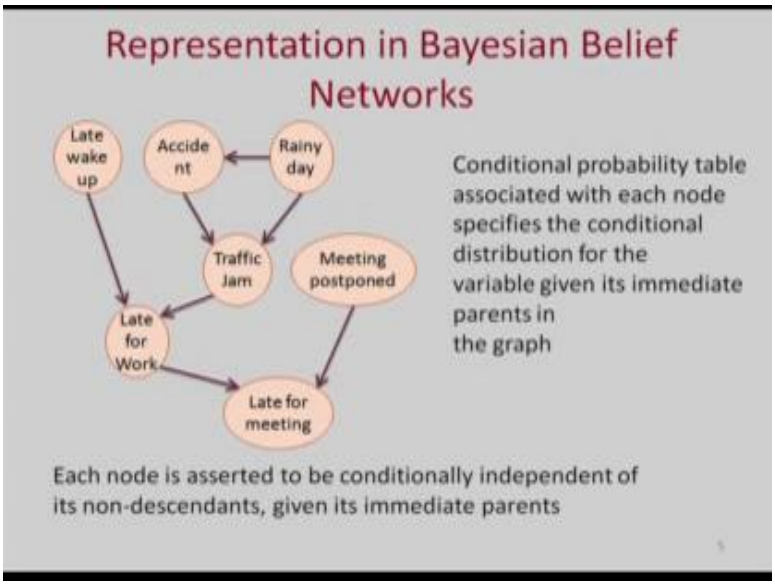
(Refer Slide Time: 14:57)



Rainy day also does not have any parent we need 1 probability value to be associated here accident has 1 parent rainy day. So, what is the probability of accident being true given that it is a rainy day and accident being true given that it is a not a rainy day. So, we have 1 parent, we need to keep 2 values. Traffic jam has 2 parent accident and rainy day. So, accident can be true, rainy day can be true; both can be false; 1 can be true, 1 can be false and so on. So, there are 4 possible combinations of the parent. For each of them we can find out what is the probability of traffic jam happening, the probability of traffic jam not happening can be computed by 1 minus that value.

So, for this node we require 4 values. Similarly, late for work also require 4 values, meeting postponed 1 value, 4 values for late for meeting. So, the total number of probability values that we have keep is 1, 2, 3, 4, 8, 12, 13, 17. So, we required to store 17 probability values in contrast to the fully connected Bayesian network where we have to keep 2 to the power 7, actually 2 to the power 7 minus 1 value. So, a Bayesian network is a compact representation of the joint probability distribution.

(Refer Slide Time: 16:40)



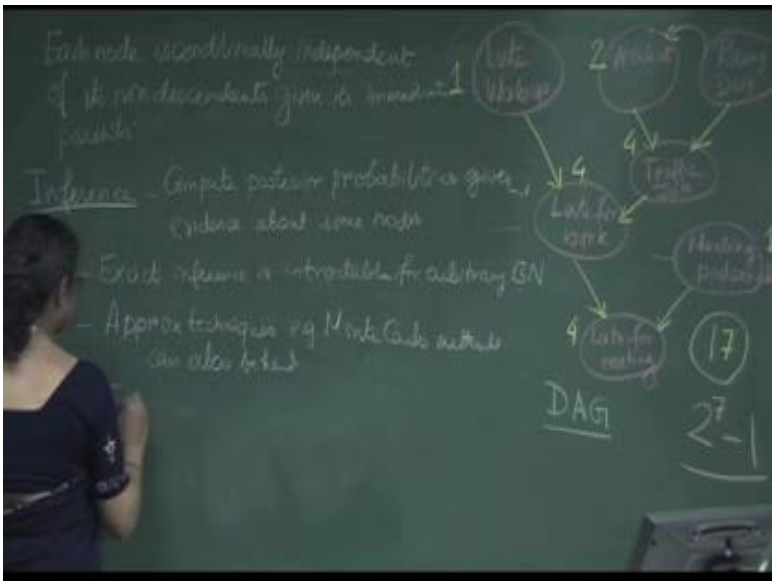
Representation in Bayesian Belief Networks

Conditional probability table associated with each node specifies the conditional distribution for the variable given its immediate parents in the graph

Each node is asserted to be conditionally independent of its non-descendants, given its immediate parents

Now, what we like to do is look at an example of; this is the example of bayes network that we have looked at and corresponding to this example, we have conditional probability table associated with each node, the number of values of the conditional probability table given that all the variables are Boolean I have already written here.

(Refer Slide Time: 17:26)



Each node is conditionally independent of its non descendants given its immediate parents

Inference – Compute posterior probabilities given evidence about some nodes

– Exact inference is intractable for arbitrary BN

– Approx techniques eg Monte Carlo methods can also be tested

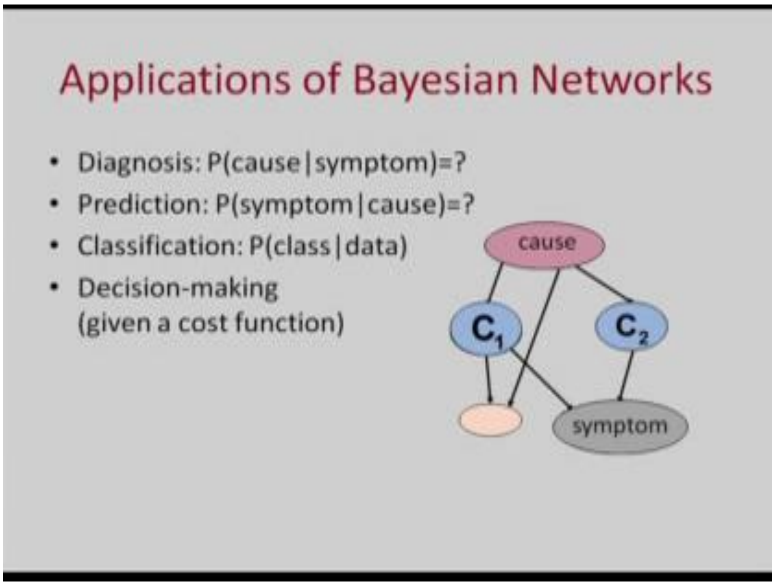
So, each node is asserted to be conditionally independent of its non-descendants given its immediate parents. So, let us write this down. The bayes network says that each node is conditionally independent of its non-descendants given the value of its immediate parents. So, late for meeting is independent of late wake up given whether you are late for work. Meeting postponed is independent of late for work given the value of late wake up and traffic jam. So, these are examples of conditionally independence that we can read off based on this statement that I have made. Now, what we can we do when we have a Bayesian network?

So, once we represent the local probability tables we can use the bayes net for making inference. So, inference means you want to compute certain probabilities of certain probability distribution of certain variables given certain evidences, you know the values of some variables and you want to find out probability distribution of some other variables which are of your interest. So, inference you compute posterior probabilities given some evidence. Evidence means some of the nodes you know the values of the nodes and in order to do inference efficiently you can exploit probabilistic conditional independence that is encoded in the belief network.

Unfortunately in general, exact inference is intractable for arbitrary base network. If have an arbitrary base network the exact inference is intractable. For certain special types of base network which have some special structure exact inferences tractable, but even when exact inferences intractable there are certain methods by which you can do inference in a tractable type. So, there are approximate techniques for inference, e g Monte Carlo methods, we will not talk about these methods or in fact, any inference method specifically in this class.

These are also can be heard for certain type Bayesian network, but in practice many of these methods are useful, but you can have efficient algorithm the leverage the structure of graph. In the last class, we talked about naive bayes and we said that naive bayes represents certain independence conditional independence assumptions. We can draw a naïve bayes the probabilities in the variables involved in naive bayes in this graphical structure and that particular structure inference efficient. There are other types of structures like poly trees structures which are more general structure for which inferences efficient, but it is not true for other structures, but there are different algorithms for dealing with poly trees and certain variations of poly trees, but those details we will not talk about in this class.

(Refer Slide Time: 22:23)



Applications of Bayesian Networks

- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: P(class|data)
- Decision-making (given a cost function)

Let us look at, what are the situations, what are some standard situations where we can apply Bayesian network? Bayesian networks can be applied in many scenarios, for example, it can be applied for diagnosis. Suppose, you know the symptoms of a disease you want to identify what is the possible cause.

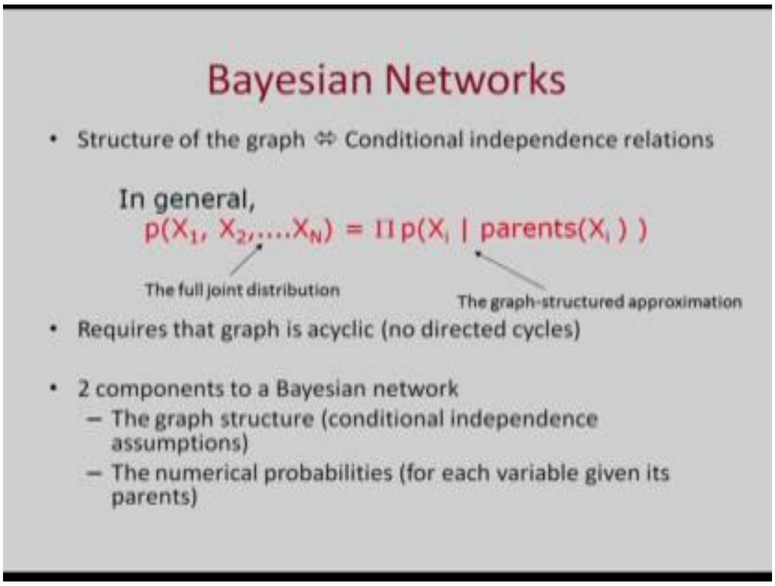
(Refer Slide Time: 23:07)

For example, suppose a cancer causes tumor and then there are other causes for a tumor. A cancer may also cause weight loss and so on. So, if you are given the symptoms you can try to infer what is the possible cause, we saw an example of this type of inference where we looked at using bayes theorem and. So, from the symptoms you can find out what is probability of the specific causes. So, if you are late for work what is the probability that there is traffic jam? So, given the symptom you want to find out what is the possible cause. So, this is for diagnosis you want to diagnose what happened given that you can see the symptom.

Suppose, some machine is malfunctioning you observe certain things you want to understand, what is the possible cause this is for diagnosis a bayes network can also be used for prediction. If you have cancer, what is the probability of weight loss? What is the probability that you have a tumor? So, it can be used for prediction given the cause, find the probability of the symptoms?

Bayes network can be used for prediction, find the probability of class given the training data? This is how it is used for basic supervised machine learning you find a probability of class given the data. It can also be used for decision making if you are given different cause functions you want take different decisions, so that your utility is maximized and for such applications also bayes network can be used.

(Refer Slide Time: 25:31)



Bayesian Networks

- Structure of the graph ⇔ Conditional independence relations

In general,
$p(X_1, X_2,\ldots X_N) = \Pi\, p(X_i \mid \text{parents}(X_i))$

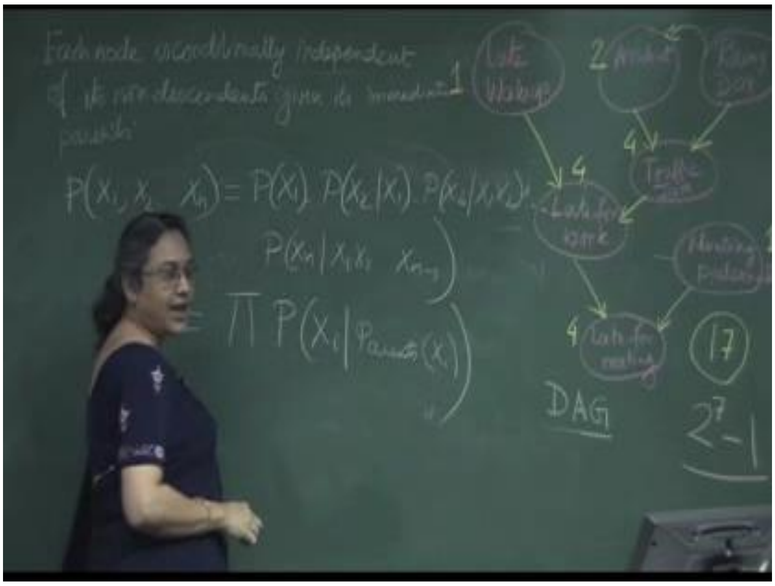
The full joint distribution

The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)
- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

Let us look at how to define a bayes network formally. As we have already said the structure of the graph denotes the conditional independence relationships. In general, the joint distribution of X 1, X 2, X n can be written as the product of probability of each node given its parent.

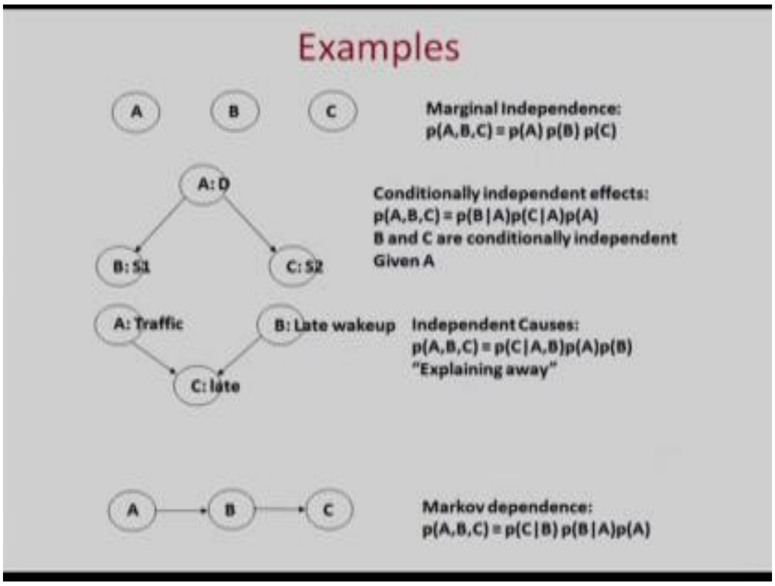
(Refer Slide Time: 26:00)


Each node is conditionally independent of its non-descendants given its immediate parents
$P(X_1, X_2 \ldots X_n) = P(X_1) \cdot P(X_2|X_1) \cdot P(X_3|X_1X_2) \ldots P(X_n|X_1X_2 \ldots X_{n-1})$
$= \prod P(X_i | Parents(X_i))$
1
Late Wakeup
2
Accident
Rainy Day
4
Traffic
4
Late for work
Meeting
4
Late for meeting
17
DAG
2^7-1

In general, in most cases if you have to find out probability X 1, X 2, X n we can apply the chain rule by which we get; this is equal to probability of X 1 times probability X 2 given X 1 times probability of X 3 given X 1 X 2 dot, dot, dot probability of X n given X 1, X 2, X n minus 1. This is the normal chain rule applied to a joint probability distribution, but in a bayes distribution you can say that this is equal to the product of probability of each node given just the values of its parents.

So, in this way it is a compact representation of the joint distribution. So, the graph is required to be acyclic. So, there are 2 components to a Bayesian network, the graph structure and the numerical probabilities or the conditional probability table associated with each node.

(Refer Slide Time: 27:25)



Examples

Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)

Conditionally independent effects:
p(A,B,C) = p(B|A)p(C|A)p(A)
B and C are conditionally independent
Given A

Independent Causes:
p(A,B,C) = p(C|A,B)p(A)p(B)
"Explaining away"

Markov dependence:
p(A,B,C) = p(C|B) p(B|A)p(A)

Now, we will look at some examples of Bayesian network. So, here is a situation the first example, we have 3 variables A, B and C then there are no edges. So, they are completely independent. Therefore, probability the joint distribution of probability of A, B, C is simply probability of A times probability of B times probability of C. This is the simplest case, where there is no relation among the variables that completely independent.

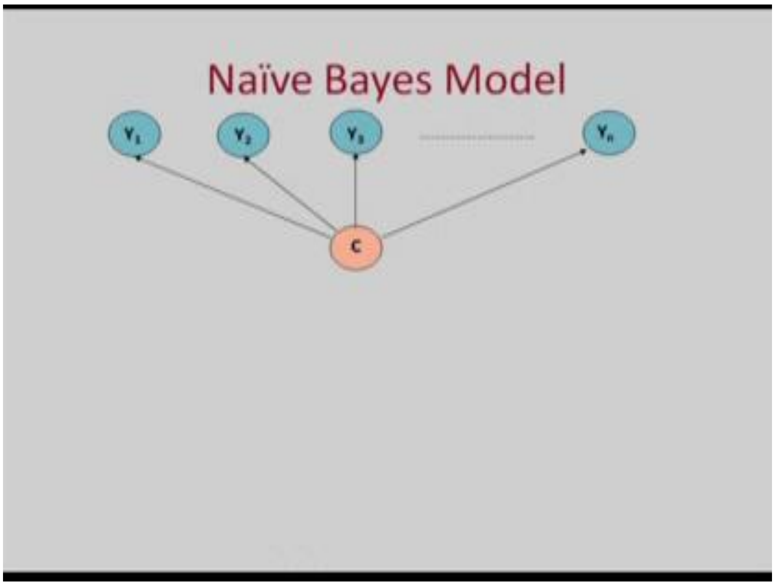
Look at the second example, we have suppose A is a disease, B is one symptom of the disease, C is another symptom of the disease. The conditional independence of the bayes network means the joint distribution probability of A, B, C is given by probability B given A times probability C given A times probability of A. So, B and C are conditionally independent given.

Third example we have A, and B are 2 causes for C, suppose C is late for work, A is traffic jam, B is late wake up. So, late wake up and traffic jam are independent causes of being late. So, traffic jam and late wake up are independent if you do not know whether it is late, but they are not independent if you know whether it is late because one can if you are late, one of them can explain the reason for other. So, if you are late and there is traffic jam the probability that you have woken up late is less, but if you are late and there was no traffic jam it is more highly probable that you have woken up late.

The fourth example, we have these variables A, B, C and we have no relation from A to

B, B to C and this represents Markov dependency. So, C is independent of A given B. So, if A, B and C occur at consecutive time step, A at time t minus 1, B at time t, C at time t we can say that probability of C depends only on the current state B it is independent of the previous stage. So, this is the Markov assumption which is often used. So, these are 4 examples of Bayesian network.

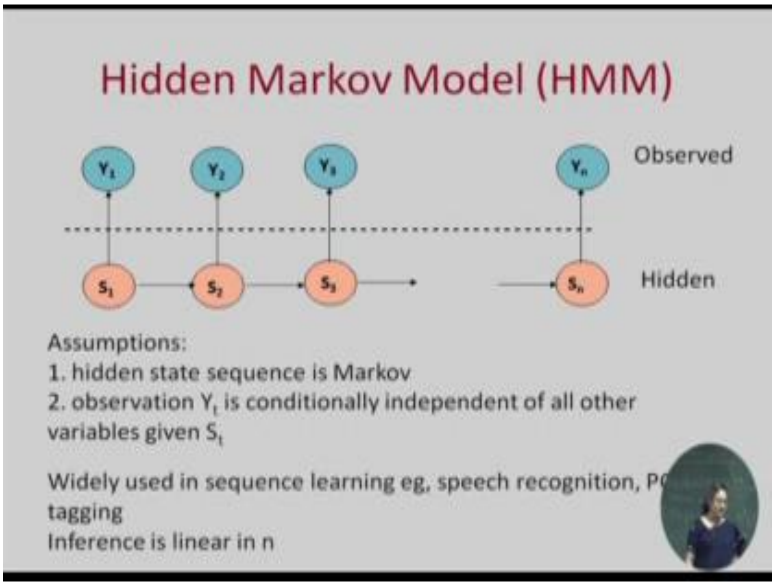
(Refer Slide Time: 29:57)



Naïve Bayes Model

And we have already talked about the naive bayes model in the earlier class where we have certain attributes these attributes are denoted here as Y 1, Y 2, Y 3, Y n and C is the class. So, in naive bayes we assume that Y 1, Y 2, Y 3 all of them pair-wise, they are independent of each other given the class and the class determines the probability of Y 1 if class is 1, the probability of you say that probability of Y 1 given class 1 probability of Y 1 given class equal to 0. So, these are the particular relationships that you have to study in a naïve bayes model.

(Refer Slide Time: 30:39)



Hidden Markov Model (HMM)

Assumptions:

1. hidden state sequence is Markov
2. observation Y_t is conditionally independent of all other variables given S_t

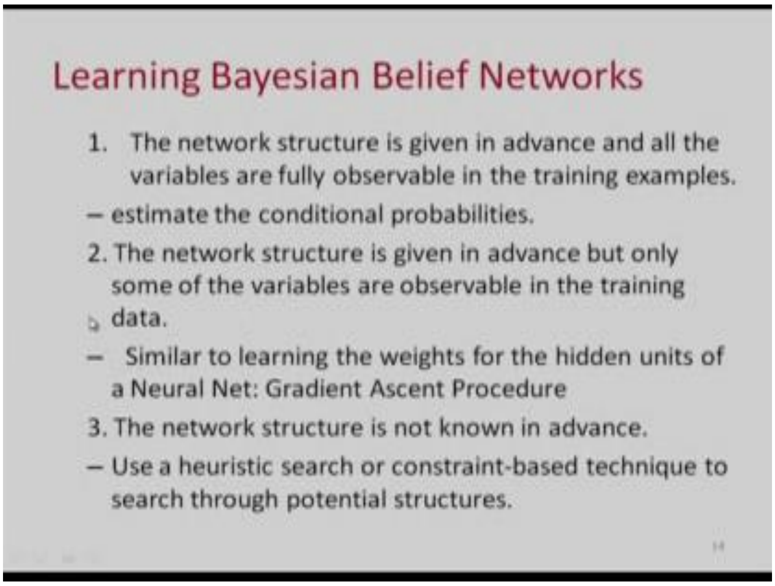
Widely used in sequence learning eg, speech recognition, P tagging

Inference is linear in n

And lastly this is a model for Hidden Markov model. So, Hidden Markov model is another graphical model, we will not unfortunately study it in detail in this class, but in a hidden markov model, there is the underline state of the system which follows a Markov model. So, S 1, S 2, S 3, S n are the states of the system at different time instances and the Markov assumptions means that S 3 is independent of S 1 given S 2, S n is independent of the previous states given S n minus 1 and these states are hidden, but you have an observation and the observation what you can observe depends on the current state.

So, hidden Markov model is depicted by this type of graphical model and efficient algorithms for inference and learning of hidden Markov model. They have wide spread application in speech recognition, part of speech tagging, genes sequence modeling and other such areas.

(Refer Slide Time: 31:47)



Learning Bayesian Belief Networks

1. The network structure is given in advance and all the variables are fully observable in the training examples.

– estimate the conditional probabilities.

2. The network structure is given in advance but only some of the variables are observable in the training data.

– Similar to learning the weights for the hidden units of a Neural Net: Gradient Ascent Procedure

3. The network structure is not known in advance.

– Use a heuristic search or constraint-based technique to search through potential structures.

Now, how is learning helped by belief networks? So, we will just mention that there 3 cases, 3 ways in which you can use belief networks for learning. In case 1, the network structure is known to you network structure given in advance and from your training examples you can learn the conditional probability table. So, you can estimate the conditional probability tables, the network structure the variables the edges are given to you and in the training example values of all the variables are known to you and you have to estimate the conditional probability, and after that you can use the bayes network for making inferences.

In the second case, the network structure is given, but only the values of some of the variables are to you in the training data. So, the entire network structure is given, but you know the values of some of the variables. Here, you have to also estimate the probabilities corresponding to the hidden variables for that some framework like gradient ascent, etcetera can be used. Again, we will not talk about the details in this class.

The third case is whether network structure is not known. Here, you have to learn the network structure using some heuristic search or constraint-based technique. This is slightly advanced and we will not talk about this. With this we come to the end of this lecture and this module.

Thank you very much.