NPTEL

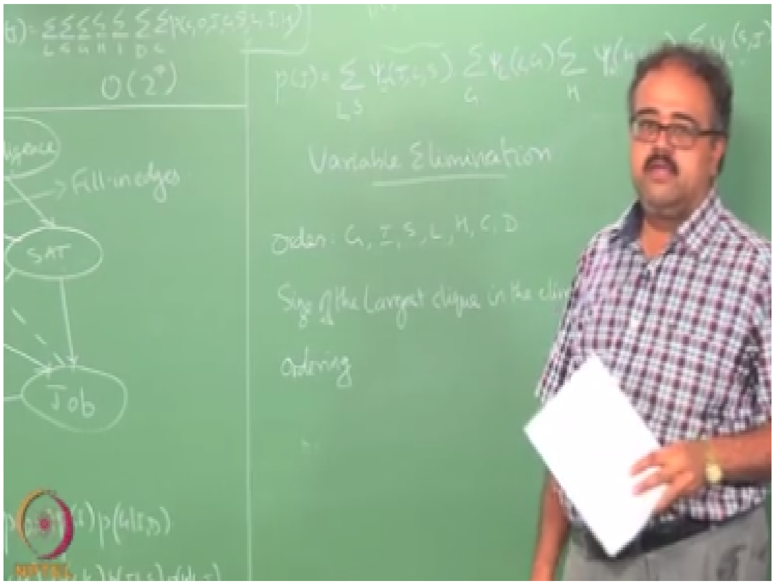
NPTEL ONLINE CERTIFICATION COURSE

Introduction to Machine Learning

Lecture-68
Variable Elimination

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

(Refer Slide Time: 00:14)



Right, so we are looking at graphical models we looked at both directed and undirected models right. And I said the thing of interest was, so there are two things that are of interest. The first thing is given a model right, how do you do inference using the model right. So what is the inference question, inference question is trying to answer queries on marginal’s right. So I give you a very complex joint probability distribution, I want to know what is a probability that there is an earthquake, yeah that is not a very complex system, but anyway.

So what is the probability that there is an earthquake right, I can also ask for conditional marginal’s given that John called what is the probability that is nice case. So these are things we looked at right, so it turns out that this itself is a hard problem and for large graphs you will have to come up with ways of approximating given this right. So I will kind of motivate why there is a hard problem in a minute.

And the second problem that we are interested in this what is the first problem sorry. So what could be the second problem? No find the model then, how do you derive the model right, but you have closed right, so how do you find the model right, some the raw data may I will give you training data, I will give you a lot of data how do you find the model right. So the Bayesian network structure learning itself is a hard thing.

So the simple problem is even in the structure learning they split it into two things right. So I should probably put this down, the first problem is inference right. So here there are two components with, so given the graph right. So I will give you the graph find the parameters and in the directed case that would be finding the conditional probability distribution. So once I give you the graph I know exactly what are the conditional probability distributions I need, I can just go to the data count and find it out right.

And in the undirected would be find the potential, so given the graph find the potential right. As soon as I give you the graph you know what are the potentials that you need to estimate right. So you will have all these H potentials, you will have node potentials, and you have click potentials so you will know what are the potential that you are estimating right ,and you just go and estimate the potentials right.

So this is essentially the learning problem given. And the second problem would be find a graph right. So one of the things you should look at finding in trying to find a graph essentially you would need to find that graph structure right, that supports the in conditional independence that is present in the data, it is directed graphs or undirected graphs whatever graph structure you are learning.

So you have to infer what are the conditional independence at expression present in the data, and you have to find a graph that will support that. So essentially you will have to, there are many ways of doing it people start off with a completely connected graph, and then they start knocking off edges right. And then you can do some kind of cost complexity pruning like you do in decision trees right.

So you could have a much more complex graph that then you can try to prune things down so that you can do a tradeoff between the number of edges you have. So the variety of algorithms that people are proposed for graph structure learning. So this part is easy right given graph fine parameters is easy, how will you do that just like conditional probability distribution estimates right.

So you can very easily do that for directed graphs just counting look at the data, see how many times Mary called when there was an earthquake right, or when the alarm rang how many times Mary called and then you can essentially fill in this conditional probability right. So those things we can do a straight forward right. But learning the graph structure is little bit involving to get into that, because it is a lot of, you know lot of structure that we have to build in before you can.

So I am going to now go back to inference, so inference is the interesting part right. So let me start off with an example right. So I am taking this example from, so for a long time we did not have really good book on graphical models, and then Koller and Friedman wrote this over complete book on graphical models I mean it is like it has everything that you would need to know about graphical models and more right.

So it is like this huge stone right, but it is a fantastic book, it really is a good place to start right. So why I am saying it is a good place to start this, this is still a very active area of research right, probabilistic graphical models and every year newer techniques, newer breakthroughs keep coming up. So it is like, it is not like you can write a book and say okay everything you need about graphical models is captured in the book right.

So because it is still evolving field right I am going to draw a really large graphic here okay, it is not a small thing which Daphne Koller came up with to capture some fraction of her interaction with students right. So depending on the difficulty level and the intelligence of the student okay, the student will get some grade intercourse right. And the difficulty level of the course depends on how coherent the teacher is right.

So the coherence influences the difficulty level okay, and then the difficulty level intelligence influence the grade right. And so depending on whether the student got a good grade or not in the course, the teacher might give him or her a letter right letter of recommendation if the grade is bad, then the probability of getting a letter is very small, as the grade is good the probability of getting letter is very high right, even there that happens.

And whether they get a letter of recommendation from the teacher or not right, it influences whether you get a job, and whether you get a job and whatever grade you did influences whether you are happy or not right, this is like, so sometimes you might be very happy for having done very well in the course even if you even though you do not find a job right. So maybe that is also possible right.

I am just giving you the structure here, because this is sufficient for us to talk about some of the difficulties in the inference process right. When you are actually solving problems in this you would need the probabilities, but we are not going there right. So they just give you the structure. Suppose I am interested in answering a, let me write this out now, probability C, D, I, G, S, N okay.

So you people see what I have written if you cannot, you can write it from the graph directly so you do not really need me to write this out right. So right, so this is a probability of coherence times, the probability of difficulty given coherence times, the probability of intelligence then the probability of great given intelligence and difficulty, so on so forth, I have just written out the joint distribution you can just look at the graph and you can write out that yourself easily right.

Now I am going to ask the question I need more space, so I am going to do this here what is the probability that a student in this universe will get a job, in this universe I mean universe is captured by this way. So what is the probability that person will get a job right, so what will you, how will you go about doing this essentially this will be okay, right. So if you think about it is essentially order of 2^7 computation if everything is Boolean right.

So it looks all right I mean so it means running this over the entire table running the summation over the entire table is not correct. So the whole idea of us doing inference was doing this factorization was to make this computation simpler right. If I did not have the factorization right I essentially would have had to do this computation. So yeah, so this is some set of running over this very large table right.

So now what we are going to try and do is try to make the summation simpler by pushing in some of the seven sums right, pushing it in to the maximum extent possible so that what I sum over okay, as smaller table as possible right. Right now and all my seven sums are running over the entire joint distribution right, I want to rearrange this in such a fashion that each sum runs over as smaller setup as possible right.

So how will I do that, I will just come to that in a minute, I have to make that yeah, yeah okay. So I will move from the conditional distribution to the potential formulation right, but you know what this means this is essentially the conditional distribution here so you can actually think of that having been represented as an undirected graph also, we can use the same technique that I am doing here even with undirected graphs right.

So that is the point I am to make that point I just switched over from the this notation to this notation. So in this particular case these factors happen to be conditional distributions, but they could be factors that you get from here. So in which case you probably have to have some kind of normalization going here right. So if you are going to use this as an undirected model then you have to have some normalization to take care of, so is it correct, so what I have written is correct right.

So the notation I am doing here is essentially this it takes in JLS as arguments again returns a distribution over J. So that is what the J here stands for, so it takes JLS as arguments and returns the distribution over J right, or some function over J, this takes L&G has arguments okay and return something over L. So that is what this is right, so this is essentially probability of a given L, S or something like that the equivalent to that in my potential notation.

So that is the thing I am marking here okay, is it clear? So now you can think about it, so the C runs once over only those two tables they are small tables, so C has just one in two entries in it right ΨC will have only two entries in it right, whether the teacher is coherent or the teacher is not coherent right. And Ψ D, C will have how many interest in it, four entries in it right, how many independent entries in it yeah, two okay lovely two independent entries, not three right.

Because given the course is not, given the teacher is not coherent what is the probability it is difficult. So automatically 1–z gives me the probability it is not difficult right, even the teacher is not coherent what is the probability is difficult and 1-z gives me the probability that it is not difficult right. So I only have two parameter, so you can see that I am reducing the parameters tremendously.

So hear what would I have had, I would have had 2^{8-1} parameters right, the full joint distribution right, if I specify 2^{8-1} parameters and 1 minus the sum of that will give me the last one. But here look I have tremendously cut down, so this has one parameter, this has only two parameters right. So likewise this is going to have four parameters that is for every combination of ID you are going to have one possible outcome for the other 1-z right.

So for every combination of ID you need to have one parameter so you will have four parameters, so likewise here you will have one parameter again, here will have two parameters so like that so you are reducing, if you take the product is much, much smaller than the 2^{8-1} that we had right. So that is the power of doing the factorization, so the number of parameters you need for specifying the joint distribution comes down significantly.

And you can do this as well right, you can start pushing the sums in, so that this sum runs over only if small number of elements right. Likewise this sum runs over a small number of elements and so on so forth, and then I can complete the entire joint distribution right. So this kind of an approach right where you push the sums in is known as variable elimination right. So for small graphical models okay, this is a good way to do inference right.

It is not an approximate way of doing inference, it is an exact way of doing inference right it gives you the same result as you would have gotten if you had summed over the entire distribution okay. So it is called variable elimination, and so the advantage is like I said they have amount of computation that you are doing you will be minimizing right. So how much computation would you be doing, what will be the maximum, what will be the largest table that you are summing over, exactly.

So it depends on how much you are able to compress the things and how much are actually able to eliminate the variable. So the more variables are supposing variable eliminate the faster will be your computation right. So think about what you are doing here, the first step is marginalizing over C right.

So I am going to say that you marginalize over C right, and you end up with a factor over D right I am going to call it some τ and D right so what will τ and d look like, that is τ and D right next what do I do what I am I marginalizing over marginalizing over D right, so this guy this whole thing I am marginalizing over right I am going to call that factor τ 2and what will be a function of G and I and that will be equal to right.

So I keep doing this next time eliminating I, so what we will end up with the factor over G & S ray yeah then what we will end up with I will have this guy as it is right am eliminating H right so there is no H here so τ 3 gs will continue propagating beyond this point right but I will also introduce a new factor called tau for which will have h you can see that right at this point I just trying to trying for you to get an appreciation of what the computation is happening right at this point you will have τ 3 you will also have τ 4 right.

So when you compute it till τ 3 you have eliminated you eliminate a τ 1 you eliminated τ2 because you have rolled up everything into τ 3 but we do τ 4you are not able to eliminate that right so τ 3 is still carries on to the next level. Now we eliminate G, so what you get at τ5 so eliminate G so we will have J left right we will have s left and L will now get added here finally you will get depending on what order you do this thing in here I will first sum over is I get this then sum over l I get my fear okay so this is essentially the or how you will be doing the elimination.

So as and when you are doing the elimination and you are creating this new factors so what we should be thinking office it is as if you are adding a new potential it is as if you are changing the dell changing the graph right so when it did this right well I did not let us certainly really add anything new d is already there right what about this. So now I create an edge between G and I and or a big day G&I already existed right but what about this now I create a potential between G and this right.

So when I come to this point so it is like I am adding a another connection between G and S right so likewise anything else is happening anything else J & L is already there JSL, JSL right I need to have a click for me to have a potential JSL I need to have a click, so I am essentially like matting H between S and L right. So you can think of the way we are doing this is essentially like we are making this larger some of these potentials are making larger and larger right.

So in this case it turns out that luckily none of the intermediate steps that we are creating makes a large table right none is nothing is larger than any of the existing tables right so we could choose a bad elimination ordering I can choose a different order, so here the order we chose was CD

IHG SL okay, so that is order in which we eliminated the variables started off with the right-hander at CD IHGL okay, suppose I did this they start off eliminating G right.

So I can sum over G and I have to put in all the factors that have G in it right, so what are the factors at RG in it I sum over G I will do right from this side right so I will have φ l l, G φ h when I summed over G over all these factors. So now I am going to create my new τ 1 right so I will call it τ 1' so τ 1' will be a function of everything in that that is not eliminated right. So G is been eliminated so what it will be so l h j i D ouch. Now I created a 5 a table there by choosing to eliminate G first right I have created a 5 a table so that is a large table and now I am going to sum over this.

So now will be summing over a table which has 2^5 entries right so that is a bad thing right so next one what I have eliminated next try to eliminate I next so what I will do that so I will have τ 'of l h j ID said any other factor that has I φ I this doing right and what will this do it eliminate I right but it will add s to the factor so my τ 2 prime will be a function of L h j d s now I have another five factor table is in fact this is the worst possible elimination order okay to give you the give you the really bad picture right that is the worst possible elimination auditory then I eliminate s.

So what do I do in that case well I add JLS also to the mix right a large JLS also to the mixer alienate is that but J and L are already there in the factor so in fact this will come down so my τ 3 will have only L h j d because I eliminated s right then I will eliminate l right nothing a new gets added that the only thing that is left out to see right yeah so by the time I come to yeah so everything else will get eliminated.

So finally I will be left with a factor that contains only D&J and then finally eliminate d so what will happen when I eliminate yes we are done to yes okay what happens on eliminate l I will end up with a factor that has HJDS then what happens if we eliminate H I will end up with a factor that has JD, JD what L right no L is already gone L&H and just end up with the factor that has JD I have a factor that as JD I will also have the this is the C's the last two factors will still be there the φ C and φ DC that those two factors will still be there right everything else will get eliminated and then what I eliminate C that means those everything all those factors will get eliminated I will be left with a factor that has only doing.

And finally eliminate T okay but what I had done along the way is that I have created a big click herewith five variables in is right, so if you notice as we went along so even though this looks like a click of four variables okay it was never created as a clicker for variable said that at best I only did a clique of three variable just two different cliques of three variables it looks like a clique of four variables but we never generated the click right but in this case we actually generate a clique of five variables so it can become very large right.

So it turns out that the complexity can be related to this the complexity of running inference on this graph can be related to the size of the largest click you generate along the way right, so these kinds of wedges that we generate like this right are called fill-in edges yeah this one G&S this one yeah also eliminating I right. So when you have this thing I mean silence well I did not want to erase everything but when you add this filly niche that essentially when you remove that so this not really a click.

So this is not really a click this is only a this the edge is the maximal click in this case, so your question is I do not have a potential that says I GI and S right, so when we did the original ordering we never did a GIS potential that is because what you pointed out. So I was eliminated and therefore we only have GI was already existing right we have a potential corresponding to GIS in the beginning.

Now why should we have one corresponding to GIS no we do not need one corresponding GIS so we do not need one corresponding to GIS, so you do not need one at all in the inference also when this fill in it is added that those things are not there right. So we only have to worry about those filling edges which actually leave you with a click is what I am writing size of the largest click in the Elimination ordering is called the induced width of that ordering.

IIT Madras Production

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

www.nptel.ac.in