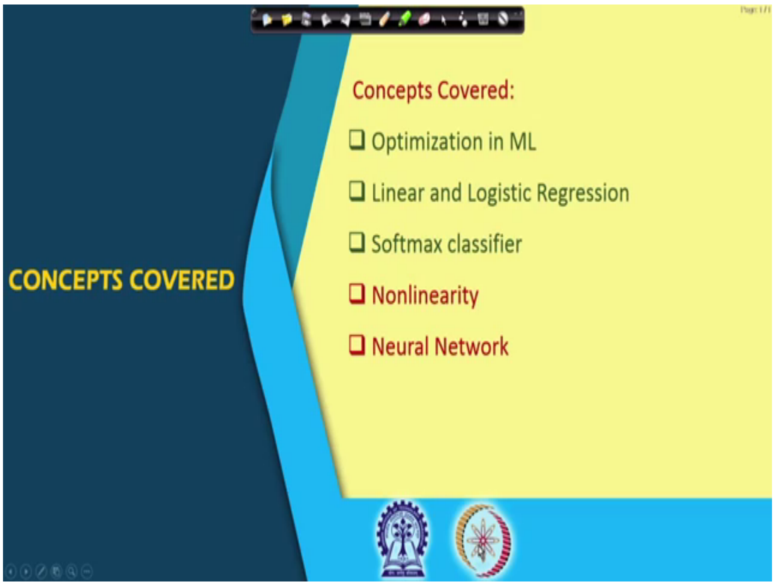
Deep Learning
Prof. Prabir Kumar Biswas
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 18
Nonlinear Functions

Hello welcome to the NPEL online certification course on Deep Learning.

(Refer Slide Time: 00:37)



CONCEPTS COVERED

Concepts Covered:

❑ Optimization in ML

❑ Linear and Logistic Regression

❑ Softmax classifier

❑ Nonlinearity

❑ Neural Network

So, you have seen that in previous lecture we have talked about the optimization problem, particularly optimization in machine learning and we have discussed how optimization in machine learning is different from the general optimization tasks. We have also talked about the linear and logistic regression.

So, in case of linear regression we have seen that a variable y, our dependent variable y is predicted using a linear function of the components of the feature vector y or feature vector x or we have written y in the form of y hat. The predicted value of y as y hat as w transpose x where x is the feature vector and w is the weight vector and in case of linear regression we have said that based on this predicted value of y, you take certain decision and we have also seen, we have also discussed that this linear regression at w transpose x gives you an idea of what is the distance of feature vector x from the separating plane, that is a plane having equation w transpose x equal to 0.

And more the distance of the feature vector x is from the separating plane, more confident our decision is. That means we know that the feature vector x is well within the region given to the corresponding class. In case of logistic regression, we have seen that this distance measure can be interpreted as a probabilistic measure. That means, more our confidence is that a feature vector x belonging to some class say y the probability of that class y given x and the parameters w, the parameter vector w should also be high.

So, as the distance goes on increasing the probability asymptotically reaches to one or in the other case as the distance goes on reducing in the negative side, the probability of the other class goes on increasing. So, that some of the probabilities belonging to the two classes always becomes equal to 1 because the data has to belong to either of the classes; either class omega 1 or class omega 2.

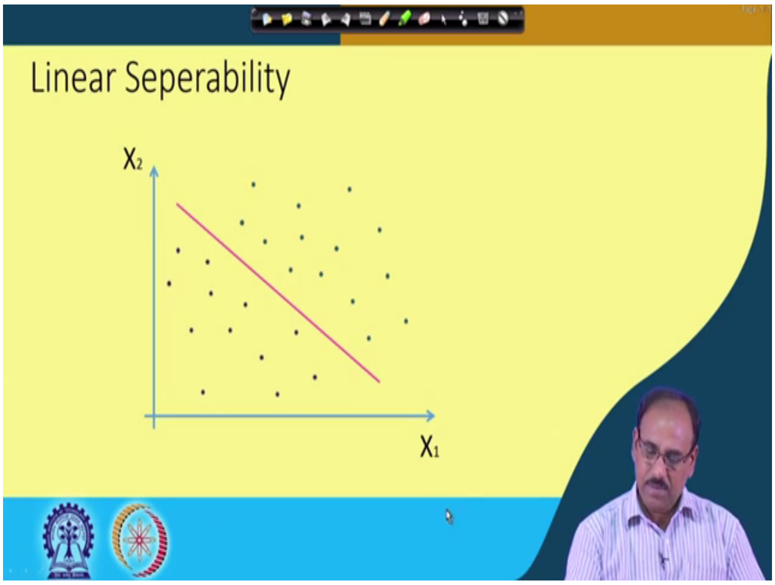
So, both this linear regression and the logistic regression we have discussed with respect to the problems which are two class problems or binary classification problems, then we have generalized this logistic regression which is a probabilistic measure to a class of classifier, set of classifiers which are known as soft max classifier.

So, the soft max classifier is a classifier which deals with multi class problem. So, earlier we had seen that in case of multi class problem we could have a linear machine or we could have a multi class support vector machine where the linear machine of this multi class support vector machine gives you a class score or it outputs a k dimensional vector or k is the number of classes and the linear machine or the support vector machine till or multi class support vector machine gives you a score for every class. And then for whichever class the score was maximum, we could classify the vector to that corresponding class.

So, as in case of the logistic regression for a two class problem or a binary problem that the distance measure is converted into a probabilistic measure. In case of soft max classifier, the class score can also be converted to a probabilistic measure. For that what we have done is given the class score plus cause the class y i to be s y i and class score for every class j to be s j. We have converted this class score into a probabilistic measured that p of y i given x w is equal to e to the power s y i upon sum of e to the power s j where the summation is taken over all j that comes into a denominator and that is what becomes normalized probabilistic measure and that is what you have done in case of softmax classifier.

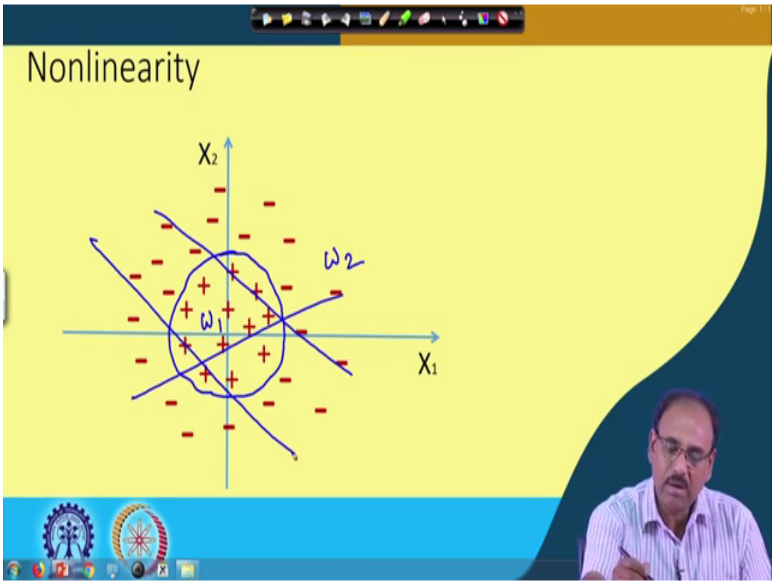
In today's lecture we are going to discuss about the non-linearity, how considering non-linearity is important for machine learning techniques and then we will extend our discussion to neural networks or deep neural networks and here it is very very important because when we talk about deep learning. The entire deep learning algorithm, entire set of deep learning algorithms are based on neural network architecture only. So, in some cases the deep learning is also known as deep neural network algorithms or deep neural network architectures.

(Refer Slide Time: 06:03)


Linear Seperability

So, first let us talk about non-linearity. So, again I am repeating the same figure that earlier we had seen that if the classes are linearly separable, I can pass a straight line or a plane or a hyper plane in multi dimension between the vectors given for two different classes or in other words that this linear function or the hyper plane can separate support way can separate the feature vectors belonging to class omega 1 and omega 2 and I can pass I can have such a hyper plane without any error.

(Refer Slide Time: 06:46)



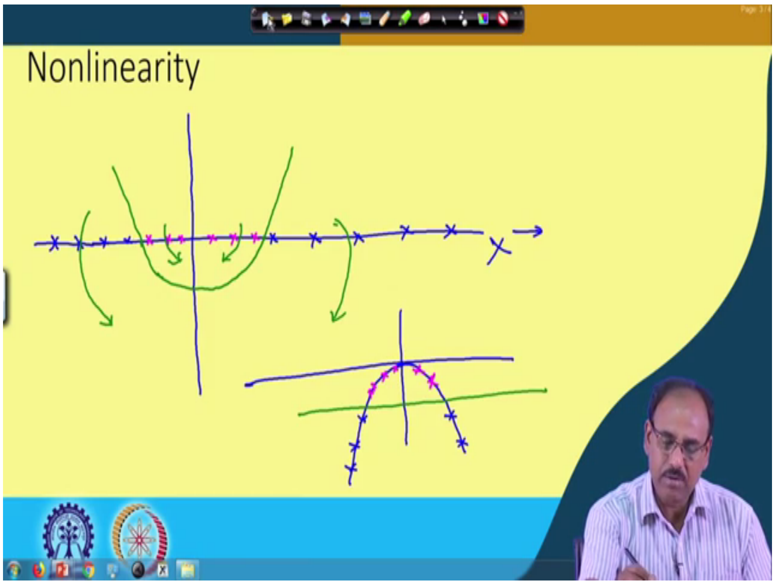
Nonlinearity

But you consider a case says something like this if my feature vectors are obvious form. So, here all the feature vectors marked as plus say these are the vectors belonging to belongs to one class. Let us call it class omega 1 and all the vectors which are marked as minus they are the feature vectors belonging to class omega 2. So, I have the distribution of the feature vectors belonging to classes omega 1 and classes omega 2 as shown over here and you can well imagine.

So, this is the case in our two dimension. In three dimension I can have a spherical distribution of feature vectors belonging to one class and outside the sphere I can have feature vectors belonging to another class and I can have many such complicated distribution other feature vectors. So, given this you can well imagine that here it is not possible to have a linear separator between the classes omega 1 and omega 2, ok. What whichever state line I form, whichever I take I will always have some misspecification.

So, this is a problem which is linearly non separable problem. I can I cannot separate the classes using a linear function or this problem cannot be solved using linear function. So, what we can do in such cases let me just simplify the problem instead of taking vectors in two dimension. Let us consider that I will take vectors in one dimension or the scalar features.

(Refer Slide Time: 08:40)



So, I take a set of features. Let me draw it properly. So, as I said that I have the features on in one dimension. So, I have a set of samples which are like this that belongs to one class and I can have another set of samples which I am marking in pink which belong to a separate class. So, given this you will find that all the samples which are marked in pink are sandwiched between the samples which are marked in blue. So, given such a case I cannot draw a straight line or I cannot separate the samples belonging to these two classes using a single point on say line X.

So, this is the feature direction. I cannot have a single point on this feature line x which separates the samples belonging to class omega 1 and the samples belonging to class omega 2. I can of course separate them if I take a curve of this form or in other words, I have got two points on this feature line using which I can separate these two classes. So, this becomes a non-linear problem. It is not a linear problem anymore.

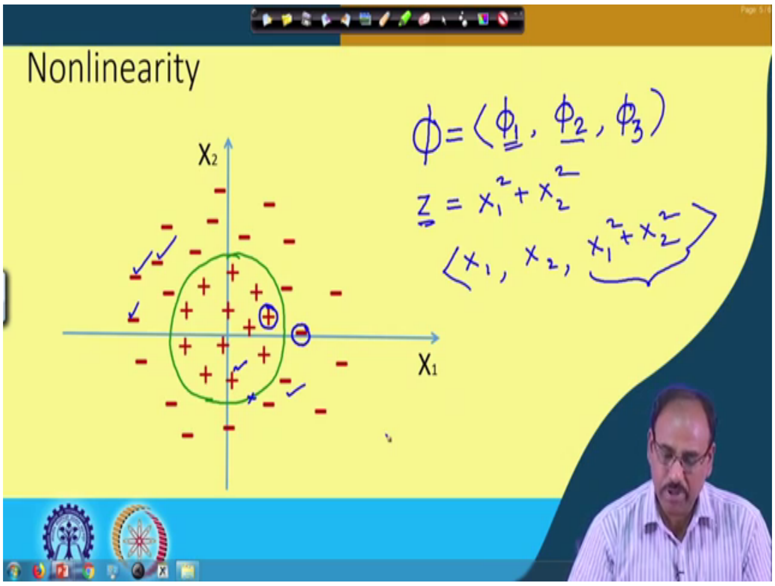
So, how to solve this? I can still solve this by having some non-linear mapping of the feature vectors right. So, what I will do is I will have a non-linear mapping of this feature vectors in such a way that these feature vectors will be remapped in this direction. So, as a result the feature vectors, all the feature vectors belonging to class say omega 1 let me draw a, redraw this figure.

So, all the feature vectors which are marked in pink they will be mapped over here and all the feature vectors which were marked in blue they will be mapped over here. So, you find that I had feature features in one dimension now they are mapped in two dimension. So, the first thing that I am doing is, I am increasing the dimension of the feature vectors and how I am doing it? I am doing it by using a non-linear mapping. I will come to what form of non-linear mapping we can do.

So, I am doing it using non-linear mapping and once I have this, now you find that I can pass a straight line over here which separates the feature vectors which are pink from the separate feature vectors which are blue. So, one way we can tackle the problem of non-linearity that is if the feature vectors are mixed.

So, that problem can still be solved using a linear classifier, but for that instead of trying to use a non-linear classifier you apply some non-linear mapping on the feature vectors and once the feature vectors are non-linearly mapped possibly on a higher dimensional space in that higher dimensional space, the feature vectors can be classified using a linear classifier. So, what we will do in that two dimensional example that we have just shown.

(Refer Slide Time: 12:52)



Nonlinearity

$\phi = \langle \underline{\phi_1}, \underline{\phi_2}, \phi_3 \rangle$

$\underline{Z} = x_1^2 + x_2^2$

$\langle x_1, x_2, \underbrace{x_1^2 + x_2^2} \rangle$

Let us go to this two dimensional example again. So, the example was like this ok. So, what I can do here is you find that this is something like a circle at the center and the feature vectors belonging to class omega 1 are within this circle and all the feature vectors belonging to class omega 2, they are outside the circle.

So, I can have a non-linear mapping of this feature vectors. So, it is a non-linear mapping using a non-linear function say phi. How this non-linear function will work? You find that I have feature vectors in two dimensional space which is given by X 1 and X 2 if I introduce a third dimension. So, I want to have a value of z where z will be equal to X 1 square plus X 2 square right. So, if I do that then you will find that for any feature vector which is outside the circle.

So, we are here. The value of X 1 square plus X 2 square will be more than the value of X 1 square plus X 2 square for any feature vector which is within the circle because for any point on the circle x square plus 1, X 1 square plus X 2 square is nothing, but square of that radius of the circle and because all these feature vectors which are negative are outside the circle for them, the square of the distance will obviously be higher than the square of that radius of the circle.

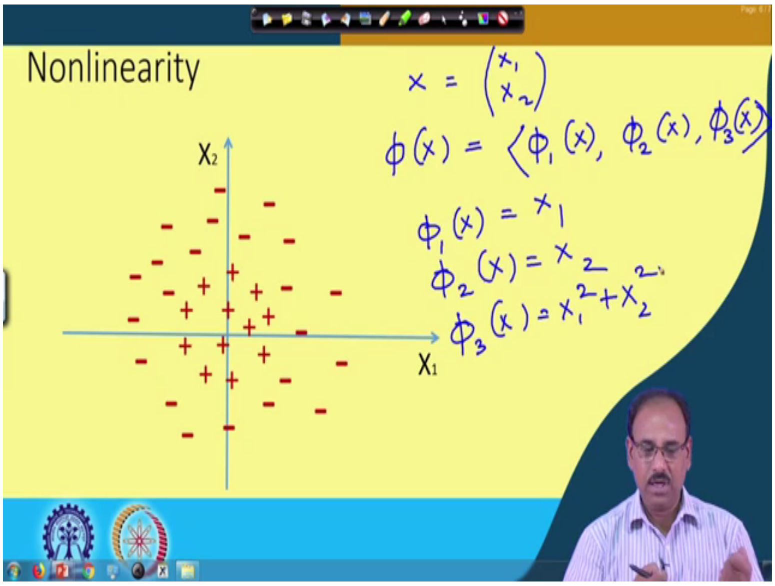
So, for all these feature vectors which are outside the circle X 1 square plus X 2 square or the value of z will be more than the value of x square plus y square for any feature vector which is within the circle. So what I will do is, I will now represent all these feature vectors. So, in the original form the feature vectors are two dimensional having components X 1 and X 2. Now I will add a third component z which is X 1 square plus X 2 square.

So, now my feature vectors will be X 1 X 2 and X 1 square plus X 2 square. So, this becomes the feature vector. So, you find that I have applied a non-linear function which non-linearly maps these feature vectors from a two dimensional space to a three dimensional space and when you consider in the third dimension that is in the dimension of z for all these feature vectors which are within the circle having marked positive for them, the z value is less than the z value of the feature vectors which are marked negative.

That means when I consider in the z dimension, all the feature vectors which are negative is somewhere over here and all the feature vectors which are positive is somewhere over here. So, once I have this, then I can always pass a plane in between which is passing through the feature vectors belonging to one class and the feature vectors is belonging to another class.

So, to by this non-linear mapping in the feature vectors I am converting on linearly nonseparable- set of feature vectors to a linearly separable set of feature vectors and we can say that this is being done by a non-linear mapping phi and this phi is nothing, but a collection of three functions phi 1 phi 2 and phi 3 where phi 1 and phi 2 phi 1 works on X 1 X 2. So, I will write at in this way. Let me clear this.

(Refer Slide Time: 17:04)



Nonlinearity

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$\phi(x) = \langle \phi_1(x), \phi_2(x), \phi_3(x) \rangle$

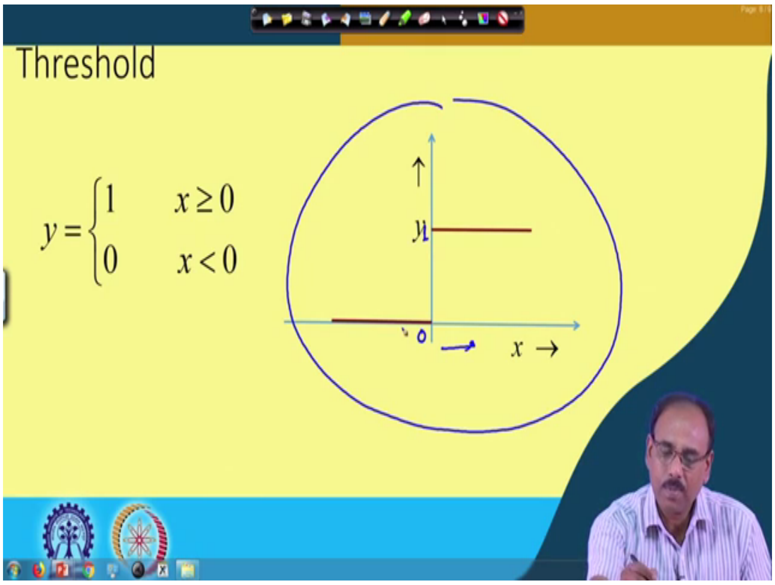
$\phi_1(x) = x_1$

$\phi_2(x) = x_2$

$\phi_3(x) = x_1^2 + x_2^2$

So, X is my feature vector which is having two components X 1 and X 2. So, phi of X, I can write it as phi 1 X phi 2 X and phi 3 X what is phi 1 X phi 1 X gives you X 1 phi 2 X gives you X 2 and phi 3 X gives you X 1 square plus X 2 square and that is what is the non-linear mapping that we are going to have. So, you find that now I have converted this feature vectors from X 1 X 2 plane from two dimension to three dimension where now the dimensions are given by phi 1 phi 2 and phi 3. So, I can simply write this as the space.

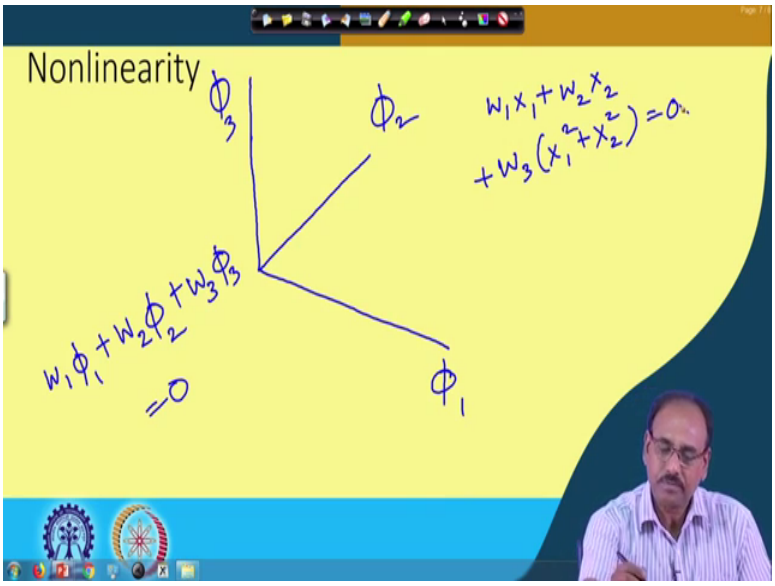
(Refer Slide Time: 18:15)



Threshold

$$y=\begin{cases}1 & x\geq 0\\ 0 & x<0\end{cases}$$

(Refer Slide Time: 18:20)



Nonlinearity

$$w_1x_1+w_2x_2+w_3(x_1^2+x_2^2)=0$$

$$w_1\phi_1+w_2\phi_2+w_3\phi_3=0$$

Now I can represent as I am having three dimensions. I can put it like this is phi 1, this is phi 2 and this is phi 3 and once I have this, you find that the hyper plane in this phi 1 phi 2 phi 3 space can now be written as say W 1 phi 1 plus W 2 phi 2 plus W 3 phi 3 that equal to 0 right.

And now, if you expand this phi 1 phi 2 and phi 3 what I have phi 1 is nothing, but X 1. So, this simply gives you W 1 X 1 phi 2 is nothing, but w X 2. So, that gives you W 2 X 2 plus phi 3 is X 1 square plus X 2 square. So, I have W 3 X 1 square plus X 2 square which is equal to 0 that is the equation of the plane that I have now I find that. I can have a twelve interpretations of this equation in the sense if I consider this equation where my variables are X 1 and X 2, the equation is a non-linear equation right whereas, while training I am given the training vectors. That means, for all those training vectors X 1 X 2 are fixed.

So, I can also consider this equation to be a equation in W 1 W 2 and W 3 where W 1 W 2 and W are the variables, but because the feature vectors given for training are fixed, so I can also consider X 1 and X 2 to be constants. So, if I consider X 1 and X 2 to be constants, then this equation is a linear equation. So, if I consider this equation to be a equation in X 1 X 2, it is the non-linear equation which gives you a non-linear mapping of the feature vectors.

So, the feature vector X 1 X 2 is non-linearly mapped into phi 1 phi 2 phi 3 domain and the classifier of the separating plane is a plane where W 1 W 2 W 3 are variables and X 1 and X 2 are fixed. So, by this non-linear mapping you are mapping the feature vectors into another space and possibly a higher dimensional space and in that higher dimensional space the feature vectors are linearly separable.

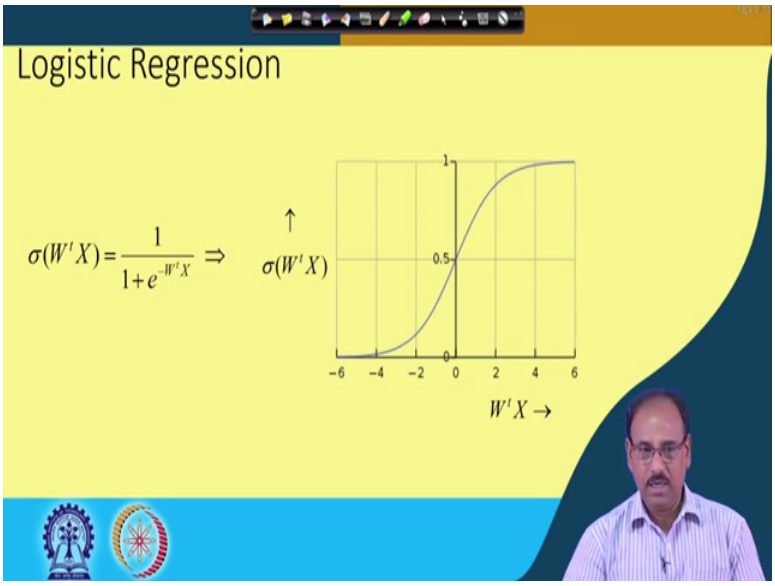
So, you will come to ah, so this is a very simple example where using a simple mapping I can convert a non-linearly linearly non-separable set of feature vectors into a linearly say separable set of feature vectors, however the type of non-linearity may not be so simple. There can be complicated non-linearities and those are the non-linearities which have to be solved by the neural networks that we will see later and the neural networks are to be trained using the training vectors. So, I will come to those problems later.

So, what are the different types of. So, it is quite clear that when I have a linearly non-separable problem, I have to have non-linear mapping of the feature vectors from one space to another space and in most of the cases this mapped space is of higher dimension than the original space and theoretically it can be proven that if I map feature vectors to infinite dimensional space, then whatever be the complexity of the non-linearity every non-linear problem can be solved as a linear problem when you are increasing the dimension to 0 and for such non-linear mapping, I obviously I have to have some non-linear functions.

So, now, let us see that what are the different kinds of non-linear functions that can help us in such a non-linear mapping. So, one of the non-linear function which we will use is what is known as threshold function. It is a very very simple function. So, the threshold function says that if y is a function of x, then output y the dependent variable y will have a value of 1 if x is greater than or equal to 0 and it will have a value of 0 if is x is less than 0.

So, here this non-linear function is depicted in this form. So, as x is greater than 0 value of y is 1. So, here it is 1 if x is less than 0, the value of y is 0. So, here it is 0. So, this is the simplest kind of non-linearity which is a threshold function that we can have and we see that use of this sort of non-linearity when we start our discussion on neural network.

(Refer Slide Time: 24:00)

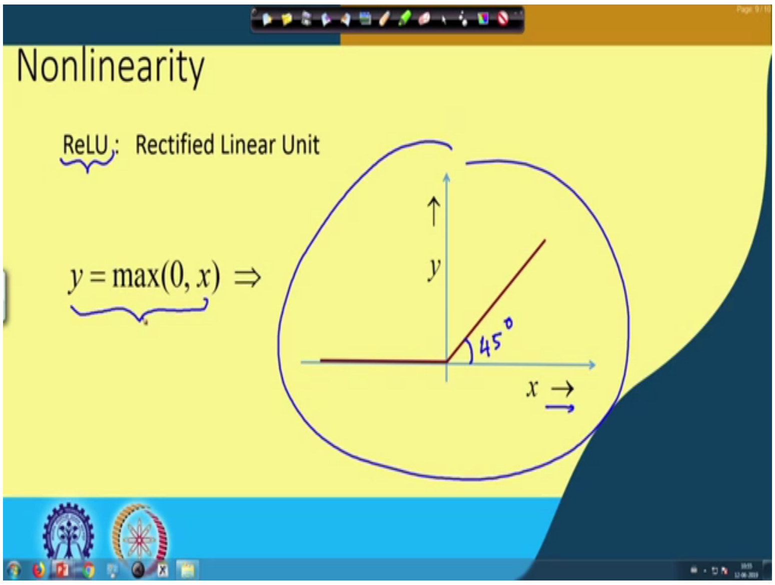


Logistic Regression

$$\sigma(W'X) = \frac{1}{1+e^{-W'X}} \Rightarrow$$

The next type of non-linearity which we have already discussed is Logistic Regression. You find that this mapping that we have done from W transpose X to sigma W transpose X is nothing, but a non-linear function. So, this sigma W transpose X which is 1 over 1 plus e to the power minus W transpose X, this is a non-linear function. So, this function also helps us to transform a non-linear problem to a linear problem.

(Refer Slide Time: 24:33)



Nonlinearity

$\underbrace{\text{ReLU}}$: Rectified Linear Unit

$\underbrace{y = \max(0, x)} \Rightarrow$

The other kind of non-linearity which is widely used in case of neural network or modern deep neural network is what is known as ReLU or Rectified Linear Unit. So, this they rectified linear unit simply puts as if y is a function of x, then y will get a value of maximum of 0 or x. So, naturally you find that if x is greater than 0, y gets the value of x and if x is 0 or less, y will have a value of 0. So, for all positive x y and x are same for 0 and less a value of x is 0 or negative, then y will assume a value of 0.

So, I can depict this function in this form. So, here you find that as x is greater than 0, y is equal to x. So, if you take the gradient of this slope of this line, this slope is nothing, but 45 degree. So, as y is greater than 0 value as x is greater than 0, y becomes equal to x. If x is less than 0 or 0, then value of y is 0. So, this is a linear rectified unit a rectified linear unit which is in short it has written as ReLU and obviously you find that it is also unknown linear mapping.

This also gives a non-linear mapping because if y is equal to x or y is some constant a times x that is a linear function, but the moment I put y in this form it is a non-linear function. So, in case of neural networks we will see later that in simpler cases we can use threshold functions in most of the popular neural networks the non-linearity that is used as a sigmoidal function, but in modern neural networks in deep neural networks instead of sigmoidal function people preferred non-linearity. The reason is a threshold function is a simple short of non-linearity for simple networks.

We can use that, but the problem with the threshold function is it is non-differentiable. I cannot differentiate into threshold function and we have seen in training earlier that in most of the cases training uses a gradient descent approach where gradient is nothing, but our differentiation, right differentiation in multiple directions.

So, because threshold function is cannot be differentiated, it is a non differentiable function. So, there it leads to problem in case of training of the neural network because gradient descent cannot be applied on that. So, that problem is slightly overcome if we use sigmoidal functions non-linearity. Of course, sigmoidal function is not the only one. The other kind of non-linearity which people also have tried is what is known as tan hyperbolic non-linearity where you have variation of the output from minus 1 to plus 1. In case of sigmoidal function it is 0 to plus 1.

So, tan hyperbolic kind of non-linearity has also been used, but there again the problem means when the value of W transpose X is very high or the value of the argument is very high, the gradient is very very slow right because your curve is almost parallel. So, gradient almost vanishes and because of that your training or the learning algorithms becomes very slow which is solved by This ReLU because in case of ReLU as long as x is greater than 0, the gradient of the function is unity right. So, here the gradient does not vanish.

So, that is an advantage of ReLU you and when you have modern deep neural networks where you have large number of nodes large number of layers hidden layers, then they do becomes more advantageous over sigmoidal function. So, today we have discussed about non-linearity and will continue with these discussions in our next lectures.

Thank you.