

Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 06

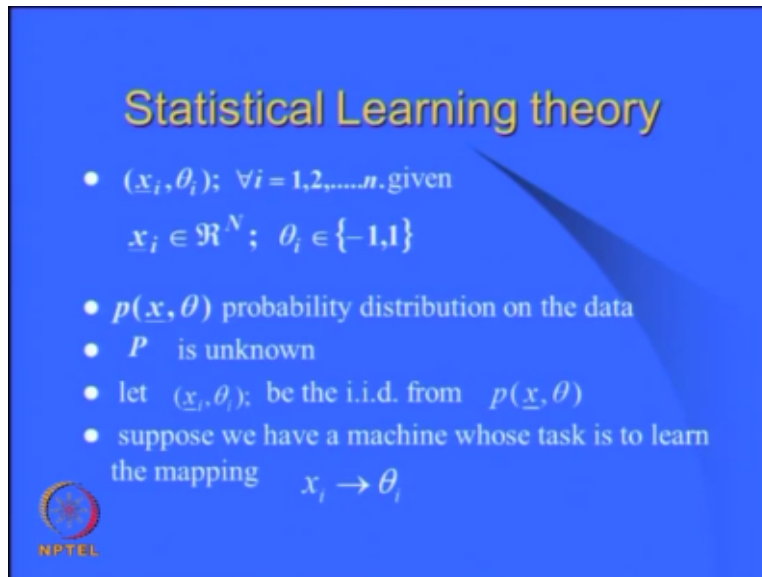
Lecture 06

Support Vector Machine [SVM]

Prof. C.A. Murthy
Machine Intelligence Unit.
Indian Statistical Institute. Kolkata

This is basically a lecture on support vector machines though you would see on the screen the title statistical learning theory.

(Refer Slide Time: 00:24)



Statistical Learning theory

- $(\underline{x}_i, \theta_i); \forall i = 1, 2, \dots, n$. given
 $\underline{x}_i \in \mathfrak{R}^N; \theta_i \in \{-1, 1\}$
- $p(\underline{x}, \theta)$ probability distribution on the data
- P is unknown
- let $(\underline{x}_i, \theta_i);$ be the i.i.d. from $p(\underline{x}, \theta)$
- suppose we have a machine whose task is to learn the mapping $\underline{x}_i \rightarrow \theta_i$

NPTEL

Let me tell you a bit of history there were two statisticians named ethnic and turban Incas they are they are the main persons who created this subject statistical learning theory there are there are statisticians from Russia you know there was a cold war period between America and Russia and these people have done their work basically during the Cold War period and after the Cold

War was over after then when there were communications between Russia and America they became normal.

These people they went to United States and they presented the statistical learning theory in a conference in computer science in a computer science conference they presented this one the basic problem that they attempted to solve here is when you design a classifier you have a training set using the training set you design the classifier then using the test set somehow you measure its performance and then if the performance on the test set is also satisfactory then you say that fine everything is fine with the classifier.

But is it really fine with the classifier even if it does well on the test set how do you say that your classifier is generalizable the performance of the classifier that you have somehow got how do you say that it has generalization capability is there any mathematical way of expressing edge and if you express it mathematically is there any way of obtaining it, and if you also obtain it then for the different classifiers that we are using is there any way in which you can calculate the generalize ability of this classifiers.

So this is the basic question that they attempted to solve there since they are statisticians and they attempted to solve the whole thing using statistical language, so they coined the term statistical learning theory support vector machines which probably you have heard from many people it is sort of a byproduct of statistical learning theory sort of a byproduct of statistical learning theory this is the basic history and you will find a book by Vapnik on statistical learning theory which is basically a book on statistics.

And you will find support vector machines being considered a part of neural networks you will find them to be considered a part of machine learning and data mining and of course since we are talking about classifiers and their performance you will consider them to be a part of pattern recognition, so you will find support vector machines almost in all these fields you will find support vector machines in all these fields and the generalization of support vector machines like kernel machines etc.

This is sort of the basic little bit of history now let me try to explain the basic terminology you look at your screens first one is you are given points in N dimensional space okay, you are given small end points x_1 x_2 x_n you see the very first step you are given small endpoints x_1 x_2 x_n

they are in N dimensional space θ_i denotes the label of the class label of the point x_i I am assuming that you have two classes only and the class labels are given as -1 and 1 and $p_{x|\theta}$ this is the probability distribution on the data.


That means there is some probability distribution this is the p is the density function and P that is the actual P of a is equal to $\int_a P$ over a of p , p is the density function and P is the actual probability, now these points are x_1, x_2, \dots, x_n and the corresponding θ_i they are assuming they are assumed to come from the distribution $p_{x|\theta}$ where $p_{x|\theta}$ is not known and here it is written they are I independent identically distributed, now in classification what exactly is the problem the problem is you are given n points n points.

And you have a corresponding class labels somehow you need to find the function from x_i to θ_i are you understanding it somehow you need to find the function if you find the functional form which for every x_i , if it gives you if you find you find the functional form where for every x_i the function gives the value θ_i then you are done you need to find the corresponding functional form that is the basic problem of classification, these F you can call them as you can have many names for it.

Okay and some FS when for one class it will give you plus one another class they are going to give you -1 and at some place they will get the value 0 , if then the value is 0 then you call it as the separation between the class 1 and class -1 right, when the value is given as 0 then you call it as separation between the class plus 1 and between the class $+1$ and -1 .

(Refer Slide Time: 07:31)

- let $\mathfrak{F} = \{f(x, \underline{\alpha}) : \alpha' s \text{ are adjustable parameters, } f \text{ is a function}\}$ for a given input x and choice of $\underline{\alpha}$, $f(x, \underline{\alpha})$ will always give the same output. A particular choice of $\underline{\alpha}$, generates a "trained machine". A neural network with a fixed architecture, with $\underline{\alpha}$, corresponding to the weights and biases is a learning machine.
- Risk $R_f(\alpha) = \int \frac{1}{2} |\theta - f(x, \alpha)|^2 dp(x, \theta)$
- Empirical risk $= \frac{1}{2n} \sum_{i=1}^n |\theta_i - f(x_i, \alpha)|^2 = R_{emp, f}(\alpha)$
- choose η such that $0 \leq \eta \leq 1$



Now what is it that we are given we are given a set of functions script F that is the functions are f xs are the input α is the parameter set α bar it is written it is the vector form, so all these α are adjustable parameters and f is a function actually let me just try to explain it to you I hope all of you know what multi-layer perceptron is you have an input layer and you have some hidden layers you have an output layer when you are going from input layer to hidden layer you have several connections and you have connection weights.

You start with some initial connection weights from input layer to hidden layer one hidden layer one if you assume to hidden layers and hidden layer one two hidden layer two then hidden layer two to output layer at between every two such layers you have many connections and you have connection weights, put all the connection weights together and write it as a vector form that vector you take it as α that vector you take it as α okay and the xs are your input okay given an x and given an x given an x.

And given an α and you have the usual neural network which is feed-forward neural network it will give you an output the output is this F okay, given the set up in given the given the set of input $x_1 x_2 x_n$ input vectors given α digestible parameters then, if you apply your neural network methodology the output is f okay these are the adjustable parameters and f is a function, now if you changes the α if you changes α the corresponding f is going to I mean the values are going to be changed if you change your network architecture then f is itself is going to be changed.

So for a given input x and choice of α f of x α will always give the same output if you have a function f and if you have specified your α f of x α will always do the same output a particular choice of α generates a trained machine by this particular choice when you are training a neural network you start with some choice and you go on changing, it till by the end of your I mean you have given some rule for termination and when it terminates you assume that you have got in nice I mean values for α .

So you are training the machine to get nice value for this α in neural network with a fixed architecture course with α corresponding to the weights and biases is a learning machine now when you have a function f when you fix α what is the exact risk that you are taking the risk is it is observed is f of x α expected is the actual one is θ take the difference take the difference modulus and $\int x \theta$ do the integration over all these x is that will give you sort of errors or risk what is it that we are calculating.

What we are calculating is θ for the i^{th} point for an α θ I is the targeted output f of x_i α is your observed output the difference I is equal to $\frac{1}{n} \sum_{i=1}^n$ and $\frac{1}{2n}$ this is empirical risk this is what actually we are calculating what we are supposed to calculate is this here I needs to tell you one thing I need to tell you one thing all these things are explained here using the sign modulus and the similar results actually you will get when you take the square terms when you take the square terms.

Which in neural network when you try to minimize the error you take the square terms and then you take the double \sum you use some gradient descent and then you do that, I to do the minimization okay there you take the square term for the error okay any of the results are similar okay, so here everything is explained using the sign modulus this is one thing that I am telling you there is another one you will get them all this material from a famous lecture notes or I do not want to call it lecture notes it is tutorial it is written by Christopher budgets it is available on internet okay a tutorial on support vector machines whatever I am going to tell you about the support vector machines.

Most of it you will find you will get from that particular tutorial okay, now so this is the risk that we are calculating and the actual risk that we are supposed to calculate is $R_f \alpha$ now choose η , so that $0 < \eta \leq 1$ here I am sorry this less than or equal to sign should not be there it should be strictly less than 1 okay in fact I would prefer that the other


less other equality also should not be there 0 strictly less than E than strictly less than 1 this equality sign should not be there okay.

(Refer Slide Time: 14:39)

Then

$$R_f(\alpha) \leq R_{emp, f}(\alpha) + \sqrt{\frac{h(\log(2n/h) + 1) - \log(\eta/4)}{n}} \dots\dots(1)$$

with probability $(1 - \eta)$



Then what was proved by Vapnik was the actual risk that we are supposed to calculate it is less than or equal to the empirical risk plus square root there is an H here $\log 2n$ n is the number of points divided by H + 1 - $\log E \eta$ by 4 by n with probability $1 - \theta$ that means this relationship holds with probability $1 - \theta$ this was what was shown by Vapnik this was I think in the year 1983, 84 R is it 93 94 I am not exactly sure 83, 84 or 93, 94 this thing was shown, now if you look at this expression note that in neural networks we are trying to minimize this empirical risk.

We are trying to minimize this empirical risk but what we are supposed to be doing is we are supposed to be minimizing the actual risk actual risk if we minimize it actual risk, if we minimize it then that is the thing that we want to do it but by minimizing the empirical risk are we actually able to minimize the actual risk the problem is that after minimizing the empirical risk still this much term is there still this much term is there and the actual risk is less than or equal to this Plus this relationship is holding with probability $1 - \theta$.

Now if we want this relationship to hold then probably we need to take the value of θ to be very, very small we take it to be 0.05 then $R_f \alpha$ less than or equal to empirical risk plus this that will happen with probability 0.95, so usually people take the value of theta to be either 0.05 or some 0.01 some very small value.

(Refer Slide Time: 17:15)

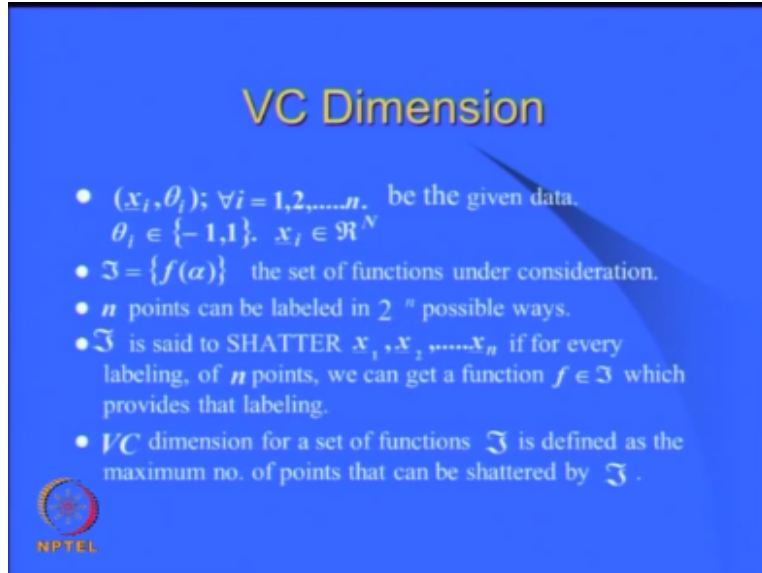
- h is a non negative integer called VC dimension (Vapnik Chervonenkis). h provides "capacity".
- η is small, say 0.05.
- Let us denote $\sqrt{\frac{h(\log(2n/h)+1) - \log(\eta/4)}{n}}$ by ξ .
- ξ is independent of the distribution P . ξ is called "VC-confidence".
- If we know h , we can compute ξ .
- Learning machine is another name for a family of functions $f(x, \alpha)$. We take that machine which minimizes the right hand side of (1). This gives the lowest upper bound on the actual risk.

NPTEL

Now there is an unknown term here that term is H n is the number of points η is this parameter 0.05 or mean point yeah 0.05 0.01 n is the number of points and there is this H what is this H , is a non-negative integer this is called VC - dimension Vapnik VC for Vapnik turbine in case okay this is a non-negative integer H is called VC - dimension it has to be always an integer it cannot take fractional values this H provides what is known as capacity we will come to what this H is slightly later η is a small value say 0.05, now let us denote this term this one square root of $H \log 2n$ by H this whole term let us denote it by inside okay this guy is independent of the distribution P , so in this one in this ξ η is a constant that we have already fixed n is the number of points so the worldly term is H this H is a non-negative integer in fact H is independent of distribution what is this VC- dimension we will define it slightly later this is something independent of this is independent of the distribution of the points.

So the whole ξ is independent of the distribution p_z is called we see conference, now if we know H we can compute inside now learning machine is another name for a family of functions if we take that the machine which minimizes the right-hand side of one this is the right we actually we minimize this is something independent of that independent of the independent of the distribution and we minimize this and by minimizing this, we hope that $R_f \alpha$ and empirical α they are somehow very close.

(Refer Slide Time: 19:58)

A blue slide with the title "VC Dimension" in yellow. It contains a list of five bullet points in white text. The first bullet point defines the data points (x_i, θ_i) for $i = 1, 2, \dots, n$, where $\theta_i \in \{-1, 1\}$ and $x_i \in \mathcal{R}^N$. The second bullet point defines $\mathcal{F} = \{f(\alpha)\}$ as the set of functions under consideration. The third bullet point states that n points can be labeled in 2^n possible ways. The fourth bullet point defines that \mathcal{F} shatters x_1, x_2, \dots, x_n if for every labeling, there is a function $f \in \mathcal{F}$ that provides that labeling. The fifth bullet point defines the VC dimension as the maximum number of points that can be shattered by \mathcal{F} . In the bottom left corner, there is a small circular logo with a red and blue design and the text "NPTEL" below it.


Now let us see what the VC dimension is VC dimension it is actually a nice quantity $x_i \theta_i$ are the given point they belong to \mathcal{R}^N and our number of point it is they take values $-1, 1$ and these are the family of functions under consideration, now if you have small n points in how many different ways you can label them to power in different ways do you agree to that if you have small n points you can label them in to power and different ways that is all the points you put it in class one that is one way one point you put it in class one the rest $n - 1$ point you put it in class -1 $n - 1$.

You put it in one point all the points you put it in class one that is one way $n - 1$ points you put it in a class one point you put it in class two then $n - 2$ points you put it in class one you put it in class two if you do liked it you have to power indifferent ways in which you label end points.

(Refer Slide Time: 21:19)

VC Dimension

- $(x_i, \theta_i); \forall i = 1, 2, \dots, n.$ be the given data.
 $\theta_i \in \{-1, 1\}, x_i \in \mathbb{R}^N$
- $\mathfrak{F} = \{f(x)\}$ the set of functions under consideration.
- n points can be labeled in 2^n possible ways.
- \mathfrak{F} is said to SHATTER x_1, x_2, \dots, x_n if for every labeling, of n points, we can get a function $f \in \mathfrak{F}$ which provides that labeling.
- VC dimension for a set of functions \mathfrak{F} is defined as the maximum no. of points that can be shattered by \mathfrak{F} .

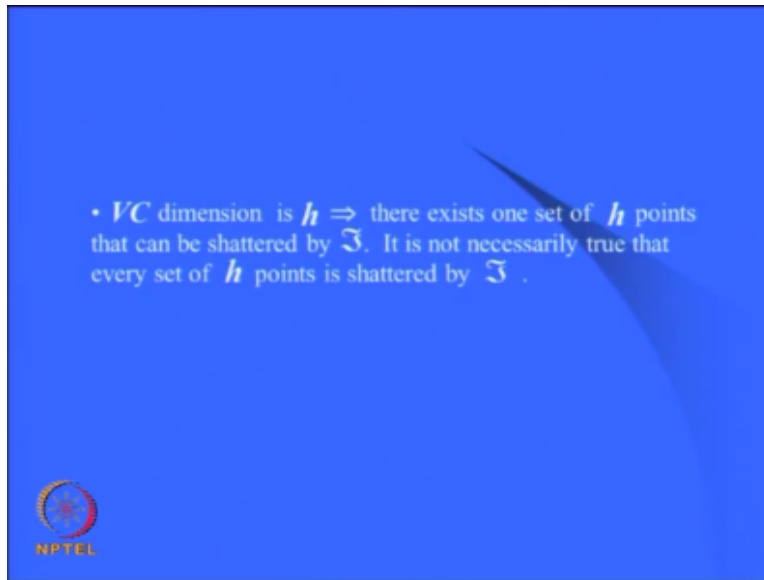


Now a set of functions τ is said to shatter end points a collection of end points if for every labeling of these end points we can get a function f which provides that labeling is this clear to you let me try to explain you have you have a set of functions you have a set of n points this set of n points can be labeled in 2^n ways for every labeling you need to get a function note that when I started this lecture I asked you what is our aim from the set of points you need to get a function to θ_i .

The corresponding labels once we get a function then we are through our aim is just to get that function okay, now you have got 2^n different labeling is possible okay for each labeling if you have a function with use that labeling then we say that this set of functions is said to shatter n points, I will explain it to you slightly more D in a more detail after a few minutes now we see dimension for a set of functions is defined as the maximum number of points that can be shattered by this VC dimension is the maximum number of points that can be shattered by suppose the maximum number of points is 10.

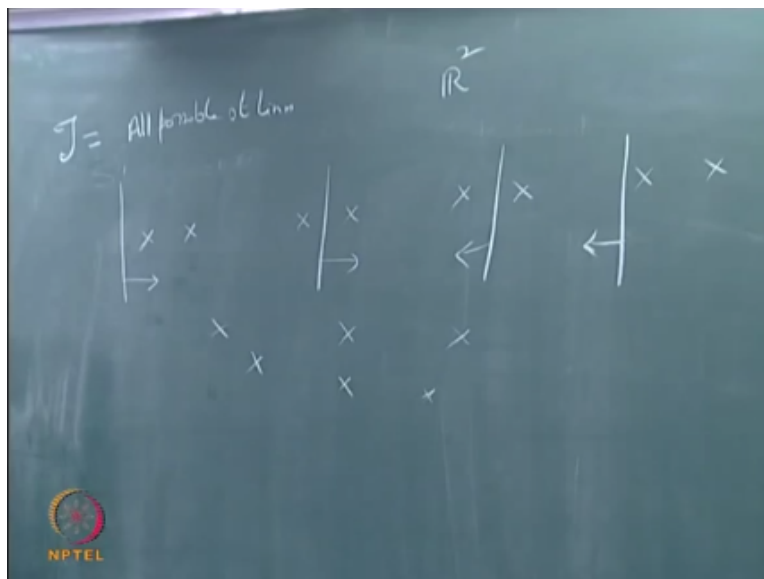
That means a set of 10 points if it is shattered by the collection of functions and no set of 11 points 11 points or twelve points or 13 points or 14 points no set of 11 points or 12 points or 13 points or 14 points can be shattered by the set of functions then the VC dimension then the VC dimension of the set of functions is that value 10 VC- dimension for a set of functions is defined as is defined as the maximum number of points that can be shattered by τ .

(Refer Slide Time: 24:03)



VC - dimension is H implies there exists one set of H points that can be shattered by τ but it does not mean that every set of H points can be shattered by it does not mean that every set of H points can be shattered by it I will explain all these things by using an example, now I can do the example on the board I will do the example on the board okay I can do it on the board suppose you take set of straight lines suppose you take set of straight lines and you take two points you are function now.

(Refer Slide Time: 24:51)

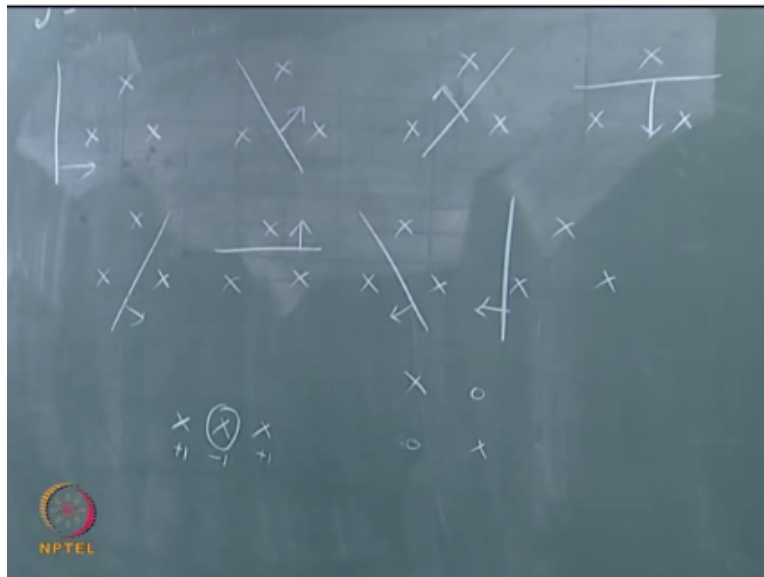


You are set now is all possible straight lines all possible straight line okay and let us say we are in two dimensional space let us say we are in two dimensional space, now you take two points how many labeling are possible you have four labelings okay, now you take one straight line here and this arrow denotes they are the given the sign positive sign that means all these both the points they are in the class plus one this is the straight line that is giving you this now the second one for the same two points this is another straight line.

That is putting this point in class 1 this point in class -1 for the same two points we have a straight line which puts this point in class 1 this point in class -1 and you have the fourth one both the points are in class-1 and this side is class +1 it is clear, so two points they can be shattered by the set of straight lines I have taken these two points like this I could have taken them like this I could have taken these two points in this way I could have taken these two points in this way I could have taken these two points in this way.

In any way I take okay if it is set of two points it can be shattered by set of straight lines am I correct every set of two points, now what I will do is that instead of two I will take 3.

(Refer Slide Time: 27:27)



I will take three this is 1 2 3 4 5 6 7 8 here first I will put all of them in class one all the 3 of them in class one then I will start putting two points these two points are in class one these two points are in class one and then these two points are in class one okay, now I will put one point in class one this is in class one this is in a class one this is in class one then I will put low point in class

one for this one no point in class one, so here I have taken a set of three points this is one set of three points this is shattered by all the lines he is shattered by the set of possible lines is this clear.

Now let us see whether every set of three points can be shattered the answer is not the answer is no you take three points on a single line, let us say this point goes to -1 these two points are $+1$ can you get a single straight line which gives you this result no okay, so now here you have a set of 3 points that can be shattered by straight lines, now you take any set of 4 points take any set of four points no set of 4 points can be shattered by straight lines no set of 4 points can be shattered by straight lines.

You have this famous example you remember this example now you are seeing the connection between this one and neural networks many persons when they introduce support vector machines they I mean they into banner introduce this one they first tell this example and then they go to all these shattering x' and other things in fact this can be proved mathematically that no set of 4 points can be shattered by straight lines, so in R^2 for the set of straight lines the VC dimension for the set of straight lines is 3.

Let me repeat in R^2 the VC dimension for the set of straight lines is three because there exist a set of three points that can be shattered by the set of straight lines a set of three points okay and no set of 4 points are anything more than 4 can be shattered by set of straight lines okay, so VC dimension for the set of straight lines is this is the one.

(Refer Slide Time: 31:53)

Oriented hyper-planes and shattering of points in \mathfrak{R}^N

- Let $N = 2$.
- \mathfrak{T} consists of all straight lines

NPTEL

So in two dimensions τ consists of all straight lines this is the example.

(Refer Slide Time: 31:59)

- VC dimension of straight lines is ≥ 3
- Note that VC dimension of straight lines is not 4. So VC dimension is 3.
- It can be proved that VC dimension of hyper-planes in \mathfrak{R}^N is $(N+1)$.

NPTEL


VC dimension of straight lines is greater than or equal to 3 they note that VC dimension of straight lines is not four because no set of 4 points can be shattered by this one, so VC dimension is three now it can be proved that VC dimension of hyper - planes in \mathfrak{R}^N is $n + 1$ VC dimension of hyper- planes in \mathfrak{R}^N is $n + 1$ that means, if you are looking at \mathfrak{R}^N and if you are looking at all possible hyper planes then you will be able to get a set of $n + 1$ points which can be shattered by these hyper- planes.

And no set of $n + 2$ $R^N + 3$ $R^N + 4$ points can be shattered by it why this shattering is important the shattering is important because note that in MLP we assume an architecture okay, we assume an architecture and then we make it learn, now the moment you have Schumann architecture you have assumed certain functional form right you have assumed certain functional form, now with that functional form by varying all those α if the given set of $x_1 x_2 x_n$ that, but if the given set of points if you are not able to forget about given set of points if you can if you are in a position to shatter.

At least a set of n points then probably we can think about getting the classification properly for that given set of n points let me repeat, if the with the function under consideration if you are able to shatter at least a set of n points say you are given smaller number of points then we can think of whether we can shatter the given set of n points the given set of n points it has two classes some labeling is there and you are assuming a functional form by Schumann architecture you are assuming a functional form.

By assuming an architecture you are assuming a functional form whether this functional form whether it at all can it shatter at least a set of n points if it is not able to shatter it whatever you do I mean it is not going to I mean if the VC dimension is less than that then you have a problem are you understanding me if the VC dimension is less than the value is smaller than you do have a problem.

(Refer Slide Time: 35:26)



- It is not necessarily true that learning machines with more parameters will have a high VC dimension, and learning machines with less parameters will have low VC dimension. (Examples exist in literature for the above statement.)
- A family of classifiers will have infinite VC dimension if they can shatter n points, however large n may be.
- Examples exist in literature where a set of functions has infinite VC dimension but not able to shatter finitely many points.

Now there are some comments it is not necessarily true that learning machines with more parameters will have a high VC dimension and leveling machines with less parameters will have low visualization examples exist in literature. The second one is that a family of classifiers will have infinite VC dimension if they can shatter a set of n points however large n may be okay, now examples exist in literature various set of functions has infinite VC dimension but they are not able to shatter a set of four points.

Here I wrote set of finitely points finitely many point the example that was given there was four points it has infinite VC dimension but on the other hand a finite point set having just four points it is not able to shatter, why the problem is that if you have a set of n points that can be shattered by the given function set then the VC dimension is at least equal to that value smaller, a set of n points.

We are not saying that every set of n points is to be shattered so VC dimension by the very definition it is a very weak one I hope you are understanding this it is very weak because you are satisfied if it shatters a set of n points one set of n points if it shatters you are satisfied, but then our given point set that also has n points but then one set it can shatter but this meet may not be able to shatter then you have a difficulty here right.


Then you have a difficulty here this is one of the problems that is actually I mean because VC dimension as per definition it is a very weak one yes, if VC dimension is ten means at least one set of ten points it can shatter so if it is something 11, 12 or 13 you know that you probably may not get the what is that you may not get the classification that is there. But VC dimension 10 means 89876 whether you can get the classification of this 876 point that is not clear.

Something more you know that you cannot get it but something less than that you do not have an idea that is the basic difficulty with VC dimension that is the place where theory needs to be developed it should be something more strong than that, that a set of points can be shattered.

(Refer Slide Time: 38:56)

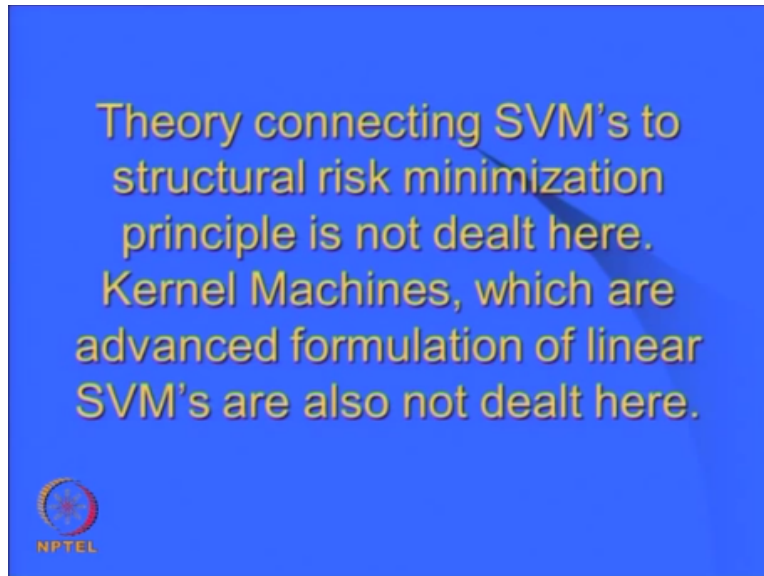
Minimization of Risk

- $R_f(\alpha) \leq R_{emp,f}(\alpha) + \xi$ with prob. $(1 - \eta)$
- ξ depends on the chosen class of functions.
- $R_f(\alpha), R_{emp,f}(\alpha)$ depend on f too.
- Divide the entire class of functions into nested subsets $\mathfrak{F}_1, \mathfrak{F}_1 \subseteq \mathfrak{F}_2 \subseteq \mathfrak{F}_3 \subseteq \dots$
- VC - dimension of \mathfrak{F}_i , is h_i
 $h_1 < h_2 < h_3 < \dots$



This I do not want to dwell into these things okay,

(Refer Slide Time: 39:03)



Theory connecting SVM's to structural risk minimization principle is not deal there so I do not want to deal with this thing.

(Refer Slide Time: 39:12)

Maximum Margin Classifier

$(x_i, \theta_i); i = 1, 2, \dots, n; \theta_i \in \{-1, 1\} \forall i$
 $x_i \in \mathbb{R}^N$.

Data is linearly separable


i.e. $\exists w \in \mathbb{R}^N$ Such that

$w \cdot x_i > 0 \quad \forall i$ for which $\theta_i = 1$
 $< 0 \quad \forall i$ for which $\theta_i = -1$

or $\theta_i w \cdot x_i > 0 \quad \forall i$

If there exists one such vector w for which $\theta_i w \cdot x_i > 0$ then there are infinitely many such vectors.

How does one choose one optimal classifier ?



These are extremely, extremely highly mathematical subjects these are highly mathematical subject I do not want to go into all that mathematics atleast now. Now maximum margin classifier so I do not want to go into the connections between the VC dimension theory and SVM's I am directly coming to SVM's, so you have excise $\theta_i = 1$ to n belongs to 1 to $+1$, $x_i \in \mathbb{R}^n$.

Now I am assuming that data is linearly separable that means there exists a hyper plane which gives you on one side of the hyper plane you will get the positive points $+1$ points and on the other side of the hyper plane you will get negative points that is -1 point. Now there is a basic theorem here if the given data set is linearly separable the given data set is linearly separable if and only if the convex hulls of those things do not intersect.

I hope you know this result okay, I will repeat it the data is linearly separable that means the $+1$ point set and the -1 point set they are linearly separable that there exists a hyper plane where on the positive side of the hyper plane you get all the $+1$ points on the negative side of hyper plane you got all the -1 points this is possible if and only if you take all the positive points can construct its convex hull take all the negative points it is convex hull and this convex hulls they do not intersect this is different only if that is if the convex hulls do not intersect then you get a hyper plane.

And if you get a higher plane then the convex has do not intersect both these things are satisfied. So data is linearly separable so there exists a hyper plane w , so $w \cdot x_i$ greater than 0 for all i for

which θ_i is equal to 1 less than 0 for all i for which θ_i is equal to -1 or you multiply by θ_i then θ_i times this is $w' x_i$ is greater than 0 for all i .

For when θ_i is equal to 1, 1 times $w' x_i$ that is greater than 0 when it is -1, -1 times this is also going to become greater than 0. So if there exists one such vector w for which this is greater than 0 this place this i and this i they should be replaced by prime transpose then there are infinitely many such vectors how does one choose one optimal classifier, I hope this is known to all of you if you have one hyper plane then you are going to have infinitely many hyper plane.

You are going to are infinitely many hyper planes if you look at the basic the hard limiting simple perceptron okay, simple perceptron then in the convergence theorem in the simple perceptron you assume the linear separability of the classes and you assume a hyper plane and you go on changing it till you go on changing it and then you can prove that as the number of iterations increases the error actually I mean it goes to actually 0 that can be shown mathematically that is called perceptron convergence theorem.

And so and you have too many hyper planes it will go to one of the hyper planes. Now here the question is how does one choose an optimal classifier, optimal from the point of view of what.

(Refer Slide Time: 43:44)

Problem Formulation

If $\exists \underline{w}$ such that $\theta_i \underline{w}' x_i > 0$ then for every $\delta > 0$
 $\delta \underline{w}$ also satisfies $\theta_i (\delta \underline{w})' x_i > 0 \forall i$

We shall set the "margin" (min. distance of the hyper-plane to the +ve points = min distance of hyper-plane to the -ve points = 1) to one and achieve it with minimal weight

i.e. $\min_{\underline{w} \in \mathbb{R}^N} \frac{1}{2} \underline{w}' \underline{w}$ where $\theta_i \underline{w}' x_i > 1 \quad \forall i = 1, 2, \dots, n$

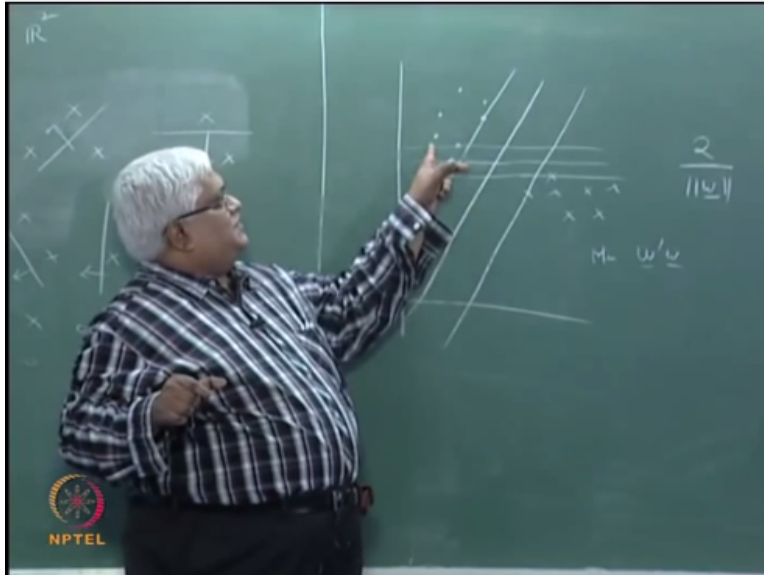
This is a QP problem



Now if there exists w such that $\theta_i w' x_i$ is greater than 0 for all i then you take any δ multiplied by any δ then δw also satisfies this condition okay, then what we can do is that what we can do is that we shall set the margin that is minimum distance of hyper plane to the positive point same as minimum distance hyper plane to the negative points that is positive points and negative points we shall set the margin as 1.

And achieve it with minimal weight now let us see the meaning of that let us see the meaning of that.

(Refer Slide Time: 44:43)



Now you please look at it here you have two classes in this class 1 2 3 4 5 6 points in this class 1 2 3 4 5 6 7 points, two classes. Now this is one hyper plane with respect to this plane this hyper plane you take the distance of this hyper plane with every one of the point find out the one that has the minimal distance the minimal distance is this.

Again with respect to this hyper plane find out the distance of the hyper plane with everyone of the points find out the one that has the minimal distance the minimum distance is this, okay. now you can choose this hyper plane in such a way that this is the shift is 1 and this shift is 1 okay, this shift is 1 and this shift is 1 so that this totally it becomes 2 and the distance is actually two by norm of the way to at norm of the vector it is 2 by a norm of the vector.

Now basically if you take this distance is some value but then you look at this hyper plane this hyper plane when you take the distance of this hyper plane with every one of these points the one that has the minimum distance is this and for this hyper plane again you do the same thing here the one that has the minimum distance is this now this is more than this, this distance is more than this.

So basically what we would like to do is that we would like to choose a hyper plane in such a way that for that hyper plane with the same shift that you get a negative point and the same shift then you should get this positive point so that then basically you are going the distance between these two is 2 by actually norm of F norm of W , where W is the W gives you the equation for this hyper plane that $W \cdot x$ okay W gives you the equation for this hyper plane.

Similarly in this case also we use the equation for this hyper plane okay, now you are what we would like to do is we would like to maximize this or another way of putting it is normal w is same as $w'w$ and you take the square root then you will get the norm so you like to maximize this because you want to take the distance to be same and distance to be the maximum if maximization of this is same as minimization of this you want to maximize the margin, maximize the distance between this changes the distance between this and this is taken as the margin and this classifier there is another name for it that name is maximum margin classifier.

The word margin is used as the distance between this hyper plane and this hyper plane the distance between this hyper plane and this hyper plane here if you take the distance between this and this hyper plane this will give you some margin here and this will give you another margin this margin it is maximum of all the possible margins that we can have so this is maximum margin classifier.

So a way of saying it is you get the margin you maximize it are you minimize this or you minimize this if you write $1/2$ here it does not matter, because $1/2$ is a constant you minimize this or you maximize this. So minimization of half of $w'w$ where $\theta' \theta_i$ into $w'x$ is greater than 1 for all i is equal to 1 to n . Now this is what is known as QP problem quadratic programming problem.

Many results in fact there is quite a bit of literature on convex optimization there is quite a bit of literature on convex optimization the functions under consideration here they are all they are mostly convex functions in fact I am W that is a convex function do you know the meaning of a convex function, a function is said to be convex a function is said to be convex a function is said to be convex, f is said to be convex if for every x and y these are vectors $f(\lambda x + (1-\lambda)y)$ is less than or equal to $\lambda f(x) + (1-\lambda)f(y)$ as an example you please look at this say this is your x this point is x say this point is y okay.

And this is your function now $\lambda x + (1-\lambda)y$ is a point here $(\lambda x + (1-\lambda)y)$ this is the point in between now this is $f(x)$ this is $f(y)$ right, $\lambda f(x) + (1-\lambda)f(y)$ this is if you vary λ over all 0 to 1 this is for all λ belonging to 0 to 1, if you vary λ in the interval 0 to 1 then you will basically get this line segment. Now you consider every value of the function in this interval that value is less than the corresponding value here so this is convex it is clear.

You take any value of the function here and this is less than this value $\lambda f(x) + 1 - \lambda f(y)$ so this is a convex function there is quite a bit of literature on convex optimization and this quadratic programming problem of how to get this w 's that is basically solved by using many results that are available in convex optimization many results that are available in convex optimization. As you can see the main problem here is a quadratic programming problem.


I hope all of you know what linear programming means, linear programming means you have constraints linear and the function that is to be optimized that is also linear then the problem is called linear programming problem. In quadratic programming problem the function to be optimized that is quadratic as you can see $w' w$ that is quadratic that is why actually it is called quadratic programming problem QP problem.

(Refer Slide Time: 54:26)

If the data is not linearly separable, make a "soft" formulation of the problem

$$\min_{w \in \mathbb{R}^N} \frac{1}{2} w' w \quad \text{such that} \quad \theta_i(w; x_i) \geq 1 - \gamma; i = 1, \dots, n$$

N.P. Complete



And so we have assumed that the data is not linearly separable now if it is not linearly separable then what people would do is $\theta_i w'x$ is greater than or equal to some $1-\gamma$ or you can take this γ to be dependent on i , you can take this γ to be dependent on i that also you can have it either you can have something fixed but usually if you look at the literature you have this thing γ as dependent on i $1-\gamma_i$.

So minimum of again $w'w$ subject to these constraints now this is going to be an extremely and in fact it is an extremely complicated problem to solve and this is when the classes are not linearly separable then we make what is known as a soft formulation of the problem.

(Refer Slide Time: 55:39)

Lagrange Multiplier

$$J = \frac{1}{2} w'w - \sum_{i=1}^n a_i [\theta_i w'x_i - 1]$$

(a_i 's are the Lagrangian multipliers)

$$\nabla_w J = w - \sum_{i=1}^n a_i \theta_i x_i$$

$$\nabla_w J = 0 \quad \text{gives} \quad w = \sum_{i=1}^n a_i \theta_i x_i$$

With a dual formulation of the QP problem, the minimization can be achieved.

NPTEL

Now though it is quadratic programming problem that is true but then when people try to solve this thing quadratic programming is coming slightly later okay, and you see the optimization function that is under consideration is $1/2 w'w - \sum a_i [\theta_i w'x_i - 1]$ where this a_i are the Lagrange multipliers use Lagrange multipliers the basic problem is quadratic programming problem because the constraints does the function to be maximized is quadratic that is $1/2 w'w$.

Now you when you want to solve it one of the ways of that you do it is by using Lagrange multipliers this is our Lagrange multipliers now you do differentiation partial differentiation so you get this as one so you take this thing to be equal to 0 that gives your W as this now with a dual formulation the minimization can be achieved that is true but it is quite intensive in programming minimization can be achieved which is true.

But then the programming part of that thing is not actually a simple one is not actually a simple one, so when this is generalized this is generalized to a case where the data sets the true class problem is not linearly separable which we had discussed in the previous slides then you take $1-\gamma$ or $1-\gamma_i$.

(Refer Slide Time: 57:33)

Lagrange Multiplier

$$J = \frac{1}{2} w' w - \sum_{i=1}^n a_i [\theta_i w' x_i - 1]$$

(a_i 's are the Lagrangian multipliers)

$$\nabla_w J = w - \sum_{i=1}^n a_i \theta_i x_i$$

$$\nabla_w J = 0 \quad \text{gives} \quad w = \sum_{i=1}^n a_i \theta_i x_i$$

With a dual formulation of the QP problem, the minimization can be achieved.

NPTEL

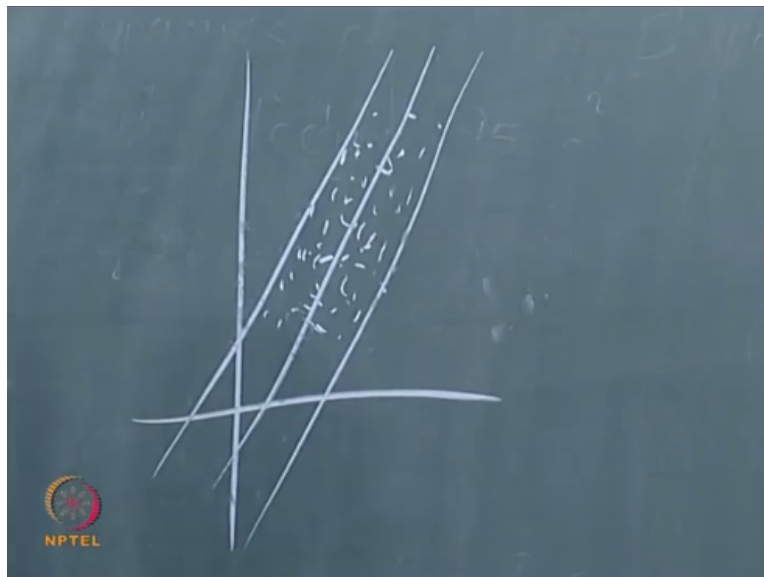
The second one is that suppose you have more than two classes suppose you have 3,4,5 classes and then the problem becomes more complicated there are two ways in which people have tried to solve it one is one against the rest it belongs to class 1 and not belonging to class 1, belongs to class 2 and not belonging to class 2, class 3 and not belonging to class 3 so one against the rest that is one way and the second way is you take every pair 1 2, 1 3, 1 4, 1 5 and for each pair you try to get the linear boundary are the soft boundary for each pair you try to get either the linear one or the soft one.

Now the problem formulation in all these cases it becomes the solution of the problem becomes extremely complicated and this has given rise to another class of problems which are known as

support vector regression problems. You see in this one this is the actual decision boundary this is the one that is giving you the maximum margin, now this decision and this one it is passing through this point which is a data point this one it is passing through this data point what is the support vector.

In fact these two points are actually known as support vectors because this is the actual decision boundary and then if you give -1 it is coming here if you $+1$ is coming here and the $+1$ line is passing through this one the -1 line is passing through this now these are known as support vectors. Now what is the usual regression problem, the usual regression problem is I will do it here.

(Refer Slide Time: 01:00:01)



The usual regression problem is you have a data set suppose you are looking at linear regression then you are you would like to approximate this data set by a line like this, now you have a line draw two parallel lines with the same distance in such a way that on this side there are no points and on this side there are no points all the points are lying in between these two lines. Now getting hold of this line amounts to getting hold of these three lines and which amounts to the problem of support vector machines.

Where in support vector machine that problem formulation the points are lying either on this side of the line or on this side of the line two classes here all the points are lying in between are you understanding me, it is basically the complementary one here all the points are lying in between.

So when you have all the points lying in between when you get this flight this is basically your regression line, which best approximates this.

So it gives it has given rise to what is known as support vector regression and regression has too many applications any forecasting problem is basically a regression problem or let me say most of the forecasting problems are regression problems for the past 20 days thus the price of the stock is so and so tomorrow, what is the price so what line approximates this what curve approximates this, this regression.

So and regression has too many applications even classification has too many applications regression has too many applications and quite many people are working on support vector regression there is one another common that I would like to say I have been talking about linear boundaries when linear boundary is non-existing then what I am saying is that I have put a margin there soft margin.

Now there is an extension of this one to non linear boundaries where people actually consider kernels people consider quadratic kernels are some other kernels for getting for obtaining boundaries nonlinear boundaries. You will find several topics are several subjects named say one subject is kernel machines you might have found a book titled kernel machines it is basically extension of SVM's to non-linear when you have non-linear boundaries then you basically use kernels to obtain them to at least try to obtain the non-linear boundaries.

So this is another extension from support vector machines one is linear but not exactly error zero and another one that is soft margin another one is non-linear where you consider kernels and another one is support vector regression. So lot of work is going on in all these fields and the work on these fuses nowadays is termed as machine learning so with this I stop the lecture.

**End of
Module 06 – Lecture 06**

Online Video Editing / Post Production

M. Karthikeyan
M. V. Ramachandran
P. Baskar

Camera
G. Ramesh
K. Athaullah

K. R. Mahendrababu
K. Vidhya
S. Pradeepa
D. Sabapathi
Soju Francis
S. Subash
Selvam
Sridharan

Studio Assistants

Linuselvan
Krishnakumar
A. Saravanan

Additional Post – Production

Kannan Krishnamurty & Team

Animations

Dvijavanthi

NPTEL Web & Faculty Assistance Team

Allen Jacob Dinesh
Ashok Kumar
Banu. P
Deepa Venkatraman
Dinesh Babu. K.M
Karthick. B
Karthikeyan. A
Lavanya. K
Manikandan. A
Manikandasivam. G
Nandakumar. L
Prasanna Kumar. G
Pradeep Valan. G
Rekha. C
Salomi. J
Santosh Kumar Singh. P
Saravanakumar. P
Saravanakumar. R
Satishkumar. G
Senthilmurugan. K
Shobana. S
Sivakumar. S
Soundhar Raja Pandian. R
Suman Dominic. J
Udayakumar. C

Vijaya. K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers
K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

Funded By
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in

Copyrights Reserved