NPTEL

NPTEL ONLINE CERTIFICATION COURSE

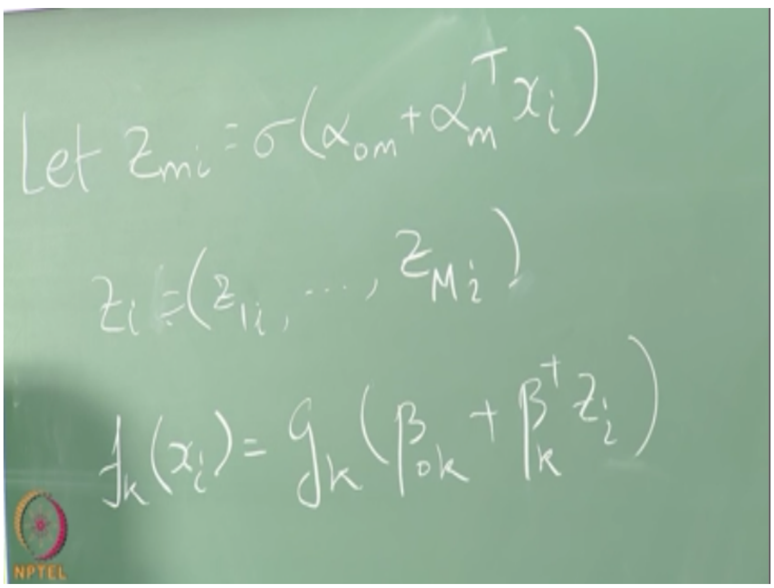
Introduction to Machine Learning

Lecture 34

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

Artificial Neural Networks III-
Backpropagation Continued

(Refer Slide Time: 00:33)



So let Z may correspond to the output of the M_{th} unit in the hidden layer corresponding to the i_{th} input this is not the i_{th} component of the input corresponds to the vector x_i right so it is the i_{th} input in my training data of n elements okay and I am going to say that and Z_i corresponds to the, the entire activation of the hidden layer for the i_{th} input it will find so far right.

So now we got rid of over what did we get rid of here the T right so this is what I was saying in regression Gk is linear and typically d is acting on TK right and so this is acting only on TK so this whole thing is so because we are only talking about the question I got rid of that this will make our life a little simpler when we write the right the gradient so I am going to take the basic

I am going to use gradient descent right so I have squared error I am going to use gradient descent.

So I am going to take the derivative of the error with respect to the single output layer weight okay this is a weight that runs from some noon on Mzm right to some output K right so that is β okay just this one, one weight I am taking here right I am taking the derivative of R with respect to that one weight okay is the setting clear right so I am taking the derivative of R with respect to a single weight here.

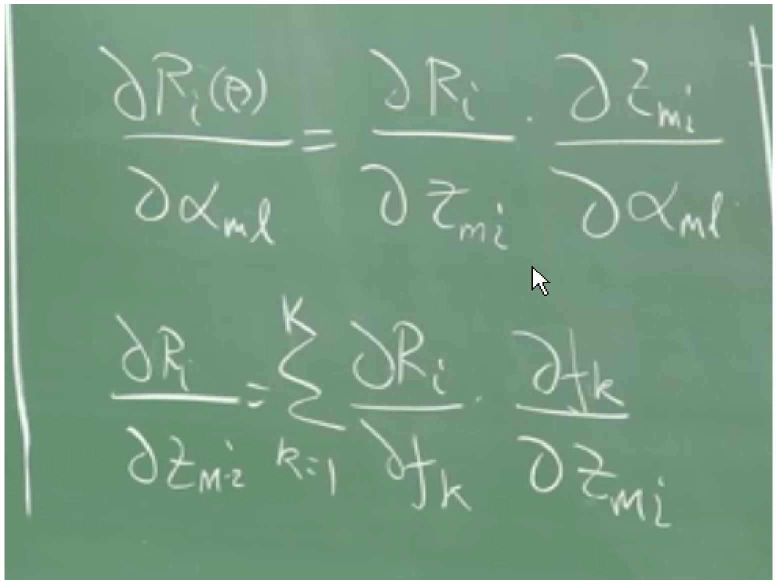
Let us just designate that as β_{KM} so what will this be equal to yeah okay let us do it in a slightly simpler fashion so I am going to assume that each term inside is denoted by RA then I just do the summation over all i okay so that way I do not have to write the summation over all a everywhere so I am going to say this is RA by β right if you remember the earlier that what we had the thing that they raced here there was just this right Y_I-f of X into X was what we had earlier right.

But the input in this case is actually ZM right if you think about what is there on the other end of this weight right so the input that comes from here is actually read them right so mathematically if you think about it just let them β_m okay right so that is what is happening so essentially that is what you are going to get so the i indicates that you are considering it only for the i_{th} input right this is clear so far we just then just taken a derivative right but exactly the same computation that we did earlier the only new thing here is they are the derivative of GK earlier.

We did not have that because we are assuming that GK was linear so GK is linear this will again vanish now comes the interesting part they will just disagreed some the single input layer wait we will consider that so I am calling it alpha m_l how will I take the derivative of the error with respect to α_{ml} you look at the error α does not appear directly at all it appears indirectly so what is the best way to do this name this using the chain rule.

So α is going to affect the output of the hidden layer right and the output of the hidden layer is obviously going to affect the error right so I am going to take the output of the hidden layer right so I am going to chain it through the output of the hidden layers are going to take those a them might do α_{ml} and by 2ZM right so one thing to note is that α_{ml} is going to affect the output only of ZM right it is going to affect only ZM.

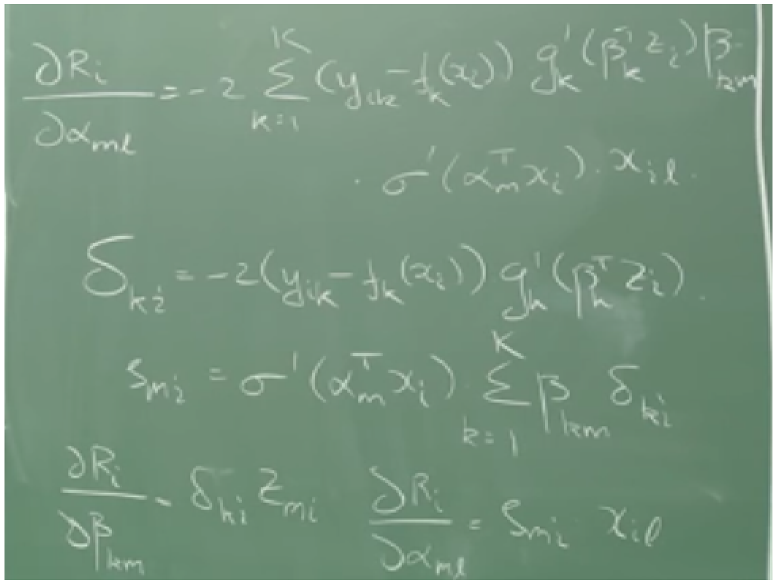
(Refer Slide Time: 08:32)



So I just need to chain through ZM okay is it clear so then let us do each one of these in turn so this is rather easy so is that you have that already so what is the what it what if they did be more consistent okay that makes sense right the derivative of ∑ yeah can you zoom in so the derivative of ∑ times x_{il} right so ∑ prime of α transpose Xi plus α^o into X al so that is essentially the, the derivative of Zma with respect to α_m it is straight forward differentiation if you are having trouble with it I do not know now is the tricky part so I am looking at ∂RA by ∂ZM right.

So what is ZM it is the output from here right but unfortunately this K goes to all the output neurons right so ZM can affect the output through all the output neurons okay so far there is been a single path that we have been considering but at this point we really have to consider all the paths of reaching the output from M okay.

(Refer Slide Time: 12:19)



So what we really have to do is look at okay so, so ZM can affect RI through FK right so the derivative of FK with respect to Zm and RI anybody with respect to FK that is a chain rule again do this over all K because I can have multiple parts of reaching the output so what is ∂R/∂K okay ∂R/∂K it may which should be able to rattle it off just the derivative of GK so putting everything together.

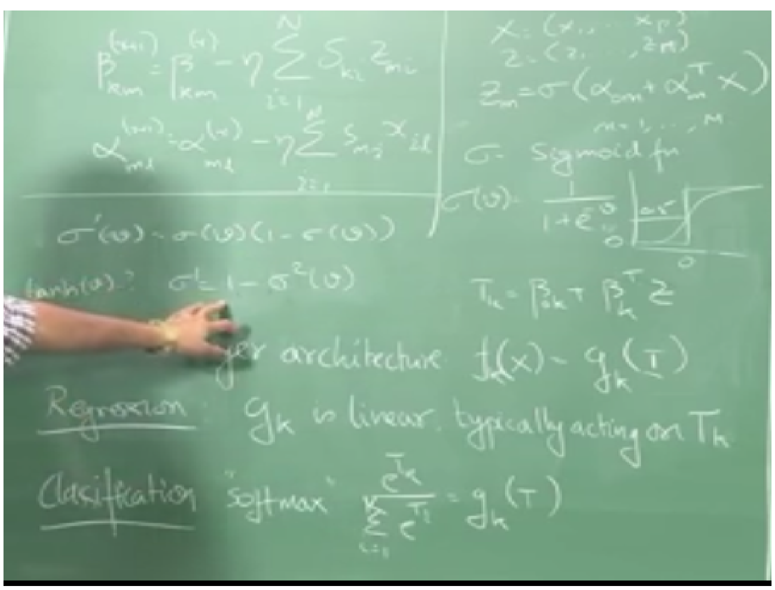
I can write that is a big expression and I did nothing I just took this and wrote it here I took that and wrote it there okay I just took the product of the two terms so what we will do now is just to introduce certain simplifying notations let us think about it I have made my job a lot simpler so that is this term ΔK which ever define so ∂RI with ∂β is essentially ΔK into Zmi right ∂RI with $\partial\alpha_{ml}$ is essentially S_{mi} into X_i that is the Δ part right.

And there you have a β and then you have your ∑prime so this all put together gives me mass S_{mi} so there nothing you just applied chain rule and done some manipulation to simplify this right if you go back and do it again okay you will find that it is very straight forward gradient computation but it took people a couple of decades to nearly a couple of decades to realize that they could do something as simple as this chain rule.

And apparently this technique which is very popularly known as back propagation so why is it called back propagation so when you take the input right and you compute the output that you are propagating the values forward through the network right but when you are updating the gradients so if you think about it so what you are doing is first you are computing the Δ’s right and then you are propagating the Δ’s back through that weights β’s right.

So essentially what you are doing is Δtimes β it like when you are going forward you do x times Δ and Z times β right so here likewise you are doing something like Δ times β right so this is something like a back propagation of this Δterm through the weights so as to update the first layer weights right so that is why it is called back propagation okay so the forward thing is whatever you do this, this is the forward pass okay and the equivalent backward passes are given by that right so the actual equations are right.

(Refer Slide Time: 18:47)



So we still left some things in there so I left a Gprime and a ∑ prime and so on and so forth so if G is your linear function great right what about ∑ prime ∑, ∑ is the sigmoid function then now you can take that derivative of the ∑ with respect to X and that is what you will get and if it is at tan H right instead of the sigmoid if I use the tan H function then my ∑ prime will be $1-\sum^2 V$ you can work it out but sadly easy differentiation always people get thrown off my back propagation but it is really nothing but differentiation and a lot of algebra right just manipulating things around it is nothing more than that everyone knows the chain rules right that is it that is it, it is just a chain rule.

IIT Madras Production

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

www.nptel.ac.in