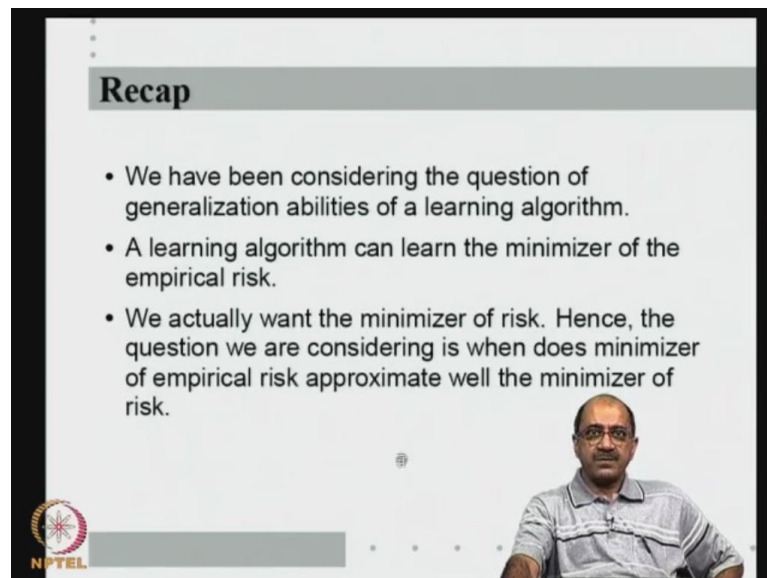


Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 25
VC-Dimension Examples; VC-Dimension of Hyperplanes

And welcome to this next lecture on pattern recognition. We have been looking at for the last few classes, on some basics of statistical learning theory. Essentially, we were asking at a theoretical level, what kind of guarantees can we give on the generalization abilities of a learning algorithm specifically, we were considering the question of.

(Refer Slide Time: 00:43)



Recap

- We have been considering the question of generalization abilities of a learning algorithm.
- A learning algorithm can learn the minimizer of the empirical risk.
- We actually want the minimizer of risk. Hence, the question we are considering is when does minimizer of empirical risk approximate well the minimizer of risk.

NPTEL

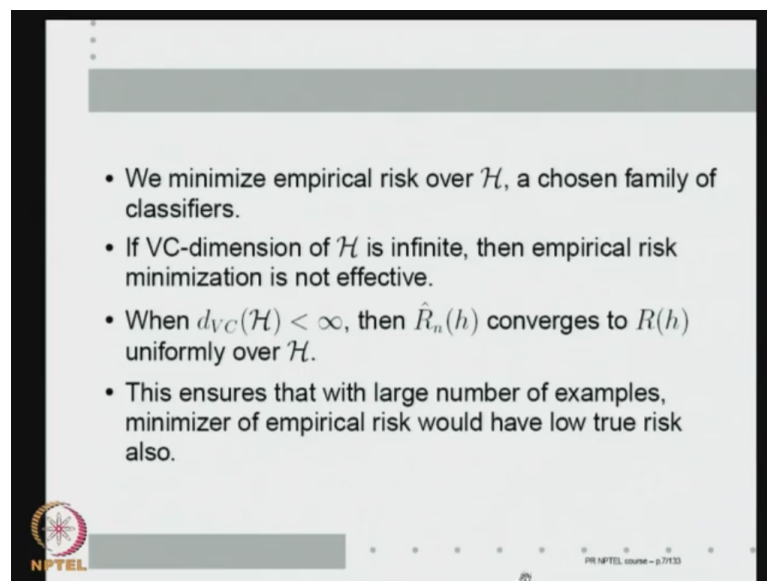
The slide features a small inset video of Prof. P. S. Sastry in the bottom right corner. The NPTEL logo is visible in the bottom left corner of the slide.

Formally understanding the generalization abilities of a learning algorithm in a classification context. So, today will be the last class, we will, rap up everything, that we have done, so far in terms of VC dimension. So, let us quickly recall what we have been doing basically, we have introduced the risk minimization at a generic frame work. In which any method of learning, from examples can be viewed that allows us to take care of arbitrary distributions with respect to which examples come different kinds of loss functions. All noise models, and so on, so risk minimization as, we have seen is a very general model. For learning from examples, and we have defined, we know the feature, feature space or what we call the input space x and the outcome space y . We choose a

convenient family of classifiers \mathcal{h} and we define a loss function and risk is the expectation of loss.

And, we are looking very minimizer of risk unfortunately as we seen risk cannot be minimized, because to calculate risk we need the underlying probability distributions. So, we approximate the expectation that is in the risk by this sample, average, that is how we got the empirical risk. So, what any learning algorithm can do, is to minimize the empirical risk all learning algorithms as we have seen, minimize the empirical risk with respect to some convenient loss function, say 0 1 loss function or square error loss function or what have you. But what we actually want, is the minimizer of the risk, the true risk that the expectation of loss. So the question, we are considering is when does minimizer of empirical risk be a good approximation to minimizer of true risk?

(Refer Slide Time: 02:34)



The slide contains the following text:

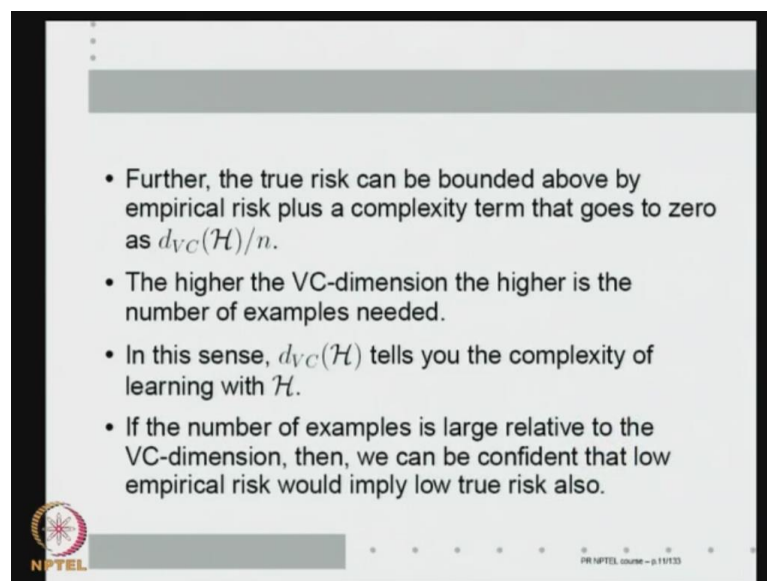
- We minimize empirical risk over \mathcal{H} , a chosen family of classifiers.
- If VC-dimension of \mathcal{H} is infinite, then empirical risk minimization is not effective.
- When $d_{VC}(\mathcal{H}) < \infty$, then $\hat{R}_n(h)$ converges to $R(h)$ uniformly over \mathcal{H} .
- This ensures that with large number of examples, minimizer of empirical risk would have low true risk also.

At the bottom left of the slide is the NPTEL logo. At the bottom right, it says "PR NPTEL course - p.7133".

Now, we minimize empirical risk over some family of classifiers \mathcal{H} , and what we seen is that if the VC dimension of this family of classifiers is finite, is infinite then the empirical risk minimization is not effective. That is the minimizer of the empirical risk can have two risk that is vastly different from the global minimizer of the true risk. On the other hand, if the VC dimension of the family of classifiers is finite. Then the finite VC dimension essentially implies, that the convergence implied by law of large numbers.

That is the sample mean expectation, sample mean approximation of expectation r_{hat} h converges to r_h . The true expectation of the risk uniformly over h and this uniform convergence in turn guarantees that minimizer of empirical risk, would also have low value of the true risk. So, essentially finally, the question boils down to if you were, if the family of classifiers h over which you are minimizing empirical risk has finite VC dimension then minimizing empirical risk is.

(Refer Slide Time: 03:44)



The slide contains the following text:

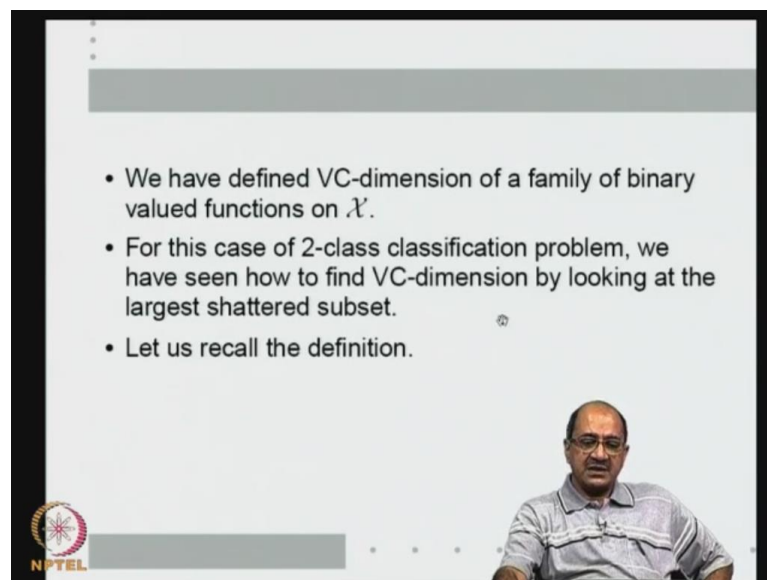
- Further, the true risk can be bounded above by empirical risk plus a complexity term that goes to zero as $d_{VC}(\mathcal{H})/n$.
- The higher the VC-dimension the higher is the number of examples needed.
- In this sense, $d_{VC}(\mathcal{H})$ tells you the complexity of learning with \mathcal{H} .
- If the number of examples is large relative to the VC-dimension, then, we can be confident that low empirical risk would imply low true risk also.

NPTEL logo is visible in the bottom left corner. In the bottom right corner, it says "PR NPTEL course - p 11133".

In addition, we also saw that the true risk can be bounded above by empirical risk plus a complexity term which goes to 0, as the ratio of the VC dimension of h by n . So, this also tells us that how many examples we actually need, before we can believe the empirical risk. So, if the empirical risk is sufficiently small, we can be confident that the true risk also be sufficiently small. If the VC dimension of h is divided by the number of examples is sufficiently small. So, higher the VC dimension higher is the number of examples needed. So, in that sense VC dimension not only tells us, whether empirical risk minimization is effective but, more importantly it tells us the complexity of learning with a particular classifiers h . So, for example, learning with a linear classifiers learning, a best linear classifier might have less complexity and learning a polynomial classifier of degree up to p . And that happens, because the VC dimension of one would be higher than the other.

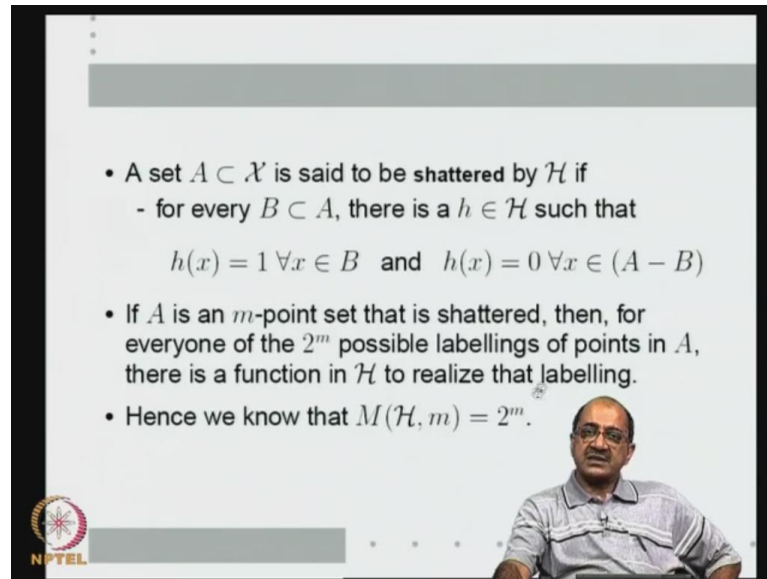
So, the VC dimension tells us, the complexity of learning with a particular h . And correspondingly tells us, what is the number of examples we need, before we can have confidence on empirical risk. In this sense, while the bounds, we got with for this generalization error or the true risk of a classifier or rather loose, it still gives us an idea. Of how many examples we need, before we can be confident that low empirical risk would imply low true risk. Essentially how many examples, we need depends on VC dimension as I said as a thumb rule we need at least 10 times the VC dimension as the number of examples.

(Refer Slide Time: 05:28)



The we have, we have defined VC dimension for the classifier case that, is when h is a family of binary valued functions on x though we are considering only two class classification problem. And, we have seen that for this particular true class classification problem. We can define VC dimension in terms of the largest shattered subset, of course, this the idea of VC dimension, the idea of bounding the true risk by empirical risk for, for say complexity terms. Holds for all kinds of families of functions, we are defining the corresponding VC dimension for other class of functions is more difficult. So, we have restricted ourselves to only considering 2 class classifiers. So, let us recall this definition of VC dimension based on the shattered subset.

(Refer Slide Time: 06:14)



• A set $A \subset \mathcal{X}$ is said to be **shattered** by \mathcal{H} if

- for every $B \subset A$, there is a $h \in \mathcal{H}$ such that

$$h(x) = 1 \forall x \in B \quad \text{and} \quad h(x) = 0 \forall x \in (A - B)$$

• If A is an m -point set that is shattered, then, for everyone of the 2^m possible labellings of points in A , there is a function in \mathcal{H} to realize that labelling.

• Hence we know that $M(\mathcal{H}, m) = 2^m$.

NPTEL

Given any subset of our input space, that is the feature space let us say A , is a subset of \mathcal{X} it is said to be shattered by the family of classifiers \mathcal{H} . If for every subset B of A , there is a h , so give me any subset B of this of A as a subset B , so give me any set of points in the feature space A . Then for every subset B of A there is a particular h in my family of classifiers, such that h takes 1 for all points in B and takes 0 on all points in $A - B$. If of course, the h that depends h that exists will depend on the B that we chose but for every subset B of A , If I can find a h that realizes that classification. By choosing a subset of B is effectively as if you are saying out of the set of points A , I have now I want to label all points in B as 1 and all points not in B as 0 that is what. And then there is a particular classifier in \mathcal{H} that can achieve that classification, if this can be done for every subset of A then, we say A , is shattered.

So, if an m point set is shattered, so if A , has m points then they, are 2^m different subsets that is 2^m different ways of labelling each point in A , with either 0 or 1, And if A , is shattered that means for each one of the 2^m possible labellings. There is a function in \mathcal{H} , there is a classifier in \mathcal{H} which will realize that labelling or which will which will classify the points as labeled. So, shattering of a set A simply means you, you do an arbitrary labelling of the points of A , with 0 and 1 then there must be a classifier in my family of classifiers. That would achieve their classification, so when an m point set

is shattered, each of the 2^m possible labeling of points in \mathcal{X} are realizable with functions in \mathcal{H} , which in turn means, what we call $M(\mathcal{H}, m)$ of the maximum number of distinguishable functions in \mathcal{H} based on all possible m tuples. Of examples is 2^m that is the reason why the VC dimension can now be defined as the cardinality of the largest shattered subset.

(Refer Slide Time: 08:28)

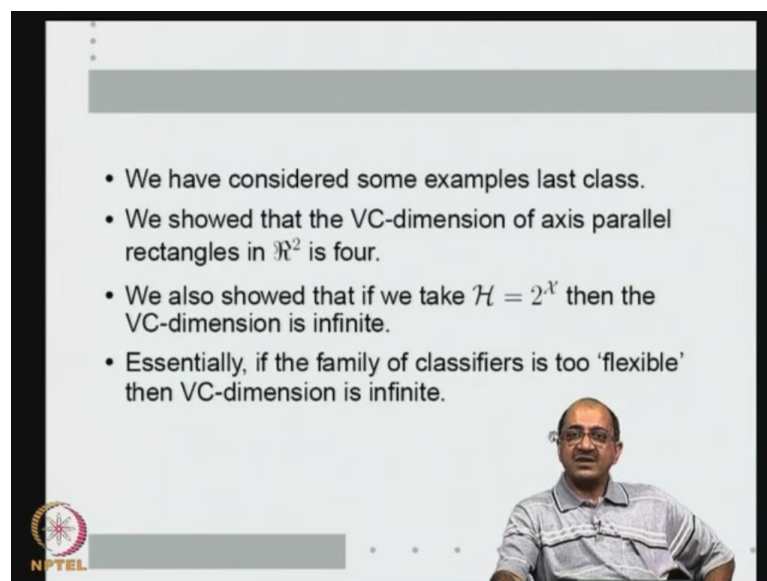
- VC-dimension of \mathcal{H} is the cardinality of the largest shattered subset of \mathcal{X} .
- If for every integer m , there is a m -point subset of \mathcal{X} that is shattered by \mathcal{H} , then, VC-dimension of \mathcal{H} is infinity.
- Note that if we have a m -point subset of \mathcal{X} that is shattered by \mathcal{H} then, for every $m' < m$, there is a m' -point subset of \mathcal{X} that is shattered.
- This is same as the earlier definition that $M(\mathcal{H}, m)$ grows as 2^m only till $m \leq d_{VC}(\mathcal{H})$.

So, we as we seen last class, VC dimension of \mathcal{H} is the cardinality of the largest shattered subset of \mathcal{X} . I will emphasize once again shattering is a property of a subset of \mathcal{X} and a family of classifiers a subset of \mathcal{X} , is shattered by a family of classifiers. So, whenever we say shattered because the \mathcal{H} is already understood. So, VC dimension is the cardinality of the largest shattered subset of \mathcal{X} , so on the when we say largest we are assuming there exist a largest. For example, for every integer m there is an m point subset of \mathcal{X} that is shattered. That mean there is no larger subset and hence VC dimension of \mathcal{H} is infinity. As I mentioned last class if an m point set is shattered by \mathcal{H} then for any m' prime less than m . There is also an m' prime point sub set of \mathcal{X} that is shattered namely an m' prime point subset of this m set.

So, a set \mathcal{A} , with m points is shattered that means all possible labellings or the m points are realizable by the classifiers in \mathcal{H} . You take any subset of that \mathcal{A} then all possible

labellings of that subset are also realizable by classifiers in H . Thus whenever, there is a m points of sets that is shattered for every m prime less than m , they will also be an m prime point subset that is shattered. So, this is same as saying that the number of maximum number of distinguishable functions grows as 2^m only till m reaches VC dimension of h . And hence defining VC dimension by the large cardinality of largest shattered subset is correct.

(Refer Slide Time: 10:09)



- We have considered some examples last class.
- We showed that the VC-dimension of axis parallel rectangles in \mathbb{R}^2 is four.
- We also showed that if we take $\mathcal{H} = 2^X$ then the VC-dimension is infinite.
- Essentially, if the family of classifiers is too 'flexible' then VC-dimension is infinite.

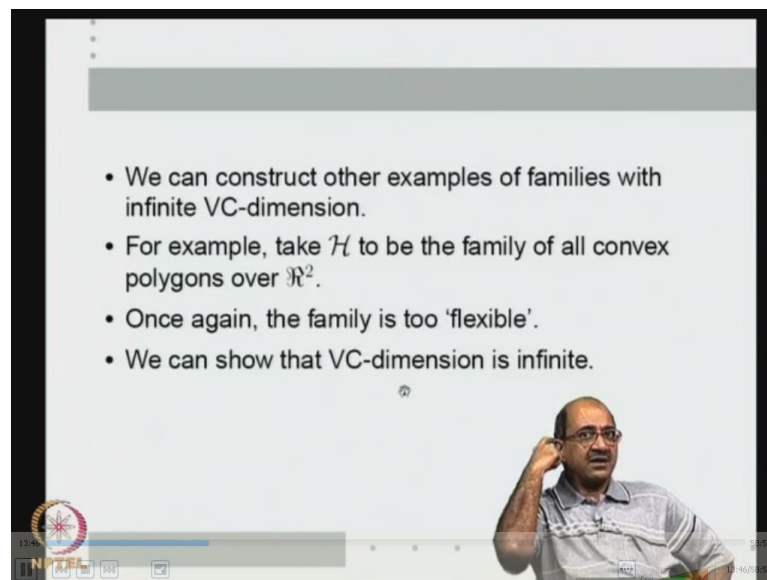
Once again let us understand what shattering means, if we find one m point subset of x that is shattered. Then, we can conclude that the VC dimension is at least m because there is at least one m point subset that is shattered. Of course, this does not mean that all m points subset are shattered. There may be other m point subset that are not shattered, but if there is at least one m point subset that is shattered. Then we can conclude that VC dimension is at least m . On the other hand to show that VC dimension is strictly less than m , we have to show that no m point set is shattered. To show that is at least m all we need to do is exhibit one example, where as to show that is less than m we have to show that no possible m points of set is shattered by H .

We considered some examples last class for example, we showed that the VC dimension of axis parallel rectangles is four and as I mentioned is also interesting to know that the

family of axis parallel rectangles can be represented by four parameters. Because each axis parallel rectangle is completely determined by the co-ordinates of its bottom left and top corners which need four numbers. we also seen that if we take H to be all possible two class classifiers H is equal to 2^x the power set of x then VC dimension is infinite. Essentially, if the family of classifiers over which you are minimizing the risk is too flexible, then VC dimension becomes infinite.

We seen both these examples last class, the the infinite VC dimension example that we considered of course, looks too drastic. We are saying if we take all possible two class classifiers then VC dimension is infinite. That is that is very obvious, because you cannot learn if you know we do not restrict our set of classifiers at all. But, that is not how VC dimension, becomes infinity is essentially this this (()) phase too flexible. We can we can think of many other families of classifier for which VC dimension will be infinite, because all those other families of classifiers are essentially too flexible in the sense. For example, it is very difficult to parameterize them with finite parameters.

(Refer Slide Time: 12:33)

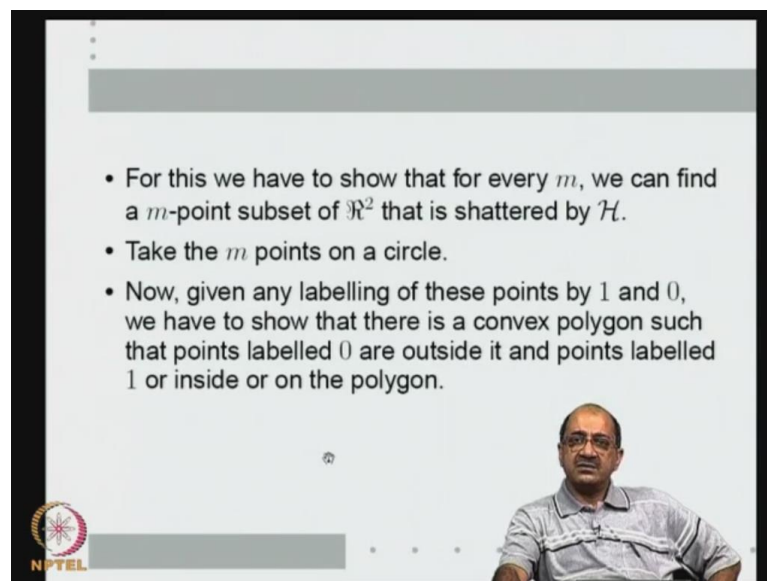


So, we will we will look at one more example, let us say H is the family of all convex polygons over \mathbb{R}^2 . For example, axis parallel rectangle is a specific convex polygon which contains only four sides even among four side convex polygons is a very specific

thing because it has to be an axis parallel and rectangle. But, instead of that suppose you take all convex polygons, of course, this is much, much smaller than smaller in the sense. There are many two class classifiers which cannot be expressed as convex polygons, because convex polygons means one class is a convex set.

The region of one class is always a convex set, so that looks like fairly restricted. But, still because we allow all convex polygons of any number of sides. Once again it is too flexible it is like fitting polynomials of any degree two points. So, if we take \mathcal{H} to be the family of all convex polygons over \mathbb{R}^2 once again the VC dimension will be infinite. Why, because the family is too flexible because I am allowing convex polygons of any number of sides. So, let us show this, so that we understand where the infinite VC dimension comes from.

(Refer Slide Time: 13:49)



- For this we have to show that for every m , we can find a m -point subset of \mathbb{R}^2 that is shattered by \mathcal{H} .
- Take the m points on a circle.
- Now, given any labelling of these points by 1 and 0, we have to show that there is a convex polygon such that points labelled 0 are outside it and points labelled 1 or inside or on the polygon.

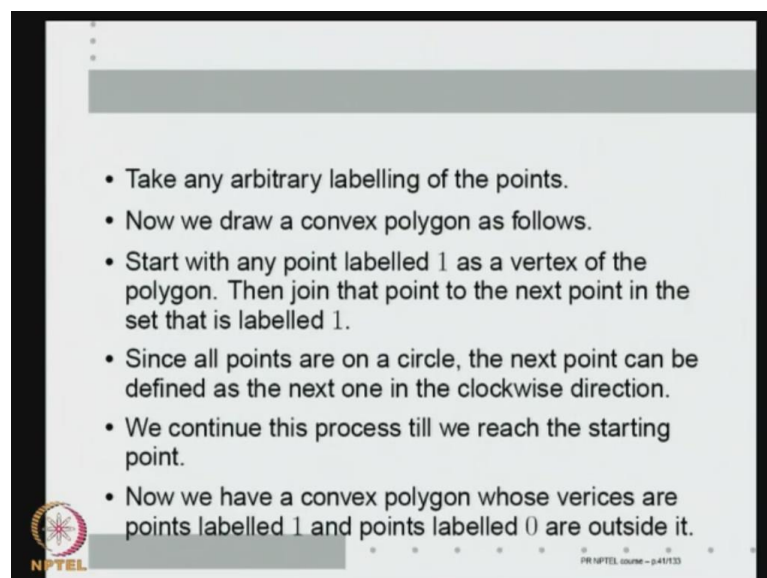
What do we have to show that VC dimension is infinite; we have to show that you give me any integer m any, any positive integer m . Then I can find an m point subset in \mathbb{R}^2 that is shattered by \mathcal{H} mind you we do not have to be able to show that every m point set is shattered. All I have to show is for every m there is at least one m points of set that is shattered. So, I have to just choose one particular or, or I have to exhibit one particular m point set for every m that is shattered. Here is how I can do it you give me any m I take

all the m points on a circle, circle put anywhere in \mathbb{R}^2 its circle and radius does not matter its center and radius, does not matter except that all the points have to be on a circle. So, I will take an m points of set or m point set such that all m points are on a circle.

Now, we are going to show that this is shattered, by the family of all convex polygons. What do I have to show that now given any labelling of this m points so you arbitrary label this m points 1 or 0. Then given any one labelling like that, we have to show that there is a convex polygon, such that the points labelled 0 are outside the convex polygon. And points labelled 1 are inside or on the polygon that been doing all alone, we, we are taking our, our functions to be such that.

Whenever, we take like axis parallel rectangles on on the polygon also as class1, it is a arbitrary thing but, let us take like that. So, we have to show that given any m and, and I take all the m points on a circle, now you can put any labelling of this points by one and 0. Now I have to show you a convex polygon, such that all points labelled 0 are outside the outside of it and points labelled 1 are inside or on the polygon

(Refer Slide Time: 15:46)



The slide contains a list of six bullet points describing the construction of a convex polygon from a set of points on a circle. The points are labeled 1 or 0. The steps are:

- Take any arbitrary labelling of the points.
- Now we draw a convex polygon as follows.
- Start with any point labelled 1 as a vertex of the polygon. Then join that point to the next point in the set that is labelled 1.
- Since all points are on a circle, the next point can be defined as the next one in the clockwise direction.
- We continue this process till we reach the starting point.
- Now we have a convex polygon whose verices are points labelled 1 and points labelled 0 are outside it.

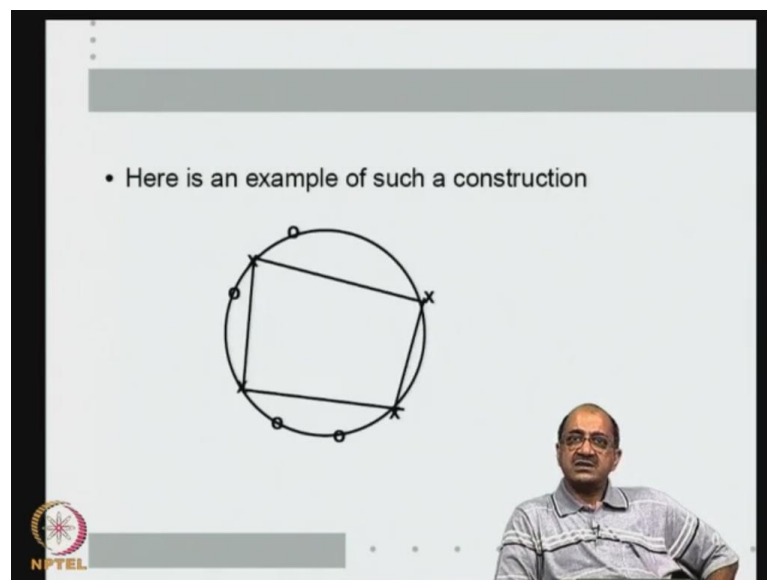
The slide also features the NPTEL logo in the bottom left corner and the text 'PR NPTEL course - p-41133' in the bottom right corner.

So, I will construct this as, as follows, so I have taken all the m points on a circle, now

you take any arbitrary labelling of these point. And what do, I do I have to draw a convex polygon for it I will draw it as follows. I start with any point labelled 1 as 1 a one vertex of the polygon. Then join that point to the next point in the set, that is labelled 1 , I will I will start with a one point then skip over all the 0. So, to say so I will find where is the next point labelled 1 and join this 1, this point labelled 1, to the next point labelled 1 what is the next point in R^2 well all the points on a circle.

So, I can take the next point to be at the next one in the clockwise direction, so keep going like this till I reach the starting point. What I have is a convex polygon, because I just joined points on a circle, and it is easy to see that, we have a convex polygon whose vertices are all points labelled 1. And all the points labelled 0 are outside it because I am drawing a cart in in the circle which is inside the circle. Between two successively two points that are successive labelled 1 all the points in between are on the arc those are the labelled 1 so they will be outside the circle.

(Refer Slide Time: 17:05)

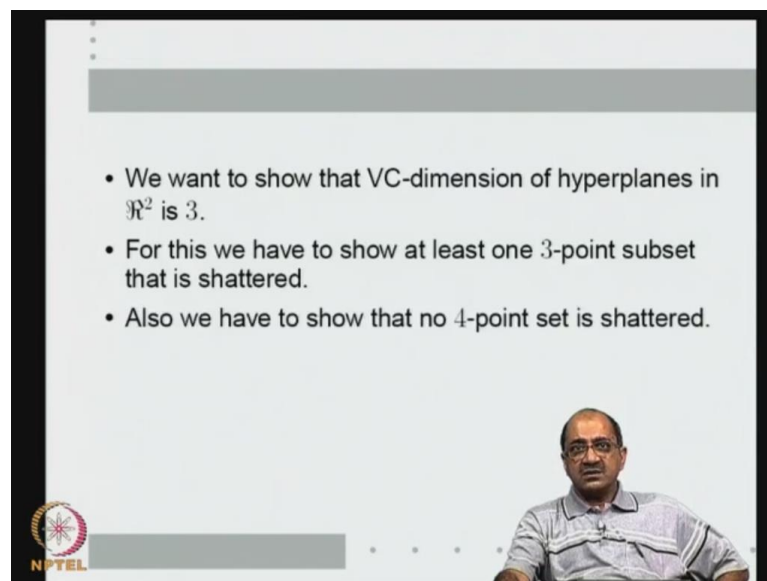


So, here is a example so I have taken 1 2 3 4 5 6 7 8 points on a circle arbitrary labelled on this is 0, this is I have, I have to, to show the labellings, I have put the points as a circular across. So, all of them are on a circle even though the, the labelling sometimes comes on it and sometimes a little bit. Because of my poor drawing, so this is label 0,

this is label 1, this is label 0, this is, label 1. These two are label 0 these two are label 1, so I start from this point, let us say, and then join it to the next point labelled 1. Join it to the next point labelled 1 join it to the next point labelled 1 and join it to the next point labelled 1.

Obliviously this is a cord, and all the points labelled 0 are in the arc of that cord, so all of them are outside the convex polygon. So, this way all points labelled 0 will always be outside the convex polygon, and all points labelled 1 will be on the convex polygon. They will be the vertices of the polygon, so this shows that any set of m points can be shattered. And hence the family, of all possible convex polygons also has VC dimension infinite. Now let us move, from these examples to one class of functions which are very important to us.

(Refer Slide Time: 18:31)



We, we mentioned last class that linear classifiers is an important special case actually we spent lot of time before, we coming to before we came to the series of lecture on statistical learning theory on learning linear classifiers, linear classifiers and linear regression functions minimizing empirical risk under square loss function. We have considered lot of algorithms and seen very many important properties of it as will become evident later. On in the course, linear classifiers are important, special case of

classifiers so let us ask what is the VC dimension of linear classifiers. So, in \mathbb{R}^d some d dimensional space we want to know, what is the VC dimension of hyperplane classifiers?

We are considering only two class classifiers any two class linear classifier is represent by hyperplane, there is a hyperplane on \mathbb{R}^d and one side is 1 class other side is other class. So, the set of all hyperplane classifiers is same as set of hyperplane functions, set of all linear classifiers is same as set of hyperplane functions. As it turns out the VC dimension of this class is $d + 1$ if you are considering feature space of dimension d then VC dimension of linear classifiers is $d + 1$. Now going to prove this, this, this is going to be the main result of this class. So, before we prove this in general d dimensions, let us first consider the case of hyperplanes in \mathbb{R}^2 .

So, in \mathbb{R}^2 , I have to show that, that, the, the VC dimension is 3, because $d + 1$ is 3. So, what do I have to show, I have to show, that VC dimension of hyperplanes in \mathbb{R}^2 is 3. As, we seen to show that VC dimension of something is m what do I have to show I have to exhibit at least one m points of set that is shattered. And show that no $m + 1$ point set is shattered. So, to show that VC dimension of hyperplanes in \mathbb{R}^2 is 3 I have to show that there is at least one 3 point set that is shattered. And no set of four or more points I say no set of four points is shattered. That is what we have to do, we have to first show that there is at least one 3 point set that is shattered. And then we have to also show that no 4 point set is shattered.

(Refer Slide Time: 20:55)

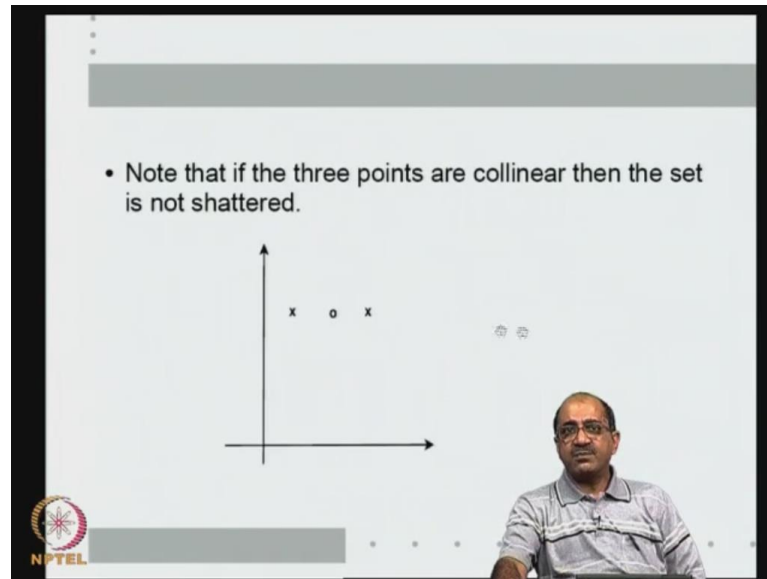
• Here is a 3-point set that is shattered.

The diagram shows a 2D coordinate system with three points marked 'x'. Three lines (red, blue, and green) are drawn such that each line separates one point from the other two, demonstrating that the set is shattered.

NPTEL

So, here is a 3 point set as a matter of fact you take any 3 point that form triangle, then you can shatter it as a matter of fact. I shown only see what do I have to show for shattering, that I can draw a hyperplanes keeping all 3 points on 1 set I have not shown that. So, that will be in the hyperplanes, so that all 3 labelled as 1 or all 3 labelled as 0 as gone. Then the remaining labelling is one of them 1 other, other two are 0, so I should be able to separate any one from the other two. So, there are 3 such cases, so here are the 3 hyperplanes that separate any one from the other two. So, for example, these are the 3 hyperplanes, that show you that the 3 points set is shattered, very simple as you already seen that one 3 point set is shattered does not mean that all 3 point sets are shattered.

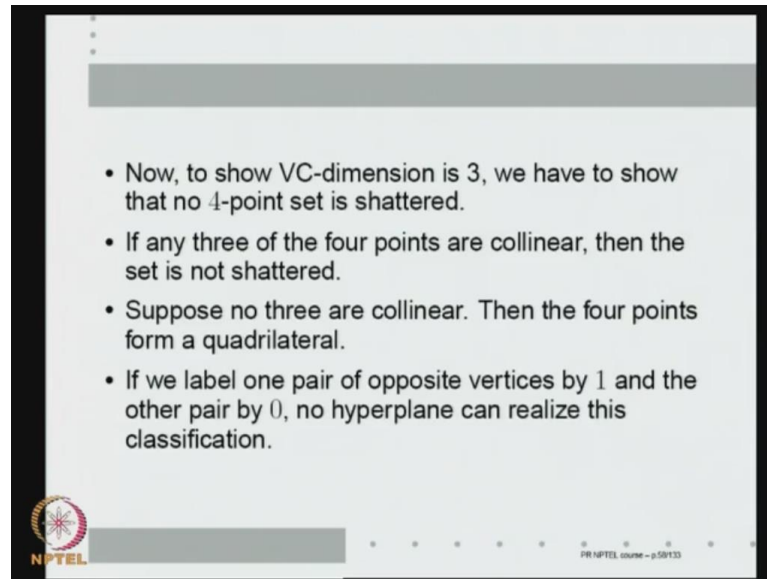
(Refer Slide Time: 21:49)



Here is a simple example, of a 3 point set that is not shattered, so if you give me 3 points in a line label the middle one as 0 the other two as 1 then I cannot draw any hyperplane. That puts the middle point on one side and the other two points on the other side of the hyperplane. So, the 3 points are collinear then the set is not shattered that is also valid. There is just one other thing I would like you to pay attention to while I drew the coordinate axis here. The coordinate axis is really useless this shattering is essentially a geometric property. Given these 3 points I am telling you I can separate them with hyperplanes whether my coordinate origin is here or here or here or here. If I move this coordinate origin anywhere, it makes no difference to whether or not a given set of points is shattered.

So, shattering is essentially property of how the points are organized in space, rather than what their actual algebraic coordinates, also coordinate origin. For example, makes no difference to shattering, the same is true of this. All I want is the 3 points in a line it really does not matter, with respect to the origin where they are so I can move the origin anywhere. But, these 3 points will not be shattered this is also important towards later on all. Now to complete the proof I have to show that no 4 point set is shattered we already know something a 3 point set is not shattered, if the 3 points are collinear.

(Refer Slide Time: 23:35)



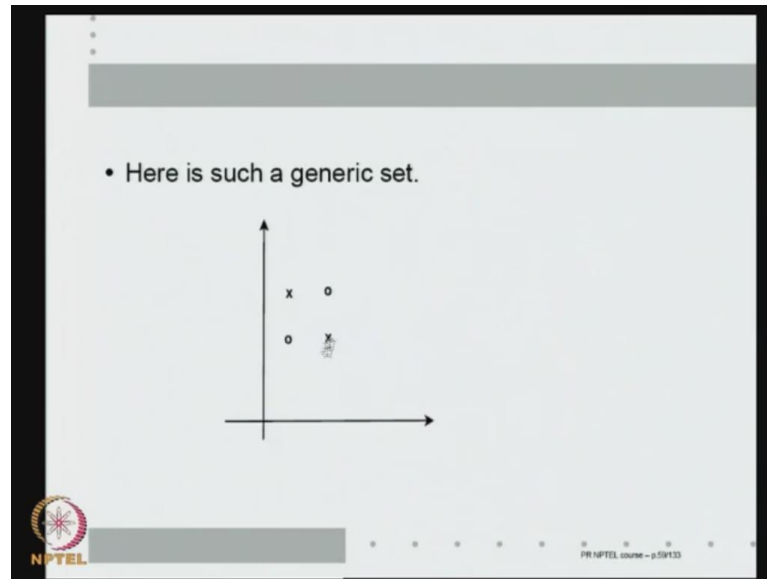
- Now, to show VC-dimension is 3, we have to show that no 4-point set is shattered.
- If any three of the four points are collinear, then the set is not shattered.
- Suppose no three are collinear. Then the four points form a quadrilateral.
- If we label one pair of opposite vertices by 1 and the other pair by 0, no hyperplane can realize this classification.

NPTEL

PR NPTEL course - p 58133

Which means give me any four point set if 3 of the 4 points are collinear. Then the set is not shattered so all possible four point sets in which 3 of the four points are collinear is anyway not shattered that is over that is shown. Now what is left I have to show for four point sets where no 3 are collinear in \mathbb{R}^2 . If you give me four points such that no 3 are collinear. Then the four points, form a quadrilateral, that is, that is the definition of quadrilateral. Given any four points in \mathbb{R}^2 if no three of them are in a line then the four points will form a quadrilateral. Because they form a quadrilateral if I label one pair of opposite vertices by one and the other pair of opposite vertices by 0. Then no hyperplane can realize this particular labelling, I hope that is clear, this is let me show the.

(Refer Slide Time: 24:41)

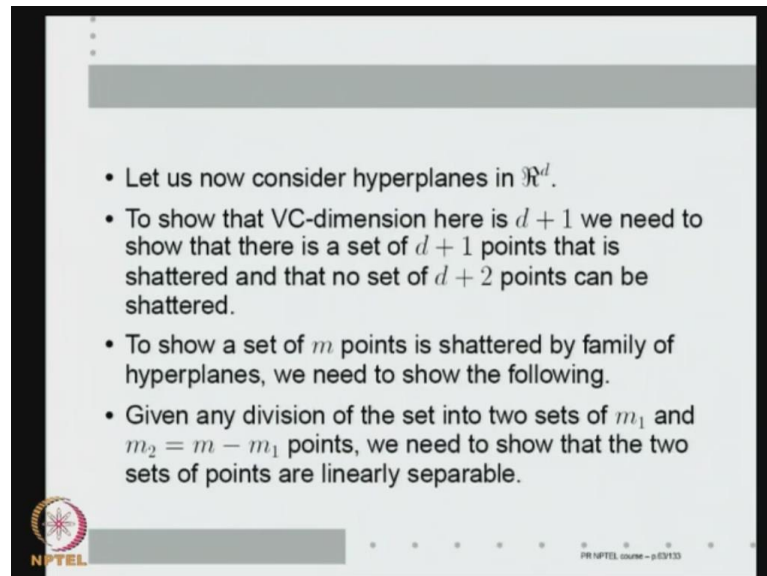


So, here is such a generic set give me four points such that no 3 are collinear then the four points form a quadrilateral. For a quadrilateral there is always, I can say which of the opposite pairs of vertices. So, this is one opposite pair of vertices, so which one diagonal will be and this is the other pair of opposite vertices. So, if I label these two by one class say one and these two by other class say 0 that does not exist a linear classifier. As a matter of fact if you remember when we did perceptron we showed one of the simple problem that perceptron cannot solve with the x r problem. This is like the x r problem as I said coordinate origin makes no sense no difference. So, I can think of let us say this is 0 and this is one so essentially these points can correspond to 0 1 and 1 0 and these points can correspond to 0 0 and 1 1. So, 0 0 and 1 1 has to give me one output and 0 1 and 1 0 has to be give me another output.

So, this is a typical exclusive or gate kind of problem, that cannot be solved by a linear classifier, because there does not exist a line to separate this pair of points with this pair of points. So, this is, this is going to be our learning example to show that, you know there are simple problems that can be, that cannot be solved by linear classifiers. Anyway in this particular exercise this shows us that no four point set is shattered because of the four points if any three are collinear then any way it cannot be shattered. If no 3 are collinear then the four points have to form a quadrilateral and if they form a quadrilateral

they cannot be shattered. Because if I label one pair of vertices by one and the other pair of vertices by 0, then no linear classifier can realize this classification. So this shows that there is a 3 point set that is shattered, no four point set is shattered, and hence VC dimension of hyperplane is 3 hyperplanes in \mathbb{R}^2 is 3.

(Refer Slide Time: 26:48)



So, now let us consider hyperplanes in \mathbb{R}^d what do I have to show we have to show that the VC dimension of hyperplanes in \mathbb{R}^d is d plus 1. Once again what does this mean I have to show that there is at least one set of d plus 1 points that is shattered. And no set of d plus 2 or more points can be shattered. I have to exhibit one set of d plus 1 points that can be shattered. And no set of d plus 2 or more points can be shattered, this is what we are going to prove now. A bit of maths a lot of equations, so let us go slowly. To show that a set of m points is shattered by a family of hyperplanes, what does that mean a set is shattered if every possible labelling of these points by 1 and 0. For every possible labelling of this points by 1 and 0 there is a classifier in by back that realizes that classification, which is same as saying the following see my classifiers are all linear classifiers.

So, a given labelling if its realized by my classifiers, that means the, the points labelled 0, are linearly separable from the points labelled 1. If I have set of points some of them

are labelled 0 and some of them, are labelled 1 and there is a linear classifier that can realize this classification. That means the set of points labelled 0 are linearly separable from the set of points labelled 1. Which is same as saying given any m point set if I divide that set into two sets one containing $m-1$ points other containing $m - (m-1) = 1$ is equal to $m-2$ points. I, i partition that set into two sets one containing $m-1$ point other containing some $m-2$ points $m - (m-1)$ points.

We need to show that these two sets of points are linearly separable so to show that an m point set is shattered by a family of hyperplanes. What do we have to show is for every possible division of this set into two subsets. The resulting two subsets are linearly separable, this is what we have to show, so to show this we have to first understand the geometry of linear separability. How can we say whether two given sets of points are linearly separable or not so to understand this geometry in a way that we need.

(Refer Slide Time: 29:08)

- For the proof we need the notion of a convex hull.
- Given $S = \{x_1, \dots, x_m\}$, the convex hull of S is

$$\text{Conv}(S) = \left\{ x : x = \sum_{i=1}^m \alpha_i x_i, \alpha_i \geq 0, \sum_{i=1}^m \alpha_i = 1 \right\}$$

- Convex hull of a set contains all points that can be written as convex combination of points in S .

NPTEL logo in the bottom left corner. Footer text: "© NPTEL course - p.00133"

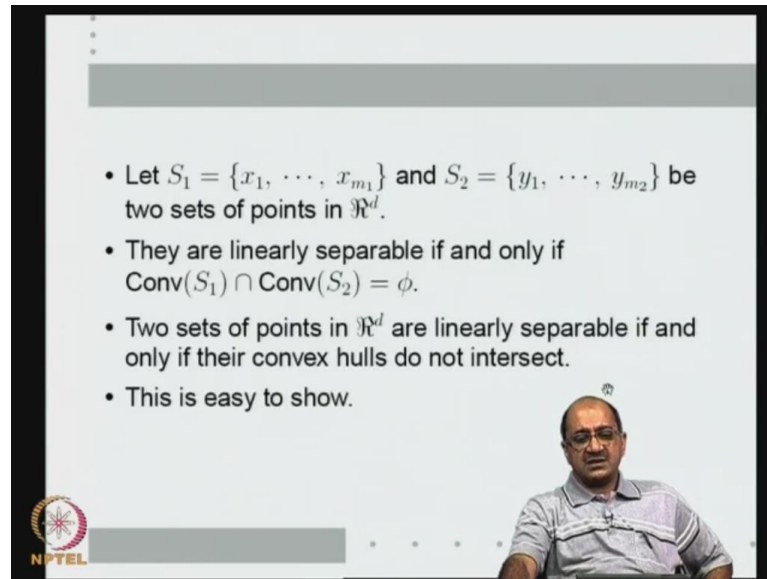
Here is a, an important concept that we need, we need, the concept of what is called a convex hull. I do not know how many people, how many of you know what a convex hull is. Let us define the convex hull I am assuming that all you people know what is a convex combination and what is a convex set I will any way I will I will briefly tell it. When I am defining a convex hull given a set S contain containing the points x_1 to x_m

these are all points in \mathbb{R}^d the convex hull of S is defined to be a set of all x . So, that x can be written as $\sum_{i=1}^m \alpha_i x_i$, where x_i are these x_1 to x_m $\sum_{i=1}^m \alpha_i = 1$, where these α_i are scalars, and the scalars α_i are, they are all non zero.

And they sum to one so given any set of vectors x_1 to x_m and set of scalars α_1 to α_m where the scalars satisfy $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i = 1$ that this vector $\sum_{i=1}^m \alpha_i x_i$ is called a convex combination of these points. Given points x_1 to x_m $\sum_{i=1}^m \alpha_i x_i$ is called a convex combination of x_1 to x_m , if the scalars α_i satisfy these two conditions. So, a convex hull is nothing but, a set of, a set that are points that can be written as a convex combination of points of S . So, given a set S of some finite points and the convex hull of S is the set of all points that can be written as a convex combination of points in S .

Suppose you have only two points the convex combination in \mathbb{R}^2 or given in \mathbb{R}^d if the line segment that joins these two points, because of this condition. For example, if I have three points not two which form a triangle, then the convex hull is the triangular disc formed by joining those three making a triangle as with those 3 points as vertices. Then all points, which are on or inside the triangle, become the convex hull, of those 3 points, a little later, I will show you some simple geometrical example.

(Refer Slide Time: 31:36)



The image shows a video frame from an NPTEL lecture. In the background, a slide contains the following text:

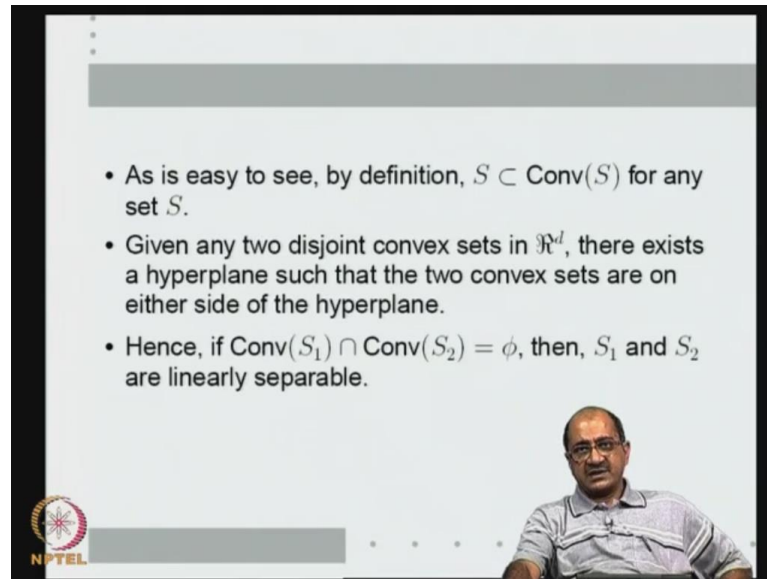
- Let $S_1 = \{x_1, \dots, x_{m_1}\}$ and $S_2 = \{y_1, \dots, y_{m_2}\}$ be two sets of points in \mathbb{R}^d .
- They are linearly separable if and only if $\text{Conv}(S_1) \cap \text{Conv}(S_2) = \phi$.
- Two sets of points in \mathbb{R}^d are linearly separable if and only if their convex hulls do not intersect.
- This is easy to show.

In the foreground, a man with glasses and a light-colored shirt is visible, appearing to be the presenter. The NPTEL logo is in the bottom left corner of the slide area.

Let us say I am given two sets of points in \mathbb{R}^d S_1 consisting of x_1 to x_{m_1} S_2 consisting of y_1 to y_{m_2} . Then a very useful result for us, is that the sets of points S_1 and S_2 are linearly separable. If and only if convex hull of S_1 and convex hull of S_2 do not intersect that is the see convex hull of S_1 , is some set of points in \mathbb{R}^d which is obtained as convex combinations of points from S_1 .

Similarly, convex hull of S_2 is that subset of \mathbb{R}^d which is obtained as convex combination of points in S_2 so the set of points S_1 and S_2 are linearly separable. If and only if the convex hulls of these two sets of points do not intersect. So two sets of points in \mathbb{R}^d , are linearly separable, if and only if their convex hulls do not intersect. It is not very difficult to show by using some well known property of convex sets, so let us show this.

(Refer Slide Time: 32:48)



Given any two disjoint convex sets in \mathbb{R}^d see, every convex set the, the, the property of convexity see, what is a convex set a convex set is a set where if you take any two points and find convex combination of those two points. All convex combinations of those two points are inside, the set, that is how a convex set is defined. The convex set a set is a subset of \mathbb{R}^d such that if I take any 2 points in that set, and find the convex combinations, find the convex combinations is just drawing a line joining those two points. All convex combinations are in that line segment on that line segment so if I take any two points the set and join a line or join them by line.

Then all points in that line are inside the set such a set is a convex set. The convex hull, because is made of convex combinations of points in it will be a convex set, given any two disjoint convex sets in \mathbb{R}^d there always exists a hyperplane. Such that the two convex sets are either side of the hyperplane, this is because any given convex set can be supported by a hyperplane, supported means there exists a hyperplane, which just touches the convex set. And all point of the convex set are on one side of the hyperplane meaning there will be a hyperplane such that the normal to the hyperplane will always make an acute angle with every point, on the convex set. That is the reason given two disjoint convex sets in \mathbb{R}^d they will always exists a hyperplane, such that the two convex sets are in either side of the hyperplane.

So, given this if convex hull of S_1 and convex hull of S_2 do not intersect, that means convex hull of S_1 and convex hull of S_2 are two disjoint convex sets. Then there will be a hyperplane, such that convex hull of S_1 is on one side of the hyperplane convex hull of S_2 is on the other side of the hyperplane. And the way the convex hulls are made S is the subset of convex hull of S for any set, because is a trivial convex combination one alpha is one and all others are 0. Because S is a subset convex hull of S if convex hull of S_1 and convex hull of S_2 can be linearly separable then $S_1 S_2$ are also linearly separable. So, this shows that if the convex hulls do not intersect then $S_1 S_2$ are linearly separable. Now, we will show the other way if $S_1 S_2$ are linearly separable then the convex hulls do not intersect.

(Refer Slide Time: 35:18)

• Now assume S_1 and S_2 are linearly separable. Then, $\exists W, b$ such that

$$W^T x_i + b > 0, \forall x_i \in S_1 \quad \text{and} \quad W^T y_i + b < 0 \forall y_i \in S_2$$

• Let $x = \sum \alpha_i x_i$ be any point in convex hull of S_1 .

• Then

$$W^T x + b = \sum_{x_i \in S_1} \alpha_i W^T x_i + b = \sum_{x_i \in S_1} \alpha_i (W^T x_i + b) > 0$$

(Note that $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$).

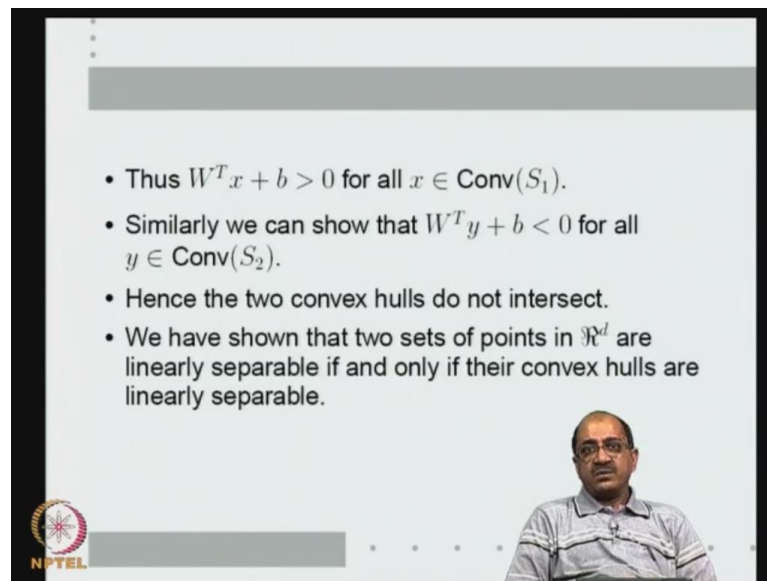
NPTEL

PR NPTEL course - p.70133

So, let us assume that $S_1 S_2$, are linearly separable, what is linear separable means recall from our lectures on p at the time of we discussed perceptron. We defined linear separability linear separable means, there is a exists a hyperplane such that all point of S_1 are one side of the hyperplane all points of S_2 are on the other side of hyperplane. So, the hyper plane is in R^d determined by W and b W is a d dimensional vector and b is a scalar. So, if S_1 and S_2 are linearly separable can there exists W and b such that W transpose x_i plus b is greater than 0 for all x_i in S_1 . And W transpose y_i plus b is less than 0 for all y_i in S_2 .

Now this is what we are given because we are given that S_1 and S_2 are linearly separable, now let x is equal to summation $\alpha_i x_i$, be any point in the convex hull of S_1 . Then suppose you take $W^T x + b$ now x is this so this will be $\alpha_i W^T x_i + b$ which I can write as summation over x_i belong to S_1 . Because this, this summation is over x_i belong to S_1 summation over x_i belong to S_1 α_i into $W^T x_i + b$. Why can I write this, because if take the second term that is b summation α_i will be equal to b because summation α_i is equal to 1. So I can put b inside the summation by multiplying α_i with $W^T x_i + b$. Now for every x_i in S_1 $W^T x_i + b$ is greater than 0. And we know α_i is positive, so the summation over x_i in S_1 α_i times $W^T x_i + b$ will also be positive.

(Refer Slide Time: 37:02)

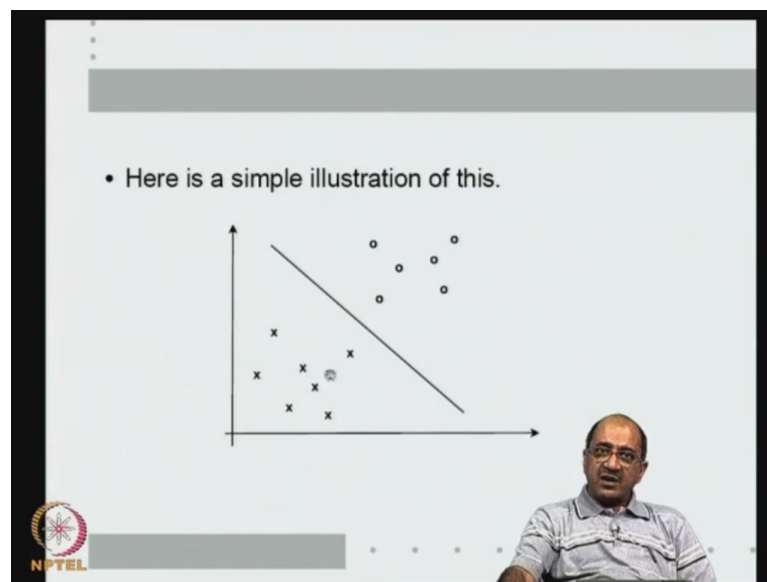


So, what it means is whatever may be the W and b that separate S_1 and S_2 are such that $W^T x + b$ will also be positive. For every x in the convex hull of S_1 not just x in S_1 but, for every x in the convex hull of S_1 , by exactly identical argument, we can show that $W^T y + b$ is less than 0 for every y in convex hull of S_2 . Which means there are some W and b such that for all points on the convex hull of S_1 in the convex hull of S_1 $W^T x + b$ is greater than 0. And for all points on the convex hull of S_2 $W^T x + b$ is strictly less than 0. Which means convex

hull of S_1 and convex hull of S_2 cannot intersect.

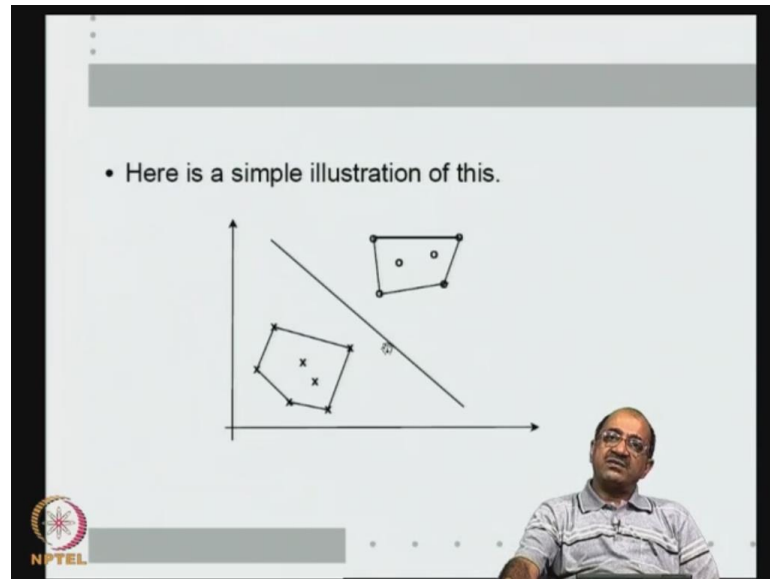
This shows that the two convex hull do not intersect and when the convex hull, do not intersect we, we completed the proof. So, what we have shown if the convex hull do not intersect then S_1 S_2 are separable linearly separable. If S_1 S_2 are linearly separable then convex hull do not intersect, thus we shown that two sets of points in \mathbb{R}^d are linearly separable. If and only if their convex hulls do not are are linearly separable, or, or the convex hulls do not intersect.

(Refer Slide Time: 38:18)



So, here is the example take this set of points, and this set of points they are linearly separable, because I can draw a line what is the convex hull of this points, and to make all convex combinations, so as I said essentially.

(Refer Slide Time: 38:35)



The convex hull will be that, I join all the, the outer lying points with lines to make a convex polygon, so all points on and inside the polygon are the convex hulls. So, for those for this set of points and for this set of points those are the convex hulls. And the set of points are linearly separable, if and only if the convex hulls are linearly separable. Because the convex hull, because the convex hull is such that you know it does not go beyond the extreme points, so to say in the set. And hence if the two sets are linearly separable, the convex hulls will also be linearly separable, this is what we just know algebraically shown.

(Refer Slide Time: 39:22)

• **Theorem:** Given m points in \mathcal{R}^d . Take one of them as origin. The set of m points is shattered if and only if remaining $m - 1$ points are linearly independent.

• For the first part we need to show:
linearly independent \Rightarrow shattered.

• Let $S = \{0, x_1, \dots, x_{m-1}\}$ be the set.
(Here, 0 is the origin or zero vector in \mathcal{R}^d).

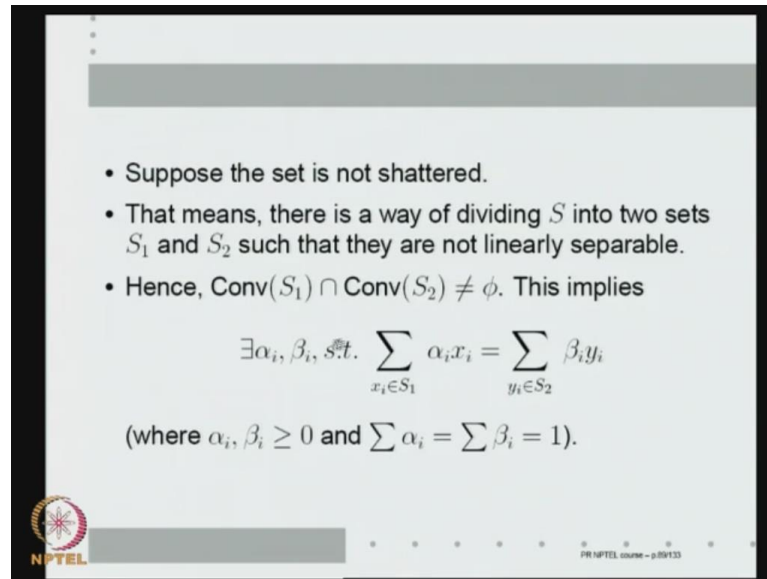
• We are given that the $(m - 1)$ points are linearly independent:
we can not have $\sum \gamma_i x_i = 0$ unless all γ_i are zero.

NPTEL PR NPTEL course - p.00133

So, now let us go back to using this to show that VC dimension of hyperplanes in \mathcal{R}^d is $d + 1$ before that we will do one theorem. Given m points in \mathcal{R}^d take any one of them as origin then the set of m points is shattered. If and only if the remaining $m - 1$ points remaining meaning the points other than the origin are linearly independent. So, showing if you give me $m - 1$ points on the origin in \mathcal{R}^d that makes m points. The set of m points is shattered, if and only if the non zero $m - 1$ points are linearly independent. So, to show this because if and only, if to show this I have to show that linearly independent implies shattering shattering implies linearly independent. So, for the first part let us show that linearly independence implies shattering.

So, we are given points a set of points 0 and $m - 1$ points let us call them x_1, x_2, \dots, x_{m-1} be the set. This is going to be the set for throughout this proof so let us remember this is the set here 0 is the origin or the zero vector in \mathcal{R}^d . We are given that the $m - 1$ points are linearly independent we are showing that given linearly independent imply shattering. So, we are given that the $m - 1$ point linearly independent which means for any scalars γ_i summation $\gamma_i x_i$ is equal to 0 only if all γ_i are 0 . Unless all γ_i are 0 I cannot have a summation $\gamma_i x_i$ is equal to 0 . For any scalars γ_i that is what a linearly independent means, so this is what we are given I have to show they are shattered.

(Refer Slide Time: 41:01)



- Suppose the set is not shattered.
- That means, there is a way of dividing S into two sets S_1 and S_2 such that they are not linearly separable.
- Hence, $\text{Conv}(S_1) \cap \text{Conv}(S_2) \neq \phi$. This implies

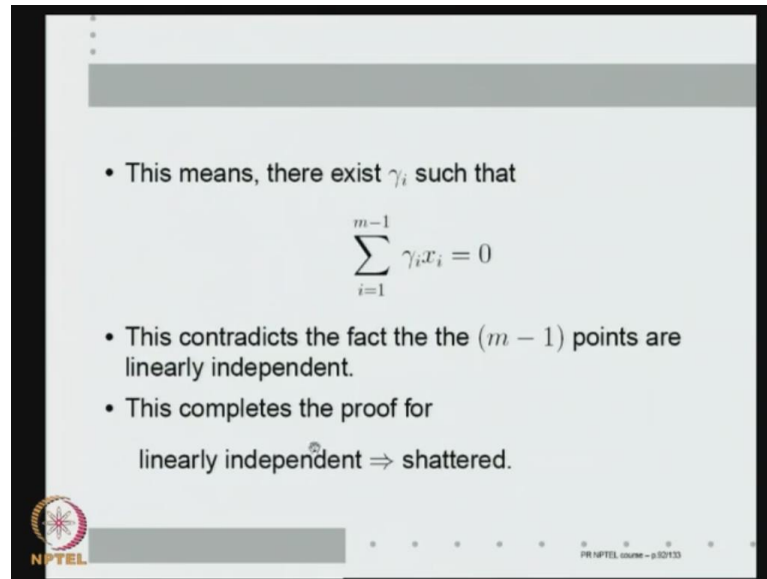
$$\exists \alpha_i, \beta_i, \text{ s.t. } \sum_{x_i \in S_1} \alpha_i x_i = \sum_{y_i \in S_2} \beta_i y_i$$

(where $\alpha_i, \beta_i \geq 0$ and $\sum \alpha_i = \sum \beta_i = 1$).

So, you have to show shattered so let us assume they are not shattered, they not shattered that means there is a way of dividing them into two sets S_1 and S_2 . Such that there are they are not linearly separable, they are not shattered means the set can be divided into S_1 and S_2 such that S_1 and S_2 are not linearly separable. Which means convex hull of S_1 intersection convex hull of S_2 is not equal to phi, because we have seen not linearly separable is same as the convex hull center set t , which means there is a point which is in the intersection of convex hull of S_1 and convex hull of S_2 any point in convex hull of S_1 can be represented as summation $\alpha_i x_i$ x_i belonging to S_1 . Any point at convex hull of x_2 can be represented as summation $\beta_i y_i$ y_i in S_2 note that S_1 and S_2 are disjoint sets.

So, the x_i is here will be all different from the y_i is here because the convex hulls intersect there is at least one set of α_i and β_i . Such that summation $\alpha_i x_i$ for x_i in S_1 is equal to summation $\beta_i y_i$ for y_i in S_2 . Were of course, α_i and β_i summed to one and they are positive. So, if I bring it the on this side call for all x_i in S_1 call γ_i is equal to α_i for all x_i in S_2 I call γ_i is equal to minus β_i . Then this is what I have so out of the m minus one points, some will be in S_1 some will be in S_2 . If I bring it this side I am considering all points all the m minus 1 points in S .

(Refer Slide Time: 42:37)



- This means, there exist γ_i such that

$$\sum_{i=1}^{m-1} \gamma_i x_i = 0$$

- This contradicts the fact the the $(m - 1)$ points are linearly independent.
- This completes the proof for

linearly independent \Rightarrow shattered.

NPTEL

PR NPTEL course - p 82133

So, essentially what I have is that there exist scalar γ_i that $\sum_{i=1}^{m-1} \gamma_i x_i = 0$. But, this is not allowed because I am given that the points are positive and not all γ_i can be zero because γ_i are obtained from $\alpha_i \beta_i$ and α_i and β_i have to sum to 1 so not all of them can be 0. So, if the set is not shattered then they must exist, scalars γ_i satisfying this which is not possible because x_i are linearly independent which means linearly independent means shattering. So, we completed the proof that if the remaining $m - 1$ points are linearly independent then the set is not shattered, then the set is shattered.

(Refer Slide Time: 43:25)

• Now we have to show:
shattered \Rightarrow linearly independent.

• We show this in its contra positive form.

• That is, we show:
not linearly independent \Rightarrow not shattered.

• Now we are given that there are scalars $\alpha_i, i = 1, \dots, m - 1$ such that

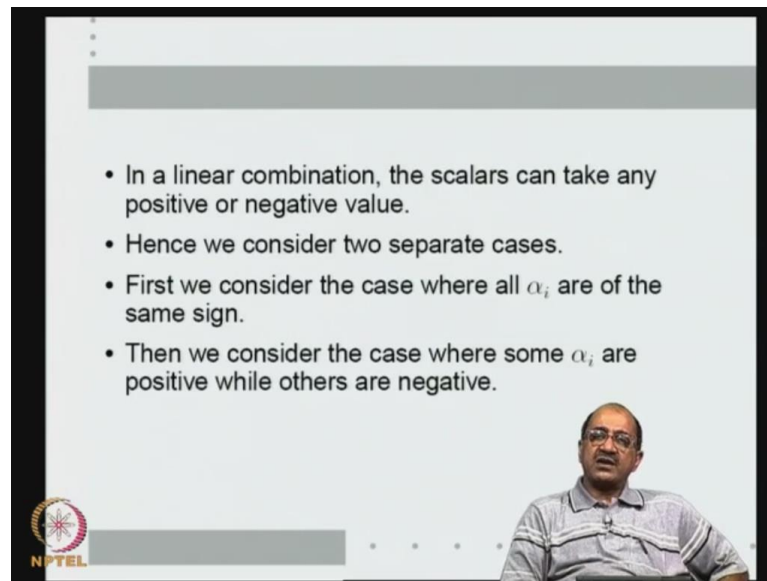
$$\sum_{i=1}^{m-1} \alpha_i x_i = 0$$

NPTEL

So, the second part we have to show that shattered implies linearly independent, we have shown that if you give me $0 \times 1 \times m$ minus 1×1 to x m minus 1 are linearly independent, then this set is shattered. Now, we are showing that if the set is shattered then x 1 to x m minus 1 are linearly independent. So, a, implies b is same as not b implies not a, that is called the contra positive form. So, we will show this in the contra positive form namely, not linearly independent implies not shattered. If they are not linearly independent means now this time there are scalars α_i , so there summation $\alpha_i x_i$ is equal to 0 not all α_i are 0.

But, summation $\alpha_i x_i = 0$, this is just a linear combination mind; this is not a convex combination. Because we are only given that they are not linearly independent, so there exists a linear combination of x_i that that is 0. So, for example some of the α_i is may be positive some of them may be negative, the α_i is do not have to sum to 1. It just show all that not linearly independent means is that there is one linear combination of x_i , that will that will sum to 0. We are given that there exist α_i such that i is equal to one to minus 1 $\alpha_i x_i$ is equal to 0.

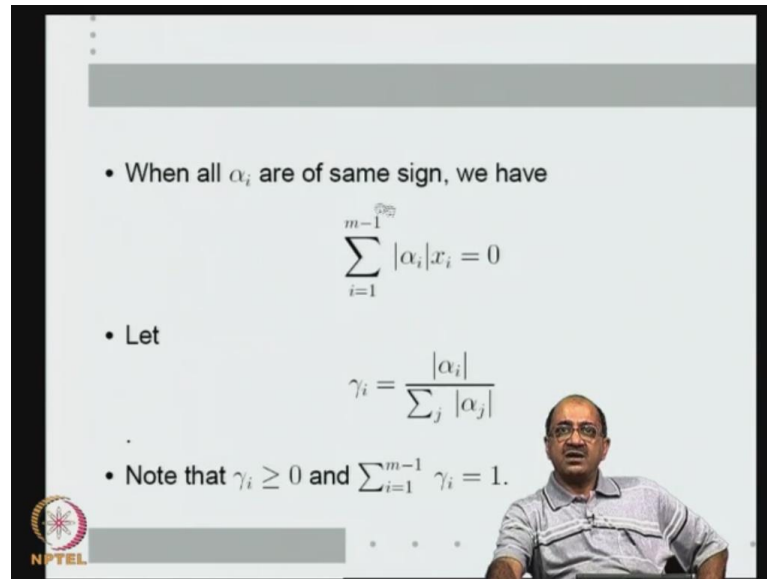
(Refer Slide Time: 44:35)



- In a linear combination, the scalars can take any positive or negative value.
- Hence we consider two separate cases.
- First we consider the case where all α_i are of the same sign.
- Then we consider the case where some α_i are positive while others are negative.

So, in a linear combination let us remember that the scalars can take any positive or negative values firstly they can be positive or negative. And there is no restrictions such as they have to sum to one or anything, so we will consider two separate cases. In the first case where alpha i are of the same sign that is the simplest case and that will also give us some idea of the proof of the general case. And then we consider the general case where there can be both positive as well as negative alpha i.

(Refer Slide Time: 45:01)



The slide contains the following text and equations:

- When all α_i are of same sign, we have
$$\sum_{i=1}^{m-1} |\alpha_i| x_i = 0$$
- Let
$$\gamma_i = \frac{|\alpha_i|}{\sum_j |\alpha_j|}$$
- Note that $\gamma_i \geq 0$ and $\sum_{i=1}^{m-1} \gamma_i = 1$.

The slide also features the NPTEL logo in the bottom left corner and a small inset image of the presenter in the bottom right corner.

So, let us suppose alpha i has this all alpha i has the same sign we have we have been given the alpha i is, are that this linear combination is 0. So, if all of them are in same sign either all of them are positive or all of them negative. If all of them are negative I can multiply by minus 1 and the equation still holds, so if all alpha i are of the same sign. Then what I am given is that summation i is equal to 1 to m m minus 1 modulus of alpha i into x i is equal to 0, like we are considering only real scalar that is why this is the absolute value.

Now, we take gamma i to be absolute value of alpha i by summation absolute value of alpha j then gamma i is, are greater than equal to 0 summation i is equal to m minus 1 gamma i is equal to 1. That is easy to see just normalize this by dividing it by some modulus absolute value of alpha j. So, this gamma i is now are greater than equal to 0 and summation i is equal to 1 to minus 1 gamma i is equal to 1. So, what do I have now if I divide this equation by summation over j absolute value of alpha j, there are some constants? So, I can divide this that then the then the factor becomes gamma i.

(Refer Slide Time: 46:18)

• Now we have

$$\sum_{i=1}^{m-1} \gamma_i x_i = 0$$

• This means that the zero vector is in the convex hull of the rest of the points.

• If we take $S_1 = \{x_1, \dots, x_{m-1}\}$ and $S_2 = \{0\}$, then, convex hulls of S_1 and S_2 intersect and hence we can not linearly separate them.

• Hence S is not shattered.

So, what I have is $\sum_{i=1}^{m-1} \gamma_i x_i = 0$ where γ_i is such that $\gamma_i \geq 0$ and $\sum \gamma_i = 1$. Which means what I have on the left side is a convex combination of points in x_i that is I have a convex combination of x_1, x_2, \dots, x_{m-1} which gives me the zero vector. What does that mean the zero vector is in the convex hull of the rest of the $m-1$ points, because the convex hull of clustered $m-1$ points contains all convex combinations of x_i , and there is one convex combination of x_i that equals 0. Which means the zero vector is inside the convex hull of the rest of the points, which in turn means suppose I am my set S is $\{x_1, x_2, \dots, x_{m-1}, 0\}$. So, if I divide into two subsets S_1 and S_2 , where S_1 contains x_1, x_2, \dots, x_{m-1} and S_2 contains 0.

Then the convex hull of S_1 and S_2 intersect there is there is a point in S_2 namely 0 of course, that is the only point in S_2 which is inside the convex hull of S_1 . So, because convex hulls of S_1 and S_2 intersect S_1 and S_2 cannot be linearly separated. Which means the original set S cannot be shattered, that means not linearly independent implies not shattered. So, what have we shown if you give me m points in \mathbb{R}^d then if I take one of them as origin and if the rest of the $m-1$ points are linearly independent, then the set is shattered, if they are not linearly independent, they are not shattered.

Of course, for the not linearly independent they are not shattered, we have done it only for the simple case. We have been considering, so far the case where all α_i are of the same sign. So, we have to consider the more general case, where the more general case is the same thing see, because all of them are same sign. I have this once are this I could normalize to make it a convex combination, once I make a convex combination I can show convex hull center set, the same thing will follow for the more general case.

(Refer Slide Time: 48:33)

• Now we consider the more general case.

• Let $I_1 = \{i : \alpha_i \geq 0\}$ and $I_2 = \{i : \alpha_i < 0\}$.

• Define $\beta_i = \alpha_i, \forall i \in I_1$ and $\gamma_i = -\alpha_i, \forall i \in I_2$.

• Note that $\beta_i, \gamma_i \geq 0$.

• Now, what we have is

$$\sum_{i \in I_1} \beta_i x_i = \sum_{j \in I_2} \gamma_j x_j$$

NPTEL

So, we consider the general case of course, I could have only considered this because this includes a special case some of them are positive some are negative. But, anyway let us say now they are both, positive or negative α_i is so let us say the set $i \in I_1$ consist of all indices. I_1 such that the corresponding α_i is are positive and I_2 contains those which are negative mind you these α_i is are those scalars which are in the linear combination of x_i is that goes to 0. Because x_i is are given to be not linearly independent there is one $\alpha_i x_i$ that is equal to zero with respect to those α_i is, I am defining the sets I_1 and I_2 . Such that I_1 consists of all indices where α_i is are positive I_2 consists of all indices that where α_i is are negative.

Now I know that over all points $\sum \alpha_i x_i$ is equal to 0, now let us say we define a β_i is equal to α_i for all i in I_1 a γ_i is equal to minus α_i for all

in i_2 . Now I have $\sum \alpha_i x_i$ is equal to 0 the, i_1 and i_2 those are positive terms i_1 and i_2 those are all negative terms. So, I can take all the negative terms on the other side if I take negative terms on the other side the coefficients becomes γ_i and the positive term the co-efficient become β_i .

So, what I have now because I am given $\sum \alpha_i x_i$ is equal to 0, $\sum \alpha_i x_i$ is equal to 0 is same as $\sum_{i \in i_1} \beta_i x_i - \sum_{j \in i_2} \gamma_j x_j$. Essentially from the equation $\sum \alpha_i x_i$ is equal to 0, I have taken all the negative terms on the other side. So, I get this equation now the equation is nice because all the scalars are positive. Now for $i \in i_1$ α_i is positive, so β_i is positive in i_2 α_i is negative so γ_i is positive.

Now this almost looks like if I take x_i in i_1 such that i is in i_1 and as one set and x_j in i_2 is another set. Then a linear combination of the points of first set is equal to the linear combination of the points of the second set where the linear combination contains all positive coefficients. So, I have to now turn it into convex combination but, I cannot simply normalize, because the summation β_i is may be different from summation γ_i is. I have to somehow just make sure that that does not pose a problem, so that is what we are going to do next.

So, given this what we doing now is, let us say $\sum \beta_i x_i = z$ and $\sum \gamma_j x_j = z'$. If $z = z'$ we are done, because it is simply says, that if I divide it by that corresponding common normalizing factor. Then this becomes a convex combination some of the x_i this becomes convex combination of the remaining x_i . So, the two convex hulls intersect, and hence this set and that set cannot be separated.

(Refer Slide Time: 51:42)

• Let $\sum_{i \in I_1} \beta_i = Z$ and $\sum_{j \in I_2} \gamma_j = Z'$.

• Without loss of generality, assume $Z \geq Z'$.

• Now we can rewrite the earlier equation as

$$\sum_{i \in I_1} \frac{\beta_i}{Z} x_i = \sum_{j \in I_2} \frac{\gamma_j}{Z} x_j + \frac{Z - Z'}{Z} 0$$

• Note that

$$\sum_{i \in I_1} \frac{\beta_i}{Z} = 1 \quad \text{and} \quad \sum_{j \in I_2} \frac{\gamma_j}{Z} + \frac{Z - Z'}{Z} = 1$$

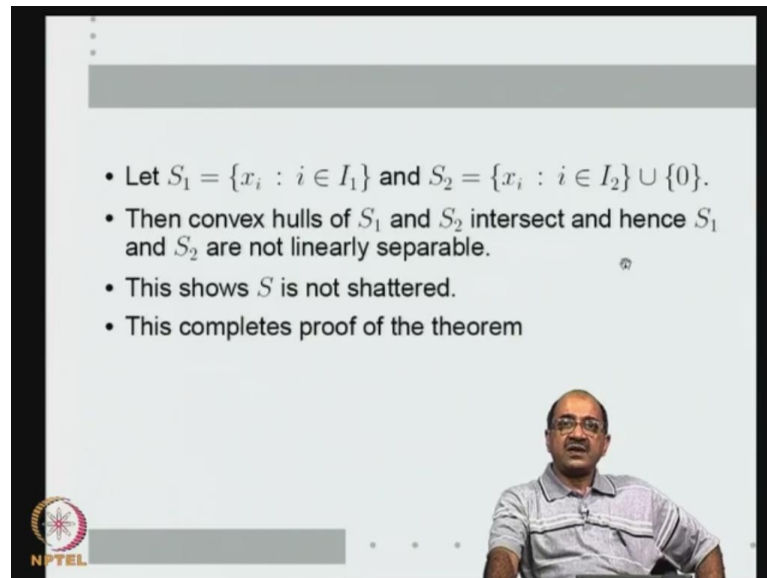
NPTEL logo and footer text are visible at the bottom of the slide.

But in general z and z prime did not have to be equal, one of them has to be greater than the other, so without loss of generality, let us assume z is greater than equal to z prime. Now I can write the earlier equation this is my equation $\beta_i x_i$ is equal to $\gamma_j x_j$ i in i_1 j in i_2 . That equation now I will write as β_i divide it by z first, so I get β_i by z into x_i is equal to γ_j by z into x_j to that now I add a 0. This is sum scalar z minus z prime by z multiplied by 0 the zero vector. So, this gives me anyway 0 so still true, I earlier have $\beta_i x_i$ is equal to $\gamma_j x_j$ divided both sides by z . So, β_i by z into x_i is equal to γ_j by z into x_j this is summation i in i_1 this summation j in i_2 . Now I can always add a 0, so I add a 0 with a coefficients z minus z prime by z , what is the purpose of this.

The purpose is now this is a convex combination of x_i with i in i_1 because summation β_i by z summed over i is 1. There are anyway positive and the sum to one, now this entire thing on the left hand side, is a convex combination of x_j such that j in i_2 plus the zero vector. Because if I sum all the co-vision for x_j j in i_2 is the co-vision is γ_j by z and for the zero vector the coefficient is z minus z prime by z . If I sum all of them summation over j γ_j by z plus z minus z prime by z summation γ_j is z prime so this becomes z prime by z add both of them I will get 1. So, what is there on the right hand side now is the convex combination of all the x_j is such that j in i_2 plus the zero

vector.

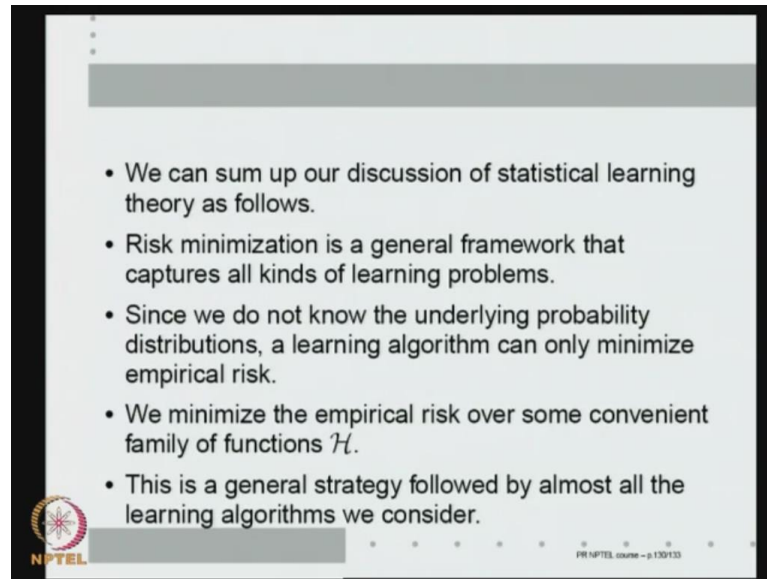
(Refer Slide Time: 53:41)



So, we take S_1 to be $\{x_i : i \in I_1\}$, S_2 to be $\{x_i : i \in I_2\} \cup \{0\}$. So, S_2 is all x_i $i \in I_2$ and the zero vector, S_1 is x_i such that $i \in I_1$. Then this is a convex combination of points in S_1 , this is a convex combination of points in S_2 , because the convex combination of x_j is such that $j \in I_2$ plus the zero vector. So, this is a convex combination of points in S_2 , if I take S_1 to be this and S_2 to be this, what we have shown is that the convex hulls of S_1 and S_2 intersect.

Because the convex hulls of S_1 and S_2 intersect, S_1 and S_2 are not linearly separable, this shows S is not shattered. So, what have we shown now with this completes the proof of the theorem, and what does the theorem say, if you give me n points m points in \mathbb{R}^d take one of them origin. If the remaining $m - 1$ points are linearly independent then the set is shattered if they are not linearly independent the set is not shattered. That is what we have shown this completes the proof of the theorem that we listed, now we have to see how this theorem tells me that VC dimension of hyperplanes is $d + 1$.

(Refer Slide Time: 55:05)



- We can sum up our discussion of statistical learning theory as follows.
- Risk minimization is a general framework that captures all kinds of learning problems.
- Since we do not know the underlying probability distributions, a learning algorithm can only minimize empirical risk.
- We minimize the empirical risk over some convenient family of functions \mathcal{H} .
- This is a general strategy followed by almost all the learning algorithms we consider.

NPTEL
© NPTEL, course - p 130133

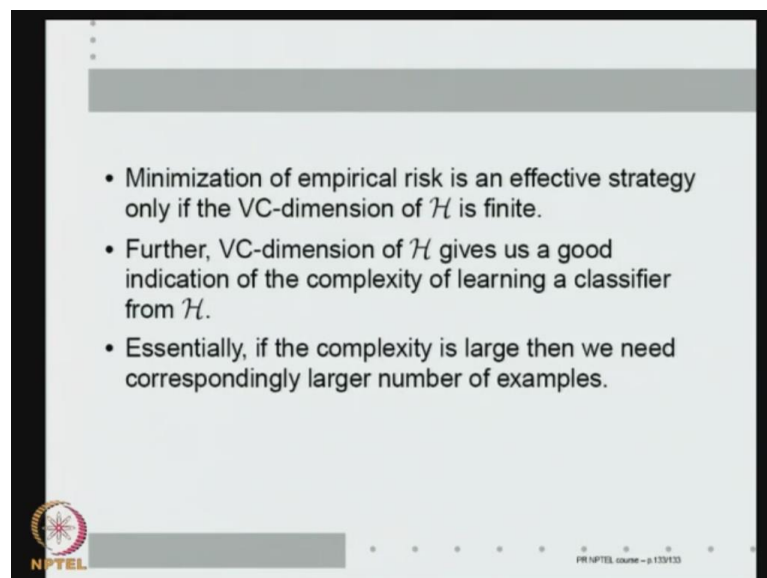
First as I mentioned already shattering of set of points by hyperplanes does not depend on coordinate origin. So, I can always take one of the points as origin by shift of origin. So, in \mathbb{R}^d , we can have d linearly independent points, so d linearly independent points along with origin will give me $d + 1$ points of the kind I want. I can take $d + 1$ points, origin plus any other d linearly independent points and such a set of $d + 1$ points is shattered. We shown, given any m points one as a origin remaining $m - 1$ points are linearly independent, then the set is shattered. So, you take 0 and d linearly independent points in \mathbb{R}^d , I can always get d linearly independent points.

So, if I take 0 along with d linearly independent points that gives me a set of $d + 1$ points that is shattered. So, we now shown that there is at least one set of $d + 1$ points in \mathbb{R}^d that is shattered. Now take any any set of $d + 1$ points, we can take one of them as the origin then there d plus 1 points. Now any set of $d + 1$ points in \mathbb{R}^d will be linearly dependent, because \mathbb{R}^d as dimension d any set of $d + 1$ points in \mathbb{R}^d will be linearly dependent. And hence given any set of $d + 1$ points in \mathbb{R}^d even, if I take one of them as origin the remaining $d + 1$ points will be linearly dependent and the set is not shattered. So, we shown that there is at least one set of $d + 1$ point that is shattered and no set of $d + 1$ points is shattered. And that shows that the VC dimension of hyperplanes in \mathbb{R}^d is $d + 1$.

So, it is a very important theorem, so essentially as the dimension grows, the complexity of learning hyperplane classifiers learning linear classifier also grows. So, from hundred dimensional space I want to learn say ten ten thousand dimensional space hyperplane. Then I need hundred times as many points, so in the hundred dimensional space if thousand examples is enough to learn hyperplanes. In ten thousand dimensional space I need hundred thousand points a hundred thousand examples to learn the hyperplane to the same accuracy. So, this is last class on statistical learning theory, so let us quickly sum up risk minimization a general framework that captures, all kinds of learning problems.

Since we do not know the underlying probability distributions a learning algorithm can only minimize empirical risk. We minimize the empirical risk over some convenient family of functions H , this is the general strategy followed by almost all the learning algorithms. That we consider, so in that sense we introduce a very nice generic framework in which all learning algorithms, can be viewed as far as the statistical properties are concerned.

(Refer Slide Time: 57:51)



And then we showed that minimization of empirical risk is an effective strategy, only if the VC dimension of H , is finite. We can only do minimization of empirical risk and

doing that is effective, if VC dimension is finite. Further VC dimension of H gives us a good indication of the complexity of the, of, of learning a classifier from H . Essentially if the, the complexity is large, then we need correspondingly larger number of examples. And the complexity also checks with our intuitive notion of complex as we seen for hyperplanes in \mathbb{R}^d . We need $d + 1$ parameters and the VC dimension is $d + 1$ for axis parallel rectangles also we need four parameters and the VC dimension is 4. And so on, so this kind of completes what we are going to do on statistical learning theory. So, from next class we will we will get back to looking at classification algorithms, for non-linear problems.