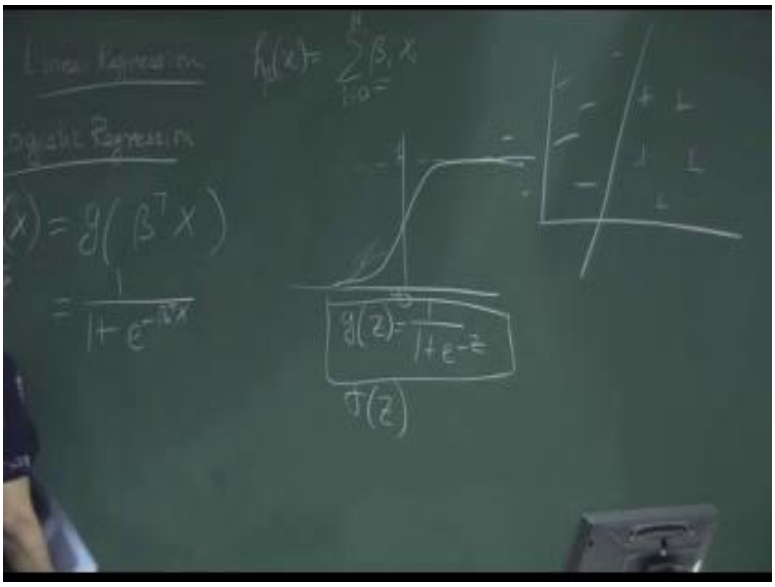
Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module - 5
Lecture - 20
Logistic Regression

Today in this module, we will study about Support Vector Machine and also before that we will have a brief lecture on Logistic Regression. So, this is part A on Logistic Regression. So, in previous class in the second week, we have talked about linear regression which is used for our regression problem.

(Refer Slide Time: 00:37)



But, if you have a classification problem you cannot use linear regression. We want to see what is a simplest way in which we can handle our classification problem? In a linear regression, we had our hypothesis function h (x) as sigma beta i x i, i equal to 0 to n when n is the number of a predictor, number of variables. So, we have h (x) i equal to beta 0 plus beta 1 x 1 plus beta 2 x 2 plus beta n x n and learning involves learning the values of this beta i in order to optimize a certain function, for example, we try to minimize the sum of square errors.

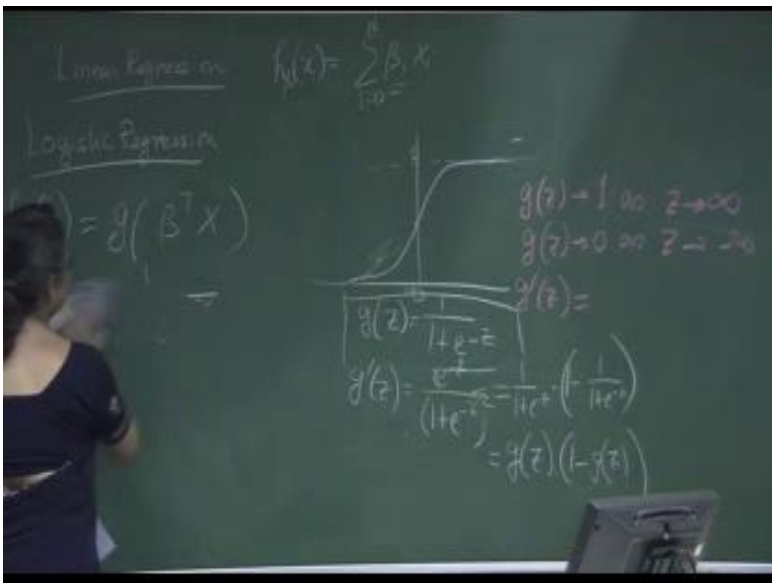
Now, suppose we have a classification problem that is we have different training points and they are positive and negative and we want to have a classification of them, we want to say when they are a positive and when they are negative.

Now, this function will give a real value and it is not appropriate for classification, but what we can do is that based on this linear function, we can apply another function on this linear function so that we can use the result for classification. So, one of the ways we can do it is, in logistic regression we use the logistic function or the sigmoid function for this task. Now, first of all let us look at what is the logistic function or the sigmoid function it is given by this formula g (z) equal to 1 by 1 plus e to the power minus z and this formula this function has the following profile.

Suppose this is 0, this function has this type of shape, roughly this as this type of shape. The value of this function varies between 0 and 1 at z equal to 0.5 the value is 0; as z tends to infinity, the values tends to 1; as z tends to minus infinity, the value tends to 0 right. So, this function gives the value between 0 and 1 and how we can use it for classification, we can say that we have a function if the output is greater than 0.5 it is positive, if it is less than 0.5 it belongs to the negative class.

So, just like in regression we use this function for classification using logistic regression we will use this function h beta x is; we will apply this function g this function is also called the sigmoid function, we can refer to as sigma z. So, we will apply this function g or sigma to sigma beta i x i. So, sigma beta i x i can also be written as beta transpose x more compactly using matrix notation, we can write h beta x for classification as equal to g beta t x which is equal to 1 by 1 plus e to the power minus beta t x. So, we can use a linear function of beta, pass it through the sigmoid function and use it as for a classification function. Now, let us looks at certain properties of this sigmoid function which makes it very attractive to use.

(Refer Slide Time: 05:28)



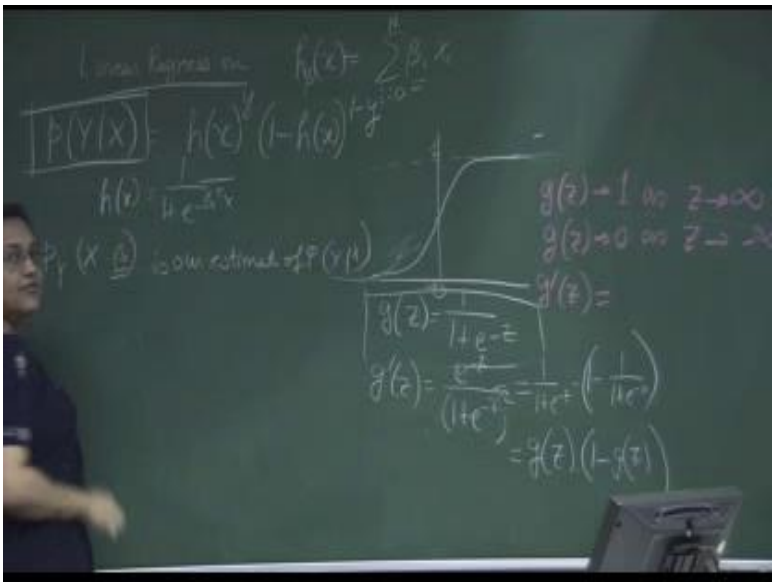
So, as we have seen g (z) tends to 1 as z is after sometime it becomes 1 as it asymptotically stays at 1. So, as z tends to infinite g (z) tends to 1 and g (z) tends to 0 as z tends to minus infinite then a very attractive feature of this function is if you take the derivative of this function . So, the derivative of this function you can take as, you have 1 by 1 plus e to the power minus z let us take the derivative. So, g dash z equal to d d z of 1 by 1 plus e to the power minus z which is 1 by 1 plus e to the power minus z whole square into e to the power minus z.

So, we do a change of variable this is 1 by x which is 1 by x square and by taking the derivative of this part we get e to the power minus z. So, g dash z is e to the power minus z times 1 by 1 plus e to the power minus z whole square and we can simplify this to write it as we can do some manipulation and we can write this as let me write it where you can see. So, this can be written as 1 by 1 plus e to the power minus z times 1 minus 1 by 1 plus e to the power minus z which is simply equal to g (z) into 1 minus g (z). So, the derivative of g (z) can be written as g (z) times 1 minus g (z). So, the derivative is extremely simple to compute and this is a property which makes it attractive to use this logistic function or sigmoid function.

So, when you are using this logistic function based on this we can look at the conditional

distribution of generating the data. So, suppose you have the input x you want to find out what is the probability of y given x, if y equal to 1.

(Refer Slide Time: 08:14)



So, we can write this as h (x). So, if y equal to 1 h (x) is the probability and if y equal to 0 the probability is 1 minus h (x). So, we can write probability y distribution of y given x as h (x) to the power y 1 minus h (x) to the power 1 minus y if y is 1, 1 minus y equal to 0 and this factor will not be there this factor will be 1. So, we have h (x) if y equal to 1 and probability y given y given x probability y equal to 0 given x is here y is 0 and 1 minus y is 1. So, we have 1 minus h (x). So, it can be given here and h (x) has we have seen equal to 1 by 1 plus e to the power minus beta transpose x.

Now, given this function we can now try to learn this function by using gradient assent just like. So, we want to maximize this function and we can use the gradient a decent method, asset method as we have used in linear regression. So, in logistic regression we need to learn the conditional probability distribution probability y given x. So, this is what we need to learn.

Now, suppose our estimate p y (x) beta are the parameters p y (x) is our estimate of probability of y given x and beta is the vector whose value the beta are the set of

parameters whose values we have got to learn . Now, what we will do is that we will do stochastic gradient descent for that we assume a single training example and with respect to the training example we will do the radiant descent. So, in order to do that we first define the likelihood of the data, what we have to do is.

(Refer Slide Time: 11:15)

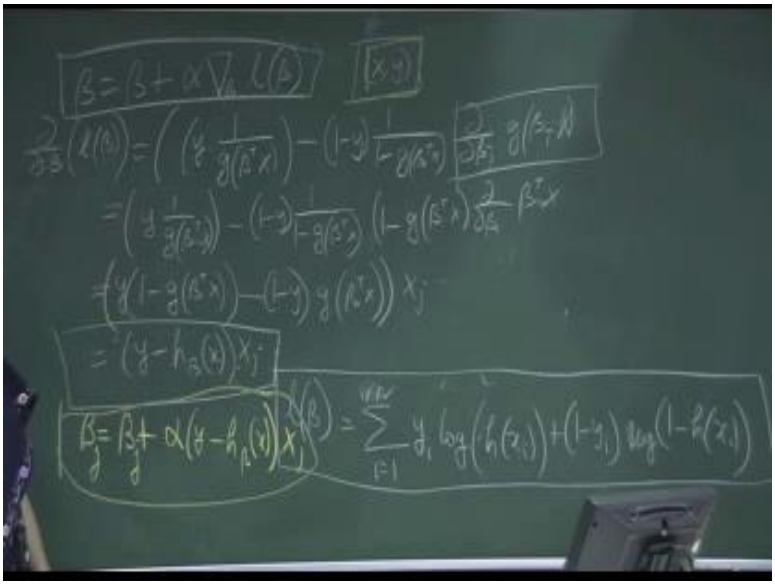
We have to learn the optimal values of beta and we use the maximum likelihood approach we find out what is the likelihood of beta. So, the likelihood of beta is the probability of observing the data given beta were the actual parameters. So, the likelihood of beta can be written as probability of the y given x parameterized by beta and because we have m training example this is product of for each training example, we find the probability of y i given x i beta.

And this as we have seen is product of i equal to 1 to m h (x) i to the power y i times 1 minus h (x) i to the power 1 minus y i. So, we have got to find beta. So, that this expression is maximized now whatever maxi. So, all the probabilities are positive. So, whatever maximizes this expression will also maximize the log of this expression and in order to make our computation simpler, we take the log likelihood of respect to beta.

So, small l of beta is the log likelihood given beta and it is given by log of likelihood of

beta, which is summation i equal to 1 to m y i log of h (x) i plus 1 minus y i log of 1 minus h (x) i. So, this is by taking logarithm of this expression we get the likelihood of beta. Now, we have to maximize this likelihood and in order to maximize this likelihood we do gradient asset. As we know that when we try to maximize this function we can take the derivative, let me rub this. So, we have this function here, which we want to maximize.

(Refer Slide Time: 14:27)



So, what we will do is that, we will take the derivative of this function and so, we will start with some initial value of beta. We will start with initial value of beta and we will update beta as follows beta equal to initial beta plus some alpha, alpha is the learning rate times the derivative partial derivative with respect to beta of the log likelihood of beta. So, this is how will update beta iteratively by using gradient asset and we can do it 1 example at a time if you are using stochastic gradient asset.

Suppose we have a single training example x y. So, x y is a single training example based on this training example and the current beta we want to find out what the next beta will be for that we find the derivative of this likelihood and we try to based, we find the derivative and we take a step towards the derivative. So, let us take the derivative of this expression. So, if we take the partial derivative of l beta, which is the function that we

have given here and what we get is y times 1 by g beta t x minus 1 minus y 1 by 1 minus g beta t x times del del of beta j g of beta t x. So, on simplification, on manipulation of this, what we get is y 1 by g beta t x minus 1 minus y 1 by 1 minus g beta t x times, expanding this part, we get 1 minus g beta t x del del beta j beta t x from which we get y times 1 minus g beta t x minus 1 minus y g beta t x times x j.

Now, we have use the fact that g dash x is equal to g (z) times 1 minus g (z) and we get y minus h beta x x j. So, upon simplification we get this as the partial derivative of the log likelihood of beta y minus h beta x x j. So, plugging in this formula here what we get finally, is beta equal to original beta plus alpha times, we had partial derivative with beta with respect to l beta which we get is this, alpha times y minus h beta x times x j. So, this is the value of the jth component of beta. So, beta j equal to beta j plus alpha times y minus h beta x times x j this is the change that we make for a single training example x j a x y.

So, given a training example x y, we do the partial derivative of this and we find out this is our like log likelihood of beta, we take the partial derivative of the log likelihood of beta and we have worked this out here after some manipulation what we get that it is equal to y minus h (x) times x j and plugging it in to this formula we get, how we can update beta j beta j is the jth component of beta beta j equal to beta j plus alpha times y minus h beta x into x j. So, this is the formula by which we can do stochastic gradient has not and we will find the right values of beta which we can use for logistic regression. With this we come to the end of today's lecture.

Thank you very much.