

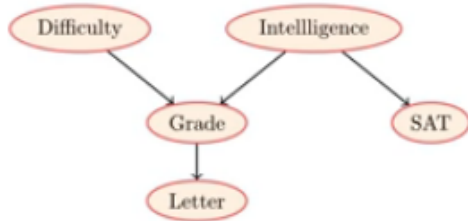
Deep Learning – Part - II

Lecture – 17.2

Factors in Markov Network

Module 17.2

Refer slide time :(0:17)

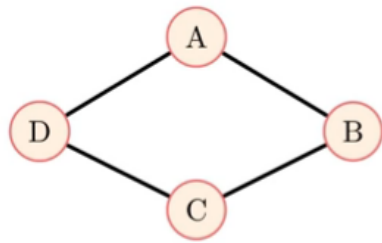


$$P(G, S, I, L, D) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)$$

- Recall that in the directed case the factors were Conditional Probability Distributions (CPDs)
- Each such factor captured interaction (dependence) between the connected nodes
- Can we use CPDs in the undirected case also ?
- CPDs don't make sense in the undirected case because there is no direction and hence no natural conditioning (Is $A|B$ or $B|A$?)

So the directed case, the factors were conditional to all the distributions and then the joint distributions, factorizes over these conditional distributions. Now each such factor, captured the interaction between the connected nodes. Can we use CPDs in the undirected case? Whenever someone asks this question, the answer is, 'Yes, we can'. No we can't. Right? Why? See CPDs don't make sense in this case, because there is not direction. What does a Conditional Probability Distribution mean? It means that, something given something. So there is already this G symmetry there. Right? You're saying that one, one random variable go in the given site and the other random variable is going to stay on the other site. There's always this A symmetrical relation is there. And the whole point of undirected graphical models, are marked on networks is that, you don't have this A symmetry. Did you get that? So now what is the factor associated with every node in a Markov Network?

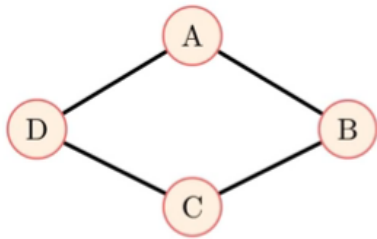
Refer slide time: (1:14)



- So what should be the factors or parameters in this case
- **Question:** What do we want these factors to capture ?
- **Answer:** The affinity between connected random variables
- Just as in the directed case the factors captured the conditional dependence between a set of random variables, here we want them to capture the affinity between them

Probability that they interact. But how do you formally say that? Then are you saying that, it should be a Joint Distribution of A & D? Okay, that makes sense. What did we actually want to capture? Okay. Factorization of the Joint Distribution. And between two nodes, what did we want to capture? Strength of the interaction. Right? So we are getting too obsessed with this, probability distribution, that's where we're thinking of answers in terms of probability distribution. But always remember that, you could always think of real numbers and then from there, getting into probability distribution, is not very hard. Why? Because always you can do, normal distribution. So let's start thinking in terms of real numbers. So every hedge let it capture, the strength of interaction between those two nodes. How to go from there. to a probability distribution? Should be straight forward, at least on paper. Do you get that? Do you get the motivation? Because we cannot use conditional distributions. Right? Okay. So now, we want these factors to capture the affinity between these two random variables. So just as in the directed case, the factors capture the conditional independence. Here you want to capture the affinity. So it's not very different. Right? So there you wanted to capture the conditional dependence, between the parameters and here you're saying, 'I want to capture the affinity between the parameters'.

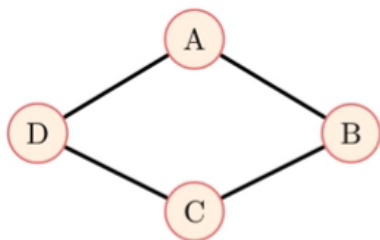
Refer slide time: (2:41)



- However we can borrow the intuition from the directed case.
- Even in the undirected case, we want each such factor to capture interactions (affinity) between connected nodes
- We could have factors $\phi_1(A, B)$, $\phi_2(B, C)$, $\phi_3(C, D)$, $\phi_4(D, A)$ which capture the affinity between the corresponding nodes.

So even in the undirected case, we want each such factor to capture the in, affinity and we could have factors of the following form. Right? We could say that, there is one factor, for AB, another factor for CD, another factor for BC and an another factor for AB. So what is the difference you see from the, connected case, from the directed case? There the factors were associated with, nodes; here the factors are associated with, Hedges. Right? Because hedges is what you want to capture not the nodes. Because if you capture the nodes then you are introducing some kind of a, A symmetry. That's what **Inaudible [3:26]**

Refer slide time: (3:29)



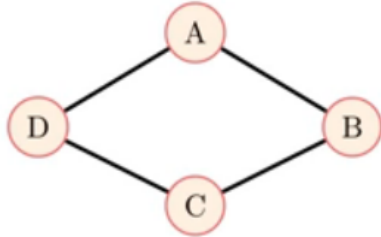
- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.

$\phi_1(A, B)$	$\phi_2(B, C)$	$\phi_3(C, D)$	$\phi_4(D, A)$
$a^0 b^0$	$a^0 b^0$	$a^0 b^0$	$a^0 b^0$
$a^0 b^1$	$a^0 b^1$	$a^0 b^1$	$a^0 b^1$
$a^1 b^0$	$a^1 b^0$	$a^1 b^0$	$a^1 b^0$
$a^1 b^1$	$a^1 b^1$	$a^1 b^1$	$a^1 b^1$

$\phi_1(A, B)$
 $\rho(A, B)$

So for every hedge you have, and we'll, re, revisit the statement of every hedge. Okay? So I could have factors of the following form. And what should each factor actually give me? So if I say, 'I want Phi 1 of A, B? Now remember and go to the discussion that we had about, marginals, joints and conditionals. Whenever I say, P of A comma B, what do I actually mean? Joint distribution, but what does that mean? I want a value for, every value that Capital A and Capital B can take. So now if I want, a factor for A, B, so what does that mean? I should give a value for every possible value that, A and B can take. Right? Does that make sense? So what I'm asking is, what's the strength of, a negative interaction between, A and B? That A also has a misconception and B also has a misconception. Right? So that's a, **Inaudible [4:26]** define it the other way. So that's A1, B1. So 1 is for misconception and 0 is for no misconception. Right?

Refer slide time: (4:34)



$\phi_1(A,B)$		$\phi_2(B,C)$		$\phi_3(C,D)$		$\phi_4(D,A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1

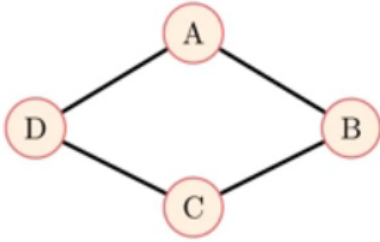
- But who will give us these values ?
- Well now you need to learn them from data (same as in the directed case)
- If you have access to a lot of past interactions between A&B then you could learn these values(more on this later)

- Intuitively, it makes sense to have these factors associated with each pair of connected random variables.
- We could now assign some values of these factors
- Roughly speaking $\phi_1(A,B)$ asserts that it is more likely for A and B to agree [\because weights for $a^0b^0, a^1b^1 > a^0b^1, a^1b^0$]
- $\phi_1(A,B)$ also assigns more weight to the case when both do not have a misconception as compared to the case when both have the misconception $a^0b^0 > a^1b^1$
- We could have similar assignments for the other factors

So for each of these things, I want to give a values, so I could give some values. Everyone can focus on this table. Right? So I've given you a factor, associated with every edge in the graph and for every factor, I've defined these values. That, A zero, B zero is 30. And you can just map it back to the conditional distributions. Infact, I can just say that, the directed graphical model is a special case of the undirected case, where these factors happen to be the conditional distributions. Right? Okay. So we'll come back to that. Let's ignore that for now. But now you have these four factors. Yes, so that's, okay, I'll come to that. I haven't **Inaudible [5:25]** so that's a interesting question. So, he's saying that, I want the strength of the interaction? Right? So they always interact in the same manner. So a zero, B zero is 100, A1, B1 is 100, A1, B zero, A zero, B1 all is 100. Okay? So that's the strength of the interaction. What's wrong with that? Think in terms of data. Think in terms of past experiences. So we'll come back to that. Okay? Who will give us these values? Who will give us these values? It's easy to write values, but who will give us these values? Text book. We'll have to 'dash' these values.

Title of the course, second word, from the title of the course. We'll have to, learn these values. Right? Always. Right? I mean, what are factors? What was the, what did I say, I'll interchangeably call them? Parameters. Right? So these factors are the parameters of your model. There are things that you will have to learn from, data. Now do you get an intuition for why this would be different? Does that make sense? Okay. Just as the conditional probability distribution, no one is going to give it to you; you'll have to learn it from the data. Similarly, these factors and values associated with the factors, you'll have to learn, from the data. How? How will you learn these from the data? How? So in principle if you have a lot of past interaction between A and B, then you could learn this from data. Does that make sense? Right? How is still not clear. And that's what we'll do over the 1 or 2 weeks. Right? So that's the purpose. That's the, eventual goal that we're heading towards. And if you're given these kind of, undirected graphical models and we'll actually focus on a very specific, undirected graphical model, from the family of graphical models. And we'll see, how to learn these parameters. Okay? But you have to understand that these factors are the parameters, of your network. So now just as in the classification case, if I ask you, what is probability of, this image, being an apple, he used all the parameters in your network, that you've learnt and finally give me that Softmax output, which tells me, what's the probability of this **Inaudible [8:25]**. Right? And he learnt those parameters, from the data. Now the questions are not very different here. You're asking me the same question. What's the probability that, A has a misconception? That's very similar to the question, what's the probability, that this image is an apple? Right? And just again as you have parameters there, you will have parameters here, also. And then you will have to learn, these parameters, from some data. How, is a big question? There's an entire course on that. So I encourage you to take it, next time. But, we'll take some parts of it and do whatever is, relevant for us. Right? So we will get there. But I want you to understand that, this is not free, this is have to be, this will have, this will come after lot of hard work. As you will see in the last assignment. Okay? So roughly speaking, 51 AB asserts that, its more likely that A and B to agree. Right? So that is what, it is asserting, that both of them being zero or both of them being 1, is more likely, than the other case. Now this is based on our past data. Right? We would have seen lot of interactions between A and B and they tend to kind of, agree with each other or convince each other and are not in this constant stage of debate with each other, that, they say that, this wrong and this is right and so on. Right? So that is why these factors, need to capture these weights. That the strength of the association is more in the positive or the agreement case, as compared to the disagreement case. The other thing this is capturing is that, it is more likely that both the misconception, rather, sorry, I, I, I forget zero and 1. But what it says is that, a zero has, no misconception. It is more likely for both of them, to not have the misconception, than have the miss, than have the misconception. Right? So what that means is that, probably A and B are independently capable of resolving their misconception. And given the fact that, they often tend to agree. Once they've cleared the misconception, they will pass it on to the other. Right? That's why in the end, both of them will not have a misconception, that's what this is saying. A zero, B zero is more likely, than A1, B1, and the case where both of them have a misconception. Right? So these are again, things which, I've just written these numbers. But, you will have to learn these from the data. It at least can you imagine that, you can learn these from the data? Imagine. Okay. Fine. And we could have similar arguments for the other factors. So these, other factors are BC, CD, and DE. Right? For all these factors, now for every assignment to A and B, we have learnt the strength of these associations.

Refer slide time: (10:56)



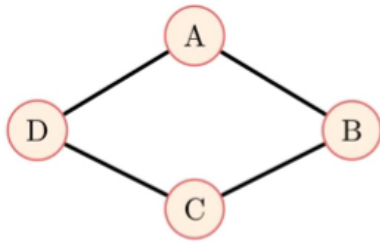
- Notice a few things
- These tables do not represent probability distributions
- They are just weights which can be interpreted as the relative likelihood of an event

$\phi_1(A, B)$		$\phi_2(B, C)$		$\phi_3(C, D)$		$\phi_4(D, A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^0	100
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100
a^1	b^1	10	a^1	b^1	100	a^1	b^1	1

Now I have this, I don't know what I wanted to notice here. Okay, these tables do not, represent probability distributions. That should be plain obvious. The values don't lie between zeros to one, they do not sum to 1. Right? And they're just weights, which can be interpreted, as the relative likelihood of an event. Right? We are just saying that it's, more likely for A zero, B zero to happen, then, any of the other cases in the table. Right? That's what these numbers are actually trying to capture. Right?

Refer slide time: (11:25)

- But eventually we are interested in probability distributions



$\phi_1(A, B)$			$\phi_2(B, C)$			$\phi_3(C, D)$			$\phi_4(D, A)$		
a^0	b^0	30	a^0	b^0	100	a^0	b^0	1	a^0	b^0	100
a^0	b^1	5	a^0	b^1	1	a^0	b^1	100	a^0	b^1	1
a^1	b^0	1	a^1	b^0	1	a^1	b^1	100	a^1	b^0	1
a^1	a^1	10	a^1	b^1	100	a^1	b^1	1	a^1	b^1	100

$$P(\cdot) = \frac{1}{Z} \phi_1 \cdot \phi_2 \cdot \phi_3 \cdot \phi_4$$

And, this is fine. For the representation of the factors, it's okay, if you're not still dealing with probability distributions. But what do we eventually want? We want a probability distribution. And what is the probability distribution, we want? The Joint distribution. Now, how do I write, the Joint distribution? How do I write the Joint distribution? Yeah, yeah, yeah. Yeah, that's fine. How will you write, the Joint distribution? Can I write it as; B of ABCD is equal to Phi1, Phi 2, Phi 3, Phi 4? Can I write it as that? And then what will I have to do? If I normal as this appropriately would you be fine? I'll include a Z. What would Z be? Summation, over all possible assignments for, Phi1, Phi 2, Phi 3, Phi 4. Does that make sense? Right? So that's one way of taking a set of real values and converting them into a distribution. And that's what I started with, that don't worry about, these values, having the properties of a probability distribution. That means, they sum to 1 or lie between 0 to 1 and all that. Just give us some values and then you can figure out, how to convert them into probability distribution. And this is one way of converting that.

Refer slide time: (13:07)

Assignment	Unnormalized	Normalized
$a^0 b^0 c^0 d^0$	300,000	4.17E-02
$a^0 b^0 c^0 d^1$	300,000	4.17E-02
$a^0 b^0 c^1 d^0$	300,000	4.17E-02
$a^0 b^0 c^1 d^1$	30	4.17E-06
$a^0 b^1 c^0 d^0$	500	6.94E-05
$a^0 b^1 c^0 d^1$	500	6.94E-05
$a^0 b^1 c^1 d^0$	5,000,000	6.94E-01
$a^0 b^1 c^1 d^1$	500	6.94E-05
$a^1 b^0 c^0 d^0$	100	1.39E-05
$a^1 b^0 c^0 d^1$	1,000,000	1.39E-01
$a^1 b^0 c^1 d^0$	100	1.39E-05
$a^1 b^0 c^1 d^1$	100	1.39E-05
$a^1 b^1 c^0 d^0$	10	1.39E-06
$a^1 b^1 c^0 d^1$	100,000	1.39E-02
$a^1 b^1 c^1 d^0$	100,000	1.39E-02
$a^1 b^1 c^1 d^1$	100,000	1.39E-02

- Well we could still write it as a product of these factors and normalize it appropriately

$$P(a, b, c, d) = \frac{1}{Z} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a)$$

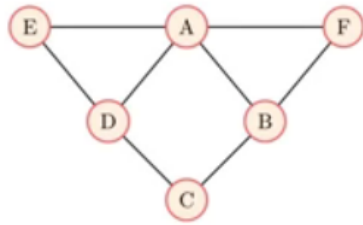
where

$$Z = \sum_{a,b,c,d} \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(d, a)$$

- Based on the values that we had assigned to the factors we can now compute the full joint probability distribution
- Z is called the partition function.

And now your joint distribution would essentially be a product of these factors that you have. Is that fine? Okay? So now once again our original goal was to take these variables, use a graph, to encode the dependencies or independencies between them. Once you have the graph, learn these factors associated with the graph and then express the joint distribution, as a product of these, factors. Okay? So we've achieved our goal. Is there anything else, we need to do? Are there, other cases possible? So let's see right.

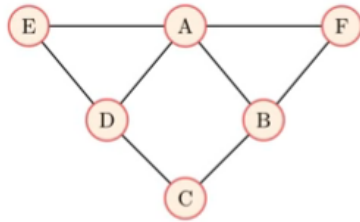
Refer slide time: (13:58)



- Let us build on the original example by adding some more students
- Once again there is an edge between two students if they study together
- One way of interpreting these new connections is that $\{A, D, E\}$ form a study group or a clique
- Similarly $\{A, F, B\}$ form a study group and $\{C, D\}$ form a study group and $\{B, C\}$ form a study group

So let us build, by adding some more variables. Now suppose, this is the graph that we have. Now we have, more students and these are the integration, interactions. Right? So can you tell me some, can you tell, explain this graph in English, to me? What does this mean? For example, what does this mean? The students A, D, and E form a study group. Right? And it also means that A and E independently also can form a study group. Okay? So in the earlier graph, my factors were these hedges. Now should I have a factor, for every hedge in this graph also? Again my quest is the same. I want to, now learn a joint distribution, over these, 6 random variables and again, whether these 6 students have a conception, misconception or not. Right? Now what should my factors be, in this case? I can of course again, do the same thing. I can again have factors for every hedge and then partition, the joint distribution, over these factors. Can I do something smarter than that? What can I consider, instead of hedges? A group. What's a group in graph terminology? What will you take from here? We'll take the 'dash' in the graph. 'Dashes' in the graph. What does A, E, D form? Yeah, it's a valid answer. In terms of graph theory, what does it form? A clique. Why is not everyone answering? Or why is that brave man, have to answer, so loudly? Oh, okay. So, okay. What is a clique? So if you have a, a, I hope I defined it properly. So if you have a set of nodes and every node is connected to every other node, then the form a clique. Right? So all the nodes in the set, are pair wise connected to each other. So that's exactly what is happening in a triangle, that all the nodes are connected, pair wise. Right? So, instead of hedges, does it make sense to take cliques? What would that actually mean, in terms of number of factors, number of parameters and so on? Reduce the number of parameters. Right? So now you will have factors associated with the cliques, instead of factors associated with the hedges. Right? So now can you tell me, again this is just a definition of the whole network, which I just explained, that these guys form a study group and so on.

Refer slide time: (16:43)



$$\phi_1(A, E)\phi_2(A, F)\phi_3(B, F)\phi_4(A, B)$$

$$\phi_5(A, D)\phi_6(D, E)\phi_7(B, C)\phi_8(C, D)$$

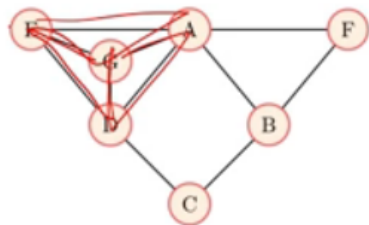
$$\phi_1(A, E, D)\phi_2(A, F, B)\phi_3(B, C)\phi_4(C, D)$$

- Now, what should the factors be?
- We could still have factors which capture pairwise interactions
- But could we do something smarter (and more efficient)
- Instead of having a factor for each pair of nodes why not have it for each maximal clique?



Now, what should the factors be? We could still have these factors. Right? I could have a factor associated with every edge and then I could normalize that to get a Probability Distribution. Right? Instead, we could do something smarter. We could have a factor associated with every maximal clique. What do I mean by a maximal clique? So remember that A and E is, also a clique. Because there are only two nodes and they are connected, so trivial E forms a clique. So when I say a maximal clique, that, it's a clique, such that, I cannot add any more nodes to it, without not making it a clique. Right? So, A, E, D, is a maximal clique. Because if I add any other node to it, then there is at least one node in A and, A, E and D, which is not connected to that. Everyone knows what's a maximal clique is? Please raise your hands? Okay. Now everyone knows. Fine. So I would, I'm saying that the factors would be maximal cliques. Infact even in those simple examples, when we had A, B, C, D, the factors were actually maximal cliques. They were just trivial maximal cliques, of size 2. But now I'm saying, I have maximal cliques of size 3. Everyone is fine with that? Okay? So now my factors are going to be, cliques, which look like these. Right? So remember that I have some cliques of size 3 and some cliques are still of size 2. Depending on the nature of the **Inaudible [18:03]** Okay?

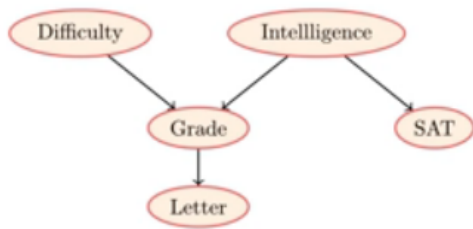
Refer slide time: (18:06)



- What if we add one more student?
- What will be the factors in this case?
- Remember, we are interested in maximal cliques
- So instead of having factors $\phi(EAG)$ $\phi(GAD)$ $\phi(EGD)$ we will have a single factor $\phi(AEGD)$ corresponding to the maximal clique

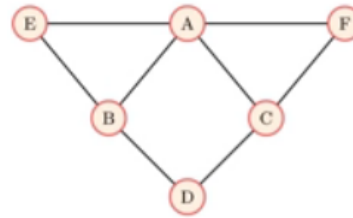
Now what if I add one more student? What would the factors be now? One factor for? Everyone is saying? A, D, G. Another factor for? A, B, F. Another factor for? B, C? One more factor for? C. So, everyone get's this? We now have maximal cliques again. Now the thing I wanted you to understand is that, I did not choose to take, each of these as, separate factors. Right? There are three, 3 sized cliques, inside that one, 4 sized clique. Right? I did not choose to take them independently. Because I was interested in maximal cliques. Okay? So, every maximal clique in the graph, I'm going to make it as a factor. Is that fine? Okay? See right now what we're doing and what we'll continue to do, for this discussion is that, we're only talking about intuitions. All these intuitions are backed with solid theory. But I don't have the time and rather the scope to cover the theory in this course. I will give you pointers; you can go back and look at those. But as long as you understand the intuitions, this suffices for this course. Right? So maximal cliques and all that, there is a solid theory behind that. But I won't get into that. Okay?

Refer slide time: (19:20)



- A distribution P factorizes over a Bayesian Network G if P can be expressed as

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | P_{aX_i})$$



- A distribution factorizes over a Markov Network H if P can be expressed as

$$P(X_1, \dots, X_n) = \prod_{i=1}^m \phi(D_i)$$

where each D_i is a complete sub-graph (maximal clique) in H

A distribution is a Gibbs distribution parametrized by a set of factors $\Phi = \{\phi_1(D_1), \dots, \phi_m(D_m)\}$ if it is defined as

$$P(X_1, \dots, X_n) = \frac{1}{Z} \prod_{i=1}^m \phi_i(D_i)$$

So now, remember that, for the directed case or the Bayesian network, the distribution factorizes over the conditionals or the local Probability Distributions. Now in the case of a Markov Network, the distribution factorizes over, what? It factorizes over some factors, where each factor is associated with a maximal clique in the **Inaudible [19:49]**. And remember that, you could choose, if you want, to not focus on the maximal cliques, but take, even sub-cliques also. But it's that, if take maximal cliques, there is no. See in the case of, that's the distinction, that I wanted you all to understand. In the case of Bayesian Networks, things were conditional to Probability Distribution. Right? So you don't have the choice of, whether I can take A, given D, comma C or A given D or whatever. Right? Whatever A depends on, you will have to take that, as a factor. Here now, you have a choice. You could say that, I would like to focus on this separately. Because, maybe learning this properly, would give me, better understanding of the Joint Distribution. Right? Or you could choose to teach the entire clique together well maybe that is more efficient to do that. So you have that choice. Probably the smartest thing to do is to, focus on the maximal cliques. But later on, I will also make assumptions, where in addition to maximal cliques, I might also be interested in some of these, local factors also. Okay? All that is valid. The only thing that we're doing is, in the end we're ensuring, that, the, graph factorizes over these factors. Sorry, the Probability Distribution factorizes over these factors and we add this partition function, which is, Z , to make sure that, at the end we get the distribution, Even though the factors are not distributions, unlike the, Bayesian case. In the Markov Network, the factors are not distributions. But we can always normalize them, by adding the partition function and get back a distribution.

Refer slide time: (21:20)

-
- Let U be the set of all random variables in our joint distribution
 - Let X, Y, Z be some distinct subsets of U
 - A distribution P over these RVs would imply $X \perp Y | Z$ if and only if we can write

$$P(X) = \phi_1(X, Z)\phi_2(Y, Z)$$

- Let us see this in the context of our original example

So I'll, I think, probably stop here.