

Lecture16.3

Can we represent the joint distribution more compactly

So, now the quest is can we take this joint distribution and represent it more compactly Okay? That's what I know I need all these values it's just that I am interested and represent these get having all these values if I want I should be able to immediately get them but I don't need to really store to raise to n minus 1 parameters. Right?

Refer slide time :(0:34)

I	S	$P(I, S)$
0	0	0.665
0	1	0.035
1	0	0.06
1	1	0.24

- This distribution has $(2^2 - 1 = 3)$ parameters.
- Alternatively, the table has 4 rows but the last row is deterministic given the first 3 rows (or parameters)

- Consider the case of two random variables, Intelligence (I) and SAT Scores (S)
- Assume that both are binary and take values from High(1), Low(0)
- Here is one way of specifying the joint distribution
- Of course, there are many such joint distributions possible



Okay? So, now let's consider the case of two random variables which is intelligence and such scores. Okay? Oh you have read the textbook. Okay? Assume that both are binary and they can take on values high and low. Now here is one they are specifying the Joint Distribution, you've specified these four values and I'm not going to say to raise to n minus 1 from now I'll just say 2 raised to n so, 2 raised to n is 4 in this case and of course there are many such joint distributions possible what do you mean by this you will read this in textbooks what do you mean by this there many said Joan distributions possibly it's two variables. Right? I mean how many what do you mean by many distributions? What is the Joint Distribution all these values. Right? So, there are many such values possible it just, just need to ensure that it sums to 1 and everything is greater than 0. So, when you say there is a family of distributions which can satisfy all this it just means, that there are many possible assignments which can actually give you this joint distribution. Right? How many parameters does it have three parameters. Right? Because, if I know three parameters the last one is just summation is 1 I know so the last one just has to be such that the sum sense to be 1. So, this distribution has three parameters, now I'm interested in reducing this three because in general this is going to be 2 raised to n minus 1 so, I want to get something which is less than 2 raised to n minus 1. Okay? Of course on this small example I'm not going to be able to do much because 3 is anyways manageable at how much do I need to really reduce it. Right? But we'll start with this small example and see even if you add one more variable how things change drastically. Okay? Now in this case notice that there's a natural ordering between these two random variables, if the random variables are intelligence and SATs code

Refer slide time :(02:16)

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

	$s = 0$	$s = 1$
$P(S I = 0)$	0.95	0.05
$P(S I = 1)$	0.2	0.8

no.of parameters=2

- What! So from 3 parameters we have gone to 6 parameters?
- Well, not really! (remember sum for each row in the above table has to be 1)

- Note that there is a natural ordering in these two random variables
- The SAT Score (S) presumably depends upon the Intelligence (I). An alternate and even more natural way to represent the same distribution is

$$P(I, S) = P(I) \times P(S|I)$$

- Instead of specifying the 4 entries in $P(I, S)$, we can specify 2 entries for $P(I)$ and 4 entries for $P(S|I)$



4:36 / 15:39

So, you could think that intelligence is what the SATs code depends; on I mean you could think that SAT score if I were to ask you a relation between them you would say that SAT school probably depends on intelligence. Right? And if I assume that way if I have to write the chain rule, there are two possibilities at P of s into P of I a given s or the one which I have already written. So, of these two possibilities you would probably choose the first one is that correct because, saying that what's the priority of SATs score given that he is intelligent or he or she is intelligent makes more natural sense it's not that you cannot ask the other question I could ask you the other question also. Right? What's the probability of someone? being intelligent given that they're SATs score is high or low that question also makes sense and we are going to see that question later on but I am just saying that it's more natural to think of it in the former way where you have the intelligence and then the SAT score given the integers. Okay? So, now from one conditional distribution one joint distribution I have splitted into two distributions. Okay? And instead of for specifying four entries, now how many entries I am specifying six empties so I've done something very smart and I had a problem with four I said six is a better number let's do six is that right wrong? So, actually by factorizing and factorizing is something good typically by factorizing I actually, increase the number of entries which means I've increased the number of parameters is that true no yes no maybe probably has a number of parameters increased no why summation one. Right? So, for the first row I actually have only one parameter once I know that probability of heads or intelligent being low or high the other one is deterministic because, I know that the sum has to be one so, how many parameters do this factorize distribution have three which is the same as the number of parameters I had in the original distribution. So, I did all this roundabout thing and I ended up with the same number of parameters, but whatever what am I achieving the process I have given you a more natural way of representing this distribution and now this idea will try to build upon this and see what happens when we have more random variables. Right? So, we still have the same number of parameters but we have seen this idea that if I could represent the Joint Distribution is a factor of some conditioners and marginal's then probably one I have a more natural way of representing the distribution Right?

Refer slide time :(04:36)

	i=0	i=1
$P(I)$	0.7	0.3

no.of parameters=1

	s=0	s=1
$P(S I = 0)$	0.95	0.05
$P(S I = 1)$	0.2	0.8

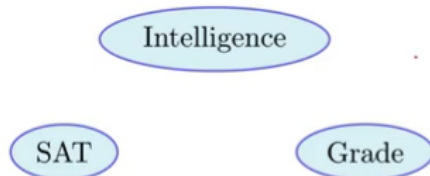
no.of parameters=2

- What have we achieved so far?
- We were not able to reduce the number of parameters
- But, we have a more natural way of representing the distribution
- This is known as conditional parameterization



So, this is what we have achieved have a more natural way, and this is known as conditional parameterizations. So, instead of specifying the Joint Distribution the parameters of the model are now conditional distributions they need to specify these conditional distributions and I am done so I don't need to give you that one table I need to give you two tables. Right? Fine Okay?

Refer slide time :(04:57)



- Now consider a third random variable Grade (G)
- Notice that none of these 3 variables are independent of each other
- Grade and SAT Score are clearly correlated with Intelligence
- Grade and SAT Score are also correlated because we would expect


$$P(G = 1|S = 1) > P(G = 1|S = 0)$$



6:24 / 15:39

Now consider the third random variable, which is we didn't have grade so far. Okay? Grade now what will happen? So, first let's notice and none of these three random variables are independent of each other is that a valid statement, for grade and intelligence it's valid I mean we would again roughly assume that grade depends on intelligence modulo copying another various other things but what about SAT score it also depends on intelligence but SAT score and grade your SAT score grade and your grading are not so, important deep learning course why would they be related are they independent of each other, yes no if the SAT score independent of the grade let's see on first page of chapter four from the textbook and not go to the second page. So, let's look at it this way Right? if I told you that the SAT score was high which of these two properties would be what do you expect to be higher? The grade to be higher or this grade to be low why is this happening you know the SAT score tells you about the intelligence which in turn tells you about the grade. Okay? So, on its own all these three variables if I look at any two pairs any pair of variables from these three variables they don't seem to be independent of each other is that fine. Okay? Fine

Refer slide time :(06:27)



- However, it is possible that the distribution satisfies a conditional independence
- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining G
- In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score
- Formally, we assume that $(S \perp G | I)$

Now however it is possible to the distribution satisfies a conditional independence what's the conditional independence that the charges face something is independent of something given something. I knew the most natural possibility can I say that the grade is independent of the SAT score given intelligence yes no maybe yes why? What does it mean actually? Why do I say? That grade is independent of a SAT score given intelligence. What am I trying to say actually? $P(G | I)$ does it make sense Right? So, if you are given that it's Intel the person is intelligent knowing his or her SAT score is not going to give you any more information wait you could determine the grade based on Sat score is that always Right? I want you to link it back to what I have been saying from day one. Right? That whatever modelling choice you make in a in the previous part of the course all our modelling choices were y is equal to f of X they always depend on our assumptions about how things work. Okay? Now here also you are doing certain assumption and I can give you an example where this assumption fails actually this may not be a correct assumption. Right? So, I'm asking you whether the grade is always independent the SAT score given the intelligence the thing of a simple case.

Right? The person is intelligent but perhaps is he or she is too lazy to write an exam in time where you just read though I know this I don't know to write the answer what's the point of writing all so, he or she may not actually be very good in time management. Right? So, the exam has to be finished in three hours but now that you know that person has done well on Sat score you know that the person has actually done well on time-bound exams that means the person is good at writing exams he can do time management also if you think of that particular view

Refer slide time :(08:27)

The slide contains a diagram and a list of bullet points. The diagram shows three ovals: 'Intelligence' at the top, 'SAT' on the left, and 'Grade' on the right. The bullet points are:

- We could argue that in many cases $S \not\perp G|I$
- For example, a student might be intelligent, but we also have to factor in his/her ability to write in time bound exams
- In which case S and G are not independent given I (because the SAT score tells us about the ability to write time bound exams)
- But, for this discussion, we will assume $S \perp G|I$

The video player interface at the bottom shows a progress bar at 9:58 / 15:39 and various control icons.

Then you could say that s is not independent of G given I because s tells you something about time bound exams intelligence tells you general about how that person can what is the capability of the person intellectual capability, of the person but together these two actually tell you whether the person can do well in a course where there are a lot of time bound exams does that make sense. So, again what is the assumption that I'm making so I could make any of these assumptions this is my modelling choice if I believe that for whatever reasons if I am a domain expert or maybe I've looked at some data or whatever I could make one of these choices I could either tell you that the SAT score is independent as a grade given the intelligence, this is what my modelling assumption is going to be or my modelling assumption is going to be that it is not independent. Right? But this decision is something that you take just as you take the decision of what is the f that you need to choose. Right? Whether you want to do is a complex neural network or a sink or SVM or a logistic regression or whatever that's the same thing. So, all these cases whenever you make these assumptions you are making whenever you are making these choices you're making some assumptions and it's up to you so, in this for this running example we are going to assume that the SATd score is independent of the grade given the intelligence. Right? For this discussion we are going to stick to this argument this is my modelling choice I may be wrong that means whatever model I make it's going to be giving me it is going to give me bad probabilities as compared to what's the true probabilities are going to

be but this is what my modelling choices in the absence of any other information this is what I am going to assume Right? Fine Okay?

Refer slide time :(09:59)

Question

- Now let's see the implication of this assumption
- Does it simplify things in any way?



Now let's see the implication of this assumption does it simplify things in any way for us.

Refer slide time :(10:10)

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no. of parameters=1

	$s=0$	$s=1$
$P(S I=0)$	0.95	0.05
$P(S I=1)$	0.2	0.8

no. of parameters=2

	$g=A$	$g=B$	$g=C$
$P(G I=0)$	0.2	0.34	0.46
$P(G I=1)$	0.74	0.17	0.09

no. of parameters=4

total no. of parameters=7



- How many parameters do we need to specify $P(I, G, S)$?

$$(2 \times 2 \times 3 - 1 = 11)$$

- What if we use conditional parameterization by following the chain rule?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

- We need the following distributions to fully specify the joint distribution

Now how many parameters do we need to specify IG NS remember that grade can take three well you had a question second oh for how many values can how many parameters we need for this Y 2 into 3 2 into 2 into 3 grade has 3 values. Right? So, intelligence and SAT score as high n low and grade as 3 values so, 2 into 2 into 3 minus 1 so you need 11 parameters for this. Okay? Now what if you use conditional parameterization for this? What would happen? So, what is the thing that you will start with chain rule? Okay we'll start with the chain rule so, we know that this Joint Distribution factorizes into this conditional and marginal. What about the first conditional Joint Distribution? What kind of a distribution is this conditional Joint Distribution Right? Okay? How does it factorize further? Does it that's what the chain rule does rate and you keep factorizing? So, you'll write it as this anything that you can simplify here in this form how many parameters do we need how many parameters you need for this term one for this term, two or three or four there are two variables. So, how many do we need in the conditional form, I was counting everything as to how many do you need in the conditional form eleven so, nothing is simplified what will simplify now? First term what happens because of the independence assumption now how many parameters do you need, one for and two what's the total number of parameters now seven so? Right? So, you see that because use of two things one is we are using a factorized form, second is the factorized form gets simplified because of dash conditional independencies. Right? Because, of these two things combined now you need one table for this which has only one parameter, you need another table for this which has two parameters because the sum of the rows is going to be 100 sorry and then you need another table for this guy where this there are going to be three parameters. Right? Is that fine everyone gets this Right? forget about the tables but at least from the equation it should be clear that you need fewer parameters now imagine this situation scale to a much larger scale where you have n very variables and you have many more conditional independencies in your graph. Right? So, every place where you had a variable of the form or term in the chain rule, of the form property of X_i given 1 to I minus 1 that means given all the other variables, this set which has I minus 1 values a lot of terms from there is going to get dropped out because, the variable X_i is actually independent of these variables given the remaining variables. Right? So, you see that writing the chain rule, gives you a factorized form the factorized form simplifies further because, of these conditional independencies and then you get

much fewer parameters than you originally had. Right? So, that's what the importance of conditional independence is, is and in whatever problem you are trying to model you would have to make these conditional independencies, known is going to give you these conditions now suppose if I go back to our original argument that we had already seen that, I was not independent of G or s and you also made an argument where G and s were also dependent, in this case how many parameters would you need? if I'd not assume the conditional independence of grade NS given high how many parameters will I need living because, this term is no longer going to get simplified. Right? So, you can write the chain rule you get, get the factorized form but the factorized form is not going to simplify, unless you have certain conditional independencies in the factorization does that make sense. Okay?

Refer slide time :(13:59)

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

	$s=0$	$s=1$
$P(S I = 0)$	0.95	0.05
$P(S I = 1)$	0.2	0.8


no.of parameters=2

	$g=A$	$g=B$	$g=C$
$P(G I=0)$	0.2	0.34	0.46
$P(G I=1)$	0.74	0.17	0.09

no.of parameters=4

total no.of parameters=7

- The alternate parameterization is more **natural** than that of the joint distribution
- The alternate parameterization is more **compact** than that of the joint distribution
- The alternate parameterization is more **modular**. (When we added G , we could just reuse the tables for $P(I)$ and $P(S|I)$)



So, now with this what do you have the alternate parameterization? So, you had this one representation for the Joint Distribution which is just one monolithic table a large table now instead of that you are giving me multiple tables so, that's a more natural way of representing of distribution why do I say it's a more natural way? Because whatever you put in a conditional is actually a natural ordering. Right? You know that SAT score depends on intelligence you know that grade depends on intelligence and perhaps not the other way around. Right? So, that's why it's more natural it's also more waters our quest for what kind of a representation were we interested in compact so, this is also more compact .Right? And this is also more modular why do I say this is more modular? when does modularity help while writing code or anything or if you want to add what if I add one more random variable to this setup would I have to change any of these three tables you just have to introduce new tables which depend on that random variable all that random variable depends on something. So, for in particular P of I you don't need to change you can just reuse it as it is P of s given I you don't need

to change in fact when you introduce the random variable G I just use the first two tables from my previous example where G was not there. Right? So, that's why modularity is also important. Okay?