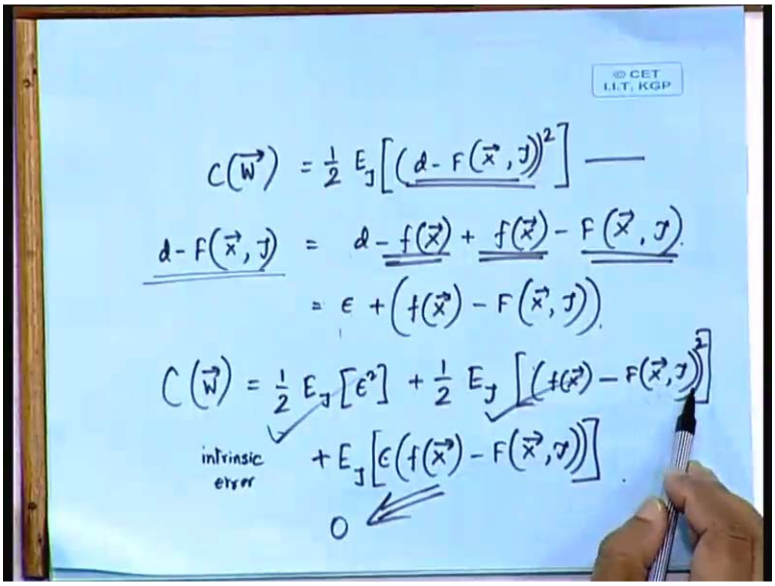
Neural Network and Applications
Prof. S. Sengupta
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 09
Statistical Aspects of Learning

Statistical Aspects of Learning, as a matter of fact in the last class itself we had introduced this topic on completion of our discussions related to the associated memory. And there, towards the end of the class we were deriving an expression that involve the cost function, our objective in the statistical aspect of learning was to get an approximation function, using the neural network. And that approximation function we are calling as f as a function of x and w.

That closely approximates the regression function which is f of x. So, we were interested in finding out that how much of fitment error is there, between the realization of actual effects and the f of x w, which you are realizing out of the neural network.

(Refer Slide Time: 02:10)



$$C(\vec{W}) = \frac{1}{2} E_{\tau}\left[\underline{\left(d - F(\vec{x}, \tau)\right)}^2\right]$$

$$\underline{d - F(\vec{x}, \tau)} = d - \underline{f(\vec{x})} + \underline{f(\vec{x})} - \underline{F(\vec{x}, \tau)}$$

$$= \epsilon + \left(f(\vec{x}) - F(\vec{x}, \tau)\right)$$

$$C(\vec{W}) = \frac{1}{2} E_{\tau}[\epsilon^2] + \frac{1}{2} E_{\tau}\left[\left(f(\vec{x}) - F(\vec{x}, \tau)\right)^2\right]$$

intrinsic error

$$+ E_{\tau}\left[\epsilon\left(f(\vec{x}) - F(\vec{x}, \tau)\right)\right]$$

$$0$$

And we were deriving the expression for cost function. So, starting with the definition of the cost function, where we were taking it as the expectation over the set of target patterns of d minus F x comma tau the square of this. So, expectation of this squared quantity over the training samples that is what we were calculating. And just for the ease of our expressiblity, what we did was that this d minus F term we had broken up into 2.

So, that we introduced the minus f of x vector in the linear regression form in the regression functional form, which we subtracted and then we added. So, that idea was to separate this term that is d minus f x and to have f x minus F x comma tau as separate entities. So, this term d minus f x was our epsilon that is error and f x minus F x tau that was nothing but, the deviation that we have between the function. between the regression function f x and the neural network realization.

And we substituted this d minus f x expression into the original definition of the C w and we got as C w this expression. In fact, we had got naturally this being a sum of two quantities, if you square it up since you have to do a square of this d minus F term, if you square it up, you get epsilon square term which is representing the intrinsic error.

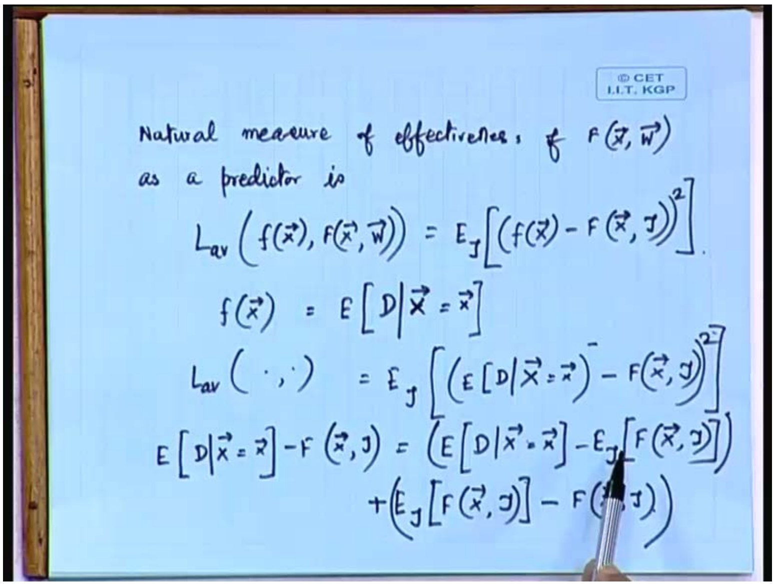
So, basically this is nothing but, intrinsic error that is nothing but, the error that we are having between the f x and it is the mean expectational error of f x. That is, what we were representing by this epsilon and then we had the square of this quantity without any problem and we had a term that involved in multiplication of epsilon with this.

Now, this term will equal to 0 and in fact, that is what we were discussing towards the end. This term will equal towards 0, simply because the intrinsic error epsilon is uncorrelated with respect to both f x and the neural network realization function f of x tau. And since, it is uncorrelated with respect to both of this, this term will equal to 0.

So, we get C w as a sum of these two expressions and out of this, this term is certain in non varying with respect to w. So, if we are doing any optimization with respect to the weight, then this term will not contribute and what contributes is only this one, that is this half of E over t. And it is the expectation or the average operator over the training samples for this f x minus F x comma tau, in fact that is what to be a interested in finding out.

So, existence of this term effectively is the neither we can call this term to be the natural measure of the effectiveness of F of x tau as a predictor, what is after all the F of x tau, F of x tau is nothing but, the predictor of the system. So, how effective this F of x tau is as a predictor, that will be represented by expectation over the training samples t of this quantity f x minus F x tau, this term square. So, the natural measure of effectiveness we can write in this manner.

(Refer Slide Time: 06:31)



© CET I.I.T. KGP

Natural measure of effectiveness, of $F(\vec{x}, \vec{w})$ as a predictor is

$$L_{av}\left(f(\vec{x}), F(\vec{x}, \vec{w})\right) = E_{\mathcal{T}}\left[\left(f(\vec{x}) - F(\vec{x}, \mathcal{T})\right)^2\right].$$

$$f(\vec{x}) = E\left[D \mid \vec{X} = \vec{x}\right]$$

$$L_{av}(\cdot, \cdot) = E_{\mathcal{T}}\left[\left(E[D \mid \vec{X} = \vec{x}] - F(\vec{x}, \mathcal{T})\right)^2\right]$$

$$E\left[D \mid \vec{X} = \vec{x}\right] - F(\vec{x}, \mathcal{T}) = \left(E\left[D \mid \vec{X} = \vec{x}\right] - E_{\mathcal{T}}\left[F(\vec{x}, \mathcal{T})\right]\right) + \left(E_{\mathcal{T}}\left[F(\vec{x}, \mathcal{T})\right] - F(\vec{x}, \mathcal{T})\right)$$

The natural measure of effectiveness, effectiveness of what of F x comma w as a predictor. So, the natural measure of this is L average, so L we are indicating by L and argument of L will be f of x since we are predicting this f of x using F of x w. So, this will be E over the training set t average of this f of x minus F of x comma t this term square expectational this.

So, this is the natural measure of effectiveness. And now, let us invoke the original definition of our f x after all what is f x, we had already learn that f x is nothing but, the expectation of the random variable D of the given the random variable x is equal to 1 realization which is x vector. So, f of x vector is equal to expectation of random variable D given, random variable x is having an instance x vector.

So, now if we use this then the L average of this I am not writing the full thing once again is equal to E t of instead of f x we have to write it as E of D given X vector is equal to the large X vector is the random variable and the small x vector is an instance of that. So, this one, so expectation of this term minus F of x comma t as it is and the square of this, so simply f x being replaced by nothing else.

Now, what we can do is that, this is the average value of the estimation error between the regression function and the approximating function evaluated over the entire training say t alright. Now, what we can do is that, this term we can rewrite in a more convenient form, what we can do is that this E of D that is the term, which is there within this E t expression E of D given capital X vector is equal to small x vector minus F of x vector comma t that can be written again as a summation of two quantities, what E of D given x equal to x vector minus E t of F x vector t.

And what is this, let us just carefully look at this term, by saying E t of this function what to say is that, this function is what, this function is our the function that we are realizing out of the neural network. And we are averaging it over the training set, because this function will have different instances of realization is not it, because we have got a large training set. So, for different training patterns it will have different functions realized and we are only taking the average over the training set.

So, that is why we are just bringing in a quantity which is nothing but, the average of this realized function, average toward what average toward the training pattern s t. So, we are just introducing this term, so that is what we have already subtracted from this, so what we have to do is also to add up this term. So, now in order to compensate for this we add up this term again which is E t of F x comma t this one minus F x comma t. So, two quantities sum of two quantities this plus this.

Now, the question is that, what is the interpretation of these two quantities, what is this one E of this expression. That means, to say nothing but, the regression functional the definition of the ((Refer Time: 12:26)) function minus the averaged value of this one. What can you call it in statistical quantity anybody can suggest a good name for that E minus the expectation of the realized function.

Student: ((Refer Time: 12:45))

Not, not the variance, variance actually we will come to this actually this is the summation of two quantities. Let us put it into this expression, because after all we are finding out this L average is not it, so we are finding out the natural measure of effectiveness. So, to find out the natural measure of effectiveness we just express this quantity as a summation of these two quantities. As since, we have to take a square of these two sum quantities a plus b, if you take this to be a and if you take this to be b, you are doing a plus b whole square.

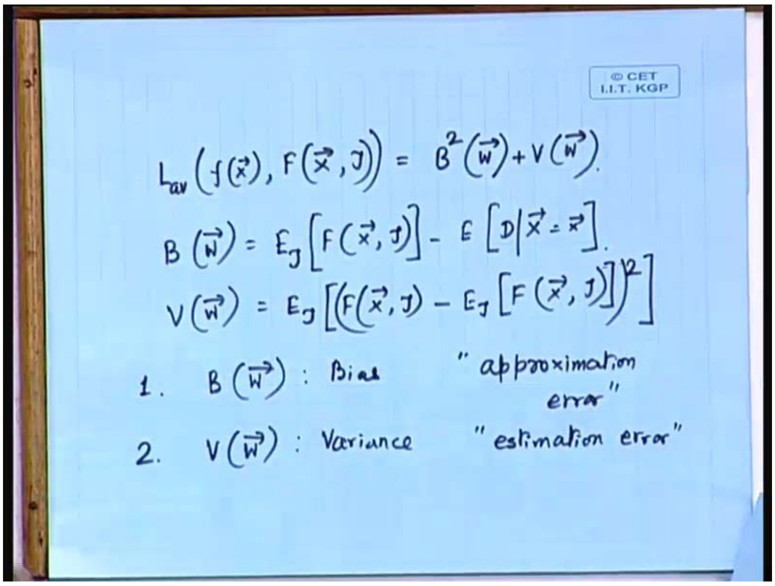
So, naturally it will lead to again three terms, one will be the a square term, that is to say this minus this whole square. This minus this effectively is what, this minus this that is expectation of D given x is equal to x; that means, to say the regression function realization minus E t of F x. That is, the averaged value of the realized function and the actual function, the difference between the two is nothing but, the bias that is the bias term.

So, if you are just putting it into this expression, you get a quantity which is nothing but, the square of the bias term. So, you get a square of the bias term and also you get the square of this quantity and this quantity means what, this quantity is yes correct this quantity is variance, the second quantity why, because we are realizing a function f of x t. Now, for different patterns in the training set our f function realizations are becoming different, but there is also an averaged F function realization which is this.

So, this is the averaged F function realization minus the f function realized out of if specific training pattern. And if we just square it up, then that leads to the variance term. So, when I am expressing this quantity, the quantity written within the bracket of this expression as a summation of two quantities, then the first term of this is the bias term and the second term of this happens to be the variance term.

So, our bias square and variance term, if you are putting in this way. And also there will be a term, which will be a multiplication of that bias and the variance and since they are independent of each other. So, that will again since we are taking the expectation operator finally, that will evaluate to 0, so in effect we are going to have the L expression the L average expression as follows.

(Refer Slide Time: 16:05)



© CET
I.I.T. KGP

$$L_{av}\left(f(\vec{x}), F(\vec{x}, \mathcal{T})\right) = B^2(\vec{w}) + V(\vec{w}).$$

$$B(\vec{w}) = E_{\mathcal{T}}\left[F(\vec{x}, \mathcal{T})\right] - E\left[D \mid \vec{X} = \vec{x}\right].$$

$$V(\vec{w}) = E_{\mathcal{T}}\left[\left(F(\vec{x}, \mathcal{T}) - E_{\mathcal{T}}\left[F(\vec{x}, \mathcal{T})\right]\right)^2\right]$$

1. $B(\vec{w})$: Bias "approximation error"

2. $V(\vec{w})$: Variance "estimation error"

So, we can just write down that L average of f x vector comma capital F x vector t. So, the natural measure of effectiveness will be nothing but, the square of the bias term, which we can write simply as B square as an argument we can only put the weight w. Because, after all that is what the only variable is and plus the variance term V w.

So, in this case the definition of the bias and the variance is the bias term is nothing but, E t F x comma t minus E of D given X is equal to x. So, this is the bias and the variance definition is pretty simple, it is E t F x t minus E t F x t you get actually this thing squared. So, we have to square it out expectation of this squared quantity, so this is the variance.

Now, the interpretation that we can make out of these two quantities is that, this bias term that is B w, this is the bias this can be interpreted as approximation error is not it yes look at it. This bias term is indicating what, the difference between the actual function and the averaged realized function. So, this is nothing but, the approximation error, whereas the second quantity that is V w which is the variance in effect that is indicating to us the estimation error.

So, the two are little difference in nature, that is what you should understand. For approximation we are directly having the error term, between the actual function and the average to realized function. And for the variance that is to say for the estimation error, it is that to what extent this realized functions are varying for different realizations. So, that is what the two things are, so the yes.

Student: ((Refer Time: 19:17))

So, that is right E tau of this quantity right, so.

Student: ((Refer Time: 19:41))

B we have defined without the expectation. In fact, B is the expectation term only.

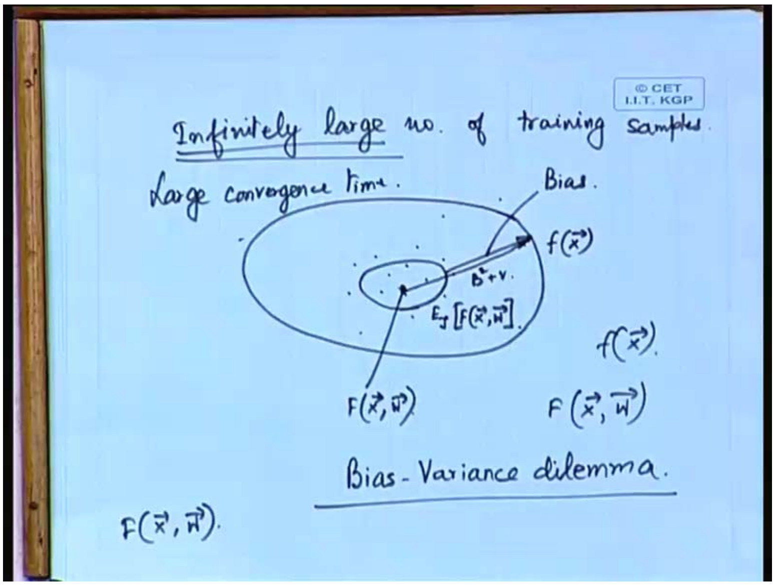
Student: ((Refer Time: 19:52))

Just a minute I will clear your doubts, you are getting a B term which is nothing but, the difference between these two, the expectation of this minus the error. So, was it our originally, our original definition was also this and in the L average, ultimately it is the expectation of these quantities.

Student: ((Refer Time: 20:22))

Effectively we just simplified this by avoiding that E tau expression outside the t expression outside. But, in effect we understand that there are two quantities effectively that is something that we have to bother with, that is one is the bias and the second is variance.

(Refer Slide Time: 21:00)



© CET I.I.T. KGP

Infinitely large no. of training samples.

Large convergence time.

Bias - Variance dilemma.

$F(\vec{x},\vec{w})$.

In fact, it can be thought of like this that, supposing this is the just pictorially I have in a two dimensional space I have a function realization, supposing this is the f x which is the actual regression function. And supposing we have the different realizations of this F of x w, now say that we have got F x w different realizations out of here and let us say that this is the averaged f x w realization. So, this one will be our nothing but, this is E t of F x w.

So, if you are taking the difference between these two, that is difference between this f x and this E t that will indicate the bias alright. Whereas, if you are taking a particular let us say that this is a particular F x w that you are taking. Then, you are saying that here this f x minus this F x w the difference between these two actually will be nothing but, the L average estimate that is what we are talking of.

So, this one if I join these then this will be our bias square plus variance this term to the power half if we are calculating the distance of this. Now, what happens is that we have two in order that our F x w is a good estimate of f x, what we have to do is that, this F x w should be brought very close to f x. That means, to say that our bias will be less and also the variance should be less.

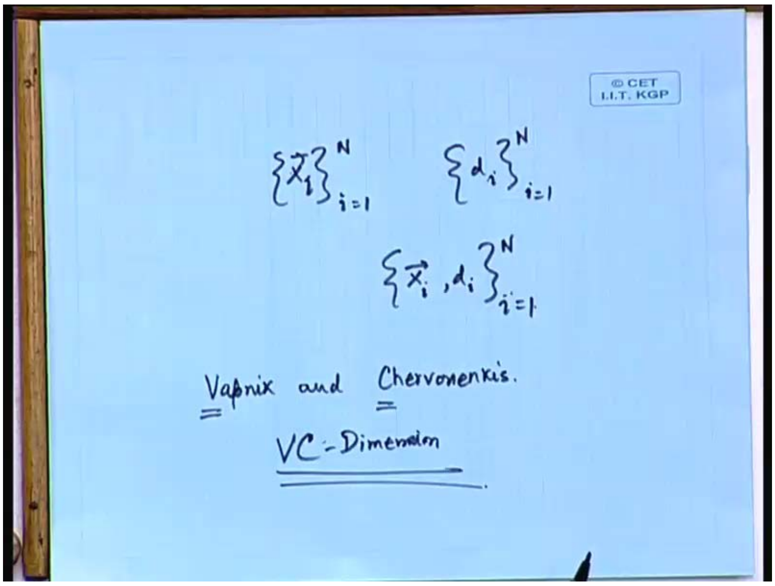
Now, it is very difficult to achieve these two things together, just simply see that supposing we have got just few training samples with us. And supposing we have got one realization here, one realization here, one here, one here, so may be that the averaged E t that we get E t of this F x w, that we get may be very close to f x no doubt. But, with such few points our variance will be excessively large, on the other hand if we try to make the variance small, then the bias also may be large.

So, this aspect is called as a bias variance dilemma. So, bias variance dilemma means that you cannot simultaneously control both bias and variance, but your objective is to ultimately reduce both. Because, in the L averaged expression you are getting the bias term as well as the variance term. So, you have to reduce both, but as the kind of argument that I was trying to give you, that with few samples with few training samples you can afford to have a small bias, but large variance or you can afford to have large bias and small variance.

But, ultimately to take care of both bias and variance what we have to do is to have infinitely large number of training samples. If we have infinitely large number of training samples only, then it should be possible for us to reduce both bias and variance. Now, if we again introduce infinitely large number of training samples that would in inevitably mean that our convergence time will be very large.

So, if you want to avoid bias variance dilemma, then you have to accept a large convergence time with the network. And not only that, say one aspect which we are every time talking about, that is to say we are simply expressing that you have got a training set.

(Refer Slide Time: 26:23)



$\{\vec{X}_i\}_{i=1}^{N}$ $\{d_i\}_{i=1}^{N}$

$\{\vec{X}_i, d_i\}_{i=1}^{N}$

Vapnix and Chervonenkis.

VC-Dimension

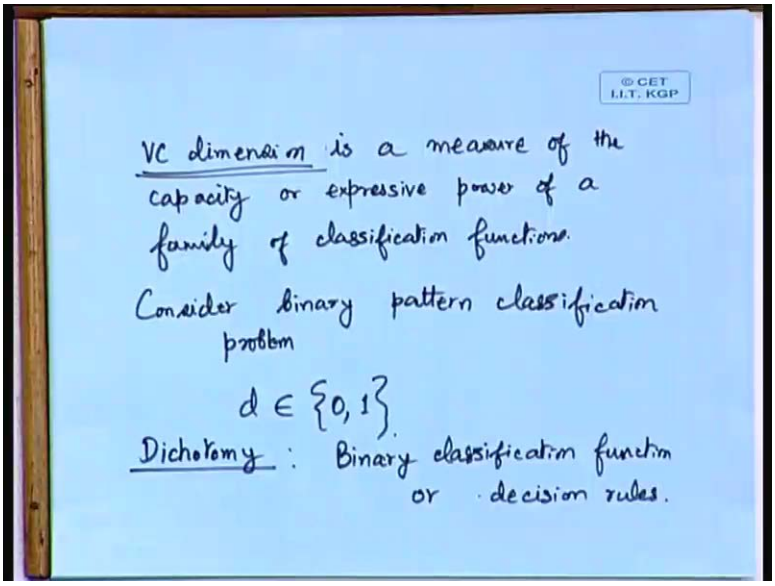
That is to say you have got a set of X i, where i is equal to 1 to N and you have got a set of X i vectors here. And you have got a set of d i i is equal to 1 to N and this X i comma d i this is your training patterns and you have got such patterns from i is equal 1 to N. Now, the question is that given a network which we are actually asking the network to learn, when a network is made to learn with these many training patterns.

What is the guarantee that this learning that it is doing is not an over learning or an under learning is there any guarantee that this much of training patterns, this many training patterns n training patterns is going to be really optimum. So, that is something that we have to really study, that what is the optimum number of patterns, so that there is no over feeding or under feeding of the data.

By again feeding of the data means, again after all what you are giving is that you are giving a set of training patterns and you are going to feed a function on that alright. So, that function feeding what you are doing should neither be under feeded nor over feeded. So, that is something what we have to analysis and to look into this before we look into this analysis there is one statistical tool that we must consider.

And that is actually named after two scientists who had proposed this theory, that is Vapnik and Chervonenkis. This two people, but I do not think that we are required to know the names of both the scientists, we can simply tell it as VC. And based on their theory we come to a measure which is called as the VC dimension and in fact, it is the VC dimension, which is a measure of the capacity or the expressive power of a family of classification functions.

(Refer Slide Time: 29:12)



VC dimension is a measure of the capacity or expressive power of a family of classification functions.

Consider binary pattern classification problem

$d \in \{0, 1\}$.

Dichotomy : Binary classification function or decision rules.

So, I can say that VC dimension is a measure of the capacity or expressive power of a family of classification functions. And in fact, based on their theory if we can compute the VC dimension of any given decision function, because every neural network as we know that ultimately is giving us a decision function. That is to say in a very simplest sense, that if it is a binary pattern classification problem alright.

In that case whatever patterns that we are feeding as an input to that network, will be ultimately classified as the class one or class zero. So, effectively it is realizing that by making use of some classification function, there is always a decision boundary or a decision function or decision boundary. That will in fact, decide that whether this particular pattern will belong to the class zero or belong to the class one.

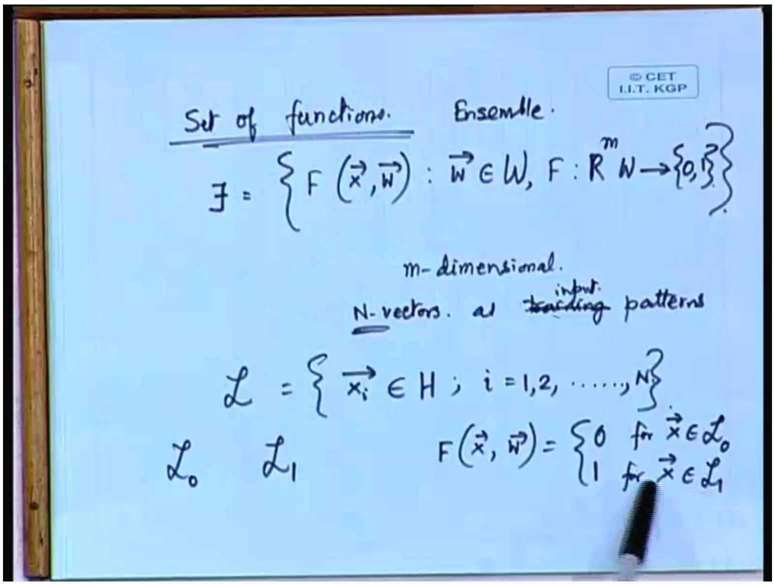
And how powerful this classification functions are up to force a problem in a different way. That given some classification functions, what is the number of training patterns that can be feed to the system. So, that the classification is possible out of that system, without making any error, that is what we are looking for and that is why the VC dimension is that is why the VC dimension has become a very powerful tool for that.

Now, let us consider a binary classification problem, binary pattern classification problem. So, we consider binary pattern classification problem and there we take that the ultimate decision it is a binary decision. So, d belongs to a set of 0 and 1, d can be either 0 or it can be 1 it is a binary pattern classification problem.

Now, how are we classify into the binary patterns, that will be based on some decision rules. Now, this rules we will be referring to as Dichotomy, so by Dichotomy we mean binary classification function or decision rules. So, in a network we could be having various Dichotomies, there could be different decision rules existing in a system. So, a particular network may be or a learning machine rather could be having a number of such Dichotomies. So, let us form an ensemble of such Dichotomy.

So, we form a set of Dichotomies let us say this is the set f that we are calling. And this is nothing but, a set of functions ultimately it is a classification function, function rules whatever way you may call it. So, ultimately the set of rules or the set of functions that we are having in the learning machine is a set of x w, where this w vector that belongs to a space w.

(Refer Slide Time: 34:17)



© CET I.I.T. KGP

Set of functions. Ensemble.

$$\mathcal{F} = \left\{ F(\vec{x}, \vec{w}) : \vec{w} \in W,\ F : R^{m} W \rightarrow \{0,1\} \right\}$$

m-dimensional.

N-vectors. as ~~training~~ input. patterns

$$\mathcal{L} = \{ \vec{x_i} \in H ;\ i = 1, 2, \ldots\ldots, N \}$$

$\mathcal{L}_0$ $\mathcal{L}_1$

$$F(\vec{x}, \vec{w}) = \begin{cases} 0 & \text{for } \vec{x} \in \mathcal{L}_0 \\ 1 & \text{for } \vec{x} \in \mathcal{L}_1 \end{cases}$$

So, this w belongs to a space W which could be actually an m dimensional space. So, this function F is belonging to in the m dimensional space W and this ultimately maps it into 0 or 1. So, this is the set of classification functions that is possible, the different classifications I things will become more clear when we discuss this with example. Right now you do not really feel afraid that it is becoming too much of abstract, it is not something that is exactly abstract.

I thing we will be able to come back to this and understand better with respect to some small examples very random version of examples that we will take. So, this is what this is nothing but, a set of functions different set of functions are there different decision functions and just an ensemble of that. So, this is an ensemble whatever functions this learning network can have I put them all together in the form of an ensemble.

That is, what I have done nothing great thing about it just a representation of that and now let us choose the training patterns. Now, again the training pattern inputs they will belong to the same m dimensional space given that we have got machines that as got m number of input all our inputs will be basically m dimensional vectors. And as a training pattern we take N such vectors as training patterns and we just form a set of such training patterns very simple.

So, we take a set L which is defined as the set of all the vectors X i alright and this X i actually belong to a space H and here i is equal to. So, this H is nothing but, an m dimensional space, so this X i belongs to an n m dimensional space and in this case how many training patterns since we are taking N. So, i is equal to 1, 2 etcetera, etcetera up to N, now all the Dichotomies are stored in this set.

And all the training patterns or all the input patterns let us say input, because when I say training that involves the input as well as the target decisions. I am saying that set of input patterns lying here, the set of Dichotomies will lie here and what are we supposed to do, we are supposed to apply the Dichotomies over these set of inputs. And ultimately there will be a, so ultimately Dichotomy will be a function that will separate this set L into two disjointed subsets L 0 and L 1.

So, ultimately we will be getting two disjointed subsets L 0 and L 1 such that F x w will be equal to 0 for x vector belonging to L 0 space. And it will be equal to 1 for x vector belonging to L 1 space is it followed.

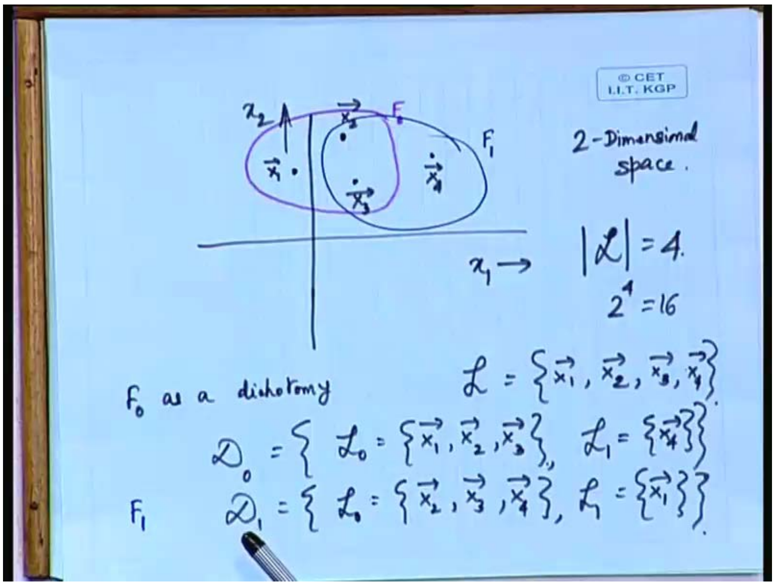
Student: ((Refer Time: 38:44))

This one, see we have got different set of rules on the system different set of decision different decision rules are there in the system. So, I have put an ensemble of all these decision rules, the decision rules are given by functions F of x w, so the set of all such functions that forms the ensemble I have put it in this set F. So, far I am only restricted to the definition of that alright and then, we have define another set which is the set of inputs.

Now, what I said that from this set of Dichotomies, let say we take any one Dichotomy out of this and that Dichotomy we apply on this set of L. So, there will be some of the patterns now this x will be, so for some of the patterns, for some of the input patterns the output will be 0 and for some other input patterns the output will be equal to 1. So, for those the patterns for which the output is becoming 0 we are calling that those patterns belong to the set L 0.

And the patterns for which the output becomes equal to 1 we are putting those patterns in the set L 1. So, this L set we are breaking up into two disjointed sets of L 0 an L 1, now it is one particular Dichotomy which has just divided into L 0 and L 1. Now, I take another Dichotomy, then some other patterns will belong to L 0, some other patterns will belong to L 1. Let us take a very simple example, I think we should go over to small example then that will make it very clear.

(Refer Slide Time: 41:10)



2-Dimensional space.

$|\mathcal{L}| = 4.$

$2^4 = 16$

$\mathcal{L} = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}.$

F_0 as a dichotomy

$\mathcal{D}_0 = \{ \mathcal{L}_0 = \{\vec{x}_1, \vec{x}_2, \vec{x}_3\}, \mathcal{L}_1 = \{\vec{x}_4\}\}$

F_1 $\mathcal{D}_1 = \{ \mathcal{L}_0 = \{\vec{x}_2, \vec{x}_3, \vec{x}_4\}, \mathcal{L}_1 = \{\vec{x}_1\}\}.$

Let us consider a two dimensional space very simple, let us not work on the m dimensional space perhaps that makes our feeling quite abstract, just take simple two dimensional space. So, that the inputs are only x 1 and x 2 mind you here purposely I did not put any vector notation, because this x 1 is in one of the direction, x 2 is a other direction, so there are two inputs x 1 and x 2.

Now, in this space I just happen to pick up four different training patterns, four different inputs and they will be vectors. So, this is let us say the position of the x 1 vector, say this is the position of the x 2 vector, say this is the position of x 3 vector and say this is the position of x 4 vector. So, what I did I only picked up four such inputs, so I have got an L set and that L set contains what, x 1 vector, x 2 vector, x 3 vector, x 4 vector.

Dichotomies we do not know yet, let us try to formulate some Dichotomies. Let us take one of the Dichotomies which puts up a decision boundary of this nature. Let say that I have defined a Dichotomy F naught that classifies the patterns as follows. So, I say that F 0 is a Dichotomy, because here lies the decision boundary means what, that whatever is inside this decision boundary they will be belonging to the two one class of patterns.

Let us say that they belong to the pattern 0 and whatever lies outside that belong to the pattern class one alright. So, I take F 0 as a Dichotomy and how is the Dichotomy defined, that Dichotomy induces two disjointed sets, because from L now I will be making two disjointed sets L 0 and L 1. So, Dichotomy F 0 will do like this, that it will make L 0 equal to x 1, x 2, x 3 alright and it will take L 1 as another set, which will be having only x 4.

So, now you get at least some idea about what Dichotomy is, this is one Dichotomy that simply specifies a decision boundary, that if it is within this it belongs to the set L 0, if it is belong, if it is outside this then it belongs to L 1. So, this way of pattern classification is done, and let us now think of a different Dichotomy, let us say that instead of this I pick up this one as a Dichotomy.

Say, now this Dichotomy I call as F 1 and F 1 is now again inducing two different sets D 1, this Dichotomy can be expressed as two sets, one is L 0 in L 0 sets we will be having now x 2, x 3, x 4, in L 1 set we will be having x 1. Now, I want to ask you a question, how many such Dichotomies are possible I have got four patterns four input pattern, can you tell me that how many such Dichotomies are possible D 0 is, F 0 is one of the Dichotomies, F 1 is another Dichotomy like this I can think of different Dichotomies.

Student: ((Refer Time: 46:01))

((Refer Time: 46:08))

Student: ((Refer Time: 46:09))

((Refer Time: 46:10))

Student: ((Refer Time: 46:12))

2 raised to the power 4 is equal to 16, that is the number of Dichotomies that we can have, even if you have it like his way. That is 4 c 0 plus 4 c 1 plus 4 c 2 that ultimately adds up to that only. So, this is a problem from combination and we are going to have

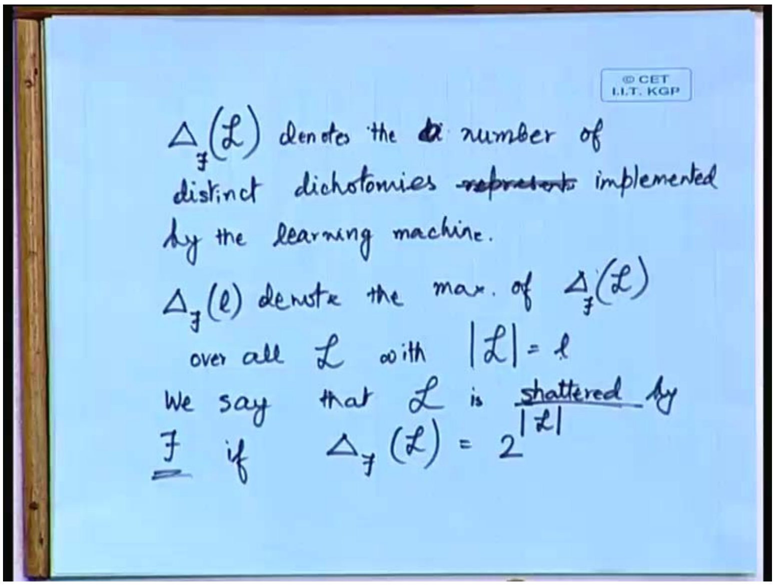
totally 16 Dichotomies, out of this why 16, because here we have taken a 4 input or rather the cardinality of this set L, that we consider that is equal to 4.

So, that is why for us the total number of Dichotomies that is possible is 2 to the power 4 which is equal to 16. Now, VC dimensions definition says, that it is the maximum number n listen carefully, it is the maximum number of n which makes 2 to the power n number of Dichotomies. That can again I am introducing one term, which requires a definition that actually shatters this L space shattering, a term that I need to define I with what to say is that, this here now that we are getting the maximum number of the cardinality of this set is 4.

And it can realize with 16 functions we can fully classify everything 16 is the ultimate number of functions, that is possible to segregate this space or to what we can say as to shatter this space with different type of patterns. So, since 4 is that number that is doing it 4 is the maximum 4 means 2 to 4 is equal to that, so in this case this particular case the VC dimension will be equal to 4.

But, here the decision boundaries are taken this way, but if the decision boundaries are taken in a different manner, then the VC dimension will also differ. But, we will come to that later on, but let us now formalize the definition.

(Refer Slide Time: 49:07)



© CET
I.I.T. KGP

$\Delta_{\mathcal{F}}(\mathcal{L})$ denotes the ~~di~~ number of distinct dichotomies ~~represents~~ implemented by the learning machine.

$\Delta_{\mathcal{F}}(\ell)$ denote the max. of $\Delta_{\mathcal{F}}(\mathcal{L})$ over all $\mathcal{L}$ with $|\mathcal{L}| = \ell$

We say that $\mathcal{L}$ is <u>shattered</u> by $\underline{\mathcal{F}}$ if $\Delta_{\mathcal{F}}(\mathcal{L}) = 2^{|\mathcal{L}|}$

Now, let us say that delta of L with suffix we are putting as f, now again lets go back to our definition. So, I think that now this definition will not seem to be very abstract, because this is now a set of all Dichotomies and this is the set of input patterns. So, this is very clear, now we are introducing a term called delta L suffix f. And this will denotes the number of distinct Dichotomies represented Dichotomies we can say implemented by the learning machine.

And let us say that delta f over small l denote the maximum of delta L suffix f over all L with the cardinality of L equal to l. So, I think that I need not have to explain this term, this you understand that this is the cardinality, so simply speaking what is that, that if l small l happens to be the cardinality of the set L, then delta l of f that represents the maximum number of delta f of l.

So, that it is there and now coming to the definition of shattering, we say that L this space L is shattered by the decision rules are there in f, if this delta f of L is equal to 2 to the power cardinality of L. So, if the number of distinct Dichotomies existing in the learning system is equal 2 to the power L, meaning what in this example if in the learning system, we have all the 16 decision rules available with us.

Then, we can say that the input space L, L is the input space that is shattered by the set of decision rules or the set of Dichotomies f any doubts, required some time to assimilate perhaps more examples.

So, we will surely consider some more examples in the next class before the concepts really goes into our mind. But, let us only think that here the number of there are a number of rules that you can put into the system. And if your set of rules actually spends all the different possibilities, if the distinct Dichotomies that you are having is equal to 2 to the power L, where L is the cardinality, then that space is shattered. And we will give several shattering examples and that will be more clear in the coming class.

Thank you very much.