**NPTEL**

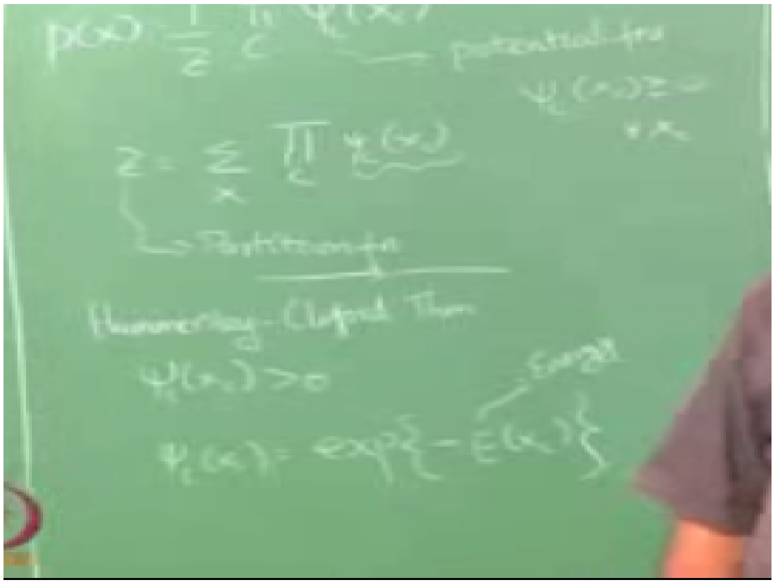
**NPTEL ONLINE CERTIFICATION COURSE**

**Introduction to Machine Learning**

# Lecture-66
# Undirected Graphical Models - Potential Functions

**Prof. Balaraman Ravindran**
**Computer Science and Engineering**
**Indian Institute of Technology Madras**

(Refer Slide Time: 00:19)



So we sometimes call this sometimes call this ψ as potential functions right, so ψC is a potential function associated with the click C right and XC is the set of variables that are participating in click C right, so this is a product over all of these clicks okay there is a small problem here right, so we have to make sure that whatever we are writing is a probability function right is how do you make sure of that right, so you basically and have some kind of a normalization factor where Z will essentially be whether the integral or the sum or whatever if you're looking at.

Looking at discrete values we have been talking about binary so far right so if you are looking at binary value at the variables essentially this will be sum over all values that X can take right suppose you have n binary variables the sum will run over $2^n$ entries right sounds like a really bad idea right anything where ever you're summing over $2^n$ elements seems to be a bad idea and it is so the most the biggest difficulty in using undirected graphical models. so why did not you have this in the director graphical models we chose the factors cleverly right.

The factors were chosen to be conditional distribution, so when it took the product it was it is guaranteed to be a distribution but here we have no such restrictions on the size right this is what makes it even more confusing I have no restrictions on the size I can be anything right so I can go run up till three million I do not care right so I can be any function right so I can become negative oops can it no right the only condition I have is ψ has to be non-negative yeah okay that is a good question this ψ have to be positive or non-negative mar-right.

So it is okay for me to have zero probability for some configuration right, so ψ can be so ψ non-negative then that is all the condition I need, so the only condition I need is that so for all values that XC can take so ψ has to be non-negative right so Z this also sometimes called the partition function which is the terminology that comes from physics right, so I am not going to get into the explanation of it but sometimes also not a separation so if you reading up something and somebody mentions partition function okay essentially the Z that they are talking about.

So that makes it a little tricky right, so we are saying that your this thing is not restricted in the interpretation right here, so I see is not restricted interpretation it can be anything and as long as you can do this normalization will get a probability okay, so I am going to write down a very powerful theorem right, so the Hamersley Clifford theorem or the Clifford Hamersley theorem it says that for any probability distribution right for any probability distribution that is consistent with this kind of a factorization over a graph right.

So any probability distribution that is consistent with this kind of a factorization or graph right the condition here is a little stricter, so the condition says that it cannot be 0 right so the condition says it has to be positive not non-negative so what the Hamersley Clifford for theorem says is if a probability distribution okay that is consistent with this kind of a factorization that is means of such a factorization exists okay then that probability distribution can also be expressed by using factors at our of this form.

So that makes our life easier, so then now my energy function or what they the e function is called the energy function all of this comes from physics right, so you will see all this energy and other things here energies and potentials you will see that so the energy function right can be anything, now so as no restrictions can be negative it can be positive right well as long as it is real I suppose not complex right, so if the energy function can be anything so this is known as the so essentially what the Hamersley Clifford theorem tells us is that.

So if you if you write your probability as a product of Exponentials right offset these kinds of factors then there exists a graph representation where this kind of a factorization can be obtained in like ways if you are able to write a factorization like this on a graph then you can have it is expressed as a product of Exponentials right, so each factor is an exponential so essentially my probability will be the product of exponentials right, so this is actually a very powerful result because it allows to simplify a whole bunch of computations right I do not have to consider any arbitrary form for my size.

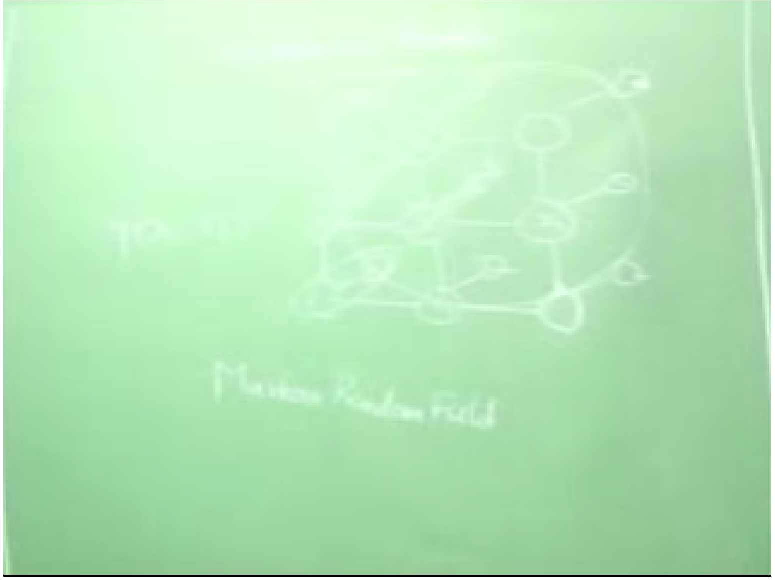
Yes immediately you see this right I do not have to consider an arbitrary form for $\psi$ it is just an exponential okay, now I have to consider an arbitrary form for my energy but now that we started talking about it as energy, so we can start applying our intuitions from physical systems right, so what should be a state with a high with a high probability exactly we know that right the state with the high probability should have low energy so when you start looking at the data right I find that configuration XC which is most popular most prevalent in the data right and I am going to assign that the least energy.

And I do this for every click the power of nomenclature, so I call it energy and now everybody understands what the graphical model is doing right, so if the energy is very high so it is going to be $e^-$ the energy so the probability is going to be low or rather the this factor will be low right and this $\psi$ c is going to multiply herein your numerator right therefore the probability that you assign will be low right if the energy is very low the e power that will be high relatively right and therefore the probability you will assign will be higher.

So that is essentially what we are going to do right, so as far as undirected graphical models is concerned so how do you decide what these energy function should be it depends on you are the prevalence of that particular configuration in the data right, so what will be the energy of $x_1$ is 1 and $x_2$ is 0 and $x_3$ is 1 okay how often did the combination occur in the input right and what should I do with that nothing I can just use the count as the energy right because energy is unrestricted do not have to worry about normalizing it or anything right., I just use the count as the energy for that counting the data I can use that as energy, so the higher the count the higher the energy and therefore more prevalent the more probable that configuration will be sorry higher the count okay then one by the count is energy then use one by countless energy right, so the higher the count the smaller the energy and then the more likely the configuration will be right sorry yeah, so that is the CC enough to do all right.

So I wanted to look at before I go on to do some inference thing I want to look at one ugly or too popular graphene be a close to end yeah so this can be zero so that can be at most one yeah in we are interested only doing probabilities here right, so that is not that is not that much of a mess of a problem right then again that is a beauty of the Hamersley Clifford theorem that is essentially it right it tells yeah it is fine you still can represent any probability distribution you want by looking at as this product of experience right. So that is the 3 thing now I will leave this since all of us now understand the undirected graphical models I start with the simple undirected graphical model right.

(Refer Slide Time: 01:49)



And right so this kind of a simple lattice like structure right so undirected graphical models are also sometimes called Markov random fields just like directed graphical models are also called Bayesian networks right undirected graphical models are also called Markov random fields right so people if you have heared of the term Markov random field somewhere right so one of the most often new structure in the with Marco random fields is like it is kind of a lattice structure okay.

So what this is really tell you it tells you that this variable okay is independent of everything else in the network given the four neighbors right then this variable is independent of everything else in the network given only these two neighbors I mean these lattices can run for like no 32 cross 32 or sometimes 256 cross 256 people typically use these kinds of lattices for modeling images okay.

So it is random you agree with the right, so all of these are random variables right so I have this collection of random variables okay it is called Markov because in this particular case right given the immediate neighbors right I am independent of everything else right so if you think of what the Markov assumption is in a probabilistic models right so sarcastic models with Markov assumption says that given the immediate predecessor for independent of the past right, so that is a normal Markov assumption right.

So given the immediate predecessor you are independent of the entire in the past so here what I am saying is instead of the predecessor because there is no notion of direction here so I am since there is no predecessor here, so instead of that I am saying given the immediate neighbors I am independent of everything else that is why it is called Markov right, so let us take XI right so XI is independent of everything else given all right so now what people do is they try to use this for by making all kinds of predictions right.

So I am interested in labeling every pixel in an image, so I have a big image right I want to label every pixel in an image give me an image labeling task I want to label every pixel are say that foreground or background and this is the guy standing there or is it the tree behind him right so I want to label it as foreground or background right, so now it is a to label task right so each of these random variables will take one of those values what are the values it will take it will take whether it is a foreground or is a background okay.

Now here is the additional assumption I am going to make I am going to make the assumption that the value of the pixel I am going to see the value of the pixel I am going to see depends on whether it is a foreground pixel or a background pixel right nothing else like the value of the pixel I am going to see it depends on whether is a foreground pixel or a background pixel and nothing else right now essentially what I will do is I am going to assign more random variables right.

Each one of them stands for the individual pixel each one of the stands for an individual pixel so now what will I do is I will observe these pixels right I initially will observe the pixels, so what will be my potentials here what are the size how many how many sides do I need one for each edge right, so that is the maximal click here I cannot do anything better than that so for every edge in this graph I will need a side but for every edge in the graph I will need a side right.

So what I will do is I will observe these pixel values okay, so some something some values I will observe okay then what I can do is I can figure out what should be this level of label of this pixel right just with this knowledge alone right because of there is this potential right I can kind of convert that into a potential on the node alone you see that because I have observed the values for the pixels right I can essentially take that entry from that thing right.

So there will be one column associated with that pixel value rights so the people will give me blank stares we are talking about a function of two variables right so I call this XI will call this y I the pixel value XI YI so I will tell you what YI is right then what will have we left with a function on XI right if I tell you what YI is I will be left with a function on XI correct so I can convert given an observation right I can convert these potentials right into potentials on XI alone.

YI is a part of the graph but the way use it will always be that I am given the YI so here is an image give me the labels which is the foreground which is the background, so I will always know this YI right so given the YI I can convert this edge potentials into node potentials right, so this essentially from now I from the function of XI, YI it will become a function of XI alone okay so if you look at many such I mean graphical model applications right, so you will actually find that they will always reduce this into node potentials and H potentials.

So it looked like they are defining something called a node potential okay which is like a potential function on single variables and then they will be defining H potentials which are potential functions on parts of variables okay in reality something like this would be happening for you to assign node potentials okay the node potential are essentially marginal some kind of information you have about the marginal’s okay, so in this case I am telling you how the marginal information comes.

It is not complete it is not the complete marginal okay given the pixel you can make some guess of work what the XI should be so that is my node potentiate so I already can reduce all of these H potentials between a node and pixel okay between a label and a pixel I can reduce it to a single potential on the label right.

So now having done that right what can I do so that is there will be some potential for a label here okay there will be some potential for label here right and there will be some potential for these two labels happening together right, so essentially this is telling okay this is a background okay what is the pro likelihood this is also a background if this is a background what is it likelihood this is also a background okay if this is the background what is the likelihood that is a foreground.

So like that right so for each of these edges so I have where the edge can change where the label can change that information I have, so finally when I assign the final labels to this so what will I do I will find that configuration of labels okay that gives me the lowest energy and essentially that would be in that this potential should be low right if I suppose this I say is background and this I say is foreground then when I say label is B here and label is F here that entry in the potential function should be low right people get that.

So that entry in the potential function should be small so like that I need to do this for all the paths here so it is a very hard problem so because it severely constrains so I have consider all possible paths and I have to figure out where we low entry occurs across this, so for example so if I say this is background and this is foreground and this is B and F and this gives me a low value right but I say B and  for this and this gives me a high value right.

So you can possibly turn this into an F right but then I the whole thing might come around and then this might get changed back into an FB right so then I might go around or not have to figure out what I see right potential to pitch this and that is the inference problem, so inference problem is really hard in the undirected case right so in so much so that when you have loops like this right when you have loops in the graph like this exact inference is impossible to actually give you the right answer is impossible like quite often we end up giving some kind of an approximation right.

So where can you give exact answers when there are no loops right so undirected with no loops is a tree right so on trees you can give exact computation but as soon as you have any loops in that then you have to do some kind of approximation there are some very special cases but we will not go that right you just want you to get an intuition of what the individual factors would mean right.

So how will I determine what these factors are okay so let us look at this right I want to look at I want to look at the gate right so what am I going to do what is it really nice oh I want to assign a low energy to the configuration that have occurs most often in my data right so what I will do is look at the label data so I look at each pixel's label right the label data will tell me which pixel is foreground which pixel is background right then I look at for these two pixels in the image let us say I have only a three by three pixel image right but for each of these two pixels I look at them right.

I will figure out okay how often was this foreground and this also foreground how often was this folk alone and this background how often was this background and this foreground how often this background in this background, so all these four things I look at just count that from the data and then I will set the energy to be some inverse of that count so the one with the largest count will get the smallest energy right it pretty simple right, so like that I can do this for each and everything right.

So that is one thing so the second thing I have to do is then start looking at this thing exactly that is what I think that is what is particle sorry you are looking at my back you can see from this I said there should be a potential function for XI and YI as well right so what that what I will do I look at the data again so look at okay when the pixel value was this okay what was the label right so I will do that I will do the co-occurrence information okay now things should start looking a little fishy to you guys I mean the pixel can have a lot of values depending on how I am encoding my color or brightness or something like that right.

So that I could end up having a very large distribution there itself right so even if you assume that my pixel is going to have 256 levels of brightness right so for every value so it is to 512 probability 512 counts that I have to make right 512 counts looks suspicious right will I have data to actually make accurate estimates of 512 individual comes well I could if I have very large volumes of data but typically what you do is you do not for YI you do not do the explicit counts like this right you try to learn the factor why I buy some kind of a parameterizes function right.

So you could use logistic regression right so figuring out what the what is the probability of XI given YI right that logistically regression can tell you right, so you encode your pixel using whatever things you want right so you can look at you can look and even more funny thing so you can do funky things you can make this a function of all of these fixed cells right so because you are moving away from your Markova of this but once you start thinking of doing a distribution you can come up with very powerful classifiers right.

So typically the YI, XI probabilities right or other the Y I XI factors or learn in a different manner they just do not do the maximum likelihood estimate you do some other thing that typically the most popular choices using logistic regression right you can do other things  but then the XI potentials you can learn and when using maximum likelihood estimate simple rational like limits provided they are small enough okay so that is basically how you train this Marco random field and it turns out that they are pretty powerful in terms of working with images and in a very wide variety of setting right and people use MRFs a lot right.

And there are variants of it which is called conditional random fields so people use those also tremendously so very popular and powerful classifier and training it can be a pain right, so I just give you a simple example right value just did the counting and stuff like that when you have a very large model right very large graph right training it can be a pain but then people because the data is paucity right.

So I had to look at all possible combinations of all variables right and so it becomes a little tricky and so you have to come up with the cleaver ways of training the models okay they say it is particularly this part of it is painful right not ψ not completely Letterman just that the inference processes is hard like you are selling your head you might you may keep going in circles it may take a long time for you to actually converge to a probability and so on and so forth yeah.

So that exists a proper assignment to this so that yeah so but the finding it is hard yeah Clifford Hamersley let tells me that no I am not worried about loops right fashion is a click there is a loop sorry one potential yeah click loose but there is only one potential right, so it is fine I will be doing the inference on that potential alone I would not get into this loop business so the loop business started off saying that okay I will make one inference one assignment here one assignment here one assignment but if this had actually been a click right then I will know what is the combination of assignments to these four variables that has the lowest potential.

Now I just done right so I would not have to runaround chasing the things so that is the right so but I still have to do the chasing run because this has other things it is involved in right so the only way I can ensure that I do not do the chasing around this if I have one connected graph become a complete graph right one if a complete graph then of course there is no chasing around and then what so what is the difficulty there is no factorization I am back in the system as well not have worried about the graph right, so if I have a complete graph there is no factorization so but that is only way I can ensure that I will not get into.

**IIT Madras Production**

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in