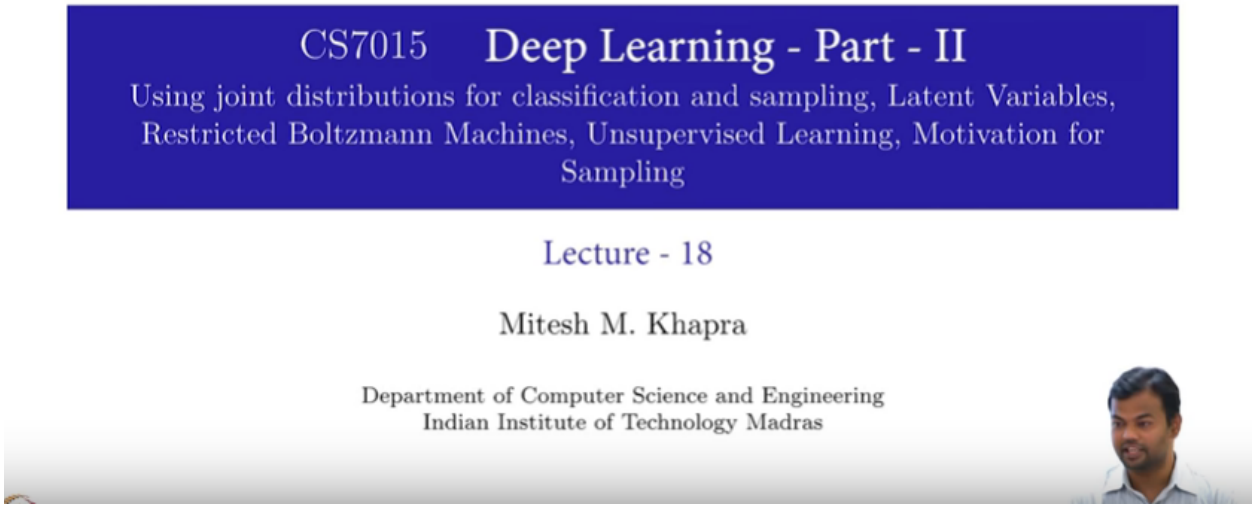
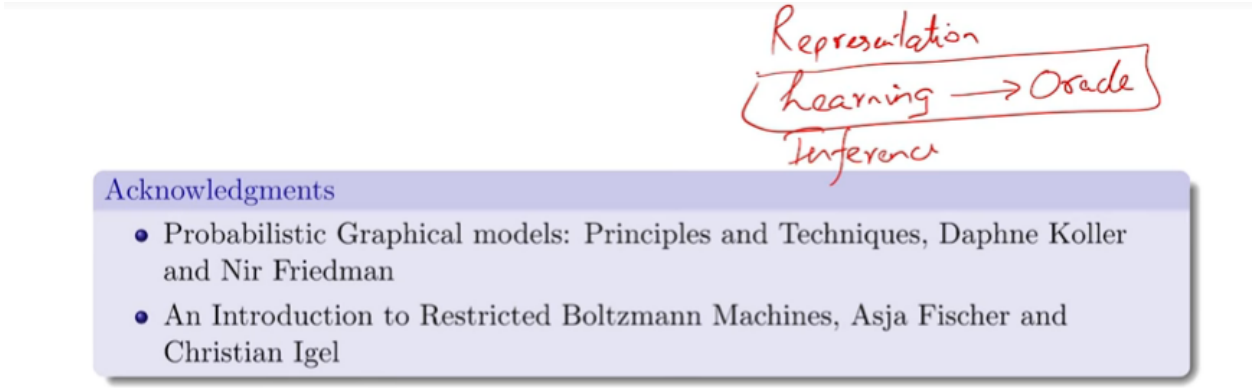
Lecture – 18
Joint Distributions

Refer Slide Time :(0: 13)



And in this lecture we'll talk about, once we have a joint distribution, what do we do with it? Then we'll introduce the concept of latent variables and then, eventually go to restricted Boltzmann machines and talk about, learning in the case of restricted bolting Boltzmann machines, which will eventually take us to sampling. Okay? So, this is what is, I don't think this is all like a homogeneous lecture, but it's put into this lecture, they'll probably split it into two different lectures later on. But, in respect of that I guess the overall story should be clear to. Okay?

Refer Slide Time :(0: 43)



So, let's just again, do a quick not a recap, but a quick contextualization of what we have done, so far. Right? So, what we have done, so far is? We are focused on representations of joint distributions. .Okay?? And after representing, we have made a case, that the number of parameters reduces, which should help learning. We have not actually done the learning part. .Okay? We have always assumed that, the learning

is either done or someone has given it to us. Right? So, representation is what we have spoken about in detail, learning is something that we have ignored and now, the question which I'm asking you is, what do you do after learning is? How to use this joint distribution? Right? So, that's a question of inference. And once I have this parameter is learned, once I have the factorization for the joint distribution, how do I make use of this joint distribution? Right? So, if you do a traditional course on graphical models, if you read the book or something, you will have three parts, which is representation? Which is what we have largely focused on, then learning, which will eventually get to? And the third one is inference. Right? so far we have said that, learning is essential Oracle, all these factors and the Clique potentials, were actually given to us by someone and you are just using that, as it is, but eventually we'll have to learn this. And this learning, will not do it in the context of general graphical models, we'll just do this in the context of RBM. Right? So, that's what we will eventually get to. And inference is something which is general, so I'll just tell you, once you have a joint distribution, what are the different things that you could possibly do with the joint distribution?

Refer Slide Time :(2:26)

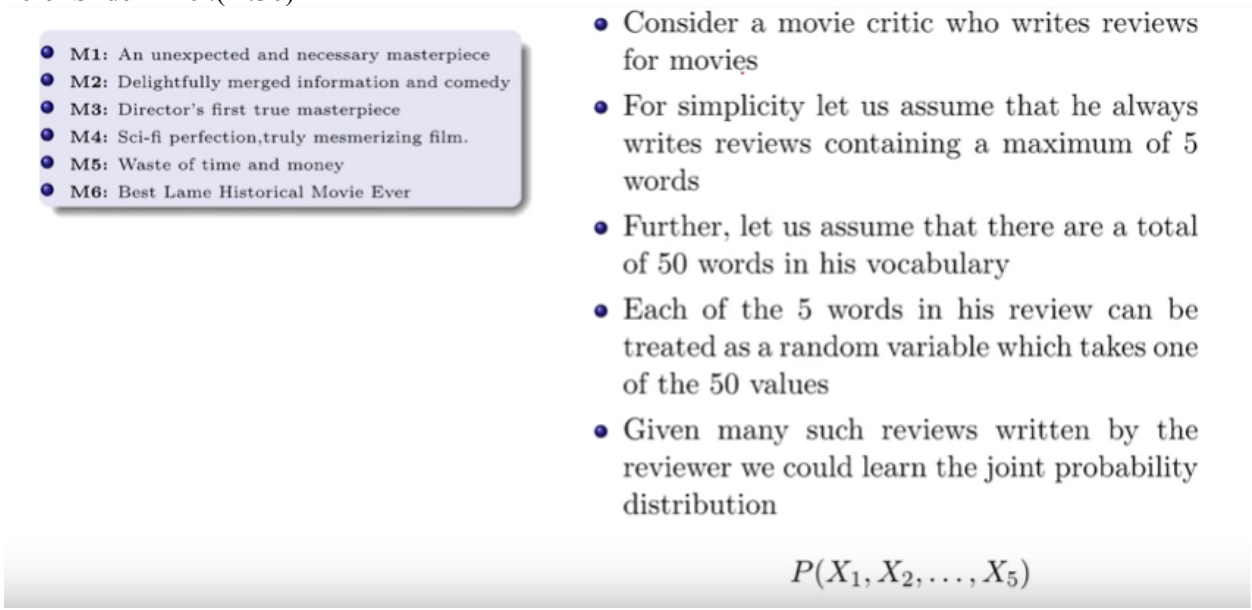
Module 19.1: Using joint distributions for classification and sampling

So, let's focus on that, in the first module of this lecture it. Using joint distributions for classification and sampling.
Refer Slide Time :(2:29)

Now that we have some understanding of joint probability distributions and efficient ways of representing them, let us see some more practical examples where we can use these joint distributions

So let us see, what we mean by that.

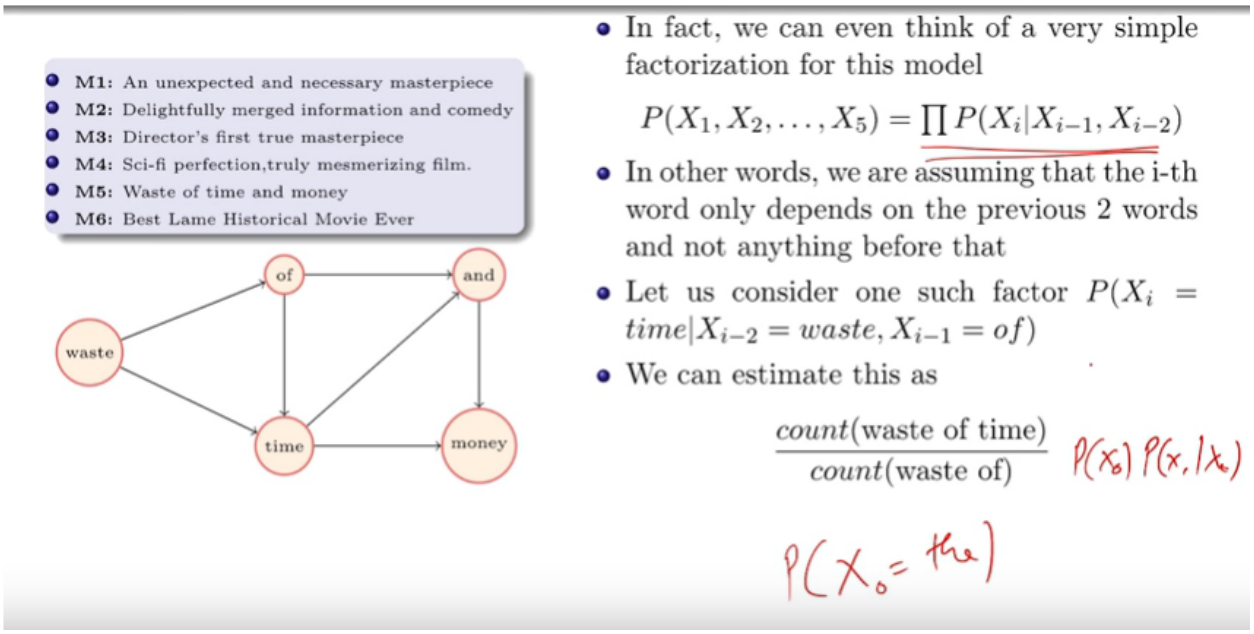
Refer Slide Time :(2:30)



So, consider a movie critic, who writes reviews for movies? Ok. And we'll make a simplistic assumption that, all the reviews that he or she writes are only these five word reviews, he or she just writes five word reviews and here are some samples given to you. So, this is your training data, which is given to you. these are the reviews written by a particular reviewer. Now, further for again for the sake of simplicity, let us assume that, this reviewer just has a vocabulary of 50 words. So, whatever reviews he or she writes, uses something from a set of 50 words that, he knows and nothing beyond that? right? so vocabulary is restricted to size 50 words. Okay? so, now how do you treat this, as how do you come up, how do you define a joint distribution here? first of all tell me, what are your random variables? so, this is where you need to put whatever you have learned to practice, right. so, I we have talked about X ,Y, Z, all this in abstraction, I'm giving you a real-world problem, tell me how are you going to use your knowledge? what's the graphical models the secondary question? what are the random variables that you can consider here? Think about the process of generating a review? right. that, the first word is a dash, the second word is a dash, the third what is a dash, where dash equal to random variable. right? so, how many random variables do you have? five. and what are you trying to learn, the Joint Distribution of P of X 1 gamma, X 2gamma, up to X 5. what does that actually mean, how many values can X 1 take? 50. how many values can X 2 take? and so on. so, what's the size of this Joint Distribution? 50 raised to or raised to 50, 50 raised to 5, just as 2 raised to5, now this is 50 raised to 5. right? so, the full distribution would have 50raised to 5 values .Okay? even for this toy example, you can see that it's kind of interactive. Okay? so we know what other random variables that we have and we are interested in learning this joint distribution X 1, X 2, X 3, X 4, X 5. Again the secondary questions are now or the second set of questions now are what is the graphical model that you are going to use? How are you going to simplify this Joint Distribution, into some factors? Right? And once you have this, once you have the factors, once you have the Joint Distribution, what are you going to do with this Joint Distribution? Can you tell me, what are

you going to do with this Joint Distribution? Go back to the title of the module, I said classification and sampling, whatever that means, how many if you are able to think about this, see you instantly of course I'll tell you things, but that's not the point. Right? I mean, you have to be able to think, on your own. others will just remain in abstraction say that, we have n random variables, each of which can take to raise to n values and so on. I do you have to apply, these to real-world problems and this is, as real as it gets. right? I mean, we all are interested in knowing, what people post on Twitter and this is exactly, what people post on Twitter. right? and I'm going to want to learn that this next guy, who is a political leader, will not be named, I want to automatically generate, tweets from that political leader. Right? How do I do that? You all know, which political leader we're talking about, I will not name him or her, mostly him. Okay?

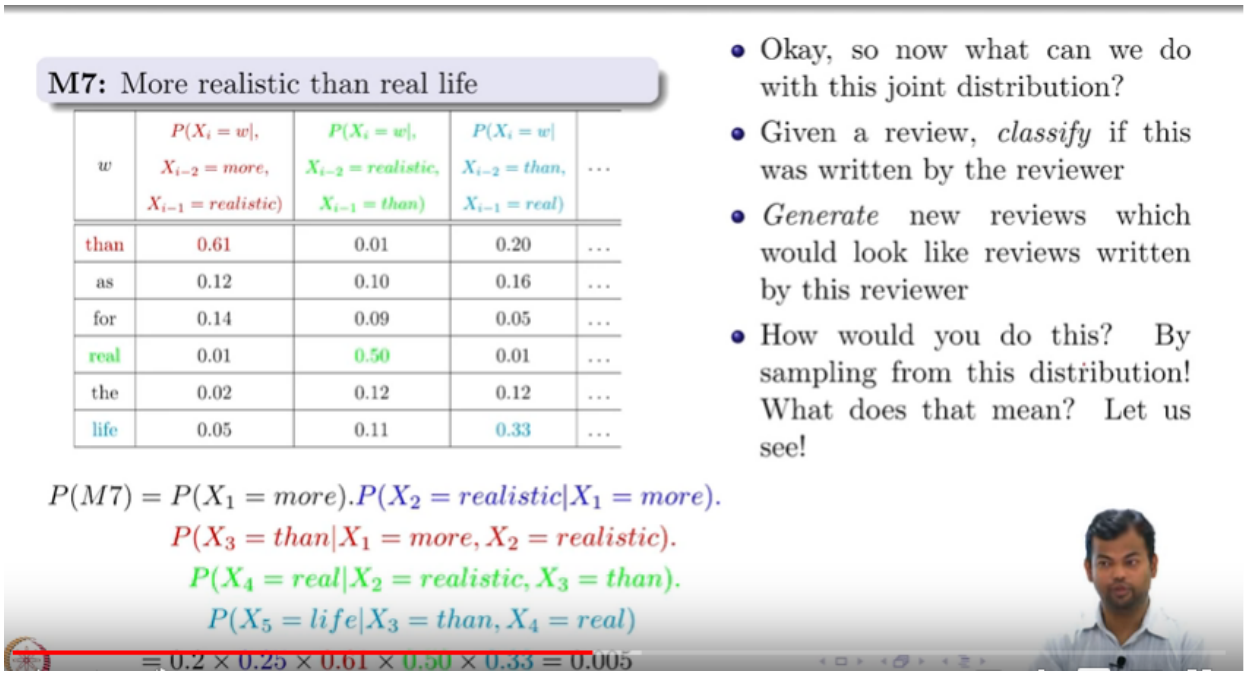
Refer Slide Time :(5:53)



So, let's first talk about the graphical model. Right? So, we are going to talk about a very simplistic graphical model, where we say that, each word depends only on the previous two words. What is this problem that I am looking at here? Have you ever looked at this problem before? Have you ever looked at this problem before in lecture 10 or 11, lecture 10, was not our intense? So there are lot of words here, what problem is this? Dash modeling, language modeling. Right? So, given the previous K words, we want to predict the next word. Right? Now, to make these connections others, what's the point it's not at lecture 19 oh yeah, like lecture 10, oh! yeah! You have to keep making these connections. Right? so .Okay? so this is the problem of language modeling. And the moment I write, the Joint Distribution as this factorization, what is the graphical model that I am considering? Can you draw the graphical model? How many nodes does the graphical model have? Five. How many edges does it have? I don't know, actually I'm trying to calculate, but you get the idea basically. Each node will have two parents. Right? How many parents would each node have? Two parents. Except for the first, two words, because you don't have, X 0 and X minus 1. Right? So, that's always the boundary case is, always the tricky one. So this is what it is right? That this is the zeroth word or the first word, then the second word depends on the first word. The

third word depends on both the first and second words. The fourth word again depends on the second and third word. And the fifth word depends on the third and fourth word. That's the graphical models, which have essentially captured. Right? So, even though, we have formally introduced graphical models only now, we have been making this assumption and this factorization, even when we are talking about language modeling. Right? So this is the factorization. And this is a natural factorization, because even if you think about, your own texting or typing. Right? Once you know, the last two words, you pretty much can predict what the next word is going to be. Or maybe if you know, the last three words, you can largely predict what the next two words can we do? What the next word is going to be? Right. So, this is not a very unrealistic graphical model that, you have chosen. So what are the factors of this graphical model? What are the factors of this Joint Distribution? How many factors does it have? Five, right. Okay? Now, how do you estimate these factors? So, I said this learning part, will do only in the context of our beams, but let's just do it for this toy example, because it's very simple to do it here. Let's consider, one such factor .Okay? I am one assignment to that factor, so one of the factors here is that X I, given X I minus 1 and XI minus 2. Okay? Now, consider one assignment to this factor, which is X I is equal to time. And the previous two words were waste off, which is not what this is. But, how would you estimate this factor? I have given you the reviews from all the past reviews for this reviewer, how do you compute this problem. it is going to be a fraction of two counts, I mean; it's going to be a one count divided by another count, what are those two counts? The numerator is numerators waste of money. Okay? And denominator is waste of .Okay? Right. so, this is how you estimate it, so now you know that, you can estimate all the factors in this joint distribution given a corpus, I've given you one example, but you can do it for all of them, in particular how would you estimate this, per your priority of X zero is, equal to the, remember this is one of the factors in the joint distribution. Right? So, this is I actually, this is not correct. Right? We should write it as, P of X 0, P of X 1given X 0 and then the product is, it clear, I mean, where you always have this first root, then the next one and then, everything else has 2 parents. Right? The first zeroth word does not have a parent, the first word only has one parent and after that everyone has, two parents is that fine. Okay? Fine. So, how will you estimate P of X 0 equal to the, will count the number of times, you count the number of times the what? Though appears as the, first word in the review. Okay? Is that fine so you can compute all these factors; it's very easy to do in this simple case. Okay? Okay? Fine.

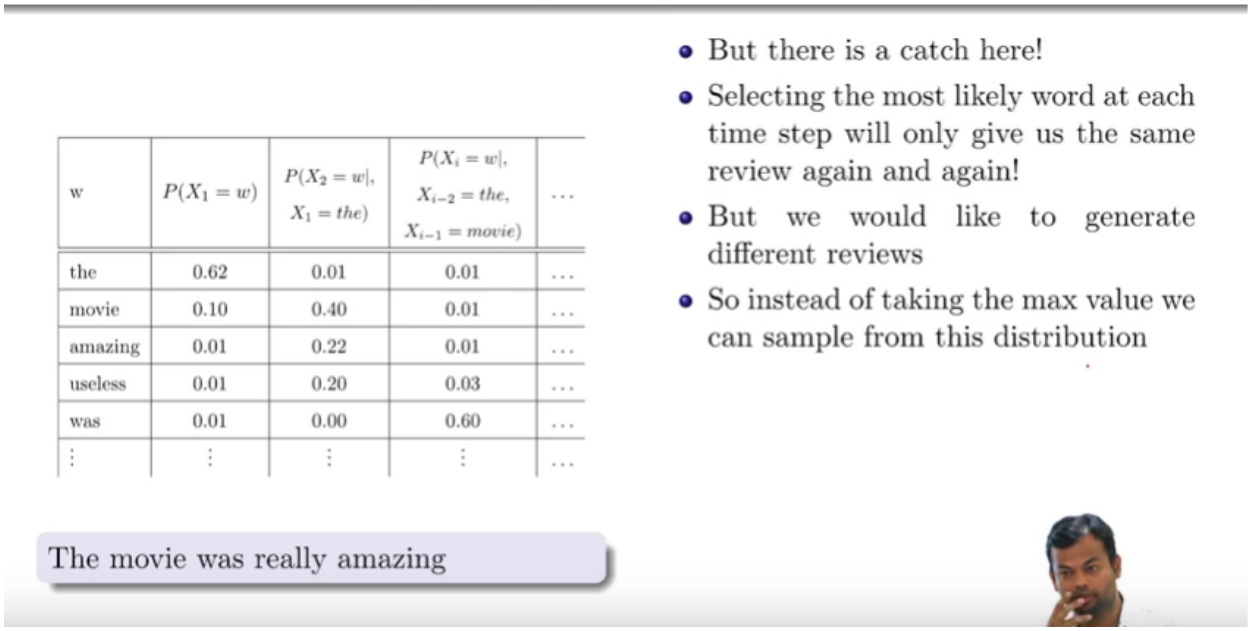
Refer Slide Time :(10:32)



So now, we have a way of representing the distribution. and we also have a way of computing this table, which every column in this table is, actually one of the factors that, we are interested, in is that fine. now, given a review, classify if this review was written by this reviewer or not, how will you do that. and this is something that I am interested in, someone tells you that, this is some statement, which someone made or this book is written by a particular author. and you want to know, whether this is true or this article was written by someone. and there, could be various other applications, of this I mean, how would you decide given this table and given a review, how will you decide, whether this review was written by this author. and all the training it had given to you was, for one author. what will you compute? so what will you compute? Joint Distribution of P of X 1 equal to, P of X2 equal to, X 3 equal to, please, have you had your breakfast. Okay? So, this is exactly what you will compute. and you know that, this Joint Distribution factorizes as, these conditional distributions and for each of these factors, you know how to compute this table, you have already computed this table, you just need to substitute, the values from that table. and the answer which you get is point zero, zero, zero five. of course it requires some domain knowledge to know, where the point zero, zero five is good enough or bad. Right? Because, it's a, it's a product of many factors. Right. So, even point zero, zero five is not bad actually, it's going to be a good value. How when you forget that? All these are decimals right, zero to one. So you're multiplying five or six values, all of which are between zero to one, so point zero, zero five is actually not bad, so perhaps this is written by that and that doesn't matter, because you'll have, what you could do is? you could compute this probability for all the reviews, that you have in your training data and then based on that decide a threshold weight, so that's not the hard part but, the main thing is that given this joint restitution you can compute, this probability and then decide whether this was written by that, that's fine, what's the

other more interesting thing that you can do? Generate, new reviews, which look very similar to reviews written by the sorter. How would you do that? That's a good but, incorrect answer and very appropriate because I'm also, going to give the same answer and then corrected of this, how are we going to generate new reviews, someone said sampling from this distribution, ignore the letter C part, mean how will you do it? How do you sample, what's the process that you use? This imagine that you are writing down the reviews, you know that you have to write five words, how will you do it? What a sample that mean, how many if you understand how you will do this?

Refer Slide Time :(13: 24)



So, now how does the reviewer start his reviews? What is the first word that the reviewer chooses? How will you get the first word? For all the first words, you have this table P of X 1 equal to W. Right? So, you know what's the priority of choosing of for all the given 50 words in your vocabulary, you have assumed that your vocabulary is only 50 words, for each of these 50 words, you know what's the probability of, X 1 equal to the, what's the probability of X 1 equal to a movie and so on. So, just pick the one word which is the most likely, word that the review is going to use as the, first word. So, what's that word? Maybe The .Okay? he'll start writing them, now how will you choose the second word, given X 1 is equal to the, find out X 2 equal to W, for all the 50 words and take the highest one. Right? This results in. the movie .Okay? and then the third word, X 3 equal to question mark, given X 2 equal to movie and x 1 equal to and again, you get some maximum word and you do that and that way you will continue. So, I've done all this complex, stuff and you have come up with one review, for the reviewer. But, your job of course is to keep posting more and more reviews from the same reviewer .Okay? So, let's generate the next review. How will you do that? Let's start the process again, select the max, the movie, was really amazing, you keep ending up with the same review again and again, no one is going to chat with your Twitter bot if you do this. Right? So, what will you do now? So, what you did was not sampling, this was just picking up the most likely outcome. So, what does sampling do actually the catch here is that, as I said if you keep doing this you will generate the same review again, again so, instead of selecting the Max, what we are going to do is we are going to sample from this distribution, what the sampling mean. Okay?

Refer Slide Time :(15: 21)

w	$P(X_1 = w)$	$P(X_2 = w \mid X_1 = the)$	$P(X_i = w \mid X_{i-2} = the, X_{i-1} = movie)$	...
the	0.62	0.01	0.01	...
movie	0.10	0.40	0.01	...
amazing	0.01	0.22	0.01	...
useless	0.01	0.20	0.03	...
was	0.01	0.00	0.60	...
is	0.01	0.00	0.30	...
masterpiece	0.01	0.11	0.01	...
I	0.21	0.00	0.01	...
liked	0.01	0.01	0.01	...
decent	0.01	0.02	0.01	...

- Suppose there are 10 words in a vocabulary
- We have computed the probability distribution $P(X_1 = word)$
- $P(X_1 = the)$ is the fraction of reviews having *the* as the first word
- Similarly, we have computed $P(X_2 = word_2 | X_1 = word_1)$ and $P(X_3 = word_3 | X_1 = word_1, X_2 = word_2)$ . . .

So, we will see that so, suppose instead of 50 just for easy ease of illustration there are 10 words in our vocabulary, we have computed the probability distribution for x1 equal to word and x1equal to does the fraction of reviews, having those the first order, that's how we have computed this probability. Right? Now, similarly we have computed all these other. Right? So, it's so on, we have computed all these other factors also.

Refer Slide Time :(15: 47)

The movie . . .

Index	Word	$P(X_i = w \mid X_{i-2} = the, X_{i-1} = movie)$	...
0	the	0.01	...
1	movie	0.01	...
2	amazing	0.01	...
3	useless	0.03	...
4	was	0.60	...
5	is	0.30	...
6	masterpiece	0.01	...
7	I	0.01	...
8	liked	0.01	...
9	decent	0.01	...

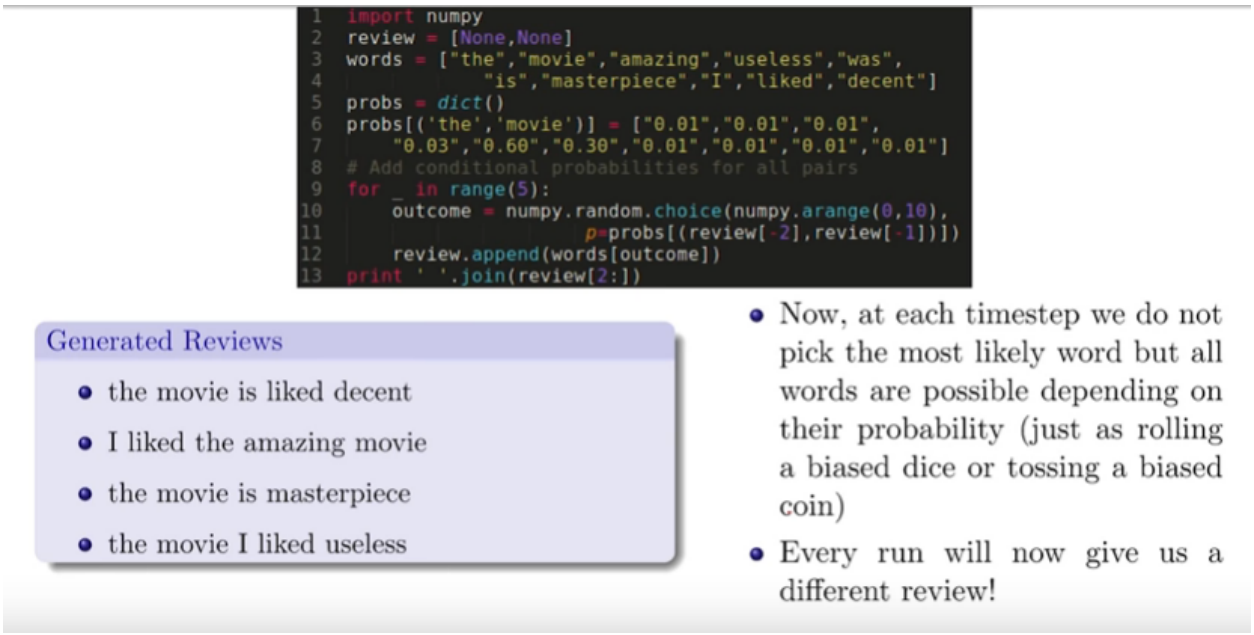
- Now consider that we want to generate the 3rd word in the review given the first 2 words of the review
- We can think of the 10 words as forming a 10 sided dice where each side corresponds to a word
- The probability of each side showing up is not uniform but as per the values given in the table
- We can select the next word by rolling this dice and picking up the word which shows up
- You can write a python program to roll such a biased dice

```
import numpy
review = [None,None,'the','movie']
words = ["the","movie","amazing","useless","was",
         "is","masterpiece","I","liked","decent"]
probs = dict()
probs[('the','movie')] = ["0.01","0.01","0.01",
    "0.03","0.60","0.30","0.01","0.01","0.01","0.01"]
# Add conditional probabilities for all pairs
outcome = numpy.random.choice(numpy.arange(0,10),
                  p=probs[(review[-2],review[-1])])
print words[outcome],
```

Now, consider that we want to generate the third, word in the review, given the first two words of the review. What are the possibilities? How many possibilities are there? 10, right? We have come down from 50 to 10. So, you could think of this, as a 10-sided dice and what you're going to do is, you're going to roll

this dice and just pick up the word, which came at the top, however unlike a fair dice, where all the sides are equally likely. What kind of a dice do you have here? You have a biased or a loaded dice where this probability of aside showing up, is proportional to the probability that you have computed in your table, does that make sense. So, you are not going to select the same face, again and again, which is the maximum face, they're just going to roll it, but you know that when you roll it, the outcomes are going to show up with a certain probability. So, it's possible that the Mac side, shows up more often than the other side's, but the other sides also have a chance of showing up and depending on whichever side shows up, you will pick up that particular word, all this is fine, how do you do this? How do you write this as a program? How do you write this as a program assume, you are given this table, how do you sample from a uniform distribution, what was asking how do you sample from a uniform distribution. So, how do you implement acorn coin toss, in Python? We can implement a coin toss, rolling a dice is just one step ahead from there. Right? How do you implement a coin toss? Suppose I tell you that the probability of heads is 0.4? Right? And I ask you, to generate a sequence of keep tossing the coin. So, that I get a sequence, heads tails, headsails and so on, according to this probability, how would you do that? You guys have a certain speech problem or something that you guys, don't know to answer. How will you do that? Generate a random number between 0 to 1 so, that's a uniform distribution, if the generated number is less than 0.4, say that less than equal to 0.4, say that the outcome was heads, it was greater than 0.4, say that the outcome was States or instead of heads tails, if you had a three sided coin, whatever that means, how would you do it? I mean head Torso entails probably that's what it means. So, how would you do it? Say the priority of heads is point four, the torso is point three and a point three five, point two five, how would you do it? Again sample from this uniform distribution, if the number is less than point four say heads, if it's between point four two point seven five say torso and the other one is still you can extend it to ten sides. Right? So, that's exactly what you are going to do here, I hope that's what is written here, you can go back and look at this code. So, here we have a conditional probability distribution, given the past two words for the movie, we have the probabilities assigned to all the ten remaining, I don't know why. I did this yeah! given the past two words, we have the probabilities that all the ten words can take. So, it's just like you had the properties for the three sides of the coin, you have the properties for the ten sides of your dives and now, I just need to generate a random number and pick the value, which falls in the right, bucket that make sense .Okay?

Refer Slide Time :(19: 06)



So, now if you run this code, at each time step you're not picking up the max value. But we are picking up a value, depending on the probability of that particular value. So, we are rolling the dice and maybe whichever side comes up it's not going to be the max always right .Okay? Now, everyone is going to get every run is, now going to give us a new review and of course some of them would not look very meaningful, because you have sample. Right? Because it's possible that you picked up, the least likely word at that point, just out of luck. Right? I mean you roll the dice and that's the word which showed up, there is a probability of even the least likely word showing up, that's why some of these reviews wouldn't make sense. But, using this, you can generate yeah! Useless reviews right sure. Okay? So, that's what I meant by classification and sampling. So, these are two things that you can do, with a joint distribution and

Refer Slide Time :(19: 55)

M7: More realistic than real life

w	$P(X_i = w \mid X_{i-2} = more, X_{i-1} = realistic)$	$P(X_i = w \mid X_{i-2} = realistic, X_{i-1} = than)$	$P(X_i = w \mid X_{i-2} = than, X_{i-1} = real)$	...
than	0.61	0.01	0.20	...
as	0.12	0.10	0.16	...
for	0.14	0.09	0.05	...
real	0.01	0.50	0.01	...
the	0.02	0.12	0.12	...
life	0.05	0.11	0.33	...

$$P(M7) = P(X_1 = more).P(X_2 = realistic|X_1 = more).$$
$$P(X_3 = than|X_1 = more, X_2 = realistic).$$
$$P(X_4 = real|X_2 = realistic, X_3 = than).$$
$$P(X_5 = life|X_3 = than, X_4 = real)$$
$$= 0.2 \times 0.25 \times 0.61 \times 0.50 \times 0.33 = 0.005$$

- Okay, so now what can we do with this joint distribution?
- Given a review, *classify* if this was written by the reviewer
- *Generate* new reviews which would look like reviews written by this reviewer
- *Correct noisy reviews* or help in completing incomplete reviews

$$\underset{X_5}{argmax}\ P(X_1 = the, X_2 = movie, X_3 = was, X_4 = amazingly, X_5 = ?)$$

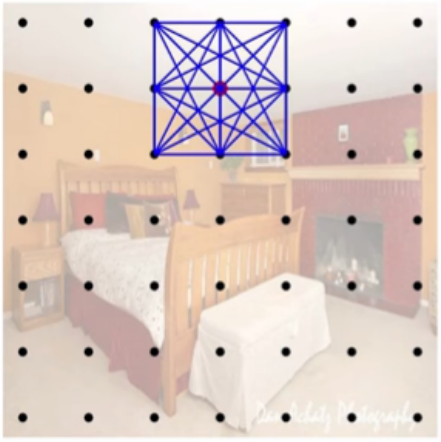
So, returning back to our story, what can we do with the joint distribution? The first thing is classification, the second thing is generation and the third thing, is actually a subset of generation, which is correct noisy reviews. Right? So, this is someone has, say taken a pic from a paper, of a review and send it to you and something in the pic is not visible, some words are not really clear. So, you want to fill in the missing words, given all the other words that you know and that's exactly asking this question, suppose the fifth word was blurred out. Right? Then this is the question that you're asking, about that given the remaining four words, what's the most likely fifth word is that fine. Okay? So, that was from the text domain.

Refer Slide Time :(20: 35)

- Consider images which contain $m \times n$ pixels (say 32×32)
- Each pixel here is a random variable which can take values from 0 to 255 (colors)
- We thus have a total of $32 \times 32 = 1024$ random variables $(X_1, X_2, ..., X_{1024})$
- Together these pixels define the image and different combinations of pixel values lead to different images
- Given many such images we want to learn the joint distribution $P(X_1, X_2, ..., X_{1024})$

Now, let's look at that example, from the other domain, which is what many people are interested in doing today. Consider images, which are in general M cross, n pixels, but we will consider their square thirty-two cross, 32 pixels. Right? And we are looking at a lot of bedroom images, because that's a data, set called by, 'Tomb Data Sets' and soon. Right? So, each pixel here is a random variable, which can take values from 0 to255 .Okay? Is that fine so, instead of that binary case, each pixel here is a random variable, which takes on values from 0 to255 and we have a total of 1 0 to 4random variables in this case. Right? And the random variables are X 1, X 2 up to X1, 0 to 4 and what are we interested in? What are we interested in? The safest answer for this segment of the courses Joint Distribution. Right? Why are we interested in the Joint Distribution? Just because we can, what will you do with the Joint Distribution/ Classify whether this image is a bedroom image or some other image, in fact these bedroom images could have been designed by a particular interior designer and you can then classify, whether this bedroom comes from the same interior designer. Right? Right? What else can you do? Become an interior designer, how I thought by giving an interest and rolling in, in tooling into a program or something. Okay? what will you do? Generate images from this distribution. Right? And if someone has given you a design and someone dropped coffee on that, some portions are blurred, what can you do? Given all the other pixel values, you can decide, what this is going to be right? The same set of things, if you are given a lot of bedroom images, just as you are given a lot of reviews by a particular author, you can learn this joint distribution, what is the factorization that you will choose for this joint distribution? In the text case in the language modeling case, the natural choice was that each word depends on the previous two words or something like that, in this case, this is pixels, you have done convolution neural networks already that was a hint. So, what's the neighborhood, what kind of a graphical model will you have? I already gave up they're not, directed or will have a mark of network. Okay? What are the factors? What are the dependencies? The dependencies give you the factors; each pixel is a random variable, what is the assumption that you're going to make? It depends only on its neighbor, that's what makes sense in the case of images. That's what was the basis of using convolution neural networks or the convolution operation. Right?

Refer Slide Time :(23: 35)



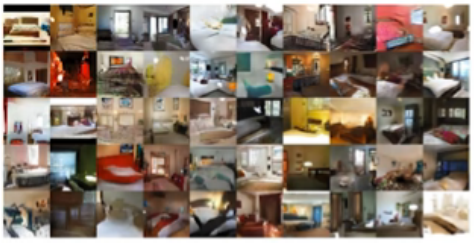
- We can assume each pixel is dependent only on its neighbors
- In this case we could factorize the distribution over a Markov network

$$\prod \phi(D_i)$$

where D_i is a set of variables which form a maximal clique (basically, groups of neighboring pixels)

So, we are going to use, the mark on that work that we'll consider is that each pixel depends on its neighborhood, the neighborhood is up to us to define, whether we want to include the diagonal pixels or not that's up to us to decide. Right? And in fact we could do it this way. So, the Joint Distribution would factorize, as these factors, each factor here, would correspond to this one maximal clique, in the diagram that you have and each clique, corresponds to all the pixels. So, I am assuming that all these pixels are each other's neighbors. Right? And that's completely as up to me how to decide, I can do the worst case assumption that each pixel only depends on its immediate, left neighbor or right neighbor. Right? That's the same as what we decided in the case of words and each word only depends on the previous words, I can do that kind of assumption also, whatever assumption I make, the distribution grid factor is accordingly. Okay? Now, what is it that I need to learn here? These factors, these factors are very, similar to conditional priority distributions, except that they are not conditional property distributions, they have taken give you the affinity, between two pixels what is the possibility that pixel one takes on the value 255 and pixel two also, takes on the value 255, that's roughly what these factors tell you assuming that the factors consider only two pixels. Okay?

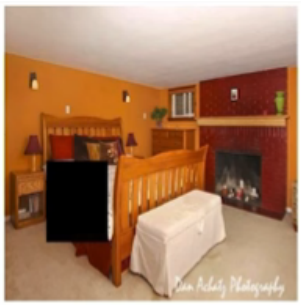
Refer Slide Time :(24: 48)



- Again, what can we do with this joint distribution?
- Given a new image, *classify* if is indeed a bedroom
- *Generate new images* which would look like bedrooms (say, if you are an interior designer)

Now, what do, can you do with this joint distribution? Once you have this, you can take an image classify fat indeed is a bedroom image. Right? And in this case I meant to denote that point zero one is actually a very low score, which I already contacted before but, assume that this is a very low, score and this is not a bedroom image, generate new images, which would look like bedrooms. Right? So, now you have, this how will you generate this now, but unlike the word case you do not have this very convenient, thing right that you sample the first pixel, then you sample the second pixel, based on the first pixel, third pixel beige on the second enforce it's not like because that's not how our factors are defined. So, are we going to sample from this situation? So, it's not very straightforward, that's the answer it for an unlighted graphical model, it's not very straightforward, to sample from this distribution and that's what we are going to spend time on when we talk about restricted Boltzmann machines, which is an undirected graphical model. Right? It's not very straightforward to do that and we will see we'd have to do something known as Gibbs sampling, to do that. Okay?

Refer Slide Time :(25: 45)



- Again, what can we do with this joint distribution?
- Given a new image, *classify* if is indeed a bedroom
- *Generate new images* which would look like bedrooms (say, if you are an interior designer)
- *Correct noisy images* or help in completing incomplete images

And again correct noisy images and this how they have written it as a separate bullet, this is essentially a subset of the generation of the imager, you just need to generate, one section of the image, instead of generating all the pixels in the image. Okay?

Refer Slide Time :(25: 56)

- Such models which try to estimate the probability $P(X)$ form a large number of samples are called generative models

So, such models, which try to learn this joint distribution P of X, from the given data, are known as, 'Generative Models'. Right? And what we are interested in is, Deep Generative Models. So, what would that mean? A very profound generative models. So, let's see I'll try to see, if you guys can answer this question, although this is going to come be way, later on, but, it really give me some satisfaction that you've understood some stuff in this course. So, what do Joint distributions have? How are we representing them, as? Product of factors, factors are the same as, parameters. We need to learn these parameters. Whenever you want to learn something. Okay, so, sorry, we want to learn these factors, whenever we want to learn something, what do we introduce? A parametric form of something. Okay? Now, can you tell me what deep generative models are going to be? Where this parametric form? So, that means for every factor, we're going to define it as some function, which has some parameters. Now, you, can you tell me, what a deep generative model would do? Would be? Where this function would be a, neural network? Right? Does that make sense? Even if it doesn't, it will become, it'll make sense after a week or so. Right? That's what we are interested, in Deep Generative Models. Hence this entire background on generative models.