Lecture -16.10

Deep Learning Part - II:

I-Maps

Refer slide time :(0:14)

We are now ready to formally define the semantics of a Bayesian Network

Bayesian Network Semantics:

A Bayesian Network structure G is a directed acyclic graph where nodes represent random variables $X_1, X_2, ..., X_n$. Let $P^G_{a_{X_i}}$ denote the parents of X_i in G and NonDescendants(X_i) denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional independence assumptions called the local independencies and denoted by $I_i(G)$ for each variable X_i.

$(X_i \perp \text{NonDescendants}(X_i)|P^G_{a_{X_i}})$

Yesterday, we were talking about, Bayesian networks and we saw they are directed acyclic graphs, where the nodes represent these random variables and the edges indicate dependence and we came up with this rule of, which the patient network actually encodes, which is any for, any given random variable, it is independent of its non different descendants, given its parents right?

Refer slide time :(0:38)

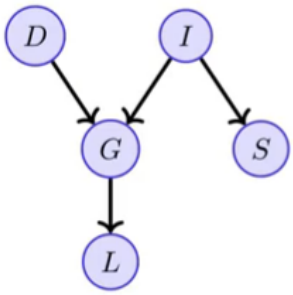
- We will see some more formal definitions and then return to the question of independencies.

So, that's the rule that we came up Right?

Refer slide time :(0:41)

Today, we move on to something known as, I maps and that's the last topic, that we'll do in directed graphical models or Bayesian networks, of course this is, I mean, directed, undirected graphical models, as a separate course in the department. So, I'm not going to cover everything, I'm going to cover the bare minimum, that we need to reach our eventual goal which is RBMS so, yesterday's lecture, was to get you started in thinking in terms of probability distributions, what are joined conditionals probably? The last two lectures are meant for that and from there we now, slowly start moving towards undirected graphical models or also known as Markov networks and from there you hopefully eventually reach RBMs. Right? So, our goal is RBMs. Our goal is not probabilistic graphical model Okay?

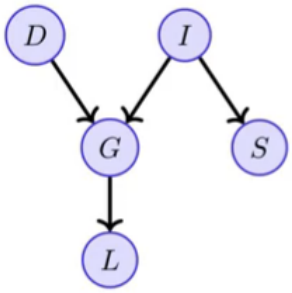
Refer slide time :(1:30)



- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$
- We define $I(P)$ as the set of independence assumptions that hold in P.
- For Example: $I(P) = \{(G \perp S|I, D), ...\}$
- Each element of this set is of the form $X_i \perp X_j|Z, Z \subseteq X|X_i, X_j$
- Let $I(G)$ be the set of independence assumptions associated with a graph G.

So let's, move on to I'm apps ,so let P be a joint distribution, over random variables x1 to xn and then we can define, IP or as a set of independence assumptions, that hold and P. Right? Or in fact not assumptions independence relations that hold in P. Okay? So, for example, IP could be G's independent of S given I, D, or G's independent of I and so on. I mean, sorry, L is independent of I N, things like that. Right? So, it could be all the independence assumptions that your probability distribution has, remember at this point, I'm not talking about the Bayesian network; I am only talking about the distribution. A distribution can exist a respect of the Bayesian network; a Bayesian network is a framework that you chose to represent a distribution. But, it's not always tied to a distribution; a distribution is independent of that. Right? And each element in this set of IP, would be of the following form ,where you will have X I is independent of X J given some Z, where Z is a subset of X, minus X I comma, X .right? And this, this I'm not sure; I think it should be, this is to the forward slashes or thing to use the right? So, it's or maybe the backward slash, just check once, how do you say X, minus X I, X, I think it's the backward slash, back Slash. Right? Yeah. Okay? So, X, minus X I, XJ, so in particulars Z could also be the empty set, in which case X I, is just independent of XJ, unconditional. Right? It just happens to be independent. Okay? Now, consider the graph G and let I G be the independence assumptions associated with this graph or Bayesian network. Okay?

Refer slide time :(3:10)



- We say that G is an I-map for P if $I(G) \subseteq I(P)$
- G does not mislead us about independencies in P
- Any independence that G states must hold in P
- But P can have additional independencies.

So, we then say, that the graph is an IMAP, for the probability distribution. If I G is a subset of IP, that what does that mean actually? The graph exactly captures all the Independence's, which are there in the distribution, is that what it means? And then, what does it mean subset. So, it means, that the graph is not going to mislead us about independence is in P. Right? It will not have any independence, which does not exist in P, it so anything that the graph tells us will hold in P for sure, but in addition, P could have additional independence is also Right? That's fine. But, at least the graph will not, come up with Independence's which do not exist in the original distribution. So, it will be faithful to the original distribution, it might have some shortcomings but, it's not at least going to be wrong in the other direction. Right? Okay?

Refer slide time :(4:01)

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- Consider this joint distribution over X, Y
- We need to find a G which is an I-map for this P

$P(X,Y) = P(X) \cdot P(Y)$

$X=x, Y=y \quad X=x \quad Y=y$

Now, consider this joint distribution over X comma Y. Right? So here's, one distribution. Now, we need to find a G which is an IMAP for this joint distribution. So, how will we go about that? So, when I say we need a G, what do I mean? A graph. Right? I want to draw a graph, which is an IMAP for this distribution, is that hard, easy, a click, for two variables.

How will you go about finding an IMAP? What does the IMAP have to satisfy? We just defined it in the previous line or what does the IMAP say, the independence is encoded in the graph, should be a subset, of the Independence's in the distribution. Okay?
So, what do you need to get first? Independence in the distribution. What are the independence is which hold in this distribution? What are the possible Independence's which can hold in this distribution? X is independent of Y, is it true? How will you, verify that find the marginal's. Okay? But, what do you do with the marginal's? What will you check front of the is equal to the joint distribution? So, what does this mean actually? I think this is, what you guys are saying .right? what does this mean for every value that x and y can take, the probability of X equal to X comma Y is equal to Y is equal to probability of, X equal to X ,y equal to Y. Right? That's what you mean, right .Okay? So, can you verify that easily? Okay?

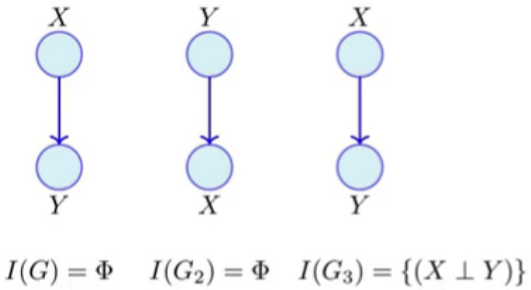
Refer slide time :(5:31)

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- Well since there are only 2 variables here the only possibilities are $I(P) = \{(X \perp Y)\}$ or $I(P) = \Phi$
- From the table we can easily check $P(X,Y) = P(X).P(Y)$
- $I(P) = \{(X \perp Y)\}$

And so here, there are only two possibilities, either the independence is associated with that distribution is X independent of Y or it's a null set, one of these two things and from the table we can easily check, that this condition holds, that's why the IP is, X is independent of Y. Now, given that this is IP, what is the graph that you will draw?

Refer slide time :(6:00)



$I(G) = \Phi$ $I(G_2) = \Phi$ $I(G_3) = \{(X \perp Y)\}$

- Since we have only two variables there are only 3 possibilities for G
- Which of these is an I-Map for P?
- Well all three are I-Maps for P
- They all satisfy the condition $I(G) \subseteq I(P)$

So, there are three options. Right? With two nodes, there are only these three options, or in fact there is one more option. Right? Oh, so what will be done? Okay? The last one should not have an edge, so the first one says X to Y, the second one says, Y to X. Okay? The third one says, there is no edge. So, which of these is an IMAP for this distribution, which one? What is the independence? Which the third one encodes? Third when there is no edge .So, what does it in encode? What about the first two? What is I G, in the first case? So, this is, this an IMAP for the distribution, no? What's the definition subset? The null set is a subset of any set .right? So, all three of them are actually I maps, alright fine. Okay? So now, this is just a toy example to falsely make you think, that for any distribution, you can easily come up with an IMAP, that's what we think. Right? Could there be cases, where we cannot come up with an IMAP why, why would that happen? Okay?

Refer slide time :(7:13)

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

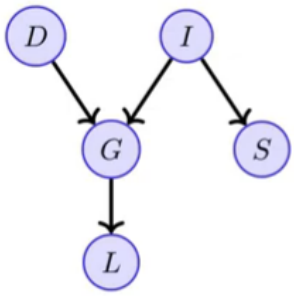
- Of course, this was just a toy example
- In practice, we do not know P and hence can't compute $I(P)$
- We just make some assumptions about $I(P)$ and then construct a G such that $I(G) \subseteq I(P)$

So, in practice we do not know P and hence we can't compute IP. Right? we computed IP, because someone had given us the table, from the table, we looked up P of X, P of Y and then found out that X is independent of Y, using the product rule. But, that's the whole point that this table is not going to be given to you, in the case of real-world examples and because, in real well you won't have two variables, you have many, many variables. So, this table is not known to us in many cases. Now, what we do in real world is that we just make some assumptions about IP. So, this is what we did in our student example. Right? We wanted to learn this Joint Distribution of student grade, Stat score difficulty intelligence and so on. And we made some independence assumptions, we said .Okay?

I think these are the independence assumptions, which hold in this distribution, which makes sense to me and we also served some counter examples, where we could have made a different set of assumptions, where we could have said that recommendation later actually also depends on the start score, it also depend on depends on the advice taken from the

colleague of the professor, all it also depends on the mood of the professor and so on. Right? So, those were all our assumptions, based on those assumptions, we first came up with the random variables, which are relevant to our world and then came up with some independence assumptions and based on those assumptions we drew a graph such that ,the graph was a subset of all the assumptions, that we had in mind. Right? So, that's the difference between, what this toy example was trying to show you and what we do in real world.Okay?

Refer slide time :(8:42)

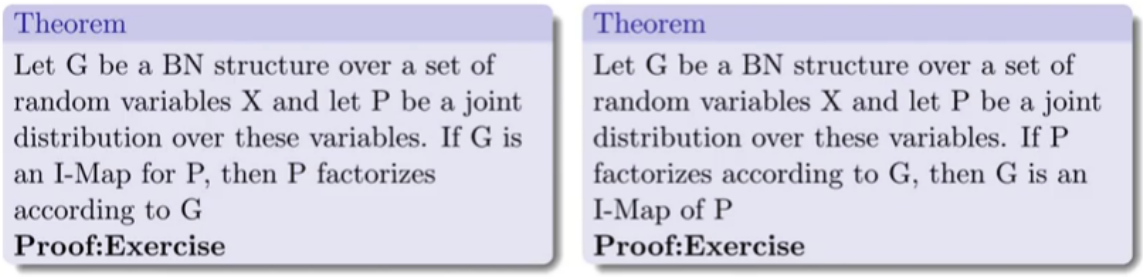


- So why do we care about I-Map?
- If G is an I-Map for a joint distribution P then P factorizes over G
- What does that mean?
- Well, it just means that P can be written as a product of factors where each factor is a c.p.d associated with the nodes of G

So, why do we care about I maps actually ,say for a very complex distribution, I gave you an IMAP, that means I gave you a graph, such that, the independence is encoded in the graph or a subset of the independence is encoded in the distribution, what will you do with that? What will that ensure? In the absence of any independence assumption, what's the factorization of a Joint Distribution? What do we always start with well factorizing as distribution? What is that one rule? So, the absence of that you always use the chain rule. Now, if I give you an IMAP, what will you do? So now, you know some Independence's and then what will happen? the factorization will, so that happens, because, if G is an IMAP for a Joint Distribution P, then P factorize is over G, that means, that P can be expressed as a product of all the conditional probability distributions which are associated with G ,conditional or marginal distribution. Right?
So, remember when I give you a graph ,a Bayesian network, I don't give you the edges, I mean, edges don't give you the nodes and the edges, there's also conditional probability distribution associated with each node, of course in practice we'll have to learn that , but we'll assume, that we have learned that somehow and once we have those conditional and marginal tables, we know that, the actual Joint Distribution factorizes over, these conditional and marginal distribution, does that make sense? There's nothing new that, I'm saying, I'm just putting it in a different light. Right? We already saw this that, if you have a Bayesian network, you can write the joint probability distribution as a factor of all the conditional probability distributions and the marginal distributions associated with the Bayesian network. Right? And now, instead of just saying that there is one unique Bayesian network, I'm saying that there could be many IMAP. Right?

Each of which, each of which have this property, that the independence is encoded in them, are a subset of the independence is in the Joint Distribution and once I tell you that, you can go ahead and boldly write, the Joint Distribution as a factorization of the conditional and marginal distributions, how many, if you get this statement? Please raise your hands.

Refer slide time :(10:40)



So here, are two theorems, on Bayesian networks, the first one tells you that, if G is a Bayesian network, over a set of random variables X and P is the Joint Distribution of these random variables, then if G is and imap for P, then P factorizes according to G. Right? And the second theorem is the converse of this that if P factorizes according to G, then G is an I Map of P. Okay? So, again get back and check these two theorems, I am not going to ask you these, just as didn't ask times, last year. So, you can go back and check, we take a look at these and it will just help you in improving your understanding. Right? So, this is why I-Maps are important, so the purpose of this module 2, was to show you that, once you have an I-Map, you can find the factorization of a joint distribution and throughout this entire discussion of Bayesian networks, there are two things ,which have been important, one is or two or three maybe, one is that the chain rule, gives you a natural ,gives you the most default factorization for a joint distribution, our aim was to simplify the chain rule and that happens only if you're billing in certain conditional independence assumptions. Right?
And these conditional independence assumptions are things that we encoded in the Bayesian network and we saw the formal semantics of the Bayesian network, which said that, X I, is independent of its non descendants, given its parents. Right? so, that one rule, allows us to simplify a lot of things in the joint distribution and that's why we always care about Independence relations, we always want to find, which are the things, which are independent and lastly we saw that, there's if someone gave us an I-Map, we can just go ahead and factorize the joint distribution, according to the I. Okay? So, these are the main things that, we wanted to learn from this lecture.