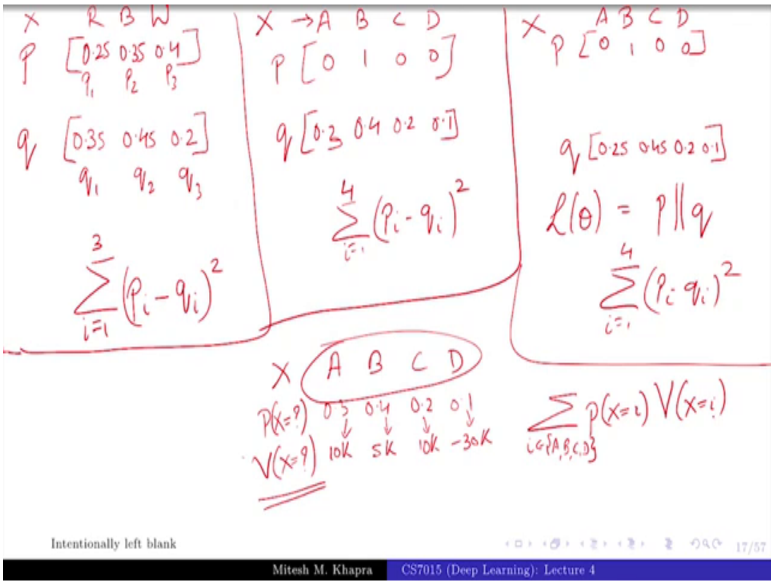
Deep Learning
Prof. Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 4.10
Lecture – 04
Information Content, Entropy & Cross Entropy

So, for the next module we need something known as Cross Entropy. So, we will just try to make some develop some intuitions for cross entropy and get to the formula for that. And then I will tell you how it relates to the problems that we deal with, ok. So, first let us start with something simple that what is it that we are trying to do, ok. So, with that I will give you an example and I will ask you a few questions and then from there we will slightly try to go towards cross entropy. So now, suppose you have an urn which has thousands of balls, and these balls are of three different colors which are red black and white see, right?

(Refer Slide Time: 00:51)



Intentionally left blank

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

17/57

So, you have an urn which has three different types of; and there are many many such balls which you have put in it. And since you have put in it you know how many red balls are there, how many blue balls are there, and how many white balls are there. So, for you it is very easy to compute the probability of each of these things right.

So, that is say our probability is 0.25, 0.35 and 0.4. Now, talking more formally what is happening here is that, you have a random variable X which can take on the values red blue or white, right. And this is the probability of each of those or the random variable X taking any of these values, ok. So, this is the setup. Now I am your friend and you tell me that you can peep into the zone, you cannot actually count take out all the balls and count them and estimate the probability. We can just take a look into this, and try to give me an estimate of what these actually; what are these probability values; that means, what is the value?

So, what is the probability that X is equal to red or X equal to white or X equal to blue. So, I just take a look at it turn it around a bit, and try to get some feel, ok. I see a lot of red balls, but a fewer blue balls or white balls and so on. And based on that I make my guesstimate, right? So, I will just say that maybe these probabilities are 0.35 0.45 and 0.2, right. So, this is actually the true distribution, I will call it as p, right, because this is the correct one. And what I have estimated, I will call it as q, ok.

And remember now p has you can think of p as a vector which has these 3 values, p 1, p 2, p 3 because there are 3 possible events here. And similarly q has 3 values, q 1, q 2 and q 3. So, in this case I clearly know that, I am wrong or when I give you these values you know that you are wrong. So, you tell me that whatever you have estimated is wrong. Then I obviously, ask you, ok. Tell me how wrong was I? So, how would you give me that number? That is the thing that we are interested in.

So, the general problem that we are interested in is that, there is a true probability distribution, and there is an estimated probability distribution, and we want to find out how bad was the estimation, ok. Now, can you tell me a simple way of computing this? It may not be correct, but still it makes sense.

Student: Squared error.

You can just take the squared error. So, what you are essentially telling me is that you could just treat these two as any other vector, right. And you could take the squared error difference between these quantities. So, what you are telling me is this, where i goes from 1 to 3, ok. So, this is one valid way of doing this, but then we are ignoring the fact that this is a distribution and hence, it has certain properties that the sum of the elements is 1 and so on all of them are positive and things like that. So, we are ignoring those kind of things we are completely ignoring the fact, that we are not dealing with a normal vector, but a spatial vector which happens to be a distribution, ok. So now, we want to find out a more principled way of computing the difference between two distributions. And in practice why are we interested in this? Because we will always have a true distribution and a predicted distribution, right. So, that is what we want to do, we have some way of computing it, but you want a better way of computing it, ok.

Now, let me make a case for why do we care about such differences right. So, let me take a simple case of a classification problem, ok. And to motivate that I will start from a different example and then I will come to the classification problem. Suppose there is a tournament going on, and there were 4 teams which leads the semifinals; let us call them A B C D, ok. Now, you were following the tournament up to the semifinals, and after that you didn't watch the tournament, and you do not know who eventually won; well, the tournament is over and someone has won it, ok.

I actually watched the tournament and I know that B has won it, ok. Now can I express this in terms of a probability distribution? Right so, first let us look at what is the random variable here; what is a random variable here? The team which won, right? So, that is my random variable and it can take one of these 4 values, ok.

Now, I know that team B won, because I saw the tournament, and I have seen that they won. So now, how can I write this as a distribution? What is the distribution comprised of? It comprised of these probabilities assigned to each of these events. And there are 4 such events here. So, how do I write this distribution? So, what you are telling me is I could write it as 0 1 0 0.

So, essentially they are telling me that all the probability mass is focused on one of these outcomes, because that is the certain outcome, that is already happened no one can change it so, that is the outcome for this tournament. So, I know that the probability of that even is event is 1 and everything else is 0. So, in other words the probability that the random variable X takes on the value B is 1 and everything else is 0, right?

So, what I am trying to tell you is that, even for a certain event, you could still write it in terms of a distribution, where all the mass is focused on that event, ok. Now again I will bring the same setup that I did not watch the tournament after the semifinals. So now, you ask me give me your prediction, what which team would win, ok. Or this is the prediction which I made before just after the semifinals or sorry, just before the semifinals that I think one of these teams is going to win the tournament, and the chance of each of them winning is something like this. So, I know the teams I follow this sport and I probably know that, ok. B has a very strong team and they have a very good record in the past few months and so on. So, maybe they have a higher probability of winning. So, these are the numbers which I assign, fine.

Now, again I have made an estimate, was my estimate perfect? When would it have been perfect? If I had predicted with certainty won that B is going to win, but I was not willing to bet everything on B. So, I said there is a very high chance it will win, but there is still a chance that there could be some surprises, ok. Now, how wrong was I? Now again tell me can you tell me what is p and what is q here? This is the true distribution and this is my predicted distribution. And what am I interested in again? The difference between them, how wrong did I go? And what again what is a simple way of doing this again, square errors. So, again this is what my formula would look like, ok.

So, this is fine in this toy case, but why do we care about in real life? Examples, that we are going to deal with in machine learning, right. So, in watching learning will deal with a lot of problems which are classification problems, ok. And in classification problems, you would again have this setup where you have a label the good thing of the label as a random variable, and it could take off one of many values right. So, I will again assume that it could take suppose you are trying to take a picture of fruits, ok. And you are trying to classify them, ok.

And I could again think that I have 4 fruits say apple, banana, cherry and dragon fruit, ok. And this random variable can take one of these 4 values depending on the image that I am seeing, ok. Now I have been given some training data. So, for every training data I have been given an image, and I have been given the correct label. So, for that training data what is this distribution? Suppose I have been given the image of a banana, what does this distribution look like? 0 1 0 0, right again I have seen it so; I know it is certainly a banana.

So, I do not have any confusion all the probability mass is focused on that. Now the same image we are going to show to one of our models, ok. And it is going to make a prediction, and will again ask it to give us a distribution, the model will give us values perhaps like this, ok. So, this is the models prediction, again the model has given us a distribution and we have a true distribution, and we are interested in knowing how long the model was. So, that.

Student: Correct the (Refer Time: 09:26).

We can correct the parameters of the model. So, this is r dash function loss function, right. So, a loss function is some notion of difference between p and q, right. And so, far we have been dealing with a very simplistic notion of this difference which is just the squared error loss, ok. And we want to do something better than this right. So, what I have told you is that you could always have a true distribution always have a predicted distribution. And you would be interested in finding the difference between them, that is the one first part; the second part is that even when you are given something with certainty, you could still write it as a distribution such that all the mass is focused on that event which was which has happened rate, which was the label was banana in this case, ok.

And then you could still predict this from your model, and now you are interested in knowing how wrong you are model wind because, that is the loss function that you will use, and then you will try to update your parameters with respect to this loss function, means that is the setup that we are interested in ok, ok. So, that is so, I made a case for why we need to find differences between 2 distributions. How to do it? In a more principled manner we have not seen that yet we will get to that, ok. So, before I get that, I also need to tell you something about expectations, ok. So, let us written to as sports example where there were 4 teams, ok. And say based on pundits and that sport, they have said that these are the probabilities of winning, ok.

And now you are into betting, and you bet place your bets on these teams and you place our bets in a way that, suppose team events then you end up winning 10 k rupees, if team B wins then probably will end up winning 5 k rupees. And if team C wins probably 10 k, and if one of the other team wins maybe will end up losing money or something like that, right. Now you want to know, what is your expected reward? So now, let us see what is happening here? This was a random variable which could take on one of these 4 values. These are the probabilities of the random variable taking that value, taking on the value

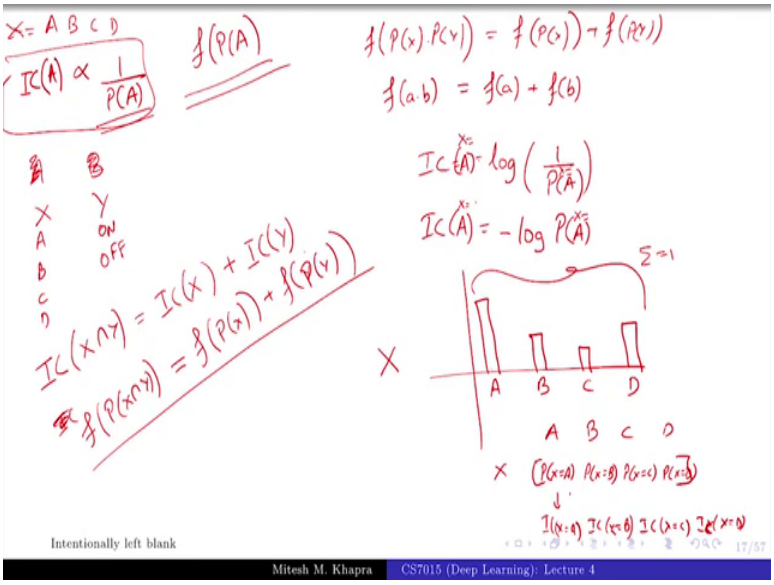
A or taking on the value B C and D, ok. This is your value or the gain or the profit associated with the random variable taking one of these values, right.

So, you have a random variable, you have a probability associated with every value of the random variable, and you have some gain or value associated with every value of the random variable. Now how do you calculate the expected gain or expected profit which is this? There is a 30 percent chance that you will earn 10 k, there is a 40 percent chance that you will earn 5 k, there is a 20 percent chance that you will earn 10 k, and 10 percent chance that you lose 30 k, right.

So, the way you will compute it is that and this is the simple expectation formula; which is the probability of now the event here belongs to ABCD, right. This is one of the 4 teams that will win, probability of that event happening in to the value associated with that event happen, ok. This is a fair computation you get the intuition that this is how you will compute the net reward that you have, ok. So, this is how you compute the expected value with respect to a particular distribution, so, this is the background that we need.

Now, I will just go on to the next slide, and now we will talk about entropy first, perhaps information content be first then entropy and then cross sector, ok. So now, what is information content?

(Refer Slide Time: 13:28)



$X = A\ B\ C\ D$
$IC(A) \propto \frac{1}{P(A)}$
$f(P(A))$
$f(P(x) \cdot P(y)) = f(P(x)) + f(P(y))$
$f(a \cdot b) = f(a) + f(b)$
$IC(\overset{X=}{A}) = \log\left(\frac{1}{P(\overset{X=}{A})}\right)$
$IC(\overset{X=}{A}) = -\log P(\overset{X=}{A})$
A B
X Y
A B C D
ON OFF
$IC(x \cap y) = IC(x) + IC(y)$
$f(P(x \cap y)) = f(P(x)) + f(P(y))$
$\Sigma = 1$
X
A B C D
A B C D
$X\ [P(X=A)\ P(X=B)\ P(X=C)\ P(X=D)]$
$I(X=A)\ IC(X=B)\ IC(X=C)\ IC(X=D)$

Intentionally left blank

17/57

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

So now, again let us take the same case that we have a random variable which can take on values A B C D, ok. Now let us what we are trying to say is that, if I know a certain thing, what is the information that I have gained? So, you and I are talking you tell me something, ok. And I want to see whether my information was enhanced, whether my knowledge was enhanced, that is how we will quantify information content, ok.

So, if you are talking to me and you tell me that my name is Mitesh, the 0 information content for me, right. Because I already know that there is no surprise in that ok, but if you talk to me and you tell me that today there is going to be a lunar eclipse, then there is a possibility there is some information content gain for me, right. Because that is not a event which happens every day if you just tell me you will see the moon today and you live in a region where it is not typically cloudy and there is no information gain there, right? So, what do you see here? When is the information gained high? When the event which happens is a very surprising event; and how do you say supplies in terms of probability.

Student: (Refer Time: 14:29).

It is a very low probability event, right? So, if there is again this tournament, and say D was the weakest team in the tournament, and a was the strongest team in the tournament if you come and tell me that D won then I would be really surprised that some information which I had gained, but if you tell me that a one then probably I already knew it at the back of my mind, right. Because A is clearly the strongest team in the tournament and there is no information gained for me.

So, one thing that we are trying to establish that the information content I see, ok; the information content of an event is inversely proportional to the probability of the event. There is a that is a fair intuition, ok. Fine, now I want am still talking in terms of vague things I am saying it is inversely proportional, but I still need an exact function so that I can compute it. So, I want something I want a function where I plug in the probability of an event, and I get the information content of the event, right.

Right now I do not have that function am just building some intuition towards that function, ok. But this is one requirement that I want the function to satisfy this is something that all of us agree with, ok. Now think of 2 events which are independent A and B, ok. So, A is the event whether the A C s on here or not, and B is a event which tells, maybe sorry sorry, sorry, sorry, sorry, sorry. So, let us consider 2 different random variables.

So, X is the random variable which can take on values 0 and 1, ok. Sorry. So, X is again this random variable which can take on these 4 values ABCD whether who won in the tournament. And Y is this another random variable which can take on the value on and off depending on whether the A C's on in this room or not, ok. What can you tell about these 2 random variables? They are independent random variables, ok.

So, this is on or off and this is which team won the tournament, ok. Now I come and tell you something about the random variable y, and I come and tell you something about the random variable X, ok. So now, I want you to tell me this, what is the information content of X and Y? I tell you something about X, and I tell you something about Y, and these 2 events are in these 2 random variables are independent, then what can you tell me about the information content? What is the condition that you would want? You gain some information by knowing things about X, and you gain some information by knowing things about Y, ok. So, what can you tell me? It should be the sum, right.

Because these two are independent events; so, whatever information I am getting from this random variable and this random variable which together enhancing my information, right. It is not cancelling out anything or is there is no common intersection there, right. If the 2 events were not independent then I would not expect this to hold because knowing something about the first event only tells me something about the second event, right, because they are dependent.

So, then that case the information gained would not be additive, ok. So now, let us see, I already made a case that this function which tells me the information gain is actually proportional to the probability, ok. So that means, this is what the input is going to be, right. And then what is the other condition that I want? This is a fair thing, right. I just replaced information content by a function, and I know that the function should depend on the probability, because that is what we have established here.

So, we know that the function depends on the probability, we still do not know what this f is exactly, but I am trying to impose some conditions on f. One condition of f, f is that this condition should hold, ok. Now let us look at this condition which I have underlined, this is f of is this fine because it 2 events are independent, you can write them as the joint probability as P of x into P of y. This is clear to everyone, right. You seem to be a bit lost Arvind clear, ok.

Now, what is happening here? I have a function f of a into b and that is actually equal to f of a plus f of b, what family of functions do you know which has this characteristic? Log, right, that is why log is a good choice for this. That is why information content is going to be the log of the probability, but I wanted to be inversely proportional, right so, it will be log of 1 by the probability, ok. So, that is why information content of this thing is; so, you see this how did we arrive at this log formula, ok.

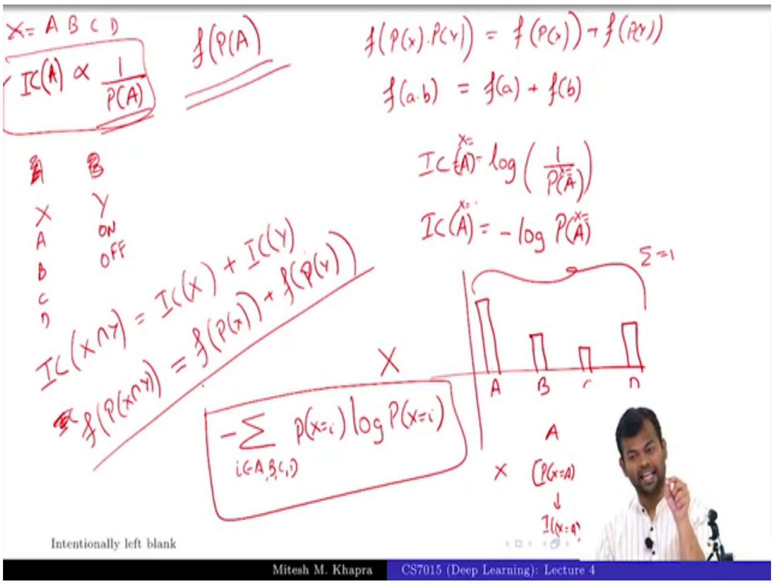
And this log can just be to any base, it does not matter, ok. So, all of you get how we arrive at the formula for information contained, ok? Now just give me a minute I need to think of what is the next thing that I have to say, ok. And so, we have found out the information content of one of these events happening, ok. Which was the x taking on the value a, ok.

Now, let us think of this random variable x. So, here actually I should have said x equal to a probability of x equal to A, ok. It makes sense because the random variable is x, and the event is x taking on the value A, how much information content is in that. So, if I know that x was A how much will I be surprised by it, ok. Now let us take this event, this random variable x which can take on values A B C and D as I said with each value there is a probability associated with it, such that this sums to 1, ok. Now, I did not need to draw this diagram, ok. I should (Refer Time: 21:01).

So, X is a random variable which can take these 4 values, which each of these values, I have a probability associated, ok. So, these are the values these are the probability values. Now what do I also have? I have the information content associated with each of these, right. And the information content actually tells me the surprise of that evening. Now if I ask you what is the entropy of this random variable X? So, remember I had this case where I was betting I am with every poor outcome I had a value associated with it? I had the same situation here, with every outcome have a probability, and I also have a value associated with this, and the value is the information content, ok?

Now, if I ask you what is the entropy or the expected information content of this random variable, then how will you compute that? I am asking you for an expectation, right.

(Refer Slide Time: 22:16)



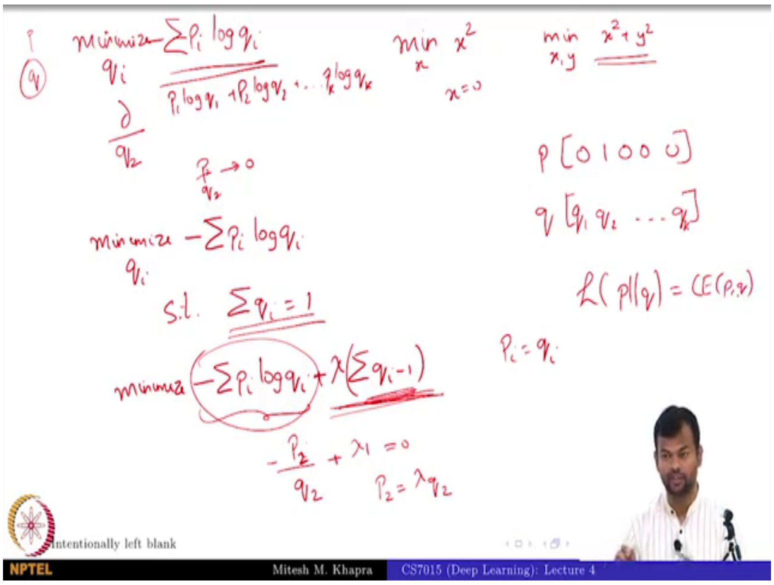
So, I will compute summation I belonging to A B C D P of x equal to i information can take what is the formula for that.

Student: minus (Refer Time: 22:29).

Minus I will just take the minus outside, ok. So, this quantity is called the entropy of the random variable, right. It is the expected information content in the random variable, ok. Now if you see what would be the expect entropy of a random variable if it is corresponding to a certain event? That means, say the sun rises always in the east, right? So, what is going to be the entropy of that 0? Why, you will have one of the sums in that summation as 1 log 1, right. And every other sum would be 0 into log of something.

So, 0 into anything is going to be 0, even though that quantity is not defined 0 into anything is going to be 0. So, the total entropy is going to be 0, ok. So, this is entropy, now what is it that we are actually interested in? Cross entropy, so, we have not gone there yet, ok. So, we need to perhaps add one more slide, so far everything is clear, ok. So now, we are interested in something known as cross entropy.

(Refer Slide Time: 23:36)


minimize $-\sum p_i \log q_i$
q_i
$p_1 \log q_1 + p_2 \log q_2 + \dots p_k \log q_k$
$\frac{\partial}{q_2}$
$\frac{p}{q_2} \rightarrow 0$
min x^2
$x = 0$
min $x^2 + y^2$
x, y
p [0 1 0 0 0]
q [q_1 q_2 ... q_k]
$\mathcal{L}(p \| q) = CE(p, q)$
minimize $-\sum p_i \log q_i$
q_i
s.t. $\sum q_i = 1$
minimize $-\sum p_i \log q_i + \lambda(\sum q_i - 1)$
$p_i = q_i$
$-\frac{p_2}{q_2} + \lambda = 0$
$p_2 = \lambda q_2$

Intentionally left blank

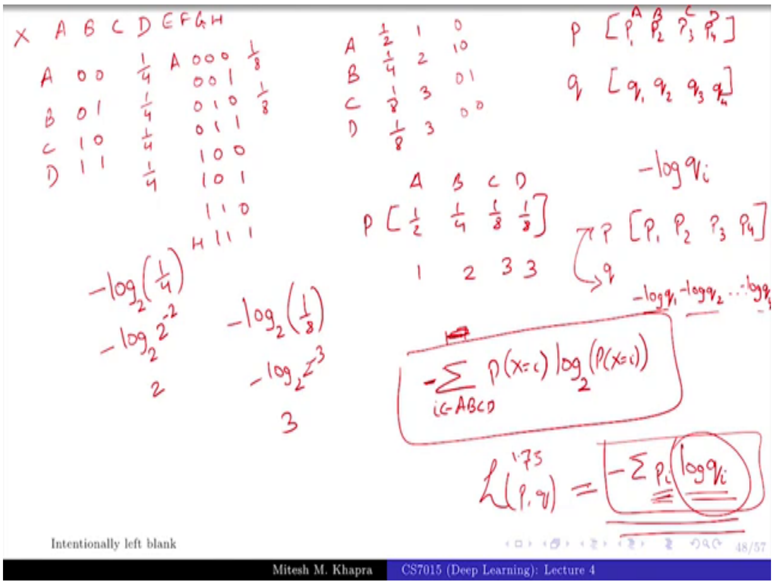
NPTEL

Mitesh M. Khapra

CS7015 (Deep Learning): Lecture 4

So, there again the situation is that there is something which is the true distribution, and something which is the predicted distribution, ok. Now, actually before going there so, let me just erase this off, how many of you have thought that entropy is related to the number of bits that you need to transmit something? Do you know why that connection exists? No, ok, now again let us think of this that you are trying to transmit a message, ok.

(Refer Slide Time: 23:59)



Intentionally left blank

Mitesh M. Khapra

CS7015 (Deep Learning): Lecture 4

And that message is again a random variable, which can take on 4 values A B C D, ok. So, think of these as 4 commands that we are trying to send to someone, right. And then based on that command someone will take some action. Now in the digital case how will you transmit this? Encode it to bits, so, what is the encoding that you will use? 0 0 yeah we will come to that, 0 1 1 0 1 1. So, how many bits are you actually using for every message?

2 bits, ok, for every bit so, maybe this is A this is B this is C and this is D. So, for every message you are using 2 bits, ok. Now, so, let us me see, ok. Let us see actually what when you are doing this, what are you actually assuming. So, actually assuming that all of these are equally likely, if all of these are equally likely can you tell me the information content of any of them? It is going to be minus log of 1 by 4, ok. That is actually equal to minus log and this is to the base 2, 1 by 4 is 2 raise to minus 2, that is equal to 2, right.

So, the information content is actually equal to the number of bits that you are going to use to transmit that message, ok. Now let us see if this is just in this special case or in a different case also. Suppose this could take 8 values, how many bits would you use? 3 bits right, so, you will have 0 0 0, 0 0 1, ok. And this would be A to H, now what are we actually assuming here? Each of these is equally likely, what is the probability? 1 by 8; what is the information content? 2 raise to minus 3, that is equal to 3 right.

So, the number of bits that you actually use to transmit something this, can you can talk of it in terms of the information content of that, right. Now suppose I want to transmit this over the long distance, ok. So, I need to bit be a bit efficient in terms of number of bits that I use right. So now, in this one of these cases, suppose it is of the following form, right. That let us look at the case where x can take one of 4 values, and say let me just put the right values. So, I will say 1 by 2 1 by 4 1 by 8 1 by 8, ok. Now what is the information content of each of these? 1 2 3 3, and this is the message that I am going to send, ok.

So, what am I doing here? I am using a different number of bits depending on the probability of that event, why does this make sense? Why is this a smart thing to do? If you want to transmit something which you are going to transmit a lot of times, you better use less number of bits for that, and this is exactly what is happening here, a was having the highest probability, and you are using the lowest number of bits for that, ok? Now what is the expected number of bits that I will use up? If it is a I will use one if it is B I will use 2 if it is C 3 and D 3.

So, what is the expected number of bits that I will use. Again I have the same situation, right, I want you to cast it into the same situation, I have the probability values, and with each of these guys, I have a cost or a value associated what is the cost? 1 bit, 2 bit, 3 bit, 3 bit. So, what is the expectation now? Can someone compute the expectation, 1.75 actually let me just write it down, it would be again I belonging to A B C D, P of x equal to i into the number of bits that you will use, right? So, that is just equal to log of log to the base 2 of P of x equal to i, minus 1, what does this quantity actually? This is the entropy we just saw this, A, this is the entropy of the random variable and what is it telling us actually that the entropy is 1.75.

So, what is the meaning of this actually? So, on average you will be needing 1.75 bits whereas, if you are assuming everything is equally likely on average you are using 2 bits, right. So, you see that on average you are making some savings here right. So, that is what the entropy tells you. If you know what the probability of these events is, then you better use that to decide the number of bits that you are going to use to send each of these ok?

So now let us complicate this a bit more, now we have the entropy. Now let us complicate it a bit more. So, there is some true distribution which exists. There is some true distribution from which these messages are coming, right? But you do not know what that true distributions, we never know the true distribution that is the entire problem that we have been dealing with in machine learning, right.

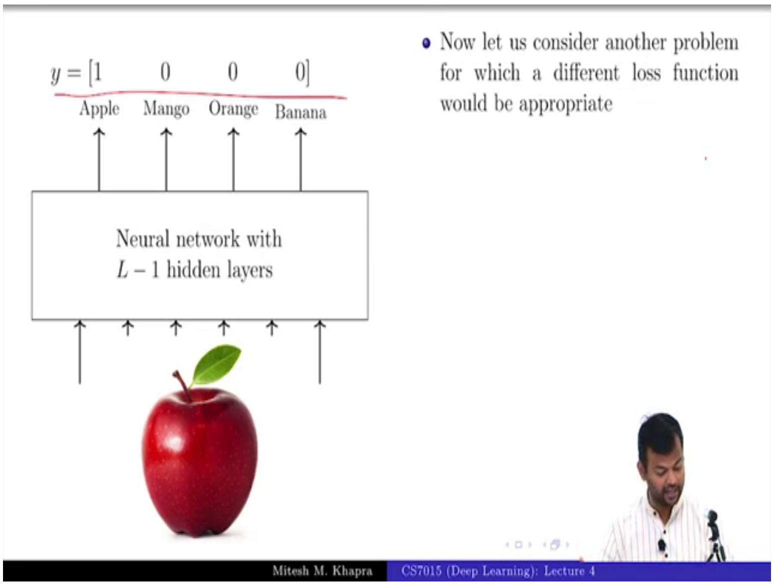
So, what you will do is we will somehow try to predict this distribution, and this then the and the recipe that you will use is the same as that I used for the example where I had an urn right. So, there are these thousands of 10,000's of messages which has going to keep coming on, you do not have access to the entire stream. But you have seen some thousand of those messages just, as I had peeped into the urn, and I had seen some balls and I had made an estimate; that I think based on these messages that I have seen so far I think these are the actual probabilities, right.

So, the true probabilities are say p 1 p 2 p 3 and p 4 corresponding to A B C and D, I do not know what this 2 properties are, but I can estimate them looking at some samples or basically using my domain all is, right. Maybe I would know that if one of these messages is stopped, right. And I am actually trying to talk to a computer or a computer program, that maybe stop is something which are used very rarely only at the end of the program or something, right? So, you have some either some domain knowledge or based on some samples I can estimate the value of this probability.

And I just try to relate it to the exact example of urns, where you had these 10,000's of balls, but you could not see all of them, you sampled some and estimated a probability. Here again there is a continuous flow of message you cannot have access to all of these, because they are going to continue. But I have seen some of those and based on that you estimate these probabilities. Now based on this estimation how many bits?

Ok, so now this is the estimation that we have, ok. Now based on this you will decide the number of bits that you will use for each of these messages, right. Because you have some estimate so, you want to be smart you do not want to keep 2 bits for all of them. So, you will just say that I will use log qi bits for the ith message.

(Refer Slide Time: 31:23)



- Now let us consider another problem for which a different loss function would be appropriate

Mitesh M. Khapra CS7015 (Deep Learning): Lecture 4

This is fair thing because I know that the information content is proportional to the probability. In fact, it is exactly given by this formula, minus log of qi. So, based on my estimated probability I am going to do this, ok. And this is the number of bits that I have resolved. Now do you see a problem with this? This is my estimation, but the data is actually going to come from the true distribution. It is not going to follow the distribution q; it is going to follow the distribution p.

So now, what is actually happened is this, right. This is the situation that we are dealing with. We have p which was a true distribution, that is the rate at which the data will come, but with each of these events the value that we have associated is now related to q, because q is what I have access to, I do not have access to p, I just have access to q so, I have associated a value based on q, does this make sense? I should have actually used log p 1 bit's log p 2 bits and log p 3 bits, but I do not know what p 1 p 2 p 3 are.

I just estimated them based on some samples so, that is q 1 q 2 q 3. And these are the number of bits that I am using. Now if I have to compute the expectation how will I do it? I have to use p because that is the true distribution from which the data is coming, right.

So, what would the expectation now look like? Everyone gets this the actual probabilities are this, but because I am poor at estimating them I ended up associating these values, which could be wrong, right. Because I would have overestimated the probability of one of these messages and hence I have reserved lesser bits for that or underestimated the probability of one of these events, and hence reserved more bits for that or vice versa, right.

I could have assigned a wrong number of bits to them, right. P no. So, do we have access to p in the sense yeah. So, someone knows that, right. I mean there is a again in the same case as in the label case, right. We have access to the 2 true p there, and we are estimating a q, when we are given these images for the training data, we know that the distribution is 0 1 0 0 if the image is B for banana, right.

Student: (Refer Time: 33:49).

Then it is validated, right. So now, this is what is this quantity called? This is called the cross entropy, ok. You get why it is called the cross entropy because now you have two different distributions involved here, ok. You have the q distribution based on which you made your decisions, you assign values to these events based on the q distribution, but the true distribution is the p distribution.

So, the actual number of bits that you use up on average is going to be based on the true probability, they try to understand that. Now what will happen is for event a you have assigned a certain number of bits. Now how many bits will get used up it depends on. The actual probability of p if that, message is repeated many times then that is how this summation would be computed, right, is that clear?

So, this is called the cross entropy, but now why is this the difference between 2 distributions? That is what we wanted. Given 2 distributions we wanted to be able to find the difference between them. Now am telling you that cross entropy is a way of finding that difference. Why is it so? So what would you want this difference to what is the property that you would want this difference to have? If p is equal to q, then if p is equal to q then;

Student: (Refer Time: 35:08).

Not 0 maybe it should take the lowest possible value, right. So, this function, right. This is actually telling you loss of p comma q, right. This is what this is and we are calling it as the cross entropy, this function you take it is minimum value when p is equal to q, right. Because now at that point you are not really making any loss that is the best you could have done. Does this function take it is minimum value when p is equal to q? Yes, why? How is that obvious, but why there could be something else which is lower than the actual entropy, right. Why, how you have to we are trying to minimize something. So, you have to give me answer, ok.

So, yeah so, let us do that, ok. So, how many of you it is obvious that q is the answer? I mean the answer is p is equal to q it is not, ok. Now this is the part which am a bit worried about, but I will just do it anyways. So, this is what we want to do right. So, let me see how do I put this, so, remember that we had a p and we had a q, and we want to find a q such that this quantity is minimized, ok. That is what our objective is right.

So, we want to minimize this with respect to qi, ok. Now how do you find the minimize suppose I have this problem? How do I find the minimum value? How do I find the value of X which minimizes this, take the derivative, and set it to 0, ok. And then in this case I will get x equal to 0 is that value, ok. Can I do the same thing here. And suppose it was this so now, this is a function of 2 variables, again I could do the same thing, I could take the partial derivatives and set them to 0, right.

And I will get the minimum value. Now here this is actually a function of how many variables? K in general right, so, q 1 q 2 q 3 up to qk, ok. Now can you try doing the same thing can you can you take the derivative and set it to 0. This is again a sum, right. It is very similar to this situation, right. It is actually let me just write it down, it is p 1 log q 1 plus p 2 log q 2 up to p k log qk; ok.

Now I want to take the derivative with respect to one of these guys; say, q 2, what would it be? P 2 by q 2 is equal to what will I do? That is the derivative p 2 by q 2, I will set it to 0, do I get anything? What is it that I am doing wrong here? There is something that I am deliberately doing wrong. Is this an unconstrained optimization problem? There is a constraint on the variables, what is the constraint? So, why my true optimization problem is minimize with respect to q is; such that summation q is equal to 1.

Do you know an easy way of dealing with these problems? How many of you know the Lagrangian multiplier, ok. How will I use it here? What will my objective function become? Then summation of qi minus 1 lambda then minus ok, How many of you understand the intuition behind this, that is a good answer, ok? Now let us let me try to explain why this makes sense, right. This is the constraint that we have to operate within this constraint, ok.

What I have done is I have taken the so now if the constraint is not satisfied what will happen to this quantity if the constraint is not satisfied? That means, my summation is not equal to 1 that is what means whether the constraint is not satisfied, what will happen to this quantity, it will be a non-zero quantity, right. Fine then what will happen to my overall objective? And I think we have made a mistake this should be plus, right. I should add it, right. Should be plus no it does not oh the lambda can be, ok sorry, so, let me assume this is plus, ok.

So, what I am trying to do is that, this is my objective function which I am trying to minimize. I have added another quantity to it, if this quantity is not equal to 0, then I will not be the absolute minimum, I will be at the minimum plus something right, but if this quantity if the constraint is satisfied, then this quantity will go to 0, then am actually at the minimum of the function, do you get this, right? So, this is the function that I want to minimize, I have added some quantity to it. Now, that quantity is actually related to the constraint that I do not want to violate. If I violate the constraint, this is going to be non-negative, right.

So, whatever minimum value I achieved, I will be slightly higher than that, because some non-negative value has got added to it, ok. Is that fine? But if the constraint is satisfied, then I can achieve the minimum value. So, that is roughly the intuition behind using this Lagrangian multiplier, it is a very crude intuition, but there is of course, a lot of math behind that, but I am just giving you the intuition behind this which one.

Student: (Refer Time: 41:39).

Yeah that is what you could adjust the lambda, and ensure that it is not negatively, ok. So now, now can you do the same thing can you equate this to 0, can you take the derivative and equate to 0? What will you get now? This term will give you p I by q 2 as before, oh sorry p 2 by q 2 as before plus lambda times yeah plus lambda times one, ok. Fine so, equal to 0. So, then what will you end up getting? p 2 is equal to I think it is something wrong here, now this should be minus p 2 by k, right.

So, p 2 is equal to lambda times q 2, ok. And then further actually you can show that lambda is going to be equal to 1, how can you show that your constant is fine. So, dcr how we will get lambda equal to 1, right? So, what does it actually tell you, then p 2 is equal to q 2; that means, all. In fact, you can show that all p is r equal to q is; that means, the distribution q is equal to disturb you.

So, this cross entropy term will be minimized when your true distribution, or when your plated distribution is the same as the true distribution, and hence it captures the difference between the 2, ok. And that is exactly what we were interested in, we were interested in a quantity which can allow us to capture the difference between the true difference between a to distribution and the predicted distribution, right.

So, we have arrived with that quantity and that quantity is cross entropy. So, therefore, for all our classification problems where we have this scenario that we are given the true distribution where all the masses focused on one of the labels, and you are estimating a distribution where you could give nono quantities to many of those. And you want to find out how wrong your estimates were with respect to the true distribution, you can use cross entropy as a measure for that right. So now, your loss function which you wanted to depend on the difference between p and q, it can just be the cross entropy between p and q, ok. Is that clear?