Neural Network and Applications

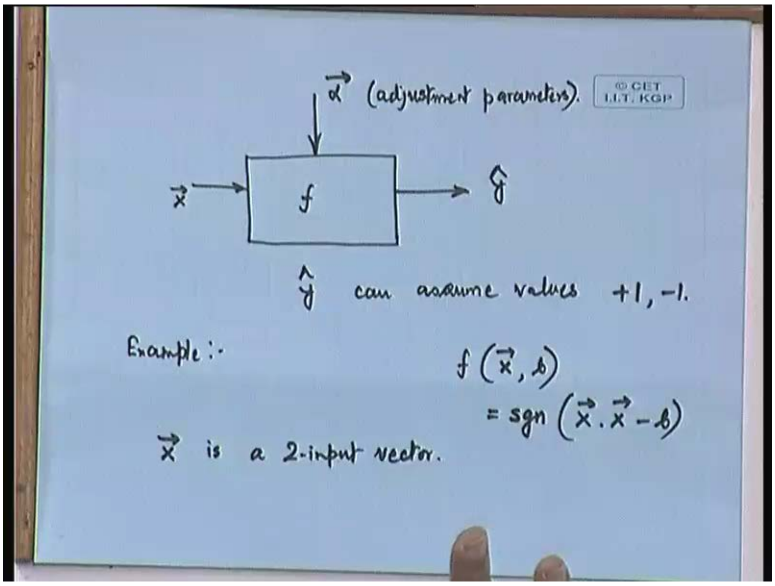
Prof. S. Sengupta
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture No - 10
V.C. Dimensions: Typical Examples

VC Dimensions and we will be considering some typical examples**.** Now, we had introduced this particular topic in the last class, but I think that many of the students felt that some more examples and some simplified explanations when necessary because, it is somewhat difficult to grasp the concept of VC dimensions immediately. But any way I mean, before we present any simplified analysis for that it is surprises to say, that the basic purpose of using the VC dimensions is that, we would like to evaluate for a particular learning network.

For a neural network I mean, be it single layer or a multi layer perception type or any other configuration, we would like to known that how many number of patterns we can reliably feed to the system, so that it can accurately classify. So, for that purpose VC dimension was providing us with some index that is, what we were looking for, but we need to understand it in a little more lucid terms, so that it is possible for us to find out the implications of the VC dimensions for any typical learning machine. Now, we are considering fundamentally the learning machines.

(Refer Slide Time: 02:44)



$\hat{y}$ can assume values +1, −1.

Example :-

$f(\vec{x}, b)$
$= \text{sgn}(\vec{x}.\vec{x} - b)$

$\vec{x}$ is a 2-input vector.

Where, the input is going to be in the form, some vector let us say that x vector that is the input to a learning machine supposing, this is the block diagram of the leaning machine and we can denoted by a function f, because after all this is the function f. Which, we are going to realize in this case, we will be approximating this actual regression function f by the neural network function. If of x w that is what we had discussed earlier and this learning machine will be having a set of adjustable parameters.

Let us say that we indicate that by alpha vector, which is nothing but the set of or the vector indicating the adjustment parameters and we will be getting the output from this network. And let us say for the sake of simplicity, we assume that we have only one output and we indicate the output as y estimated, because this is what we are estimating by using the neural network. Now our estimation that may be correct, it may not be correct. In fact, we are for the time being classifying only the binary, the predicted output we are indicating only as a binary thing.

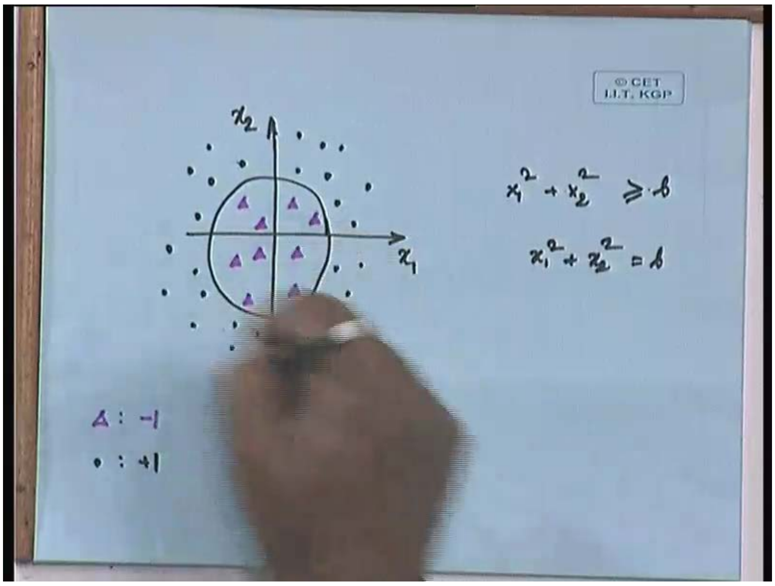
So, that our y cap or y estimate that can assume values in the binary domain that is, either plus 1 or minus 1 and so the thing is that we will be feeding the set of inputs and we will be adjusting this alpha such that, the classification that we make out of this y is going to be correct. Now, what kind of a function can this f be, we are not putting any constraint into this function f, what we simply do is that we can assume some different models of this f function.

So, the first example that we take for the function is that let us say, that we consider the function f as follows that this function f naturally, this f will be a function of x vector obviously, and it will also be a function of some b some constant or bias, which you can think of for the network. And this can be indicated by any typical function and let us say that we consider a signum function whose argument is x vector dot x vector minus b, b can be assume to be a scalar quantity.

Now, can you tell me that let us assume, that x vector is a two input vector, so this is another simplified assumption that we are making that we will be understanding the VC dimension concept from a very simplified dimensional point of view, so we are considering a two input vectors, so our x vector is a two input vector. So, that means, to say that the patterns that we are going to have, the patterns will be forming a set of points in the two dimensional space.

I mean it is easy to imagine or extent the concept to m dimension, because in that case we will be considering that, if x vector is an m input vector then we will be considering an m dimensional space. And there, one particular pattern will be indicated a point in the m dimensional space, so here we will be having the…

(Refer Slide Time: 07:14)



I mean, we will be having a two dimensional vector, two dimensional space. Let us say, that we indicate this by the x 1 and the x 2 as the variables in this two dimensions. And let us say that this is one particular pattern that we are feeding, so one particular pattern will be a point in this two dimensional space and likewise there can be several patterns into it, but if we define the function to be of this form, that it is x vector dot x vector minus b.

Can anybody tell me that what is going to be the nature of this function, see what I mean to say is that here take this argument, take the argument of this signum function. So, what does it mean that as long as this argument is greater than 0, the input will be classified whatever input you are feeding as x vector that will be classified as plus 1 and if the x vector is such that this argument becomes less than 0 then it will be classified as minus 1 simply that.

So; that means, to say that somewhere there will be a decision boundary for this function. And can you tell me that what type of decision boundary does this function realize, x vector dot x vector it is in the two dimension, mind you just a simple two dimensional example can anybody like to have a guess, Student: ((Refer Time: 9:04))

Circle, yes I mean, anybody feels otherwise, so somebody has answered that it is a circle. In fact, it is indeed, so and it is quite easy to verify that you see this x vector is consisting of x 1 and x 2, x 1 coordinate and x 2 coordinate. So, it is a two element vector and if you take any particular point let us say, you take this particular point and then you will be having x vector as this one as the x 1 component, this one as x 2 component and if you are taking x vector dot x vector.

If you are taking the dot product of that in that case what are you getting out of it, you are getting x 1 square plus x 2 square and what is the boundary equation that this argument of this equation that is x dot x minus b that should be greater than or equal to 0, for a signum function realization; which means, to say that x 1 square plus x 2 square that should be greater than or equal to b.

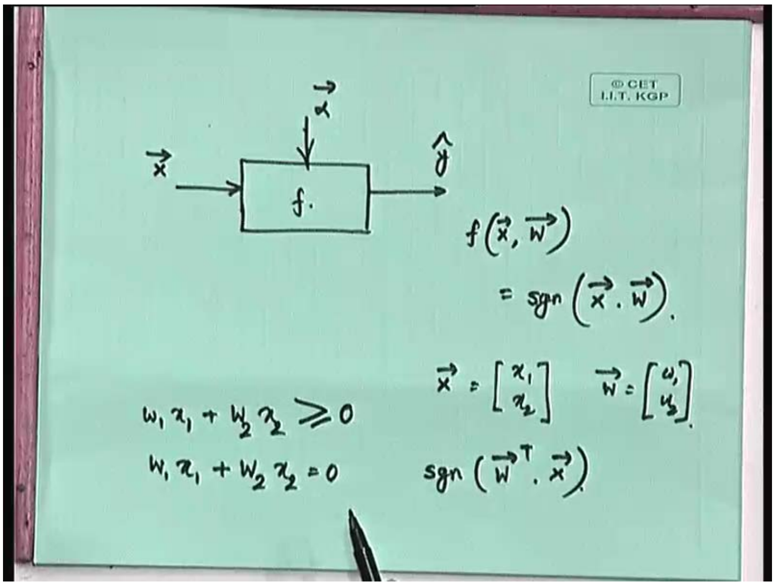
Now, this is when you take the limiting case that is, if you take the boundary of the decision then the boundary of this decision will be x 1 plus x 2 square is equal to b and this is nothing, but the equation of a circle in the x 1 x 2 space. So, here the decision function that we are getting is something like this, let us say that the decision function is realize this way of course, my drawing is not indicating a good circle, but I mean once we indicate an equation of this form.

If the function assumes an equation of this form, this will definitely form a circle like this and we will be having different patterns. Let us say that there are two classes of patterns in this, some classes of patterns are indicated by this color let us say these are the different, so when I am indicating the patterns with this color, this will indicate an output to be equal to minus 1 and when I am indicating the patterns with this color and instead of using the triangle I use a filled circle, just to make a distinction between the two categories of patterns.

If, it is like this if the set of patterns that we are feeding to the learning machine is like this, where all the delta patterns they are minus 1 and all the filled circles they indicate plus 1. Then you can just see that from the diagram that I have drawn there is a proper classification that has happened that all those which are inside all the patterns which are inside the circles. They will be indicating an output equal to minus 1 and all the patterns which are outside they will be indicating plus 1.

So, this is a decision boundary that we could show out of this, now instead of a function like this, so here the function is of this form I mean, where this is a quadratic function that we are taking, instead of a quadratic function if we take a simple linear function. Then what happens, let us say that again we take the model as follows that x is an input.

(Refer Slide Time: 13:18)



$$f(\vec{x}, \vec{w}) = \text{sgn}(\vec{x} \cdot \vec{w}).$$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

$$\text{sgn}(\vec{w}^T \cdot \vec{x})$$

$$w_1 x_1 + w_2 x_2 \geq 0$$

$$W_1 x_1 + W_2 x_2 = 0$$

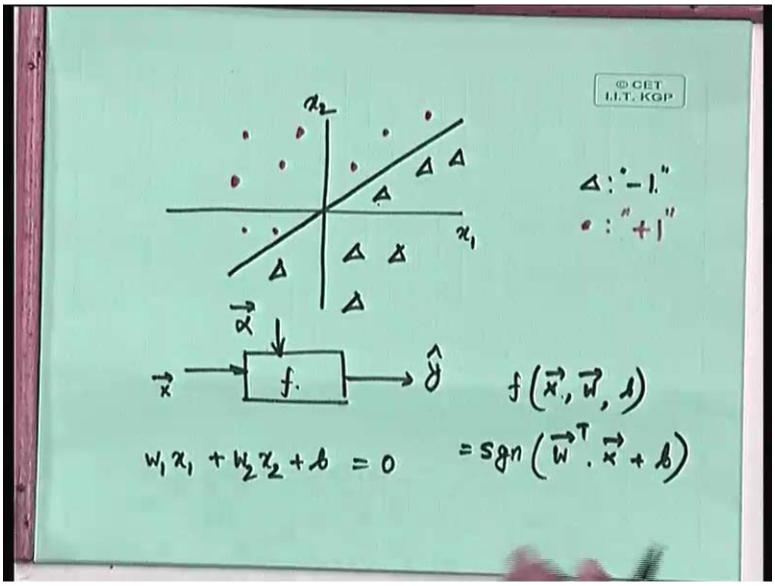
X vector the other is an input and f is the learning machine for which the adjustable parameters are alpha and y cap is the output of that, in that case we can I mean, the function f that we are going to realize is defined as follows, f as a function of w vector and x vector and w vector. We can indicate as the signum function of x vector dot w vector, where w is also here or here to be assume a two dimensional space, so x vector is two dimension even w vector is also of two dimension.

So, in that case what happens is that I mean, this is the kind of a function that we take, so if we take x vector to be x 1 x 2 and we take the w vector to be w 1 w 2. In fact, to wright it in a more precise form, we should not write it this way, we should rather say signum of w transpose dot x. So, that ultimately what we are going to have is w 1 x 1, so we will be writing w transpose as say row vector of w 1 w 2.

So, it will be w 1 x 1 plus w 2 x 2, again here the decision boundary will be such that for w 1 x 1 plus w 2 x 2 greater than or equal to 0 the patterns will be classified as plus 1 and for otherwise I mean to say the patterns will be classified as minus 1. So, here also the decision boundary that we are getting is of the equation of this form w 1 x 1 plus w 2 x 2 is equal to 0, here we assumed that the bias term is 0.

I mean we should not assume the bias to be 0 because typically some bias would be there even if we assume that b is equal to 0, let us say here w 1 x 1 plus w 2 x 2 is equal to 0 means that it is a straight line in the two dimensional space whose slope is going to be what I mean, it will be minus of w 1 by w 2 that will be its slope and intercept in this case is 0. Because this is not having any bias, so the function that we realize out of this is of this form.

(Refer Slide Time: 16:14)



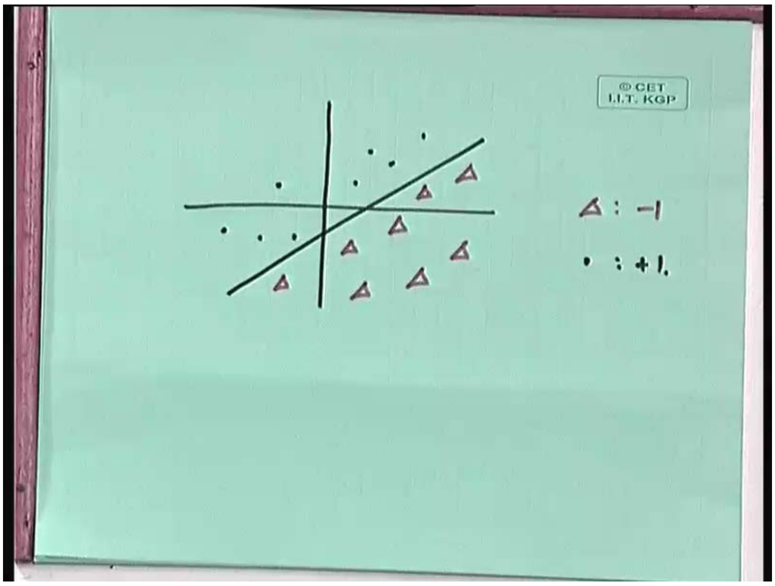
That we take x 1 space I mean, x 1 x 2 space and our decision boundary will be a straight line like this and on one side of it we will be having all the patterns, which are categorized by delta which will indicate, so here this deltas will indicate the patterns which give an output equal to minus 1 and another set of patterns, which will give us an output equal to plus 1. So, all the filled circles that I am drawing they will produce an output equal to plus 1.

So, this is the form of the equation, now just to complicate matters little bit that in the same learning machine, we take x vector and we take the function f alpha the set of adjustable parameters y cap the estimated output. We take a function f of x vector, w vector and b vector and we define the function as the signum function of w vector transpose dot x vector plus b. In fact, to this transpose type of representation would have been required here also.

So, here also we are taking the x vector definition is a column vector, so we should indicate this as x transpose x vector, so that it assumes in the quadratic form, so you can make this correction that it should be x transpose x vector I mean in the case of circle. Where as in this case it is w transpose x vector plus b, so for this the decision boundary equation will be w 1 x 1 plus w 2 x 2 plus b is equal to 0, this is the decision boundary, so greater than that will classify the patterns into this, less than will classify into this and here this b basically is an indicative of the intercept of the straight line.

So, if we vary b then what we are varying is the bias or the intercept of the straight lines. So, that is why that whereas, with 0 intercept the function f was looking like this passing through the origin in this case with intercept.

(Refer Slide Time: 19:00)



The function will not pass through the origin, rather that will pass through the origin only in a special case when b is equal to 0, but otherwise I mean with some b it will go like this. So, that the decision boundary will be such that on one side of it, we will be having all the patterns that give us output equal to minus 1 and on the other side we will be having the patterns with output equal to plus 1.

So, these are some of the typical functions that one can take, but we have to really come back to our original purpose of specifying and really understanding that what VC dimension is any question at this stage.

Student: ((Refer Time: 19:57))

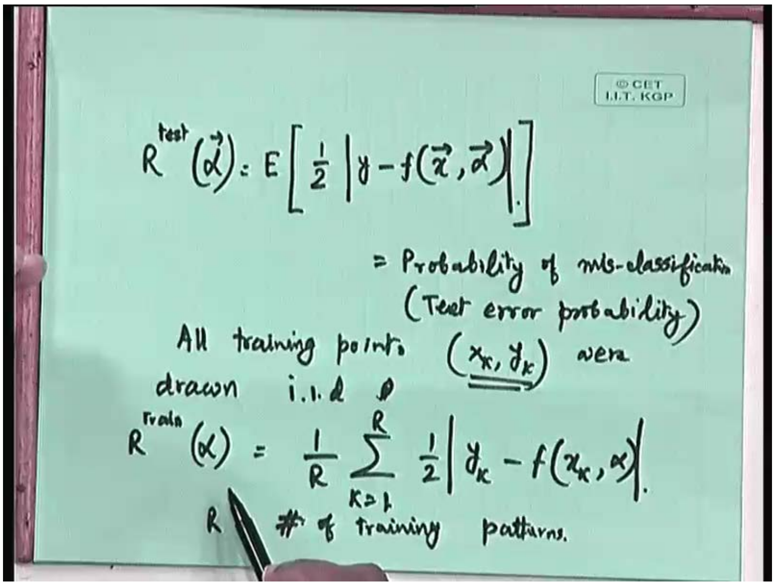
Let us see I mean, this is a good question, the question is that what is this alpha, In fact here alpha or the adjustable parameters are nothing, but w’s and b’s that is all, all alright.

So, typically as a general case I would write that f is a function of x vector and the adjustable parameters alpha vector, where alpha will indicate both b as well as the w's. Now, in this case, I took the simplest of case that we have only three adjustable parameters w 1 w 2 and b.

Now, with this type of simplified functional representation in mind, we will give some typical examples of VC dimension, but before we give that let us, define some quantities which we need to define in order to specify the limits of the VC dimensions as indicated by Vapnik, the person who purposed this VC dimension concept.

Now, let us…

(Refer Slide Time: 21:17)



© CET I.I.T. KGP

$$R^{test}(\vec{\alpha}) = E\left[\frac{1}{2}\left|y - f(\vec{x}, \vec{\alpha})\right|\right]$$

= Probability of mis-classification (Test error probability)

All training points (x_k, y_k) were drawn i.i.d

$$R^{train}(\alpha) = \frac{1}{R}\sum_{k=1}^{R}\frac{1}{2}\left|y_k - f(x_k, \alpha)\right|.$$

R # of training patterns.

Take a quantity of this form, let us take the expectation of y minus f of x alpha, x vector alpha vector and we take the mod of this and as a very typical case, I can take it to be half of this. Now, what does this basically indicate, if I take the expectation of half into mod of y minus f of this. And assuming in this case that y could be assuming a value equal to plus 1 or minus 1, that means to say that if y is equal to plus 1 and f of x alpha is also a plus 1.

In that case it is indicating a correct classification in which case this will be equal to the expectation of this in this case the error will be 0. Whereas if we have y as plus 1 and f of x alpha as minus 1 then it will be, this mod of this quantity will be 2, so that is why we are multiplying it by half so that it will be an error equal to 1. So, for some of the patterns the error will be equal to 0 and for some of the patterns the error will be equal to 1.

So, if we take the expectation of this that is something which is going to lie between 0 and 1, so we can as well indicate it as a probability of misclassification. So, by this by taking the simplest of assumption as plus 1 minus 1 categorization, we can indicate that this expectation of this quantity is a probability of misclassification for the test patterns, you see here I have purposely indicated this to be y, I did not put any y k, I did not put any training pattern, I assume that y is a test pattern.

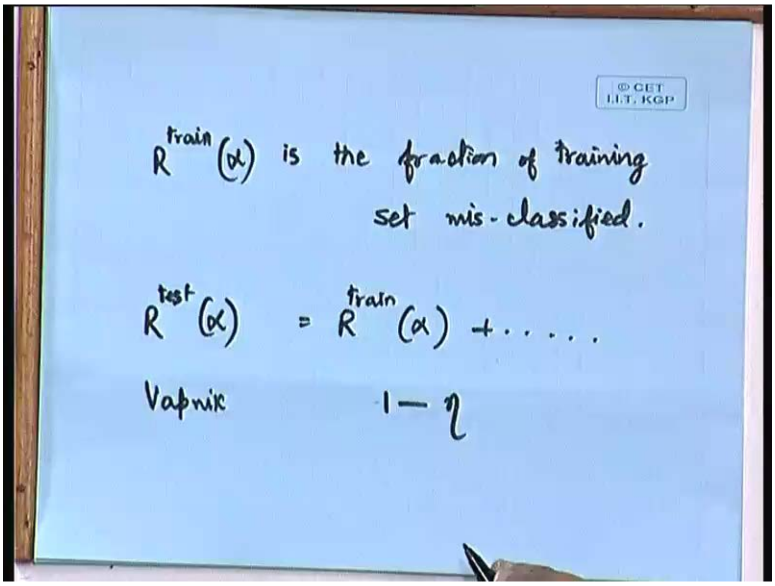
So, it is a probability of misclassification or you can also call it as a test error probability and this is defined very rightly as E of half of this quantity and I, now indicate this as R test because this is a test pattern, so this is R test as a function of alpha that is equal to E of this. Now, let us say that all the training points, so let us assume that all the training points, that is the set of x k y k they were drawn as identically independent distribution, so i i d distribution.

So, then also the future test points also will be taken from that same i i d anyway, with that assumption, we can define the training accuracy or the fraction of training set misclassified as R of train alpha, which will be equal to 1 upon R into summation half of mode y k minus f x k alpha, mod of this whole quantity where we add it up from k is equal to 1 to R. Now, please take time to understand what I have written, I have assume that there are R number of, so R is the number of training patterns, I indicate the R to be the number of training patterns.

So, there are set of training points xk yk, yk will vary from 1 to R and I have indicated the quantity here as summation of half into yk minus this and what does that indicate, so overall this when added up, when added up for all the R patterns and then we divided by 1 by R, that will indicate what the fraction of training sets which are misclassified in this. Now, where is the question of training set becoming misclassified, it can; obviously, happen you are feeding the training set and you are adjusting the parameter.

So, with some adjusted parameters or for some set of parameters, it could be that your network is not classifying the training patterns also properly. So, this R train will indicate that, what is the fraction of training set misclassified, so I should write this as.

(Refer Slide Time: 26:52)



© CET I.I.T. KGP

$R^{train}(\alpha)$ is the fraction of training set mis-classified.

$R^{test}(\alpha) = R^{train}(\alpha) + \ldots\ldots$

Vapnik $1-\eta$

R train alpha is the fraction of training set fraction of training set now, without going into any detailed regress mathematical analysis. If I ask you that out of these two parameters that we have defined just now, that is R test and R train, out of this which one would you expect to be lower, naturally you would like to have this value of R to be as low as possible is in it because you want correct classification. So, that is why your R should be low, so out of this R test and R train which one would you expect to be.

Student: ((Refer Time: 27:55)) obviously, you would expect R train to be lower because, you are adjusting all the parameters by feeding the training patterns and then only you are applying the test patterns and trying to find out its accuracy, so naturally the network is already tuned to the training patterns. So, without going into any mathematical analysis we can say, that R train is going to be less than that of R test or to.

Student: ((Refer Time: 28:25))

No, no I mean not alpha k you can say that it is again alpha vector because what we are doing is that xk yk is a pattern which is applied on the network and the networks adjustable parameters are alpha vector, so I should say x k comma alpha vector.

Student: ((Refer Time: 28:54)

Xk xk yk is a training pattern, training point.

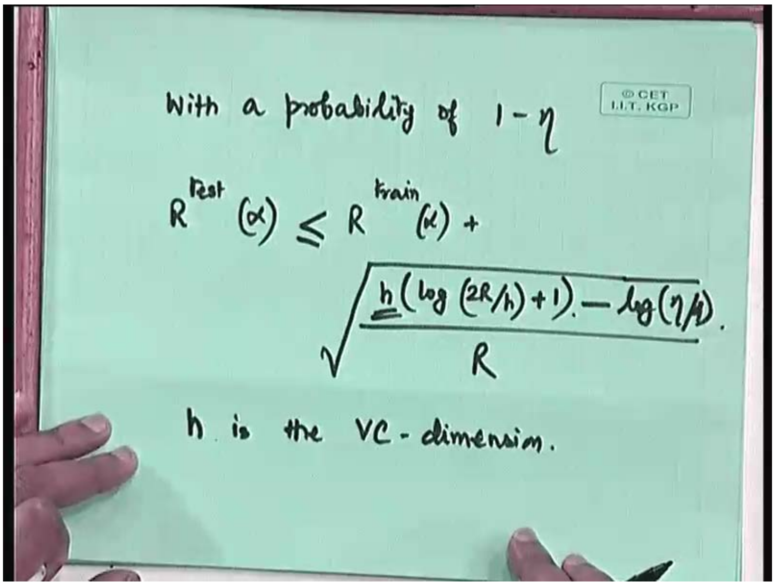
Student: ((Refer Time: 29:04))

But, no, no but the weights need not be w 1 and w 2 only I mean it could be a multidimensional weight, so whatever free parameters are there in the system, so I am indicating that all the free parameters in the systems by this alpha vector. So, the free parameters will remain as it is, I am not putting different free parameters for different test patterns for the free parameter of the systems are adjusted already.

I mean you take it that way or may be at the end of the pattern, it is not that every independent pattern has got independent this, this is the characteristic of the network whereas, xk yk is a pattern, so let us not confuse between these two. So, now; obviously, what, I mean to say is that our R train is going to be less than this, so that is why one can express the R test as R train plus something.

R test alpha it is possible to write that R test alpha will be R train alpha plus something, what is that something that we do not know, but actually it is Vapnik, he has shown analytically that with a probability of 1 minus eta, eta is in this case, should not be confused with the learning rate. Eta in this case is some constant, where it has been shown by Vapnik one interesting result. That is what I am presenting to you because the whole concept of the VC dimension actually originated from that.

(Refer Slide Time: 31:10)



With a probability of $1 - \eta$

$$R^{\text{test}}(\alpha) \le R^{\text{train}}(\alpha) + \sqrt{\frac{\underline{h}\left(\log(2R/h) + 1\right) - \log(\eta/4)}{R}}.$$

h is the VC - dimension.

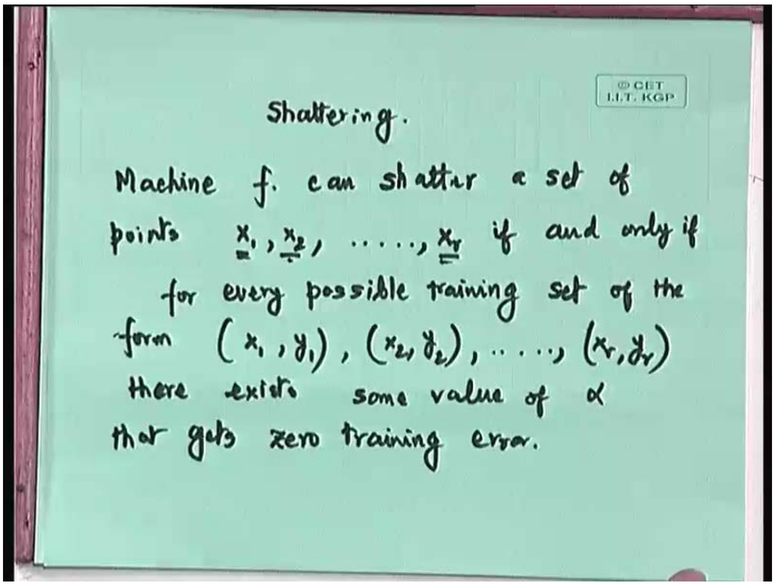
He showed that R test alpha again, I am not indicating every time alpha to be a vector, I think you please understand that, R test alpha should be less than or equal to R train alpha plus a quantity, which I am writing it is under root h log of 2 R by h plus 1 minus log of eta by 4 this divided by R, where R again is the total number of training points.

Now, in this case eta as I told is a constant, so Vapnik’s result shows that with a probability of 1 minus eta R test can be written like this.

The, what we do not know is this quantity h and this quantity h is what is the VC dimension. Here h is the VC dimension, but again this result is of no use to us, if given a network, given a learning network we do not know how to measure h, so we should know that how to get an idea about the h given a network. So, if we know what h is then naturally h will give us an indication that, how to obtain this R test and R train from a given set of training patterns.

Now, in order to arrive at this VC dimension concept, we define what shattering is, in the last class also we defined, but I think that we needed some more examples without which it is I mean, I can well understand that it will be very difficult to get a concept of shattering.

(Refer Slide Time: 33:39)



© CET
I.I.T. KGP

Shattering.

Machine f. can shatter a set of points $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_r$ if and only if for every possible training set of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_r, y_r)$ there exists some value of α that gets zero training error.

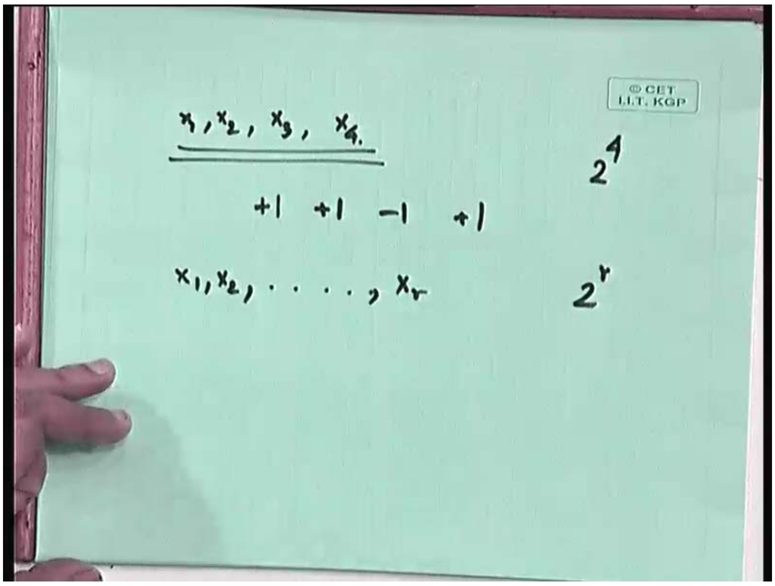
Now, today I am going to define the shattering in a more simplified way, now we can say that a machine f, what is machine f, machine f is nothing, but these things the block diagram that we have shown here, this f the function that we are taking that itself we are taking it as a learning machine f. We can say that a leaning machine f can shatter a set of points x 1 x 2 etc, etc up to x r, so we are taking r number of different training points.

So, it can shatter a set of points, if and only if, for every possible training set of the form, now how are we indicating the training set; obviously, as x 1 y 1, x 2 y 2 up to xr yr, this many training sets. There exists some value of alpha, what is that alpha again the free parameters, some value of alpha that gets zero training error. Now, you see our objective is to have as minimum of training error as possible, I mean if possible will try to have a zero error. In fact, zero training error is definitely going to be our objective.

Then, there should not be any error while at least training the system, now if given r number of patterns like this, it is possible that for some particular value of alpha, the system leads to zero training error then we can say that these set of points a x 1 x 2 up to xr has been shattered by f. Shattered or in other words fully exploited by x, so let us understand it this way.

Now, here you see that once we take x 1 to xr, here now x 1 can assume a value of I mean, because of the x 1 pattern y 1 could assume a value of plus 1 or minus 1 again for x 2 it could be plus 1 or minus 1. So, that there will be 2 to the power r such training sets to consider is in it, one of the training sets could be I mean, let us take a example of let us say four points.

(Refer Slide Time: 37:15)



$\underline{x_1, x_2, x_3, x_4.}$ $\qquad 2^4$

$+1 \quad +1 \quad -1 \quad +1$

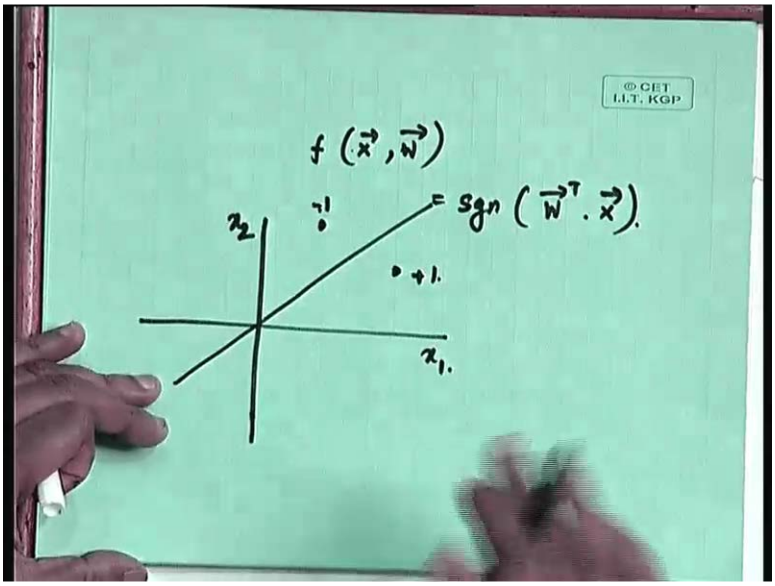
$x_1, x_2, \ldots, x_r \qquad 2^r$

Let say, x 1, x 2, x 3, and x 4, so could be that the training patterns are such that when I give x 1 as input the output is plus 1, when I give x 2 as input the output is plus 1, when I give x 3 as input the output is minus 1, when I give x 4 as input the output is plus 1, if it is like that. Then this is a particular training pattern plus 1 plus 1 minus 1 minus 1, this is one output that results out of this.

Now, there could be, so here with four such inputs we can form 2 to the power 4 or 16 such patterns is in it, I mean the space, the whole space will contain 16 patterns. So, when I have x 1, x 2 up to xr then we can have up to 2 to the power r such training sets. Now, given this example of shattering, given this definition of shattering that if you can find for a give learning machine, for a given f, if you can find some adjustable parameters or some values of alpha, which will be which will give you a zero training errors, then it is shattered.

So, let us now take some very typical example, and again in order to simplify matters I will take examples from two dimensions only, because only two dimensions I can represent on a piece of paper. Although, the examples that we will be presenting will be too much over simplify, but I think that if you understand that, you can extend this concept again to the multi dimension case. Now, let us take a function of this nature, let us take….

(Refer Slide Time: 39:12)



That f of x w to be equal to signum function of w transpose dot x, if we take it to be like this, then this is the space, can you see there is some problem with the lighting, so never mind I mean, if you can see you can continue with that, now again we take a two dimensional space x 1 and x 2. Now, if we let us say have only one point, only one point, now, is it guaranteed that it can shatter one point obviously.

This is the most trivial case because I can imagine, that this straight line passes like this or a straight line passing like this, so it is possible to find out a solution that classifies this points. So, one is actually a trivial case, now supposing I take one pattern over here and another pattern over here. Now, is it possible to shatter this two points using a straight line of this form are you sure answer is yes or no.

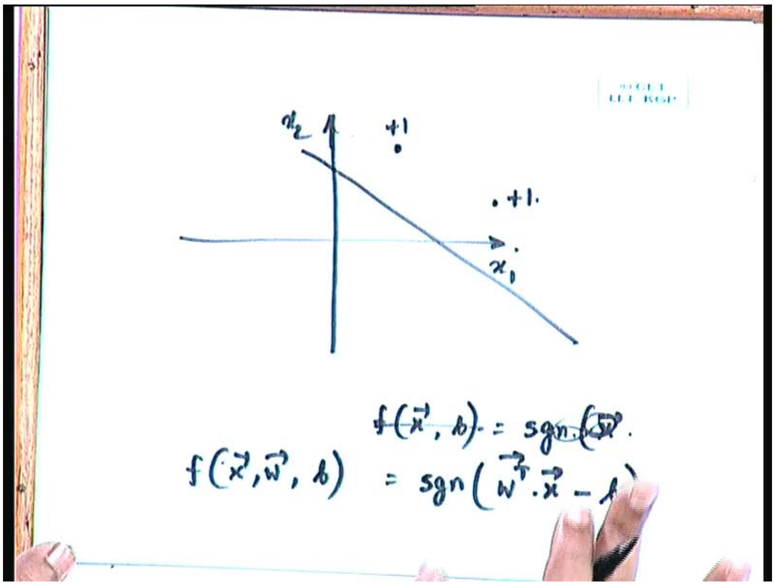
Student: ((Refer Time: 40:53))

You are saying yes, you are assuming if you assume that this is minus 1 and this is plus 1, obviously I can find the solution a straight line passing like this, but if I take this to be minus 1 this also to be minus 1 or I take this to be plus 1 and this to be minus 1, both to be plus 1 or both to be minus 1, then is it possible to find a solution, it is not.

Student: ((Refer Time: 41:30))

I mean, I can hear some very interesting arguments, some people are saying that it is not possible. In fact, that is what is correct that it is not possible, but again somebody said that we require two lines, it is not that we require two lines I mean let us, try to look at it again.

(Refer Slide Time: 41:58)

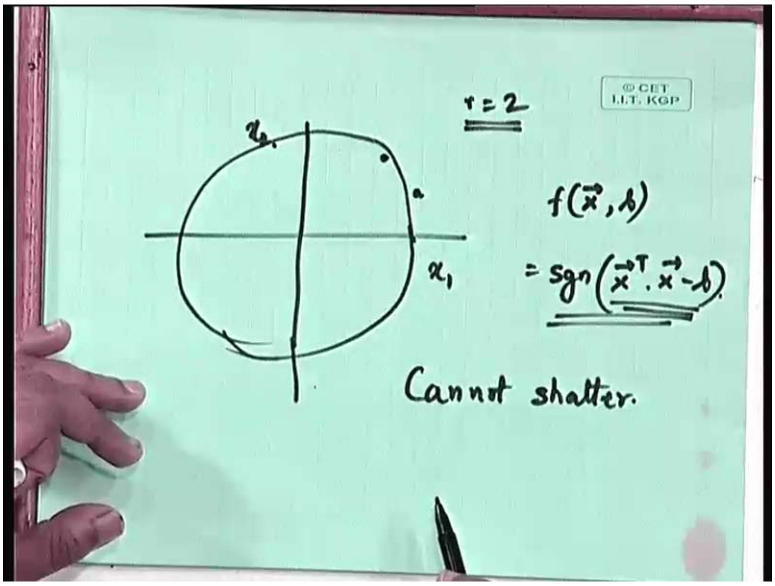

x_2
+1
+1.
x_1
$f(\vec{x}, b) = \text{sgn}(\vec{w}$
$f(\vec{x}, \vec{w}, b) = \text{sgn}(\vec{w}^T \cdot \vec{x} - b)$

We say that, we take the x 1 x 2 space and here we take this as a pattern, which gives plus 1 and this as a pattern which gives minus 1. Now, is it possible to find a line, it is possible only thing is that I mean with both as plus 1, let us say with both as plus 1 or with both as minus 1, can you separate.

Student: ((Refer Time: 42:36))

You can, what you can do is that you have to take a line may be of this form, this line can, so this lines equation will be actually of the form f x b here, we have to introduce a bias to the straight line. We should write it as f x w b as signum of w transpose x minus b. So, this is the form and with this type of a straight line, it is possible to shatter. Now, let us take the quadratic function, let us take the circular function as we had taken.

(Refer Slide Time: 43:51)



Let us consider again x 1 x 2 space and we take two points let us say, we take one point here and one point here and we consider the function f x comma b, which is defined as signum of x transpose x minus b. Now, is it possible for us to define a I mean to shatter this point using this function, shatter two points using this function.

Student: ((Refer Time: 44:42))

See, there you will have to note one thing, that the patterns could be having any 2 to the power R combination, so here I have got two points. So, my R is equal to 2, so I can have four combinations, I can have this as plus 1, this as minus 1, this as plus 1 plus 1 both minus 1 minus 1 both or this as minus 1, this as plus 1. So, it is possible for us to define the patterns in, so I am saying that for all these 2 to the power R combinations, you should be able to find out a solution, that is it, is it possible to separate out these.

Student: ((Refer Time: 45:39))

It is not possible you see that I, keep on increasing this circles radius, radius I can increase by using this b, now I can try to set a circle of this, again these two points, certainly if I have both to be plus 1 or both to be minus 1 then I cannot separate using these. So, here you see that the same number of point that is for just two points, the quadratic function is unable to separate whereas, the linear function which we considered just little while back could shatter two points.

So, here linear function can shatter two points, but a quadratic function cannot, so this cannot shatter, whereas in this particular example this can shattered two points.

Student: ((Refer Time: 46:55))

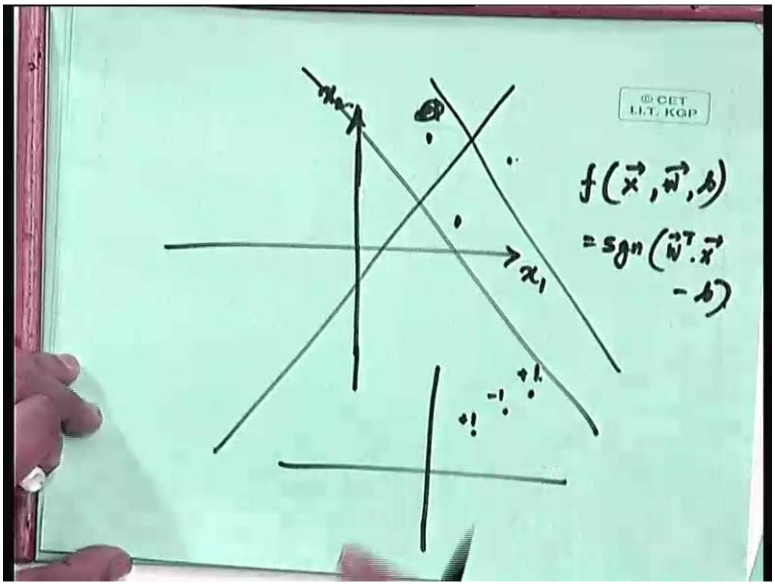
Look, for any combination I can shatter that is the that is the definition for any of the 2 to the power R, I should be in a position to shatter, you see here, no matter whether I have plus 1 plus 1 or minus 1 minus 1 or plus 1 minus 1 I mean, minus 1 plus 1. I can shatter using this equation, whereas, I cannot shatter using a quadratic equation.

Student: ((Refer Time: 47:36))

Yes, both become plus 1 plus 1 and completely no, the thing is that you have to consider all these possibilities, that it could be plus 1 plus 1 plus 1 minus 1 minus 1 plus 1, all these four combination should be possible with these two points. Then is it possible for the circle to shatter, it is not, for a circular function to shatter, it is not whereas, given two points any of this four combinations can be shattered by using straight lines, but using circles it is not.

Now, let us take the straight line case only, because straight line we can well understand that straight line can shatter two points. Let us increase it to three points.

(Refer Slide Time: 48:35)



Let us say, that we have a two dimensional space here x 1 x 2 and we take three points let us say plus 1 I mean three points, now these three points could have any combination I mean there are now how many eight combinations possible, is in it, all of these can be plus 1 plus 1 plus 1 in this case I take line like this it is possible. Now, I take let us say this as plus 1 and these two as minus 1, I pass the straight line like this, it shatters.

Now, I take two of these as minus 1 and this as plus 1, I can have, so can these three points be completely shattered.

Student: ((Refer Time: 49:41))

One is I mean, what is your point of argument is a very, very interesting case that somebody has considered all the three points are collinear.

Student: ((Refer Time: 49:59))

Plus 1 minus 1 and plus 1, you cannot you cannot yes, this is the typical case, where you cannot, When they are falling in a line yes, but except for that case you can shatter using this any.

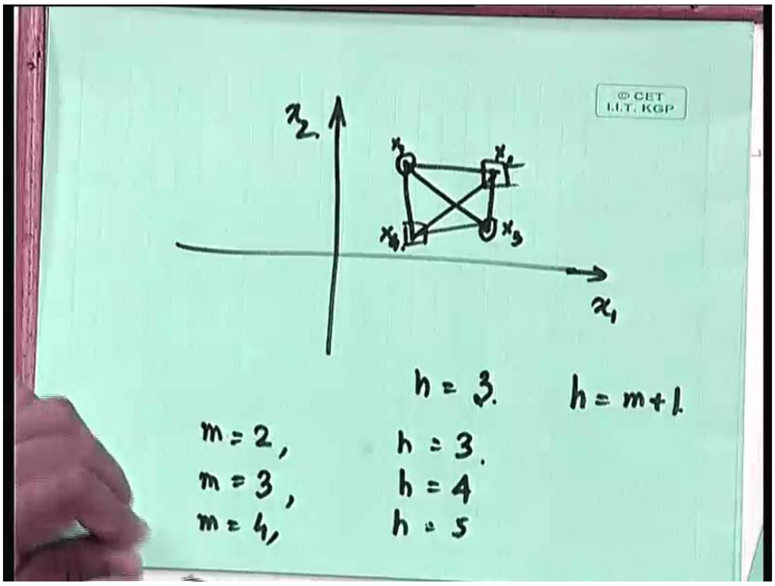
Student: ((Refer Time: 50:28))

No, when I take a function f, I should say that the function f of, f x comma w comma b. When it is of the form signum of w transpose x minus b, It can shatter a set of three points that is what I am saying.

Student: ((Refer Time: 50:58))

Obviously, because my adjustable parameters are w’s and b’s, so, what is the definition of shattering again I go back to the example to the basic definition, you see can shatter this set of points if and only if for every possible training set of this form there exists some value of alpha. So; that means, to say that if I can manipulate my w’s and b’s, such that I can steer it towards a shatter then it is possible.

(Refer Slide Time: 51:44)



So, up to three points we can. Now, let us say that, now I increase this and I take four points, X 1 X 2 X 3 and X 4, four points are there in the x 1 x 2 space two dimensional space.

Student: ((Refer Time: 52:05))

Yes, no this is a particular case this shows; that means, to say that we have to consider the case of non-collinear patterns, non-collinear patterns.

Student: ((Refer Time: 52:34))

No, non-collinear that is all once you say non-collinear, so any three non-collinear points can be shattered using the function that we had used I mean, using the linear function we could shatter three points. Now, using the same linear function can we shatter four points, I think the general consensus that is emerging out of this class is that is we cannot. In fact, you can look at this way that you have got four points.

X 1 X 2 X 3 and X 4 and you can have actually here 16 number of combinations of pattern that is possible out of it and if you say that you are having X 1 X 3 to be one class of patterns and X 2 X 4 to be another class of patterns, in that case you cannot imagine a straight line that can separate the plus 1 patterns from the minus 1 patterns. In fact, the basic philosophy lies like this that when, there are four points then by joining any two pair of points out of these four, you can have how many six such lines is in it.

And two such lines, they intersect with each other, the two diagonals, if we are taking the line joining X 1 X 3 and X 2 X 4 they intersect with each other. So, that is why if you are having patterns which give you plus 1 values at X 1 X 3 and which give you minus 1 values at X2 X 4, then it is not possible for you to pass any straight line, however adjustment you can do with w's and b's, it will not be possible for you to have that.

So, naturally four is ruled out, so how many points could we, shatter using the linear function, we could have three, three non-collinear points, so in this case I should say that h is equal to three. So, the VC dimension for this particular for a special case of two dimensional case, when we have only x 1 and x 2 as the variable, there h becomes equal to three; that means, to say that up to three patterns we could shatter or out of 2 to the power eight combinations.

I mean 2 to the power 3 that is eight combinations could be shattered or rather three points could be shattered, so the VC dimension h is equal to three, so when the dimension that is m is equal to 2, we have h is equal to 3. Now, it is possible to extend this analysis for hard dimension when m is equal to 3 then we can show that h is equal to 4 and when m is equal to 4 h is equal to 5. So, the law that immerges is that h is equal to m plus 1.

One more than what the dimension of the system is. So; that means, to say that if we are taking let us say five dimensional input X 1 X 2 X 3 X 4 X 5, that means to say that up to six patterns can always be reliably separated or reliably separated means the correct classification exists up to six patterns, but this result is somewhat worrying to us, because if we are spending our time and energy to build five input network and ultimately, we show that it can only classify six patterns.

We are expecting somewhat more than that is in it. So, that is why, we have to see that whether a single layer network is good enough or if, we go in for a hired layered network where we cascade one layer with another then whether, we can increase the VC dimension or not that is something what we have to see later on. So, that is all for today's class, so in the next class we will be discussing one or two more points, related to this VC dimension and then we will go over to the next topic, that is the single layer perceptron.

Thank you very much.