Machine Learning for Earth System Sciences

Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

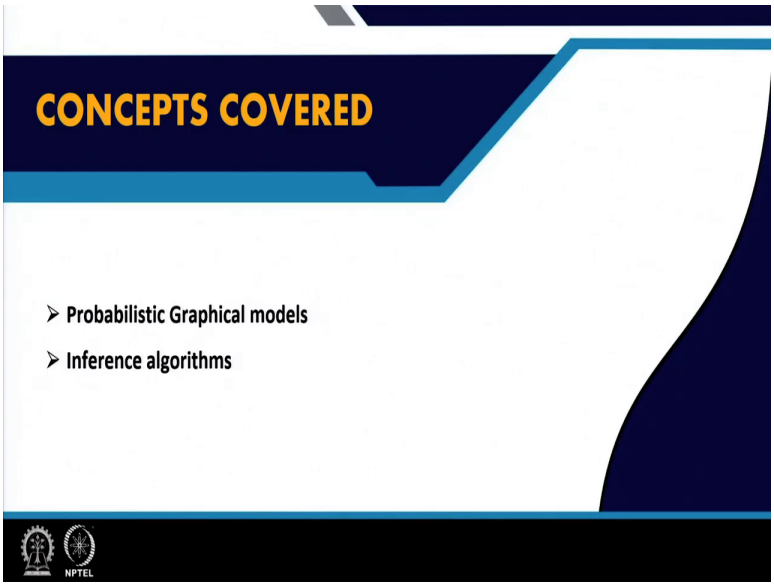
Module - 02
Machine Learning Review

Lecture - 15

Probabilistic Models for Earth System Science

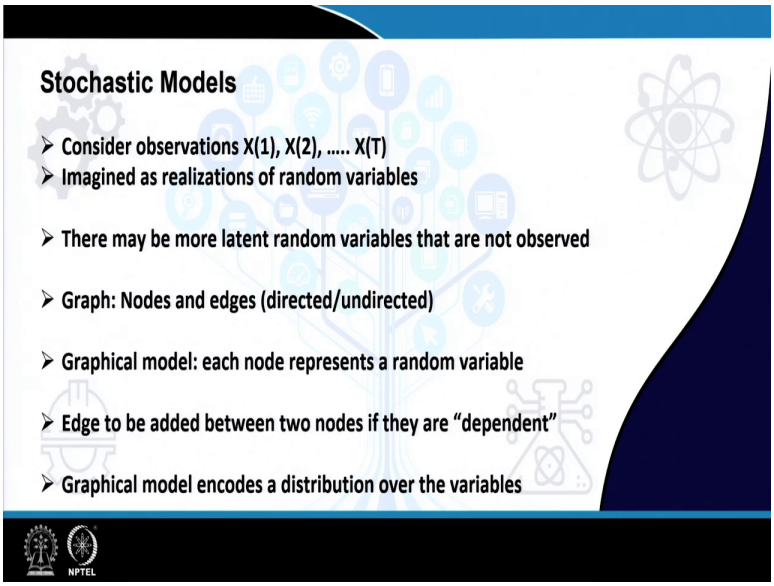
Hello everyone, welcome to lecture 15 of this course on Machine Learning for Earth System Science; today is the last lecture of the current module, which is module 2 about machine the Machine Learning Review. In today's lecture, we will discuss Probabilistic Models for Earth System Science.

(Refer Slide Time: 00:44)



The main topics which we are going to cover today are probabilistic graphical models, different types of them and inference algorithms.

(Refer Slide Time: 00:53)



So, first let us come to the topic of stochastic models. So, let us say that we have observations $X(1)$, $X(2)$, .., $X(T)$ as usual and we imagine these as the realizations of random variables. So, these observations needless to say they can be spatio-temporal in nature. Now, there may also be more latent random variables in the model, which are not observed.

So, like we like as we have already discussed in some of our earlier lectures that we developed some kind of a spatio-temporal stochastic model, which like it is where these observations are considered to be some realizations of some random variables, which are observed and then there are also some latent variables.

Now, so basically that means we have a large collection of random variables. So, now, there is this concept of a graph, the graph data structure or mathematical structure which I am sure all of you are familiar with. So, the graph is a collection of nodes and edges and these edges they can be directed or undirected.

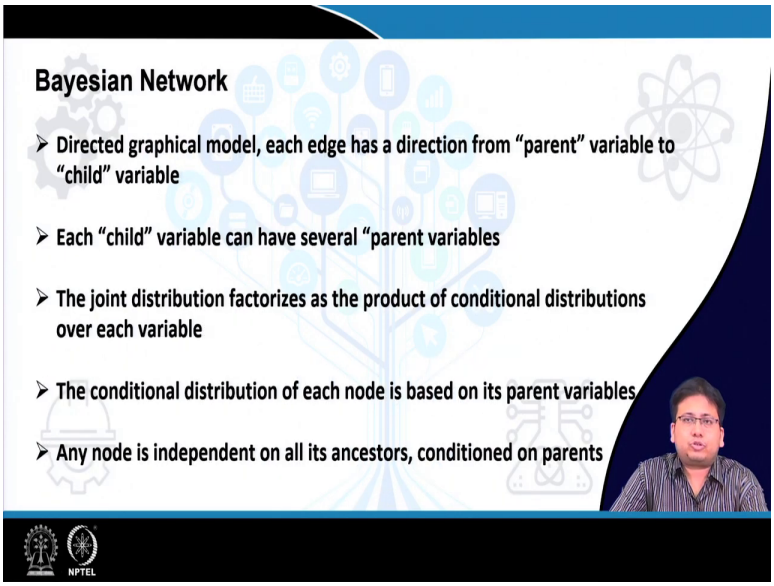
Now, there is this concept called graphical model, it is supposed to like a an may be an unification of graph theory and probability theory. So, in a graphical model, a graphical model is basically a graph, but here the each node represents a random variable. Now, an edge can be

added between two nodes, if they are if those the corresponding random variables they are somehow dependent on each other.

And so, graphical model what it effectively indicates or it represents some kind of a probability distribution over the variables. Now, if you have a set of variables, their joint distribution like we like it might be possible to factorize their joint distributions in a large number of ways; for example, if we know that all the random variables are independent of each other, then their corresponding joint distribution will be nothing but the product of their marginal distributions.

But it may a if that is the case, then we would simply have a collection of nodes with no edges among them. But in general that will not be the case, there are some of them will be dependent on each other; so there will, which means that there will be edges. So, the joint distribution then cannot be factorized in that particular way. So, it will be, but we may still be able to factorize it in some other ways. Now, the this graphical model it basically tells us how or in what ways the joint distribution might be factorized.

(Refer Slide Time: 03:35)

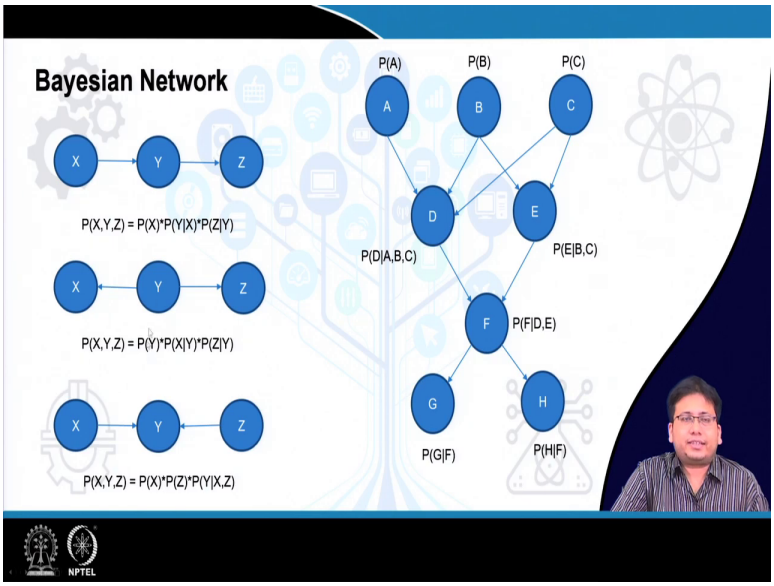


Now, there are different categories of graphical models. So, like broadly they can be divided into two categories, the directed graphical models and the undirected graphical models; the directed graphical models these are sometimes also known as Bayesian networks.

So, in this as the name suggests here each edge has a direction like. So, in a normal graph just like we talk about say source to destination or source to sink when we are talking about the direction of an arrow; in this case we say that the edge points from the parent variable to a child variable and we can say that each child variable can have several parent variables.

Now, the joint distribution factorizes as a product of conditional distributions over each variable.

(Refer Slide Time: 04:31)



So, let us see an example of this. So, like let us consider this particular graph. So, there are three nodes X, Y, Z. And as you can see the direction of these edges; so we can say that $X$is the parent of Y, Y is the parent of Z. So, the joint distribution of these three variables X, Y, Z; so

$$P(X, Y, Z) = P(X) * P(Y|X) * P(Z|Y).$$

So, in these kind of a Bayesian network, the conditional distribution of each node is based on its parent variables and the joint distribution factorizes as the product of conditional distributions over each variable. So, like for the variable Y, it has only one parent so X; so we write $P(Y|X)$.

Now, if you come to Z, it that also has only one parent namely Y. So, its conditional distribution of Z is $P(Z|Y)$. So, the joint distribution we simply write as the product of the conditional distributions of every node conditioned on their respective parents. So, x has no parents. So, I

write $P(X)$, then Y it has only X as a parent. So, I write $P(Y|X)$ and Z has only Y as the parent. So, I write $P(Z|Y)$ and I multiply all of them together that is the joint distribution.

Similarly, in this case you may note that Y has no parents; X and Z both of them have Y as the parents. So, this one in this case the joint distribution factorizes in this particular way; first we write $P(Y)$ which has no parents, then for $X,\ Y$ is the only parent. So, I write $P(X|Y)$ and for Z also Y is the only parent. So, I write $P(Z|Y)$. Now, if you come to this case. So, here you see that X and Z both of them have no parents, but Y has two parents X and Z. So, this one factorizes as $P(X)\ *\ P(Z)\ *\ P(Y|X,Z)$

So, like as I said earlier the conditional distribution of each node is based on its parent variables. So, it has two parent variables X and Z. So, I write $P(Y|X,Z)$. Now, these are of course, simpler structures. So, we can have a more complicated model also.

So, let us say this one has 8 variables and this is the particular edge structure. So, in this case also the joint distribution of these 8 random variables I can factorize it as follows; I can write first of all $A,\ B,\ C$, I see that they have no parents. So, I write $P(A)\ *\ P(B)\ *\ P(C)$. So, note that this is like there is no cycle, there is no directed cycle in this graph; had it been the case, then things would have been much more complicated.

So, let us just assume that this is a DAG that is a directed acyclic graph. Now, if you come to D, it has three parents $A,\ B,\ C$. So, its conditional distribution is $P(D|A,B,C)$; but E on the other hand it has only two parents $B,\ C$. So, I write $P(E|B,C)$. Similarly, for F that has two ancestors $D,\ E$, so we write it in this way if you consider $G,\ H$, both of them have the same ancestor F. So, I write in this way.

So, the joint distribution of all the eight variables; so we can write it as $P(A)\ *\ P(B)\ *\ P(C)\ *\ P(D|A,B,C)\ *\ P(E|B,C)\ *\ P(F|D,\ E)$ and so on, ok. So, apart from indicating the joint distribution, the Bayesian network also indicates some or informs us about which variables are conditionally independent or of which other variables. So, in general we can say that any node is independent of all it is ancestor's conditions on its parents.

So, like if you consider this case, if you for example, if you consider the node F; now A, B, C. these are all ancestors of node F, that is we have a directed path from each of them to the node F. However, D and E are their immediate parents. So, like we so, because A, B, C they are the ancestors of F; so in general we like we cannot say that F is independent of A, B, C rather they are depend, like there is dependence relation. But if D and E are both known, then we can say that give that $P(E)$ that sorry we can say that F is independent of $(A|D, E)$.

Similarly, we can say that F is independent of $(B\ OR\ C|D, E)$. So, it is like these ancestor, I mean these parents they like they insulate every node away from their like ancestors. Now, what about the relations between say F and G? So, like we can say of course, that G they are dependent on each other; that is G is indeed dependent on F, because they are like there is a parent child relations on them.

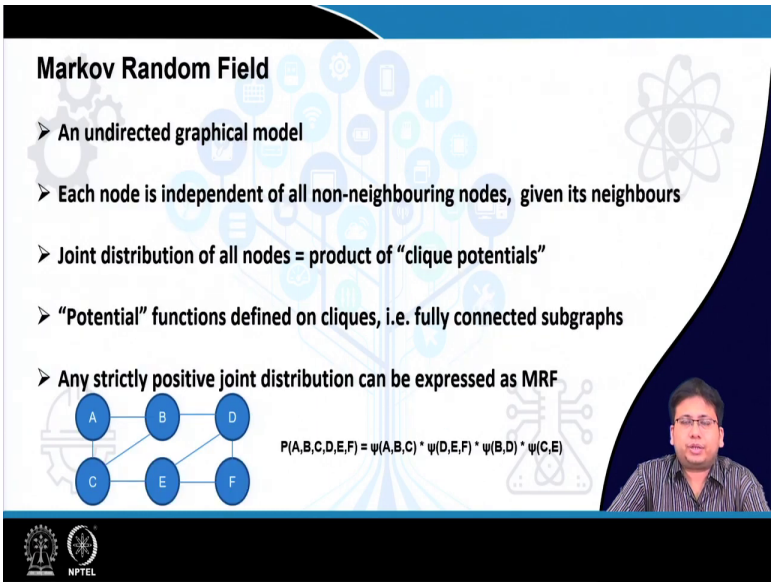
Similarly, F is also dependent on G. And in general if we in this kind of a graphical model, when we are talking about the independence between two nodes; we can actually find out whether they are going to be independent or not by using the concept of d-separation which like, which is beyond the course which is beyond the scope of this particular lecture, but I encourage you to study more about it.

So, in general when we are talking when we are given two random variables and we are trying to find out whether they are conditionally independent or not; so then basically it is a statistical test, where we have there are conditional independent tests like that, where the null hypothesis is may be that they are independent and then we see whether that null hypothesis can be accepted at some level of significance.

So, basically that is based on the definition of independence, which says the joint distribution is the product of their marginal distributions. So, by drawing samples, people try to establish or disprove that claim and based on that they find out the dependence relations.

But in case of the graphical models, once you build this kind of a Bayesian model; then Bayesian network, then using a set of concepts or rules known as the d-separation rules you are we may actually be able without doing any test, you will be able to find out, which variable is independent of which other variables under what conditions.

(Refer Slide Time: 12:00)



So, that is the story of Bayesian networks or directed graphical models. Now, coming to an undirected graphical model, there are several undirected types of undirected graphical models. So, the most basic of them is called the Markov random field; in such a case each node is independent of all non neighboring nodes given its neighbors. Now, the joint distribution of all the nodes is the product of what is known as clique potentials.

So, in the unlike the Bayesian network, where the joint distribution was the product of the individual conditional distributions; in this case what in case of a undirected graphical model what we do is, we identify clique. So, what is a clique of a graph? A clique of a graph is a like is a fully connected subgraph that is. So, if you this is a graph like with six nodes. So, as you can understand this is not a complete graph, it is or it is not fully connected; because that is we can see that the edge from A to E is missing A to F is missing, it is not a complete graph.

But if you consider these three ABC; you can see that there can be $3C_2 = 3$ edges and there are those three edges, there is AB connected, AC connected BC also connected. Similarly, DEF that is also a like that it is a complete or fully connected sub graph. But if you consider this sub graph let us say BE, that is not connected; but if you consider like even pairs like say BD or AB, they are also fully connected sub graphs, because like between B and D there are two variables, so

there can I mean two nodes, so there can only be one edge among them and that edge is there. So, BD by itself is another clique.

So, when we have this kind of a graph, we look for various cliques and for each clique we define what is known as a potential function. Now, these potential functions will have certain properties; so they especially that they will have to be strictly positive and so on. So, that is why these potential functions are sometimes written as e to the power something to make sure that they will always be positive. And so, this is the joint distribution of these variables of all the variables that can be expressed as a product of all the clique potentials.

So, in this case you we can find that there are six like cliques like this. So, ABC, DEF these are the cliques of size 3; then like we also have BD , CE, etcetera like which are cliques of size 2, which actually connect the two these two cliques, so we can include them.

So, in general we like we did not write it in this particular way; we can actually write it as, like instead of defining clique potentials we could as well define if potential functions for each edge and simply write the product as the, sorry the joint distribution as the product of all these edge potential functions. And similarly we could even have the node potential functions that is cliques of size one.

So, like we like as long as we cover like all the pairs, like all that is it is all about identifying the different cliques of different sizes like which cover all the nodes. So, as long as you do that, you can like you can there is this particular factorization can be done in different ways based on how you define your potential function; the there is a famous theorem in this matter that any strictly positive joint distribution can always be expressed in this manner as a Markov random field.

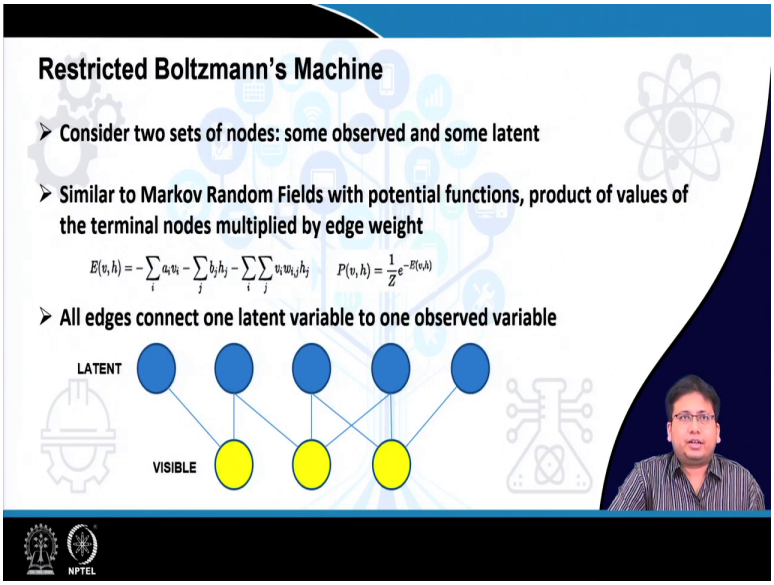
So, there is this theorem called the Hammersley Clifford theorem, which proves this. And just like we talked about some conditional independence relations in case of the Bayesian network; in case of the Markov random field also we can say something like this, we can say that each node is independent of all non neighbouring nodes given its neighbours.

So, if you consider these two nodes or these two random variables C and D. So, as you can understand they are not neighbours, because there is no edge from C and D; but its neighbours of

the neighbours of C are $A,\ B\ and\ E$. So, now, if these three nodes are observed; I mean if these like if you can condition on these three nodes $A,\ B\ and\ E$, then we can say that $C\ and\ D$ are independent or $C\ and\ F$ are independent. So, we can say. So, in general we cannot say that C is independent of D, but we can say that C is independent of D given $A,\ B\ and\ E$.

Similarly, we can also say that F is independent of A given $D\ and\ E$, right.

(Refer Slide Time: 17:20)



So, that is about the this is the basic idea of a Markov random field. Now, there are various other models undirected graphical models, which are based on the similar concept. So, there is this thing called the restricted Boltzmann machine. So, it contains two pairs of nodes some of the like one pair of nodes we the latent variables and the other pair of like other set of nodes these are the visible variables, which can be measured for which we have the observations.

And now the joint distribution of all these variables, it can be written like again in terms of these potential functions the way we defined in terms of Markov random field. So, like whenever we are considering like any edges. So, let us say that there is an edge which connects the visible node v with the hidden node edge is or let with the visible node v_i with the hidden node h_j.

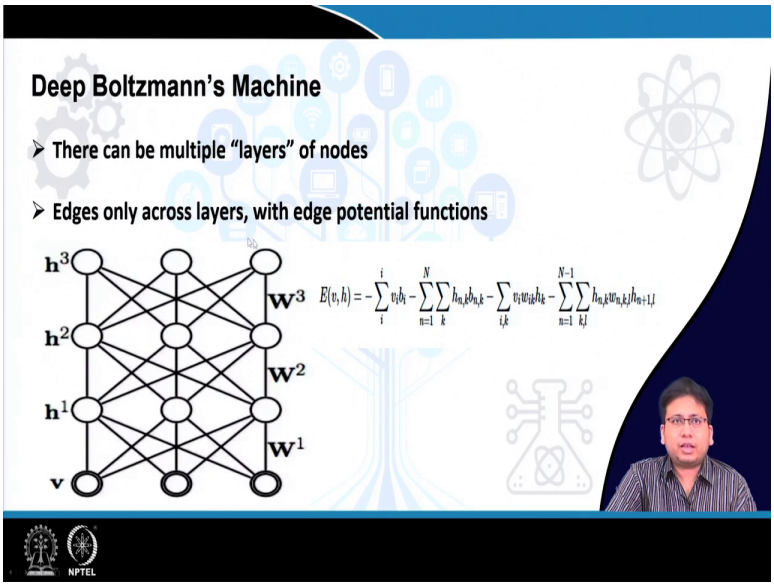
So, note that here all the edges are only like one of like are across the these two sets; I mean for any edge, one end will be latent, the other end will be visible. There is no visible to visible or latent to latent edges; that is why it is called a restricted Boltzmann machine, the edge structure is restricted in that sense. So, we when we are considering the like the edge between the i^{th} visible variable and the j^{th} hidden variable, let us say their values are v_i and h_j; of course v_i is a value which I know and h_j is a value which I do not know it has to be, like it has to be estimated.

And let us also say that the weight of that particular edge is w_{ij}. So, unlike the Markov random field and the Bayesian network, in this case we are considering weighted edges. So, we like we multiply the values of the terminal nodes or variables with the edge weight and that is like we can say the potential of that particular edge; similarly every node by itself will have their own potentials.

So, the just like in the previous case, the joint distribution is going to be the product of all these node and edge potentials; but to maintain the positivity, we usually express it in this way as a as e to the power something to make sure that it is positive and this Z is the normalizing constant to make sure that it is a valid probability distribution, that is for like for every possible values of v and h we will get some probability, but those probabilities must all add up to one.

So, to make sure that that adding up to one happens; we have this thing called Z, which is called sometime's called as the partition function.

(Refer Slide Time: 20:20)



Now, this concept can be extended to that of a deep Boltzmann machine, where we can have multiple layers of hidden variables like this. So, remember that earlier we had talked about the hierarchical models, where like we have one set of observed variables; but like there might be various hidden like we had talked about the parameter layer, the parameter model, the process model, data model and things like that.

So, these some of those these hidden variables you can consider to be the model parameters. So, like you may say that h^1 is your first one set of latent variables or process variables, h^2 are the parameters and h^3 may be the hyper parameters and things like that. And the aim here is to define a joint distribution over all of them, once again using the concepts of these edge and node potential functions.

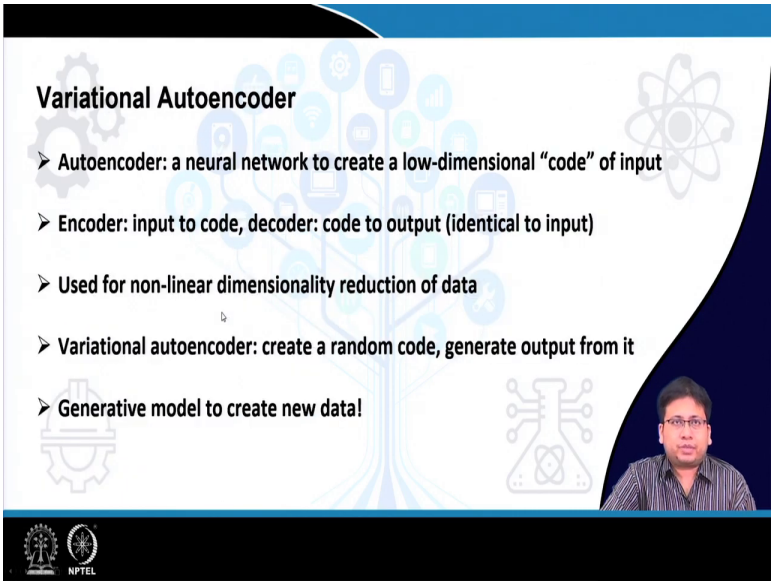
So, here also like this is the expression for what is known as the energy function? So, like in this case also in case of Markov's random field or the Boltzmann restricted Boltzmann machine also like first the energy function is defined like this and then the potential function is defined as e to the power minus the energy.

So, note that the energy function like the energy function can be defined for every node and every edge and then of course, there is a overall energy function, which is obtained by just

adding all these energy functions together and then like we the once we have the full energy function, we can convert it into the joint probability.

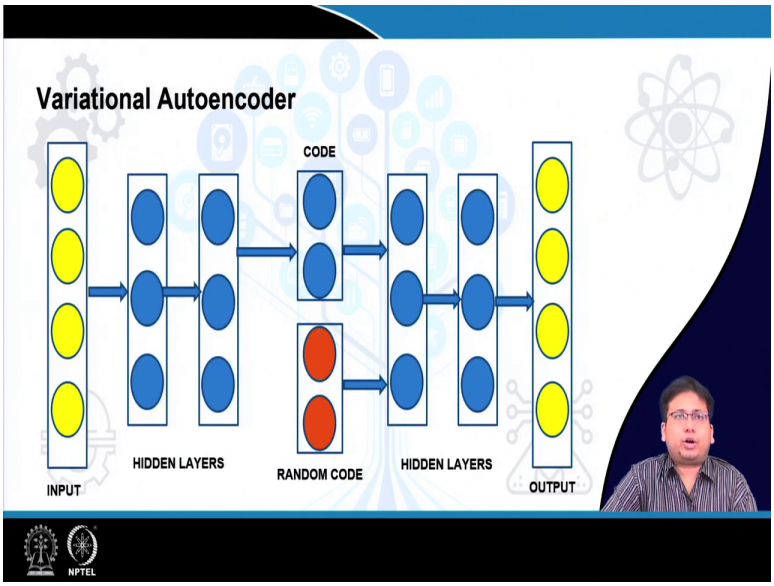
So, in this case also like we do something similar; the only difference here is that, we have multiple layers of these latent variables and this w^1, w^2 these in these are matrices which indicate or which includes the edge weights.

(Refer Slide Time: 22:18)



Now, there is another interesting generative model called the variational autoencoder. So, this autoencoder is a neural network to create a low dimensional code of the input. So, let us say the input might be something like a like a high dimensional vector or matrix or something like that now we often try to reduce its dimensionality by like algorithm such as PCA; but that we can say is the linear dimensional reduction, that is the new dimensions that are created by PCA are linear combinations of the existing dimensions. But if you want a non-linear combination, then one way to go about it is the autoencoder.

(Refer Slide Time: 23:07)



Now, what the autoencoder does? So, an auto encoder basically it takes in as an input passes that input through different hidden layers, where the usual neural network operations take place which may include convolution, pooling etcetera we had which we had discussed a couple of lectures earlier and finally, what happens is that, the input is reduced to a code. So, this part is known as the encoding.

Now, once we have the code; then there is decoding, where there are again other hidden layers which basically a carry out the deconvolution operation. And finally, gets back the out takes us to the output. And in most cases the input and output are actually identical, that is why it is called an autoencoder or the self-encoder.

The important thing here is the code that is the low it is a low dimensional representation; the input and the output are same, because like it is we are not that is we have an input vector, it is not that we try are trying to map it to something else, we are only trying to build a like a code or low dimensional representation of the input.

However, that code should be such that it should be possible to decode and get back the original thing also; that is I do not want that is of course, encoding can be done by throwing away a lot of the information, but if we throw away all the information, then we will not be able to reconstruct

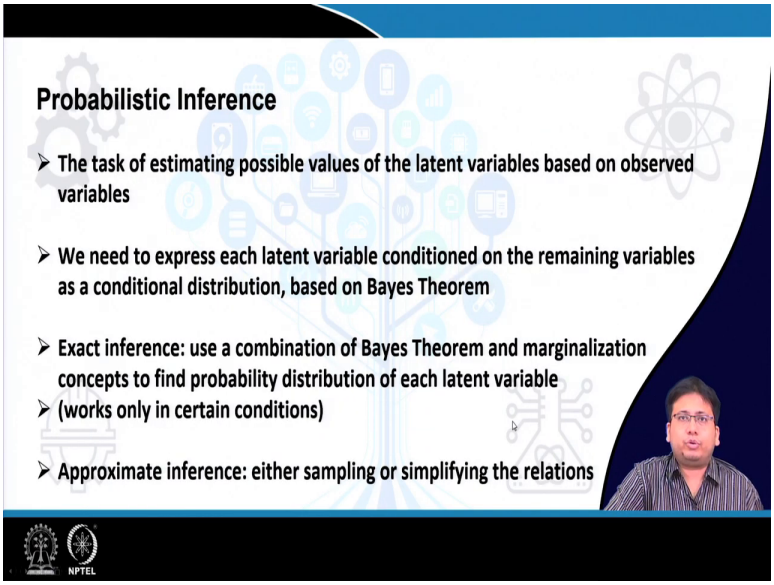
it. So, in case of auto encoders, the coding is done in such a way that once you get the code; then from the code you can again come back to the original thing.

So, it is not a lossy compression, it is we can say something like a lossless compression. Now, this is the usual autoencoder; now add the word variational in front of it and it becomes a different thing. So, here you create a random code and generate outputs out of it. And so, like instead of having a like creating a code correspond through an encoder, let us say that you just generate a code from some probability distribution and then pass it through the decoder.

So, then what now if the that probability distribution from which you are sampling the code if that is like that probability distribution satisfies the properties or is like is identical to the distribution of the or the statistical properties of the code which you get from actual data; then basically you will be able to generate new output, which has the same statistical properties as the original data set which you had, even though it may not be identical.

So, this is a an like a useful way for simulations. So, we have the domain of earth sciences we have talked about simulations especially in module 1 several times, where you want to generate new data, which is not identical to the training data; but which has similar statistical properties. So, this variational autoencoder is a very good way of doing that.

(Refer Slide Time: 26:21)



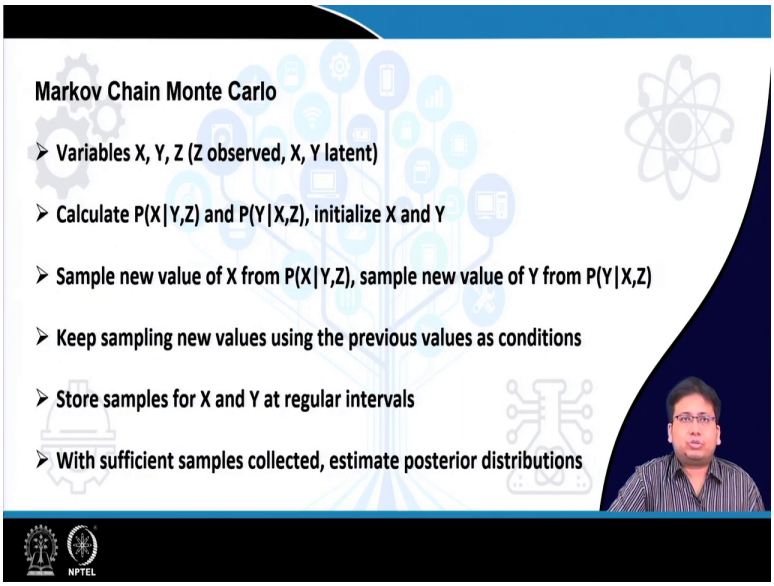
And now comes to a question of probabilistic inference. So, we have already talked about or we have been dealing with so many times of hidden or latent variables.

So, how about estimating their possible values based on the observed variables, that is called the probabilistic inference. So, for this purpose we need to express each latent variable conditioned on the remaining variables as conditional distributions based on Bayes theorem.

Now, there are two approaches; one is the exact inference, which uses a combination of Bayes theorem and the concept of marginalization, which allows us to find the probability distribution of each latent variable; that is we basically carry out there is this variable elimination algorithm, where you use the concept of marginalization repeatedly to like eliminate the other variables and come to the marginal distribution of every latent variable, that of course based on conditioned on the observed variables.

Now, these kind of exact inference it can work only in certain conditions if the graph graphical model satisfies some nice properties; the in general that will not happen, so we go for approximate inference, we either by sampling or by simplifying the relations.

(Refer Slide Time: 27:45)



So, there are two approaches; one is the sampling approach that is called the Markov Chain Monte Carlo approach.

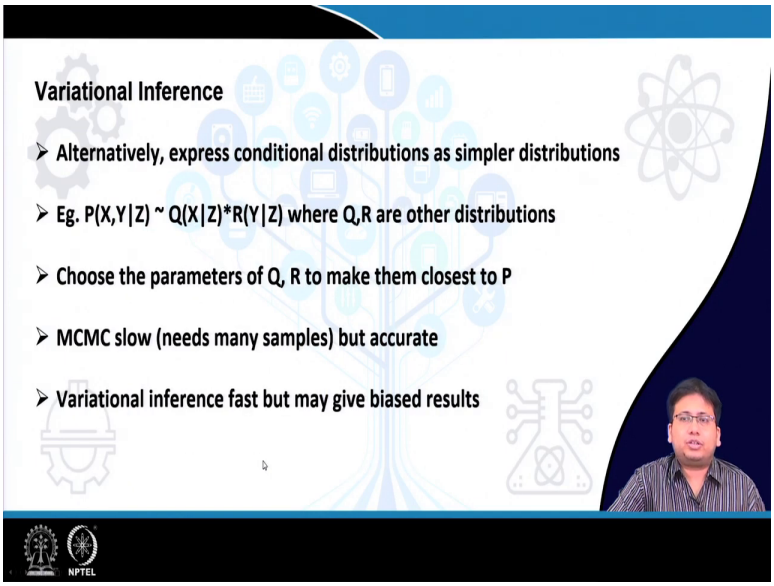
So, let us say that we have three variables X, Y, Z, where Z is observed and X, Y are latent. So, what we do first of all is to calculate the conditional distributions that is the distribution of $P(X|Y, Z)$ and that of $P(Y|X, Z)$. Now, Z is observed, $X\ and\ Y$ are latent; so we just initialize $X\ and\ Y$, that is we just assume some initial values of them.

Now, we sample new value of X from the this conditional distribution on $P(X|Y, Z)$. So, for Z of course, we know the value and for Y we consider the value which we have used for initialization. So, that way that is we draw a new sample of X and replace the earlier value which we with which we had initialized X.

Next we sample a new value from Y by sampling from this conditional distribution using the new updated value of X which we just obtained from the sample. And we keep on alternating this process once we express X, that is get a new value of X based on Y and Z and then we get a new value of Y like based on X and Z.

And this process keeps on repeating and we sample new values of the of each of the latent variables based on the previous values which are used in the conditions and like as we are generating these new values, we like store their samples at regular intervals. So, so finally, what happens is like we have a set of sample values of X and also of Y. So, once we have sufficient samples, we can estimate the posterior distribution.

(Refer Slide Time: 29:37)



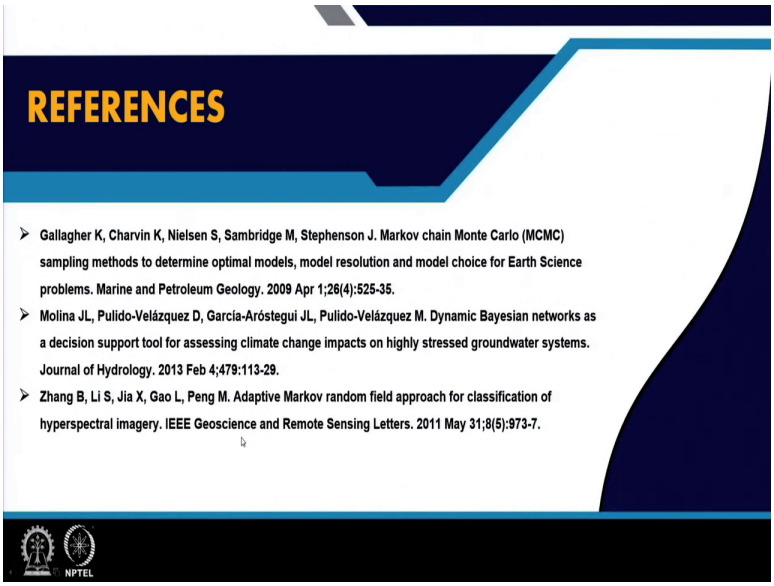
The other alternative is variational inference, where like instead of sampling; we can express the conditional distributions as simpler distributions. Say for example, if we have this kind of a joint distribution of X and Y, we can express it as a product of two simpler distributions; these like we can called as let us say Q and R, these are like proposal distributions.

So, as you can see these are simpler distributions, because they are like only on one of each of the individual variables; these need not be the actual marginal distributions of X and Y, but there are some other distributions, which have some other parameters. Now, we choose those parameters of Q, R distributions in such a way that this product is as close to this the original distribution as possible.

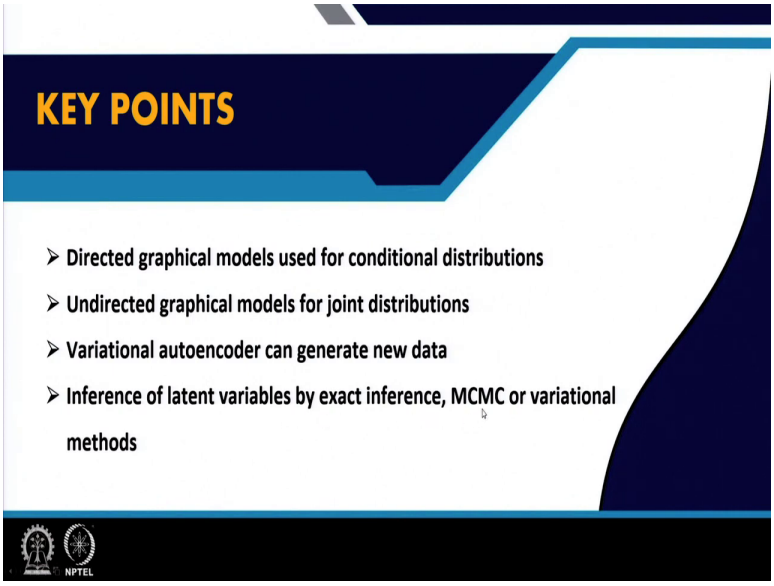
So, choosing these product, these parameter that is a optimization problem. So, basically the variational inference this approach, this is known the this is essentially an optimization based approach, which can be solved either directly or numerically as the case might be. Now, like the advantage of MCMC is that it always gives you, it will give you accurate result; but for that it will have to run for a long time, that it will have to collect lots of very large number of samples, ideally infinite number of samples. That is of course, not possible, but at least it has to be run for a very long time, so it is a slow process.

Now, variational inference on the other hand is fast, but it may like it may give biased results provided this approximation is not a good one.

(Refer Slide Time: 31:22)



(Refer Slide Time: 31:31)



So, here is a set of references, where these probabilistic models are used for problems in earth science. So, the key points to be taken home are that first of all the directed graphical models are

used for conditional distributions and undirected graphical models are used to represent joint distributions and variational autoencoder is a probabilistic model that may be used to generate new data having same statistical properties as your data set.

And when you have to do inference of latent variables, you can do it by exact inference, Markov sampling based MCMC or optimization based variational methods. So, that brings us to the end of this lecture and as well as the end of this module. So, in the next module, we will be seeing how machine learning based methods can be obtain, can be used to obtain some various new insights about various earth system processes. So, we will see you in the next lecture till, then bye bye.