Neural Network and Applications

Prof. S. Sengupta

Department of Electronic and Electrical Communication Engineering

Indian Institute of Technology, Kharagpur

Lecture - 24

Radial Basis Function Networks: Cover's Theorem

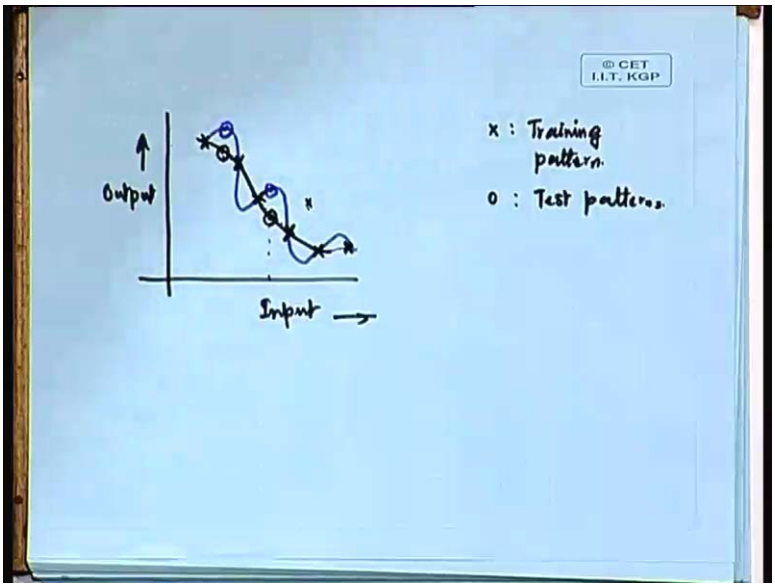
The topic for today is Radial Basis Functions. Before going into the network details, we will be talking about the very basics of what radial basis functions mean. And in that we will be covering one of the very important theorems, one of the fundamental theorems, upon which this radial basis function is based upon and that is Cover's theorem.

So, we will be treating the fundamentals as well as Cover's theorem in today's lecture. But, before we begin this topic there are one or two discussions that I wanted to have related to the back propagation algorithm. In fact, we have been seeing about a several aspects of the back propagation algorithm, so far. And one of the things which back propagation algorithm has to take care is the question of generalization.

And what we mean by generalization is that, if you have trained the multi layer perceptron with back propagation algorithm, with several input output pair of patterns. It knows that how to train the system for the training patterns, but whenever you feed a test pattern, then what is the guarantee that the test pattern is correctly. We are obtaining the correct kind of the output for the test patterns, we can only have that, because the test data was not there during the training, in general it will be so.

That you will be having a separate set of data set for the training and a separate data for the test. So, the test data does not necessarily include any training data, if accidentally there is some training data well and fine. And then, it will be very correctly classified if the problem is classification problem, but in general I think we cannot really say that. That the test data will be generating the proper output it does, if there is a proper generalization, what to say is something like this.

(Refer Slide Time: 03:23)



Like say for example, we have got some input output function, the function as such is not known to us, known to us in the sense that we are determining the function through the training process. So, here we have the output and here we have the input, although I am considering the input set of points in one dimension, but in general it will be in the multi dimensional space.

So, let us say that this is one input, this is one input, this is one input, this is one input and this is another input say this is another input like that. And for this inputs, the corresponding outputs are obtained here like if this much is the input, then the corresponding output is here, if this is the input this is the corresponding output. Another thing is that these are, all the crosses that I have given in this drawing they are the training patterns.

All these are the training patterns, where there input output behavior is known. Now, what the neural network does is that with the training it fits a curve, essentially neural network training is nothing but, a curve fitting problem. So, we feed a curve like this, if it is a good fitting, then we should be expecting a curve like this.

Now, these are all the training patterns and supposing we introduce some test patterns and let us say that one of the inputs to the test pattern is here. Now, this point is certainly not there in the training, because there is no cross at this point, so we introduce a test

pattern here, means we feed the input over here and we see the outputs. So, the output should be this, if there is a good generalization, if I have a test pattern here.

Then, given this kind of functional approximation that the trained neural network is doing, we will be getting this as output. Now, the question is that yes, if the curve is like this, then we are getting, for the test pattern also we are getting the correct kind of outputs. But, let us say that taking the same crosses, the here we remark the cross positions, that means our training pattern positions. And we have let us say realized another function, which is not that well generalized.

See training pattern you feed means that of course, the function that you are approximating it as to pass through the training pattern points. So, there is no harm in fitting a function like this, who says that it is bad, because this function which I have drawn with blue. This is also passing through all the training patterns, but if I have to accept this blue function as my approximating function, then for this input I would be getting this as the output not this.

And for this input I will be getting this as the output not this, but question is that when you see an output of this kind or an output of this kind, would say that it is generalized, it is not. Because, although we have trained it, but it lacks generalization. Now, many a times the generalization lacking is due to what is called as the over fitting of the data.

By over fitting I would like to say that you take a training set, now a training set how are you deriving, people derive it using the from the real life examples. I mean perform the actual process or from the actual data that is available, some sample datas which are available they are used for training. Now, such kind of datas are always subjected to some noise, there may be some noise that is present in the training pattern itself.

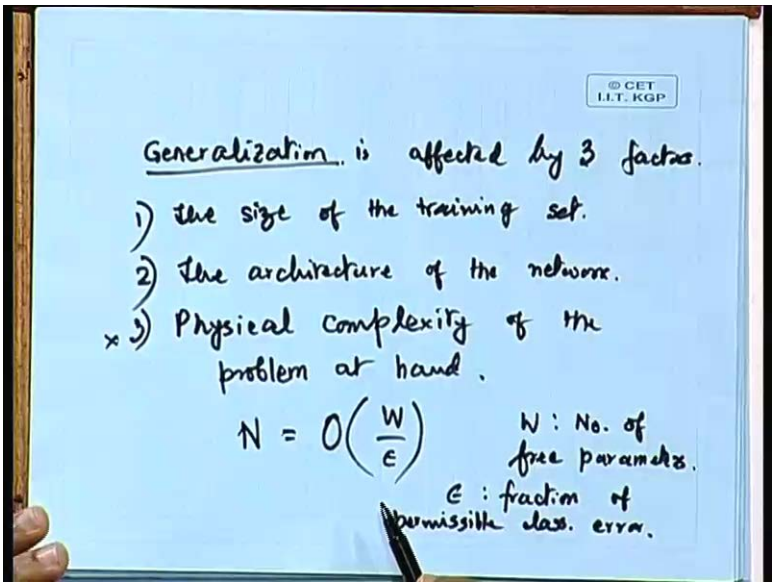
Now, if I supposing this is the set of patterns, with which I have trained. Now, supposing I have got one odd point, one outlier let us say, that supposing in this training patterns that I have got with which I have trained, supposing there is a pattern over here means this is here and then, the next is here. Now, clearly this is an outlier that it does not fit into any proper curve fitting, a proper smooth curve fitting will not feed this point.

So, this is definitely a noisy point, but if I make the neural network properly trained with all these training data’s without bothering that whether it is a noisy data or whether it is

the right kind of data, neural network trainee. And when it trains it will realize a function where such kind of generalization will be lacking, because it also learns the noisy data that is the thing.

Because, it is learns noisy data there will be errors that it will be making when the test patterns are presented. Now, the question is that what are the factors on which this generalization is influenced. Now, the generalization of the capability of the back propagation algorithm is influenced by three factors.

(Refer Slide Time: 09:20)



So, generalization is affected by three factors and what are they, number 1 very important is the size of the training set. Now, again I think this aspect we have discussed implicitly many a times, I think even during the discussions of the single layer perceptrons we have talked about this aspect. And in fact, you remember that we also talked about the VC dimension which determines, that what is the size of the training set up to which we can go.

And then, the second factor is the architecture of the network. Now, it may be that your training set is fixed, but you can go through differential architectures. Like say for example, in the back propagation algorithm only for multi layer perceptrons, you can design the architecture in a different way, you can have different number of hidden layers.

You can choose different hidden layer neurons, different number of hidden layer neurons in a hidden layer, that is to say, that you can influence the architecture of the network. And the third factor of course, it is not under anybodies control, because it is very much problem dependent, that is the physical complexity of the problem at hand.

Now, we do not have as I told you that, we do not have any control over this thing, that whatever problem we have been given to solve we have to solve that. So, what we can do is that, we can instead study these two effects that is to say the size of the training set and the architecture of the network. Now, we cannot simultaneously varied a both, so we can see that by keeping an architecture of the network fixed.

If we try to experiment upon the size of the training set, then obviously, we should get some result. And it is seen that if you fix of the architecture and determine the size of the training set, then the size of the training set that is N. If N is the size of training set that will be given by the order of W by epsilon, where W is the number of free parameters in the network.

So, W is the number of free parameters and epsilon is the fraction of permitted classification error, which means to say that if supposing epsilon is equal to point 0.01, 0.01 means that we can tolerate 1 percent of classification error. And then, W could be anything, so if it is W by epsilon and epsilon is equal to 0.01, that means to say that it is order of 100 times W, so whatever number of free parameters are. It should be hundred times of that that should be the training, the size of the training set.

So, naturally if you are wanting the fraction of permissible classification error to be very low, then you have to go in for a very large training set, for a good generalization. And the other thing that one can study is that you keep the size of the training set fixed and then, you find out the architecture. Now, lot of such studies have been made and I am not going into the details of it, because we have already covered several lectures related to the back propagation network. And in fact, I expect that those who are attending this course.

They should also have their own study, I should also encourage them to have some self study with the materials that are available in various books and journals. Just to study more about the aspects of the multi layer perceptron. So, that when it comes to the solution of your problems you can help it, you can try to use it for your benefit. So,

having said that I think we can effort to move over to the next topic that we are going to cover.

And that as I told is the radial basis function and in this, we are first going to cover a very basic theorem that is presented by cover. Now, before we going to the cover's theorem I think it is better that we spend some time, discussing about the very basis fundamentals of the radial basis functions. Now, the radial basis function structurally is multi layer perceptron, but the thing is that, whether we are calling, we are certainly not going to call every multi layer perceptron as a radial basis function.

Now, you have seen that we have gone through the back propagation algorithm for the multi layer perceptron. And the back propagation if you are looking at the theory behind the back propagation, after all what is the back propagation, you are making a stochastic approximation to the process that you are going to solve. You are essentially by leaning through the examples and adjusting your free parameters, you are basically making a stochastic approximation.

And then, after the training you are going to use it for the test patterns, so the whole approach is that please keep in mind, it is a stochastic approximation approach. Whereas, you can view the same problem of multi layer perceptron, from a different angle you can look at it purely from a surface fitting consideration. What I mean by surface fitting consideration, let me try to illustrate it a bit.

Now, let us go back to the point, where we had talked about the advantage of the multi layer perceptron that we had spoken earlier. The advantage that is there is that, if we have a set of patterns which are not linearly separable then, we could not solve it using single layer perceptron. Because, the very basic, basics of the single layer perceptron is that you have to determine a hyper plane, in the m dimensional space.

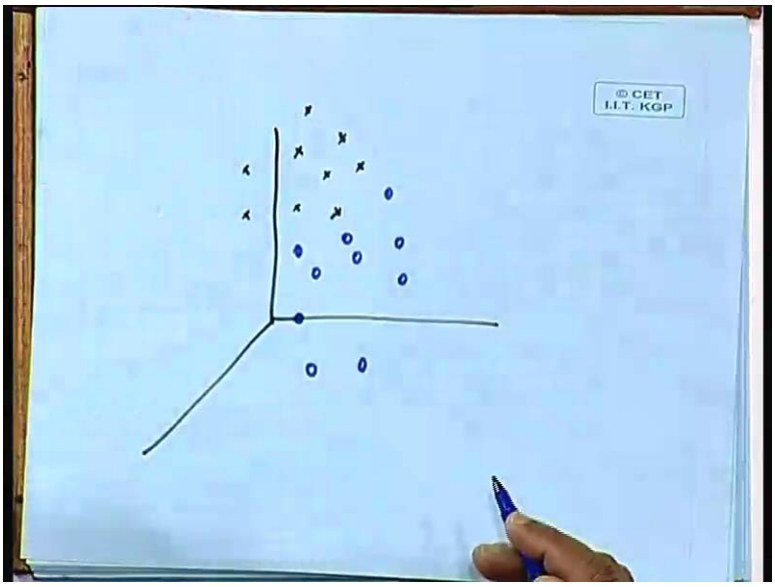
That can clearly separate the patterns from one that can clearly separate one class of patterns from the other. Now, why is it that we could solve the nonlinearly separable problems, using the multi layer perceptron, we did it. Because, we had mapped those input functions into a hidden layer and then, we went over to the hidden space, now individually I think you remember the example that I had talked of in one of the earlier classes.

Where we had shown the example of the exclusive wall, now the hidden layer which was there, those hidden layer on it is own where making a linear separability of the pattern. But, combinedly together in the combined outputs space, we had got that there was separability although with respect to the input space the patterns were not linearly separable. But, still we could do because of the in between mapping that we do the hidden layer.

Now, it is the same idea only thing is that in back propagation algorithm, you are using stochastic approximation, the debate that you can have is that rather than having that stochastic approximation. Why do not you start a fresh and try to think it in terms of a surface fitting, surface fitting like what that, let the set of patterns that we have at our disposal be not linearly separable.

Let it be in nonlinearly separable or we do not know, when we cannot pass any hyper plane we are saying that it is not linearly separable. And inevitably we are going to have that kind of situations many a times in practice. Another question is that if the pattern is not linearly separable, if you have a space like this.

(Refer Slide Time: 19:17)



Say for example, it is very difficult to draw a multi dimensional space, but let us say that here we have got pattern class one somewhere. And let us say we have another pattern class two, which are like this in a multi dimensional space. Now, they may not be linearly separable means, I may not be this is a three dimensional plane example. And I

may not be able to cut it through a plane that separates the class of this from the class of this.

I cannot pass a hyper plane, but you never know I may be able to pass a hyper sphere or any hyper quadric surface meaning that any multi dimensional surface, I could be having a surface like this, that can clearly separate the patterns. So, the thing is that if from the training set, if I am given with a training set that is nonlinearly separable. Then, my problem will be to determine a surface a hyper surface, not a hyper plane anymore.

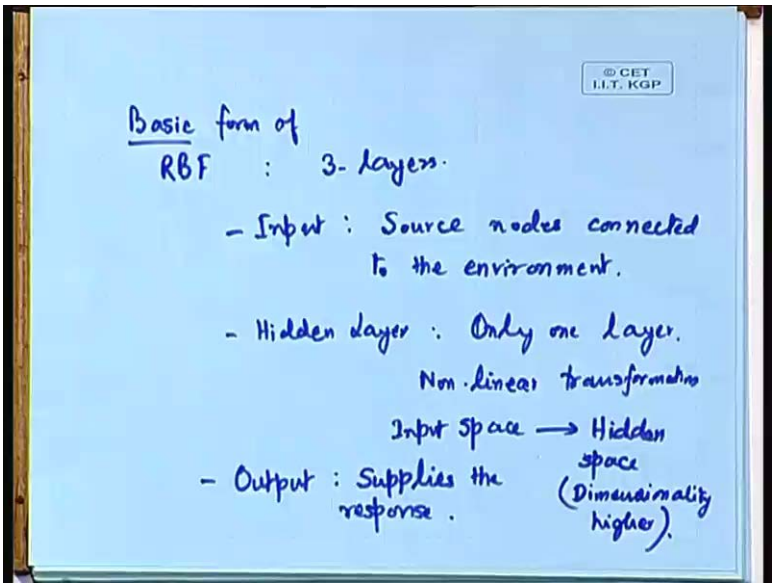
A hyper surface that can make the classification between the two competing classes. Again the fact that I have taken example of 2 classes, does not mean that it cannot be extended, it can indeed be extended to multiple number of classes also. The same idea extends, in fact we had talked about the extension into the multi class classification in the last lecture itself.

The thing is that then what happens that, then we are looking at a surface fitting problem in the multi dimensional space. So, that is what forms the essence of the radial basis function, that is fitment of the multi dimensional surface. Now, to do that what we have to do is that, just like a any multi layer perceptron, we have to have a hidden layer in between, at least one hidden layer in between. And what is that hidden layer and in a surface fitting problem, we can look at the hidden layer from a different perspective.

We can say that the hidden units, which are there the hidden layer units, the neurons which are there they provide a set of functions which forms a bases for mapping into the hidden layer space. I mean basically we are going to do a mapping from the input space to the hidden layer space. And to do that mapping we need some basis functions, which providing the neurons in the hidden layer they are providing.

So, the hidden neurons providing the basis functions and this kind of architecture is called as the radial basis function networks. Now, the basic form of radial basis function is that, it has got three layers, I am saying that it is the very basic form of radial basis function. Of course, researchers are on for multi layer things also, but of course, I think the problem the way it is solved.

(Refer Slide Time: 22:59)



It is nicely done for three layers also, so the basic form I should say, that the basic form of the radial basis function is, in it is basic form it is having three layers and what are the three layers. First is of course the input, the input of course, contains the source nodes which are connected to the environment. Then, we will be having hidden layer, in fact since we are having three layers all together, we can have only one hidden layer in the basic form.

Only one layer which does a non-linear transformation this is very important, it does a non-linear transformation from, what to what non-linear transformation from input space to the hidden space. And very interestingly it is found that most of the times, we find that the hidden space dimensionality is higher. In fact, that is based on the theory that cover had proposed, but we are going to discuss shortly.

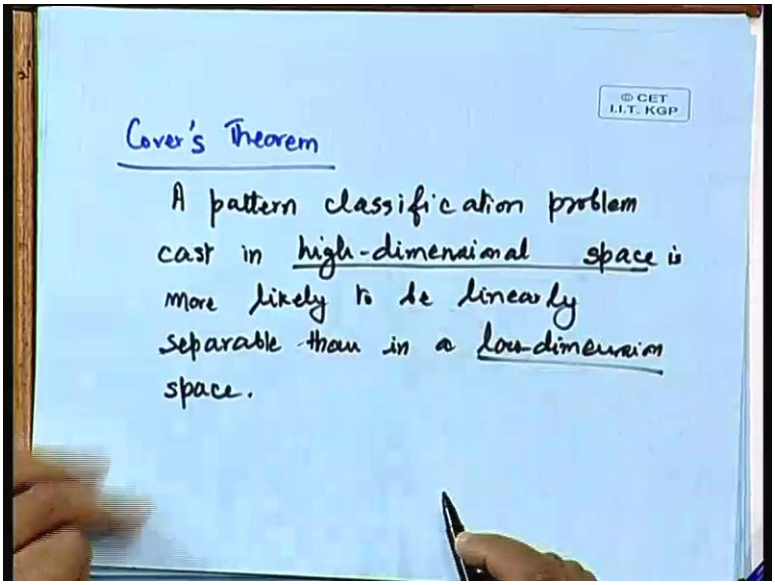
So, that dimensionality of hidden space is definitely higher, so hidden space with higher dimensionality and then, we are having the output layer. The output layer that supplies the response, so this is the three layer structure. Now, yes stressing on the non-linear and let us see that why is should it be a non-linear. You see if the input itself is a nonlinearly separable pattern, that is what we are assuming, that input is nonlinearly separable.

In that case, theoretically it should be possible for us to design a mapping function. A mapping to a higher dimensionality, where in that space, in that higher dimensional space it will be linearly separable. If we can have that, if we can design a mapping from

non-linearly separable input space to linearly separable hidden space. If in that space hyper plane separation works, then our job is done, then we can use that for the solution of our problem.

So, from the non-linear to the linear, if it has to map, then certainly it cannot be a linear function mapping. It has to be a non-linear function mapping, now with that basic form RBS we come to the Cover's theorem.

(Refer Slide Time: 26:43)

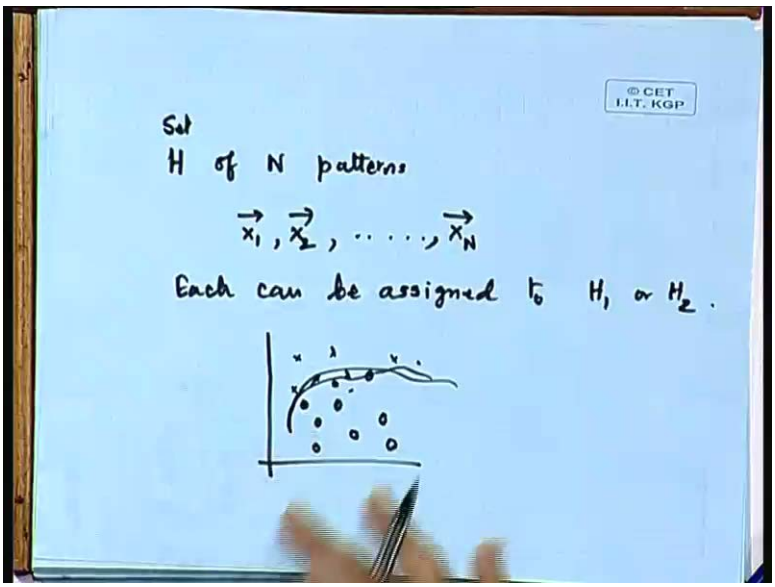


So, we first build up the basic of the derivation, although we would not cover the derivation in great details. Now, before we going to the discussion on Cover's theorem, let me tell the statement of Cover's theorem. It says that a pattern classification problem, cast in high dimensional space is more likely to be linearly separable, then in a low dimension space.

I repeat a problem is more likely to be linearly separable in higher dimensional space, than it is in lower dimensional space. So, that is why it make sense in doing a mapping from the lower dimensional input space to some higher dimensional hidden space. So, having said that I think we should write it down because the statement of that is very important. So, this is what it is, so that is why we should be considering the mapping from this low dimension to the higher dimensional space.

So, that our idea will be to make it linearly separable in the mapped space. What is non-linearly separable in input space should be linearly separable in the mapped space. Now, in order to discuss about that, let us consider a set of patterns.

(Refer Slide Time: 29:21)



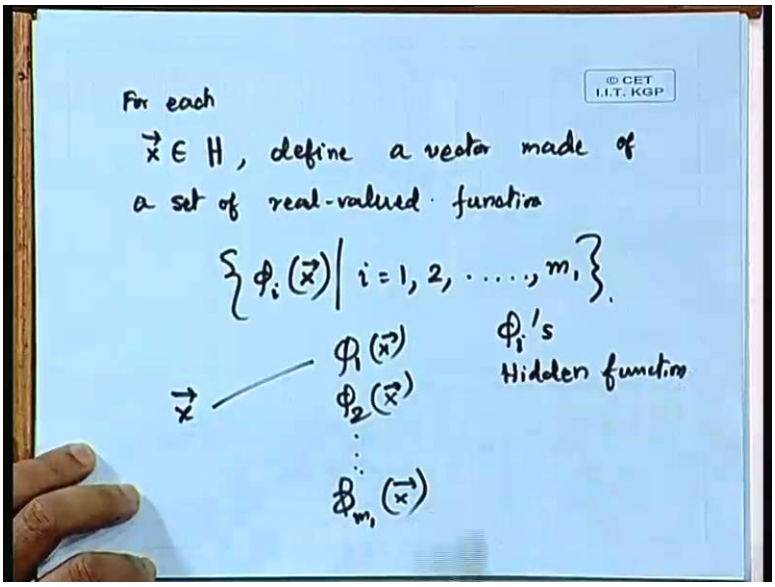
We define a set H, we define a set H or N patterns and what are those N patterns, that we are having let us denote them as X 1, X 2, up to X n. And each of these pattern each of these N patterns can be assigned, so each can be assigned to 2 classes and what are those 2 classes to H 1 or H 2, so there are just existence of 2 classes.

Now, this set of patterns, this dichotomy of patterns, because this is binary, so we can call it as a dichotomy of patterns. This is separable with respect to a family of surfaces, if there exists a surface in that family that separates H 1 from H 2. Now, again in order to explain this, let us say that we have got a multi dimensional space, where we are having patterns of one type patterns of another type.

And the thing is that, if there exists a surface that separates the patterns, this from the surface may be like this. That if a surface exists, that separates the two classes of patterns, then we can say that this dichotomy this set of patterns. This set of binary patterns are separable with respect to that surface or to say that if we have got a family of such surfaces, then if at least in one of the members of that family.

It is separable we can say that it is separable with respect to that family of classes. Now, this can be the ideas will get more clear as we proceed.

(Refer Slide Time: 31:47)



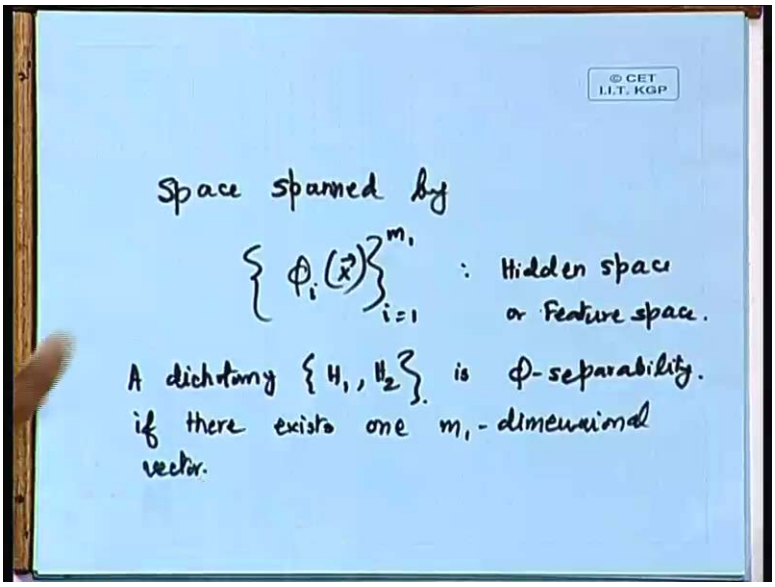
Now, what we do is that for each of the pattern, let us take a pattern X which belongs to the set H. So, X pattern X vector belonging to a for each of this pattern for each X belong to H, let us define a vector, which is made of a set of real valued functions. So, define a vector made of a set of real valued function and what are those and how are we going to write that set.

That vector will be written as like this, the set of real valued function, so we will be defining a real valued function phi i of X vector. Given that i is equal to 1, 2 up to some number let us say m 1. So, that means, to say what that we have got m 1 number of different functions, m 1 number of different real valued functions are available. Phi 1, phi 2, phi 3 up to phi m 1, there are m 1 number of real valued functions which are available with us as a set.

And with this set of functions we can map X into what we can map X into phi X. So, given sum X, it can be mapped into phi 1 X phi 2 X, so on up to phi m 1 X, so m 1 such different mappings will be possible. If we take a set of such real valued functions, now we take these, we refer to this phi i functions as the hidden functions.

So, all these phi i's, all these set of functions they are refer to as a hidden function, because it plays a roll very similar to the hidden units in the multi layer perceptron. Now, that we have got m 1 different such real valued functions available, that essentially spans a space.

(Refer Slide Time: 34:36)



So, the space that is spanned by the set of hidden functions, so the set of hidden functions I can write as the set of phi i X, i is equal to 1 to m 1 this is referred to as the hidden space or feature space. So, what was our input space, our input space is see, these are input space will be the dimensionality of this X vector the dimensionality of this X vector we have been taking as m. All the time we are taking it m, since it is essentially pertaining to the input we can call it as m 0.

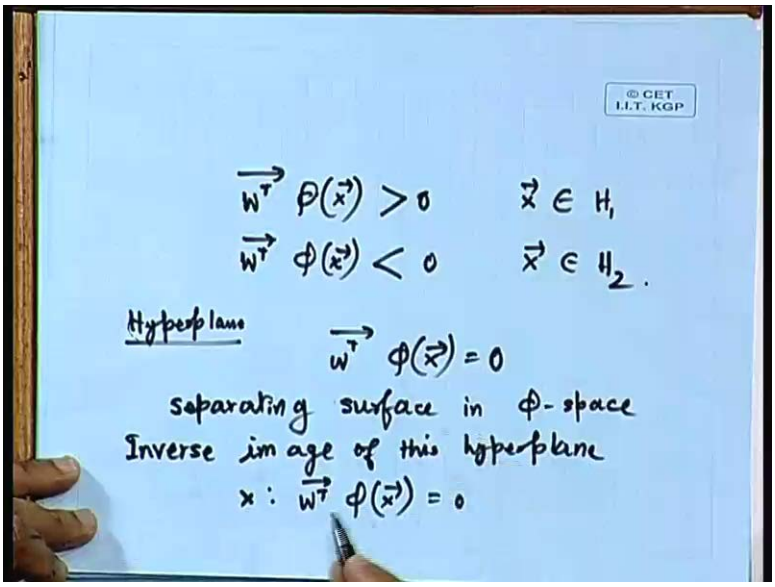
So, let us say that the dimensionality of X is m 0, so we are mapping this X into a new space, whose dimensionality is m 1, because there are m 1 number of such real valued functions. And what Cover is going to show is that m 1 should be high as possible, preferably greater than m 0, it does not explicitly say that it has to be greater than m 0.

But, the findings of Cover is that m 1 should be, higher you choose m 1, higher you can map into this the better it is for the linear separability. Now, what happens is that, you have got this set of phi functions or what you can say as the hidden functions, spanning a hidden space of m 1 dimension. Now, we can say that a dichotomy H 1 and H 2, because we have got only 2 classes over here H 1, H 2 dichotomy is phi separable.

The definition of phi separability is what, it is just putting the mathematical words into what I had already talked up at the beginning of our discussions on Cover's theorem. That this dichotomy is phi separable, if there exists an m 1 dimensional vector, if there exists one m o1 dimensional vector. In fact, finding out once such m 1 dimensional vector itself is good enough, because if we choose a family of functions, after all we have chosen a family of functions.

There are m 1 number of such functions and if out of these m 1 functions one fulfills the separability criteria, then we can say that using the phi functions is worthy. Because, at least 1 member of this phi function is able to solve our problem. So, if there exists one m 1 dimensional vector and what is the separability condition, that is very simply our good old separability that in the phi space. In the phi space the separability condition should be that.

(Refer Slide Time: 38:12)



W transpose phi X vector should be greater than 0, if that is the case, then we can say that X belongs to one of the classes let us say H 1. If this is greater than 0then X belongs to H 1 and if W transpose phi X is less than 0, the vector X belongs to the class H 2 that means, to say that the hyper plane that we are defining. So, the hyper plane equation that is W transpose phi X vector is equal to 0.

This bring the hyper plane equation, this is the separating surface, a hyper plane is the separating surface in the phi space. So, to find a network like this, so to have this

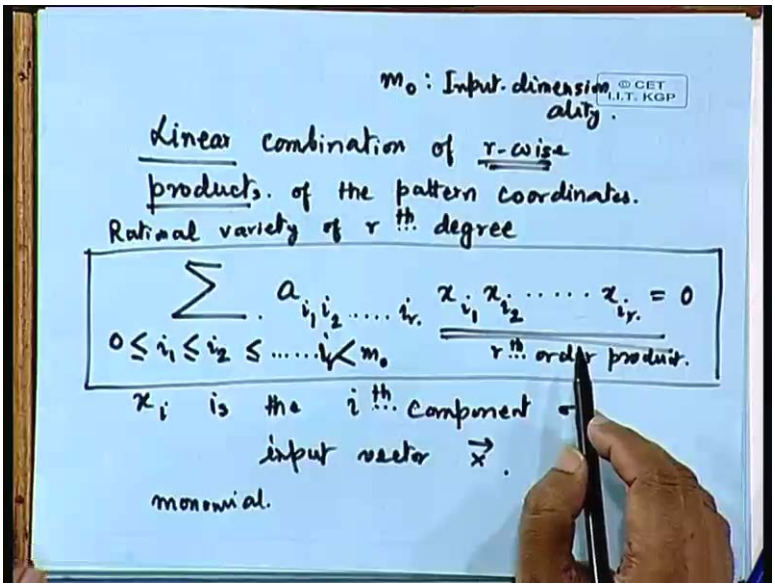
situation fulfilled, what is the condition that in the phi space the function must be linearly separable. Because, it is a linear separability condition in the phi space, but phi space is not the one, that is your input, your input is difference space, you are mapping from the input to the phi space.

So, now you have found out a good surface in the phi space, a linearly separable surface in the phi space. If you do a reverse mapping of that surface, which is a hyper plane in the phi space. If you reverse map it to the input space it will no longer remain as hyper plane, it will be a hyper surface, a general hyper surface it will be. So, the inverts image of the hyper plane if you see, that is to say that the inverse image of this hyper plane.

And this inverse image will be written as X mapped to W transpose phi X equal to 0. So, if we do the inverse mapping of this into the X space that will define the separating surface in the input space, which is not a hyper plane anymore it is a hyper surface, it is a general hyper surface. So, now the thing is that, so we are not going to, so that means, to say that the separability problem that we are addressing at the moment with respect to...

While seeing with respect to the input space is no longer remains as a linear separability problem, it is essentially a non-linear separability problem. So, let us think a fresh and try to model a non-linearly separable function. So, in the input space itself if we try to model a non-linearly separable function, then we can do it like this. So, we can express such kind of non-linear separability of patterns using a linear combination of r wise products.

(Refer Slide Time: 42:00)



So, what we are doing is that we are considering a natural class of mappings, which is obtained by the linear combination of r wise products, let me define what it is. And what is that, it says that you can express the mapping function as summation like this. I mean a linear combination, but linear combination of what terms where are r wise products means, r wise products of the pattern coordinate.

So, we have write that r wise products of the pattern coordinates, now you see the definition of the hyper plane is that when you are describing a hyper plane equation. Then the hyper plane equation will be of the form of, let us say if the elements of the X vector is X 1, X 2, X 3, X 4, X 5 let us say phi dimensional vector. Then, a hyper plane equation typically will be a 1 X 1 plus a 2 X 2 plus a 3 X 3, a 4 X 4 plus a 5 X 5 it is equal to 0.

That will be kind of hyper plane equation plus some constant term, that will be the hyper plane equation in the phi dimensional space. But, the thing is that, a here it is a hyper plane, so it does not have any term like X 1 square X 2 square X 3 squares or any cross terms X 1 X 2 or X 1 X 3 or X 1, X 2, X 3, X 4, X 5, so all such cross terms are very much permissible. So, if we entertain all such non-linear terms in our hyper surface equation.

Then, we can form a generalized hyper surface equation, where all such non-linear terms and cross terms are included. If we can do that, then we are writing a very general hyper

surface equation, so this is what we are going to write. So, a i 1 i 2 up to i r, in fact we have to say here that x i that we are writing here, I will let me first write it down then I will explain. So, it is of this form that there is term coefficient a which is written with a very long superscript i 1, i 2 up to i r and then, we have the cross terms involving this.

So, which is x i 1, x i 2 etcetera, etcetera up to x i r and that summation of this is equal to 0. And it is the summation over what space it is summation over all this i 1, i 2 up to i i r space. So, it is 0 less than or equal to i 1 is less than equal to i 2, so on, so on up to m 0, where m 0 is the input dimensionality, so m 0 is the input dimensionality. And then, what is the x i, x i is the i th component of input vector x.

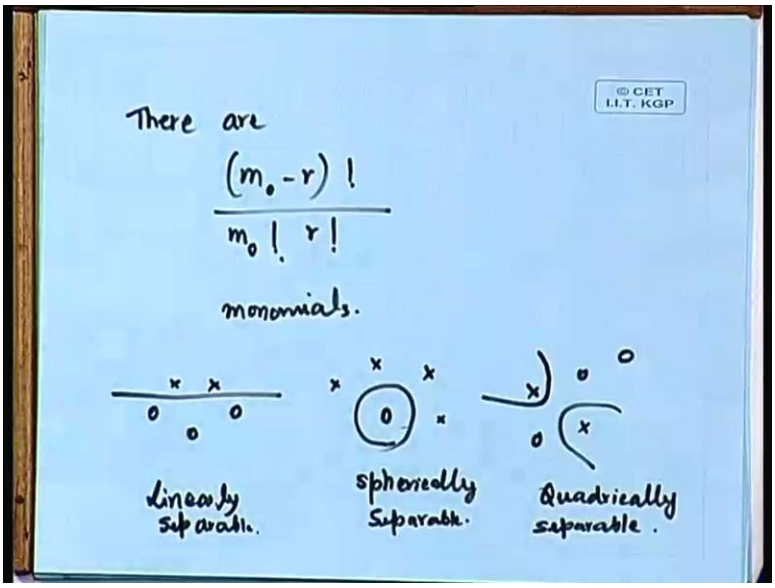
So, that means, to say that in this equation, that I have written this involves all the terms, all the higher order terms also are included in this basic definition itself. This summation, this is a very general form of representation and we call it as the rational variety of order n. So, this is called as rational variety of order r, here it is order, because we are going up to i r, so here it is r terms, so we are going up to i r.

So, this is called as the rational variety of order r degree we can say, now you see that what are these terms that we have got x i 1 up to x i r that means, to say that here I have shown r such terms. Now, this r terms that you are getting that can be picked up out of the m 0 total number of coefficients that you have got. Or rather the original dimensionality of the input vector x is m 0 dimensionality and you are choosing number r that is less than m 0.

So, and you are forming an equation like this, that means to say that it is possible for you to have different combination of this r. You have got m numbers out of which you have to pick up r numbers in different combination, so basically it is a communitarian problem. And all this r th order product of entries, you see these what I have written, x i 1, x i 2 up to x i r, this is an r th order product. So, this is an r th order product and this one is called, such kind of a product term, that is called as monomial.

So, this x i 1, x i 2 up to x i r the term that we have got this product term, this rth product term, this is called as the monomial. So, how many monomial like this as possible, we have got totally m 0 number of monomials, m 0 number of inputs available out which we have to choose r.

(Refer Slide Time: 48:58)



So, that means to say that we have got m 0 c r or rather to say we have got or rather to say we have got m 0 minus r factorial by m 0 factorial and r factorial these many monomials are possible, so I can say there are this many monomials. Now, out of this, so what you can do is that you can have different such separating surfaces now. I mean you can using this sense you are using all these r y's products, the separating surfaces that you are going to have are not the linearly separable surfaces.

In fact, you can have different type of surface separability, so you can let us say that you have chosen a dichotomy of H 1 and H 2. And let us say that we have got 5 such patterns. Now, this 5 patterns in this case, in this let us say we take some examples like this. Let us say this is example number 1, this is example number 2 and let us say that this one is example number 3.

Now, here you can see that, this is linearly separable, this dichotomy is linearly separable why, because I can find hyper plane that separates. This dichotomy that we have go this is certainly not linearly separable I cannot pass a hyper plane, but possibly I can pass a hyper sphere. So, this one is spherically or hyper spherically separable and in this example the separability is even a hyper sphere separability also is not coming in this should be separable using a quadric function.

So, this should be quadratically separable function, now some very important findings that cover had found out is that. Let us say that the activation patterns, all this X different

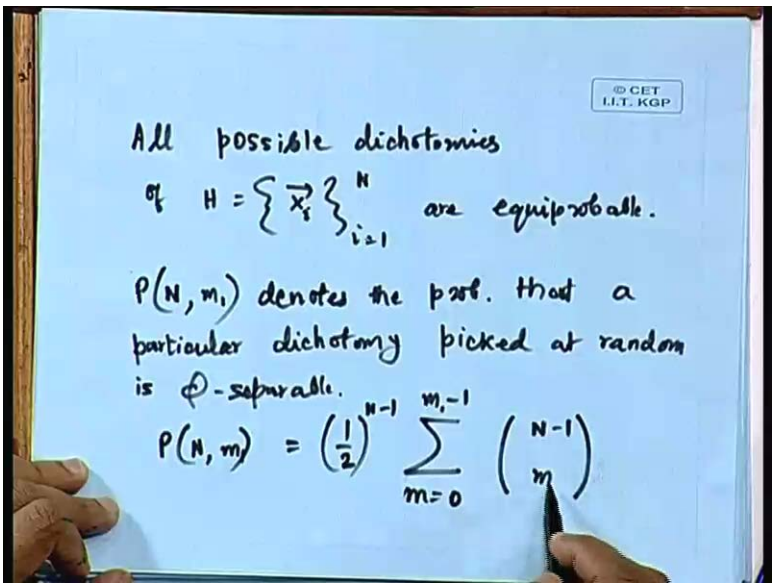
patterns that you are feeding that is X 1, X 2 up to X n, these are the different patterns that you are choosing. Now, supposing that these patterns are chosen independently. So, we want that this X 1, X 2 up to X n they are independently chosen patterns.

And also we assume that these patterns, these all possible dichotomy of these patterns they are equi probable. Let me explain what to say because you should think over it, you see that there are different patterns X 1, X 2, X 3 up to X n and they are independently chosen. And now, supposing you have decide to choose n such patterns. Now, given n patterns various kind of classifications are possible, in one of the cases there could be that one pattern is lying in H 1 class.

Few other patterns are scattered in the H 1 plus, some of the patterns has scattered in the H 2 class like this, this is possible. So, there can be various types of dichotomies which are possible out of this. So, there are let us say a large number dichotomies which will be possible and out of those dichotomies, all are not separable using even this nonlinearly separable functions.

There are only a few which has separable, so what Cover did was to study the probabilities, that such kind of dichotomies are separable. So, what he found out was that, it is based on some assumptions that all possible what are assumptions.

(Refer Slide Time: 53:53)



That all possible dichotomies of of H equal to X i, i is equal to 1 to n they are equi probable, so that is what he assumed. And then, he had defined a probability the P N comma m 1, why N comma m 1, because we have got number of patterns which are mapped into a an m 1 dimensional space.

So, if P N m 1 denotes the probability that a particular dichotomy dichotomy picked at random is phi separable. Phi separable you understand, phi separable means that when we mapping to the phi space, then it gets separable. So, the probability that a particular dichotomy picked at random will become phi separable is given by Cover's theorem.

Proof of that is very much involved and we are giving that, we are directly giving the results of that the probability in m 1 is given as 1 by 2 half to the power n minus 1. Summation m is equal to 0 to m 1 minus 1 and this is a combination the combination I am writing with this notation N minus 1 combination m. So, this is the probability and now you look at these equation, what happens when you are designing it with large value of m 1 surely the probability increases.

And if m 1 very large, then these probabilities will approach one, is it what we want. Yes, we want that because the higher the probability values are that means, to say that better separability they have got. That means, to say that definitely according to cover we can say that high probability is something that we are looking for, that is all for this class continuation in the next.

Thank you.