NPTEL

NPTEL ONLINE CERTIFICTION COURSE

Introduction to Machine Learning

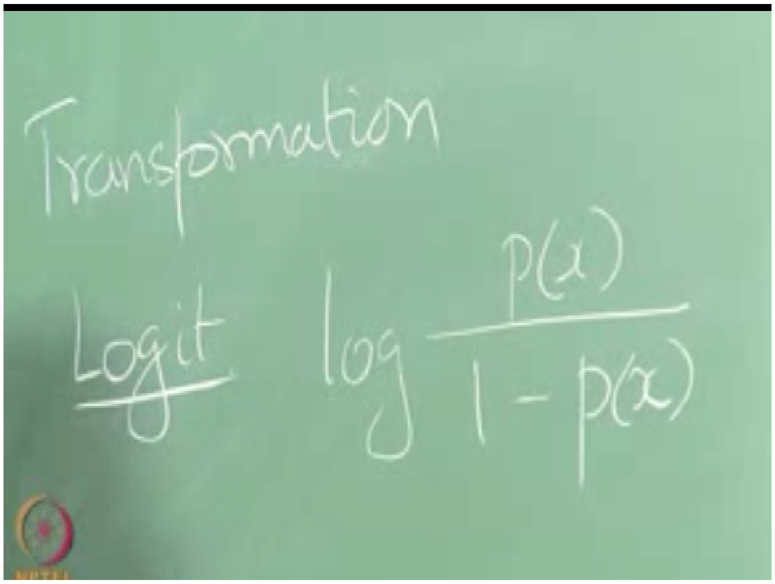
Lecture 20

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

Logistic Regression

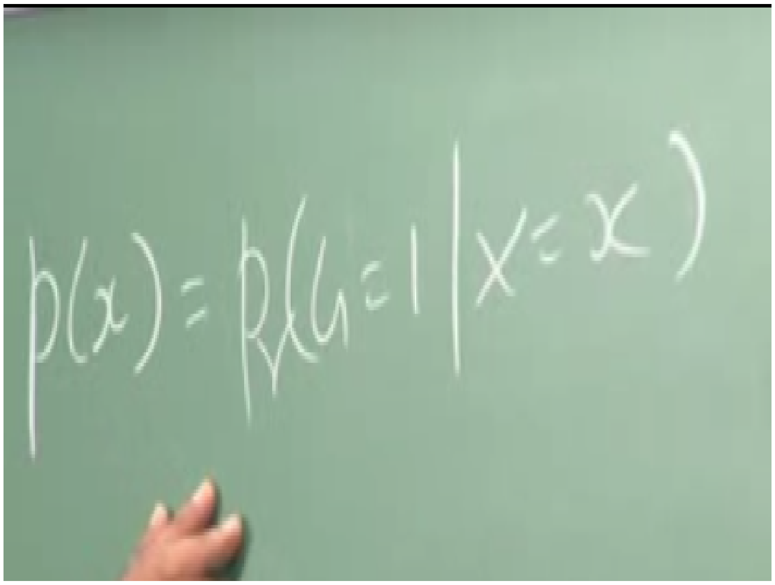
So let us go back to whatever has been bothering all of us, so what we are essentially choosing when we did regression here was we are going to make sure that the output is either 1 or 0 right and then we are trying to regression that right so what do you want our function to do basically you want your f(x) to give you probability of K given X right but then trying to do that is a little harder so what we are going to do is you are going to look at some kind of a transformation.

(Refer Slide Time: 00:49)



Of the probability and we are going to try and fit that, okay. Let me look at the logit transformation is essentially log of so if you think of P(x) and 1 –p(x) is the probability that X is okay let me put it this way.

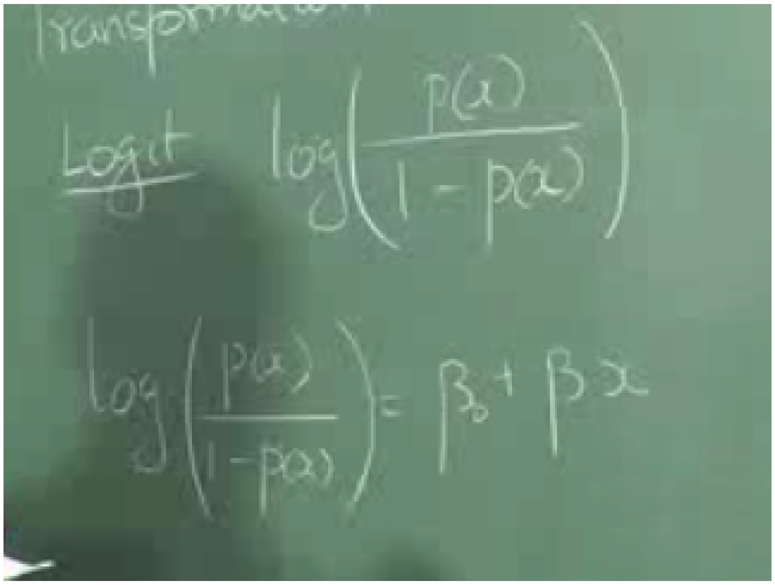
(Refer Slide Time: 01:29)



$$p(x) = Pr(y = 1 \mid X = x)$$

To make my life easier for the next few minutes I am going to assume we are dealing with binary classification okay, so the class labels is either 0 or 1 okay and P(x) is essentially okay. So when I say P(x) is essentially probability that the output is 1 given the input is X, okay. So this is makes my life a little easier when I write the next part.

(Refer Slide Time: 02:12)



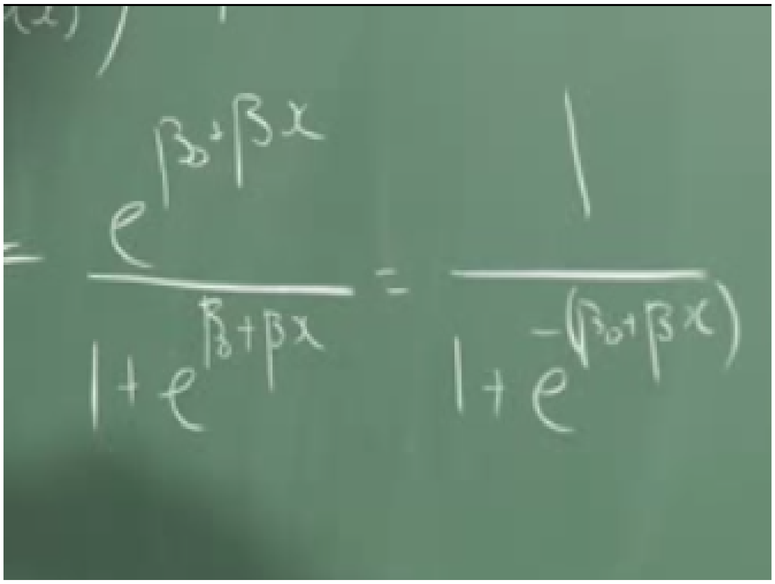
Transformation

Logit $\log\left(\frac{p(x)}{1-p(x)}\right)$

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta x$$

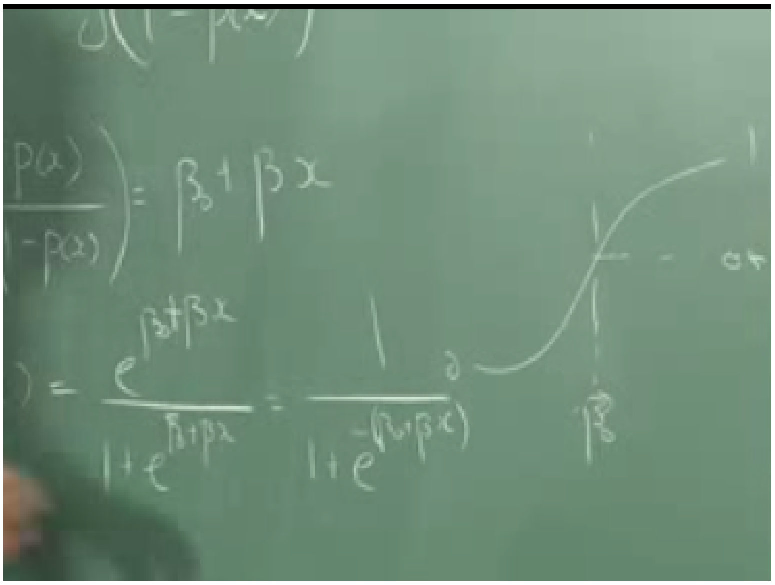
So given that P of X is the probability it is 1 what is 1 - P(X) probability 0 we are talking about binary classes right so this is sometimes called the this is you know what the probability of success divided by the probability of failure odds right, so this is sometimes called the log odds function okay or the logit function okay so this is essentially the transformation that we want to look at so what I am going to do is I am going to try and what is that okay. I am going to try and fit a linear model to the log odds, okay. So what does P(X) in this case right.

(Refer Slide Time: 03:37)


$= \frac{e^{\beta_0+\beta x}}{1+e^{\beta_0+\beta x}} = \frac{1}{1+e^{-(\beta_0+\beta x)}}$

Towards this function going to look like that is a sigmoid right so essentially we are saying that my probability of X is going to given by probability that X is 1 will be given by a sigmoid right. Term in there it is not possible it is a term in this power visible βX.

(Refer Slide Time: 04:54)



Is this term is visible, okay that is the term in the power except that here it is a minus before that okay everyone got that other room things are visible yeah because they are probably zooming in on this right, please so what do we have here so what we do if P of X is greater than 0.5 we will output this one P of X is less than 0.5 will output it as 0 right. So is it okay because even though I am doing linear regression and linear regression is unbounded I am going to plug it into this expression and therefore this will make sure that my probability is between 0 & 1, right. What is that point this 0.5 depends on what I had put for my β0 right.

No β0 is 0 then it will be 0 right –β0 yeah so what about the classified that I am learning here what is about the separating surface the decision boundary between class one and class two what is that it is complicated okay think you have X = 0.5 but what does it look at the expression that we have here right so in P of X equal to 0.5 what will this be 1 we have X of the point 5 will be 1 log of that will be 0.

So essentially the thing is β0 + Bx = 0 right that is the straight line I mean assuming X is uni-dimensional that is a straight line right so even though I did something complex looks complicated okay I used an exponential to define my probability the decision surface turns out to be still hyper plane right, so plug in P of X equal to 0.5 here so I'm going to get 0 on the left hand side I am essentially solving β0 + βx = 0 right that is just a straight line assuming that yeah it is a hyper plane maybe I should do the whole class in one dimension it makes it easier for people to visualize things anything so one thing I should point out is that logistic regression looks simple right but it yields a very powerful classifier it works very well in practice.

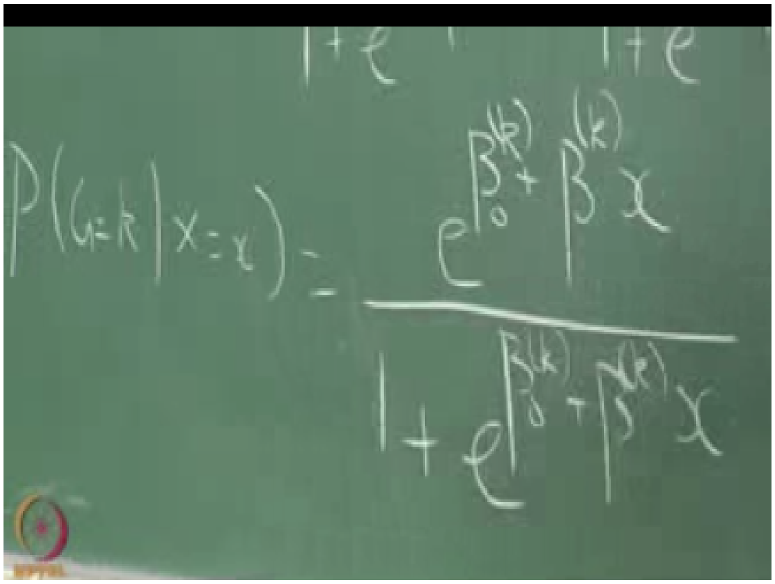
And it is used for not just for building classification surfaces okay but it is also used a lot in what people sometimes call sensitivity analysis right so they look at how each factor contributes to the output right so how each how much is each factor important in predicting the class label so for doing that they do logistic regression and then they look at the β vector and figure out how much each variable is going to be contributing to the output so people use that a lot I mean of course you can use anything that we have seen for doing this kind of sensitivity analysis I am just telling you what people use in practice okay.

So logistically aggression is something that is used vary widely in practice both by machine learning folks and by statisticians in fact when I work with few doctors right it was almost impossible to get them to accept anything else other than logistic regression as a valid classifier because they were so sold on logistic regression and with good reason because it does work very well in very well in practice right okay.

So that is for two classes, so what do you do for multiple classes? So multiple classes I'm essentially going to take recourse to this form right I am going to say keep the probability that the output is class 1 given X is given by an expression like this the probability that the output is class 2 given the input is X is given by another expression like this where which will have a different set of β0 and b β right.

Likewise for every class I am going to say is given by a different set of β naught and β so they have to do that for all the K classes I have to do it only for K minus 1 classes because the kth class probability will be automatically determined right so so I will have to have K minus 1 sets of β if I have K classes I have to figure out how to estimate those.

(Refer Slide Time: 10:52)



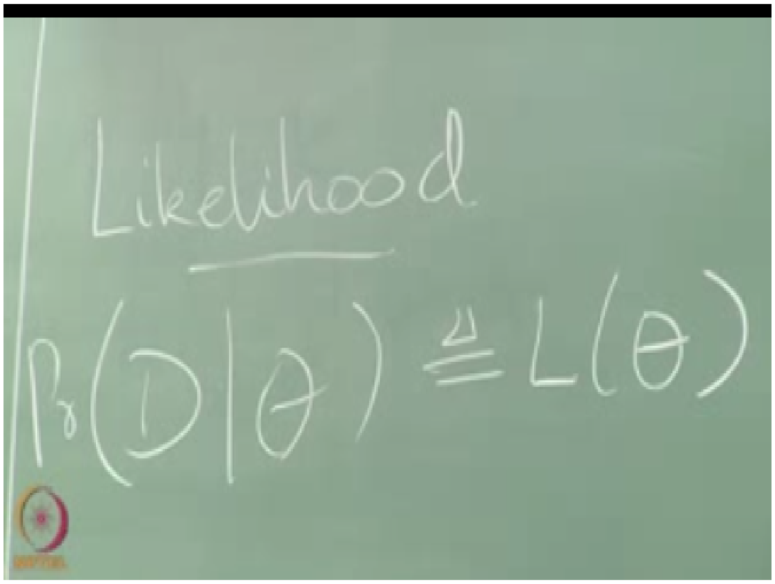
$$1+e \qquad 1+e$$

$$P(G=k \mid X=x) = \frac{e^{\beta_0^{(k)} + \beta^{(k)} x}}{1 + e^{\beta_0^{(k)} + \beta^{(k)} x}}$$

Thanks so we are going to write like this for that K minus1 classes by convention either the first class or the last class whichever arbitrary numbering you choose either the first class or the last class the coefficient is set to zero setting the coefficient to zero will essentially give you the answer that you want right, Oh sorry an I am really sorry I missed up so this is the expression that you want right so for every class K the numerator will have just that classes variables and the denominator is going to have all of it okay.

So now you agree with me setting it to zero is fine so how do we estimate the parameters for logistic regression so the little tricky since we are anyway trying to model directly the probabilities right so what we are going to try and do is maximize the likelihood okay of the data so far we have always looked at some kind of error function and we have been trying to optimize the error function within those linear regression we looked at squared error and then we did the optimization and soon so forth but here we are going to look at a slightly different criterion we are going to optimize the likelihood of the data.

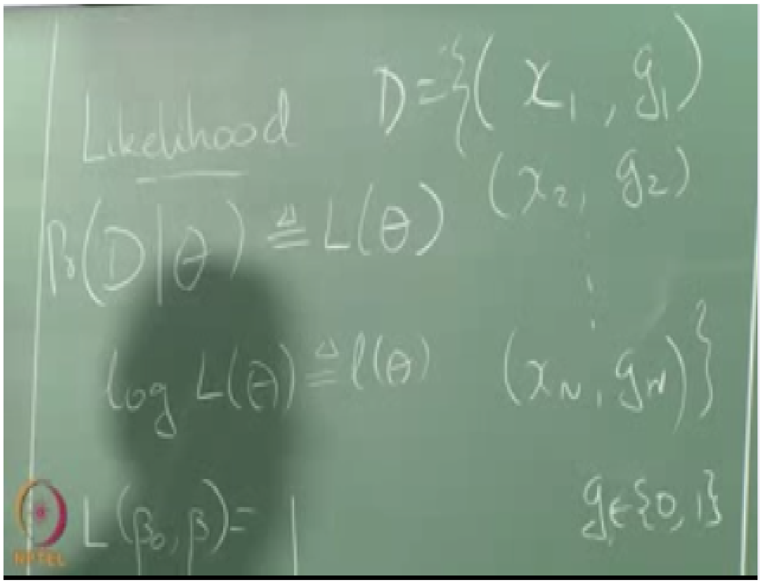
(Refer Slide Time: 13:14)



So just to keep it together I am going to do this today but I have a whole session planned on maximum likelihood and other ways of estimating parameters so when we come to that I will do maximum likelihood in more detail in a generic form right so right now I will just look at let us take an equation and maximum likelihood so what is likelihood suppose I have some training data D the training data has been given to me some training data D so the probability of D given some parameters theta the probability of D given parameters θ is known as the likelihood of D or θ okay.

So D is fixed right think about it I am given a training data D, D is fixed what is it that I am actually looking to find θ right so this I will write as likelihood of θ okay so we are always used to thinking of something that comes after the slash as the conditioning variable and the one that comes before the slash is the actual argument in this case it turns out that theta is the argument okay the probability of D given θ is the likelihood of θ makes sense right D is fixed I am really trying to find what θ is correct.

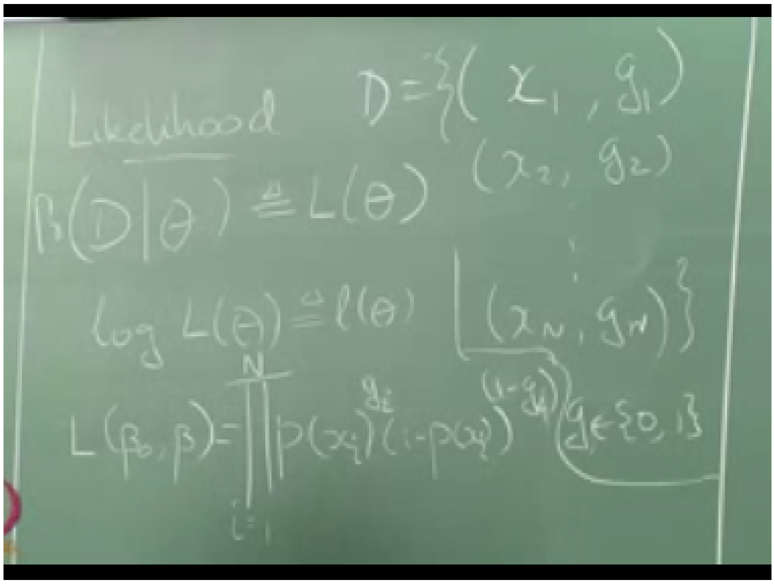
So finding the likelihood of thickness so the scoring function should be on t θ okay does it make sense and I am usually interested in the log of θ .
(Refer Slide Time: 15:05)



Because it allows me to simplify a lot of the distributions that I will be considering and we will denote this by lowercase L mostly so what is the likelihood so in our case θ is our β so my input my D is going to consist of right We are going to consist of pairs of data points like this right so X is the input G is the output we are talking about classification so G is the output X is the input right so what is the likelihood.

So I am remember I am I wanted to stay in the two class domain for y so G is either T belongs to 0 or 1 so 0 means is class 0 1 means is class 1 okay.

(Refer Slide Time: 16:41)



Likelihood

$P(D|\theta) \triangleq L(\theta)$

$\log L(\theta) \triangleq \ell(\theta)$

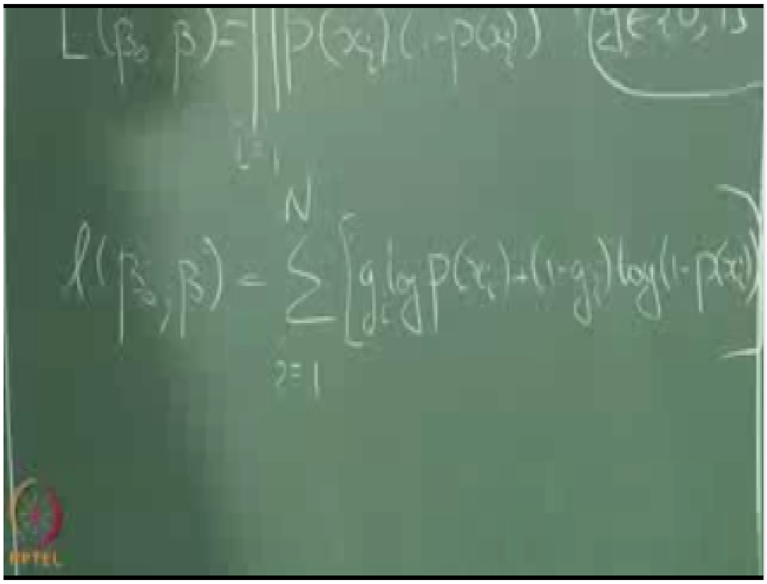
$D = \{(x_1, g_1), (x_2, g_2), \ldots, (x_N, g_N)\}$

$L(\beta_0, \beta) = \prod_{i=1}^{N} p(x_i)^{g_i} (1 - p(x_i))^{(1-g_i)}$ $g_i \in \{0, 1\}$

This is a funky expression pre-written we will come back to this we will see this again so this is the probability of one pair XG occurring okay so what is this is the probability that the X as a label 1 this is the probability that X as the labels 0 okay, and what is this this is the actual label off this is the actual label of X right if the actual label of X is 1 then this term will appear in the equation if the actual level of X is 0 then that term will appear in the equation this will become1 right so if the actual level of X is 1what should be the probability.

Probability that X equal to 1 right that is what I should be looking at so that is what this is the actual label is 0 then I should be looking at the probability that x is 0 that is what this term is right so you can see that this gives me the probability of 1 XG pair I do this for all of them assuming that they are all sampled independently right because I am assuming they are independent I can take the product. So now we know why we love logarithms right.

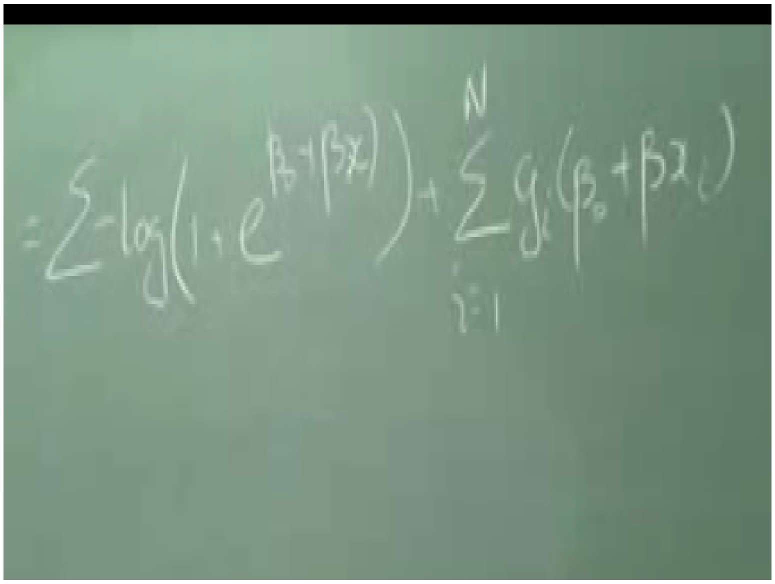
(Refer Slide Time: 18:26)


$L(\beta_0, \beta) = \prod_{i=1} p(x_i)(1-p(x_i))$ $g_i \in \{0, 1\}$
$\ell(\beta_0, \beta) = \sum_{i=1}^{N} [g_i \log p(x_i) + (1-g_i) \log(1-p(x_i))$
NPTEL

People who cannot see the board see the TV right, so that is the expression that is simple enough so now comes the interesting part we want to do what we want to maximize likelihood right so we need to take the derivative of this and equate it to zero it's fine right because log is a monotone transformation we can take the derivative of the log right nickel equated to zero and then solve for beta unfortunately life is not so simple okay.

Let us try and do the simplification which I am multiplying this out and gathering the terms okay right a multiplying this out I gather the terms here and we know what that is what is that yet we already know that so that we can insert that and simplify that there and what about this guy one minus this right I can again write it in a simpler form write log of one minus that will give me okay.

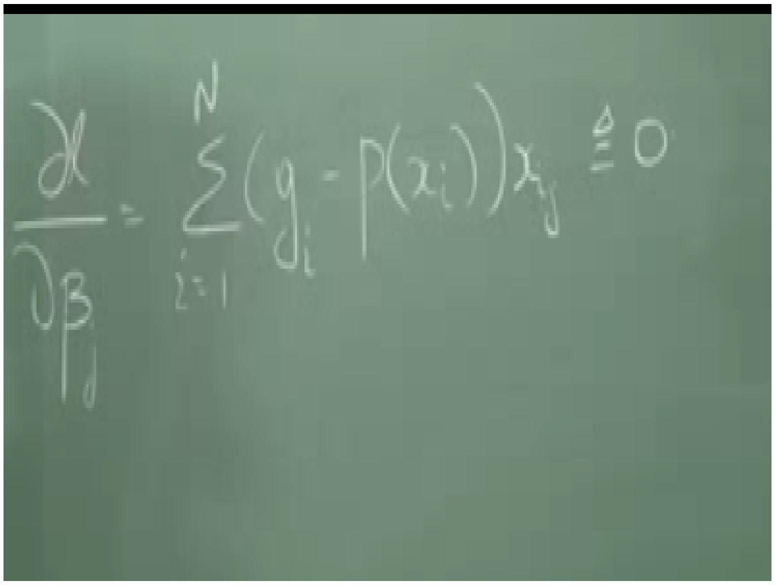
(Refer Slide Time: 21:33)



$$= \sum -\log\left(1 + e^{\beta_0 + \beta x_i}\right) + \sum_{i=1}^{N} y_i(\beta_0 + \beta x_i)$$

So now I can take a derivative of that with respect to β and equate it to zero what do I get.

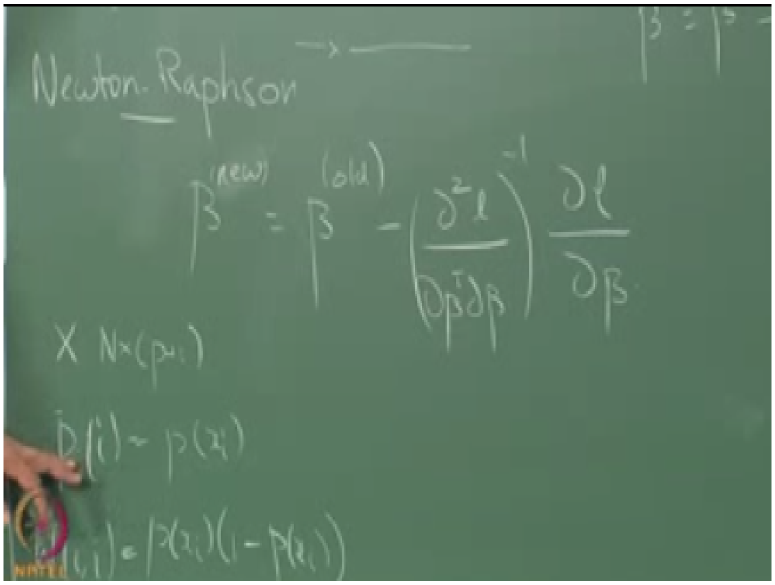
(Refer Slide Time: 22:11)


$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{N} (g_i - p(x_i)) x_{ij} \triangleq 0$

So take the derivative of this term right you are going to end up with minus P right so this will go down right and I will get e $^{\beta\ 0\ \beta\ x}$ as the numerator right I will get minus P times X since I am doing it with respect to a specific β J I will get X IJ does it make sense so this first term I will get minus P times X IJ and this one if I did I mean if I take the derivative of that I will get GI times XIJ okay.

So that essentially what I have done here looks like a nice and easy expression to solve but unfortunately it is not so because this will it is an exponential function here so it is not really easy to solve this you have to look at some other iterative method for solving this and the most popular method used is Newton-Raphson I am NOT going to go into the depths of Newton-Raphson right people are more interested you can look it up.

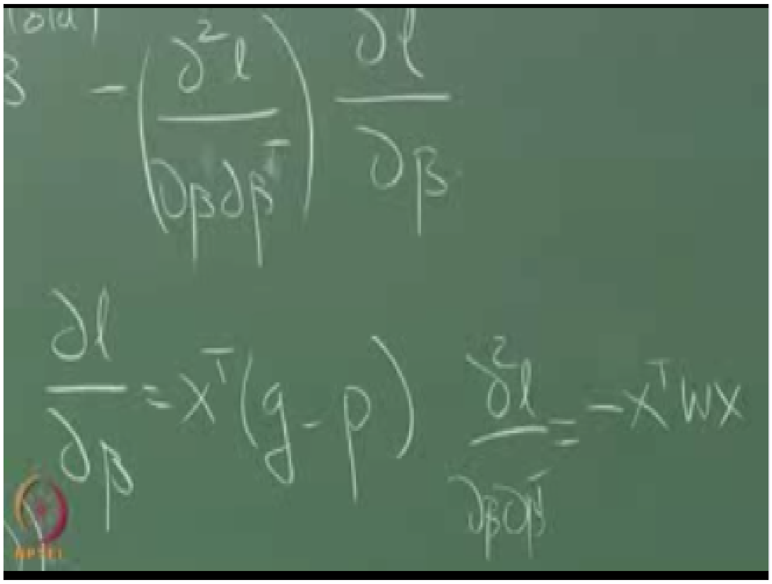
(Refer Slide Time: 24:11)


Newton-Raphson
$\beta^{(new)} = \beta^{(old)} - \left(\frac{\partial^2 l}{\partial \beta^T \partial \beta}\right)^{-1} \frac{\partial l}{\partial \beta}$
X N×(p+1)
$p(i) = p(x_i)$
$(i,i) = p(x_i)(1 - p(x_i))$
NPTEL

But the basic idea is that so people were more comfortable looking at it this way so take the whole estimate of your valves or your taking whole solution okay and look at the first derivative of the function that you are maximizing right L’/L’’ okay so you adjust this by that okay so that is essentially the basic idea behind Newton-Raphson I am just defending some terms here so X is going to be my n x p + 1 matrix as usual right my P is going to be a vector where each entry is going to be the probability of X I.

So it will be the dimensionality of P n right so it is a it is a n vector that tells me what is the probability of each X I being one right so that is that is P and W is going to be a diagonal matrix where each diagonal entry is P into 1 minus P right for that particular at a point X I so this makes it convenient to rewrite things and I am going to assume that V right is the vector of outputs like zeros and ones depending on what class it is.

(Refer Slide Time: 24:46)



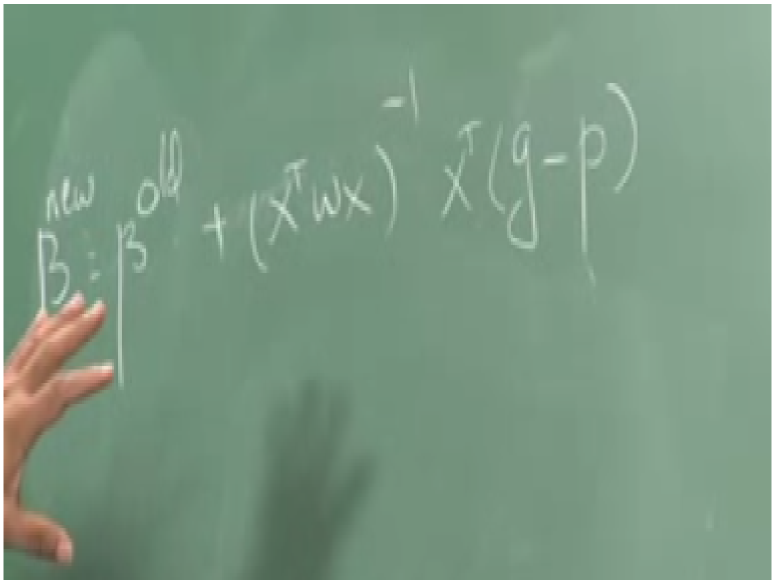
$$\beta - \left(\frac{\partial^2 l}{\partial\beta\,\partial\beta^T}\right)\frac{\partial l}{\partial\beta}$$

$$\frac{\partial l}{\partial\beta} = X^T(g-p) \qquad \frac{\partial^2 l}{\partial\beta\,\partial\beta^T} = -X^T W X$$

So I can write my ∂ l as what you know what ∂ l/ ∂ β is, it is G – P x X right in terms of the matrices it is a P is this vector Y is the vectors of zeros and ones corresponding to the class label right and X is my input right so I am just basically written this in vector notation right so you already found the derivative I have just rewritten it in vector notation that it make sense okay, so what about the second derivative so I am so I am not going to work it out.

But it is X^{-T}WX okay and so W is essentially the diagonal matrix with this entries okay so that is my second order I made my second derivative so what do I get now putting this together.

(Refer Slide Time: 28:40)



The beginning look something like regression here alright you are getting your $X^{-T} X^{-1}$ X^{-T} and all that so we just have to do a little bit more work little bit of algebra to make it look more like regression so that Is what we will do now.

I just substituted this the derivatives here okay nothing, yeah so okay so you want to solve this problem right this is what we really want to solve we want to solve when this becomes 0 right so you can see the β in here right no yes the β is in the P right so the right so I have erased the free off now but so β is in their P is the P is e $^{\beta 0}$ + β x/ 1 + e $^{\beta 0}$ β is here I really want to solve for this right I want to find the 0 of this function right.

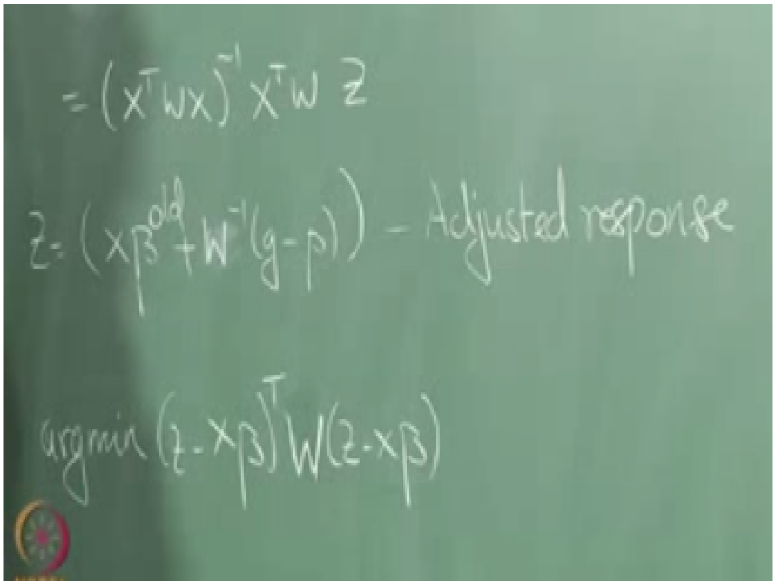
But it is not easy to do because of the fact that we have it exponential in there right so what we have to do is look at some kind of iterative method for solving this problem and so what the way we do this iterative approaches you start off with a guest called β old okay and then you do some computation you get a new guess call bit on you right so one very popular way of doing this kind of iterative thing is to do gradient following have you have you looked at that.

I mean you must have might have come across that this side so suppose I have a function like this right I am here this is my current solution right this is I will call this X old okay, so now I will compute the gradient here right and I will move in the opposite direction of the gradient to find the minimum right so instead of going all the way I can go a small step that gives me the X new right.

Normally what you will do is you will find the gradient try to equate it to 0 and get it but it can do this in iterative fashion also right you can take small steps in the direction of the gradient so likewise what we are going to do is we will start off with β whole which is some guess for this okay in fact β of all 0 actually works fine okay can start off by saying all my β at 0 okay and then try to find a β new.

So what I will essentially be doing is I will find so β 0 will put me somewhere here on the L function right I will find out what is the first order and the second order derivatives at this point with respect to β and then use that for changing my β values right so people agree with me so this is the $X^TW\ X^{-1}$.

(Refer Slide Time: 32:08)


$= (X^TWX)^{-1} X^TW Z$
$Z = (X\beta^{old} + W^{-1}(y - p))$ - Adjusted response
$\text{argmin} (z - X\beta)^T W(z - X\beta)$

If you take the product here this will be X^T W X so that is just the identity right and I take the product here I will get X^T W^{-1} back this $W^{-1}$and will get cancelled out right so I have just done some algebra to get it this way think about it what is X β whole, so what is original I mean since it is like linear regression right it is like the original response I will get if β whole is my variables and I am actually prime making a linear prediction based on X right.

So the X β whole is this right and I am essentially adjusting it you think this quantity so this is the prediction I make with my old parameters this is some kind of adjustment I am making to the prediction so this is called the adjusted response right and this turns out to be the solution of something known as weighted linear regression it is in weighted linear regression essentially what you do right.

So in linear regression that is what you are trying to minimize that is a square error right so linear regression this is what you are trying to minimize weighted linear regression you essentially have a waiting term in your error function right since I am just saying I am going to minimize the squared error for every term in the squared error I am going to assign a different weight right so for some data points I want to be more aggressive in minimizing the for some data point I want to be less aggressive.

So data points in which I have to be more aggressive I will have a higher weight for data points for which I want to be less aggressive I will have a lower weight so that will allow me to trade-off the importance of data points this is idea behind weighted linear regression right so this is essentially weighted linear regression so minimize be the minimize of this right it is this okay you can do the usual now you can take the derivative set it to zero and solve it this is easy enough to solve this is actual linear regression right.

So the minimize is $X^T W X^{-1} X^T W$ into Z right so essentially what we are saying is your β the β ν is essentially solving a weighted linear regression or weighted least squares problem okay are solving a weighted least squares problem with this adjusted response so this is called iterative rebated least squares this Is a separate algorithm called iterative rebated least squares for solving logistic regression.

But all it does this essentially does Newton-Raphson essentially is doing Newton-Raphson but the way iterative rebated least squares is described to you is okay start off with a guess for β okay form the adjusted response okay as soon as I guess as soon as I have a value for β I can find out what my P is okay so G is given to me already in the data and my W can be constructed once I know PI make a guess for β I construct my PI construct my W okay.

Form this adjusted response solve this weighted least squares problem get a new β keep repeating this until my predictions are accurate enough okay so that is basically this is it is the most popular way of solving logistic regression but there are many other ways people have come up with more efficient ways of solving logistic regression actually and but if you pickup any popular package like R or something so IRL is the base logistic regression solver that would be implemented okay. So this just to give you a flavor of how hard it can be to optimize things sometimes.

IIT Madras Production

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

www.nptel.ac.in