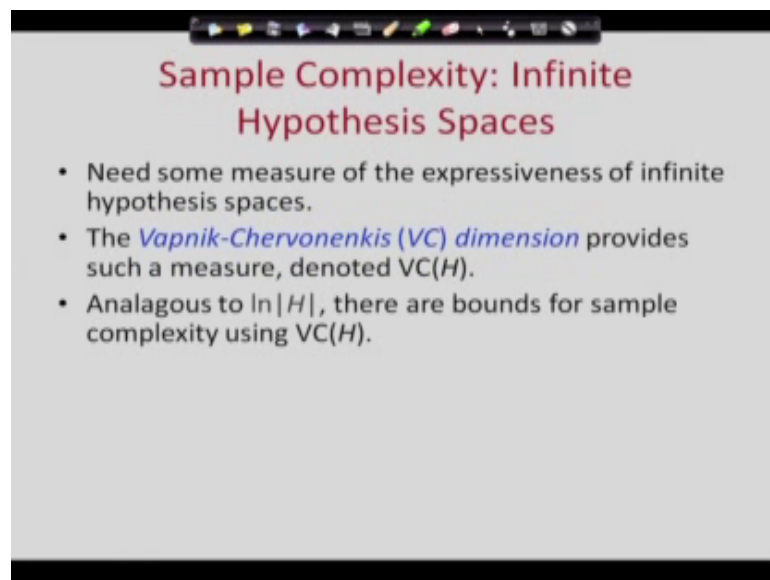


**Introduction to Machine Learning**  
**Prof. Sudeshna Sarkar**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Module - 7**  
**Lecture - 34**  
**VC Dimension**

Good morning. Today we will continue our lecture in Computational Learning Theory. And we will talk about situations where the hypothesis spaces infinite, what type of relations guaranties or theorems we have in such cases.

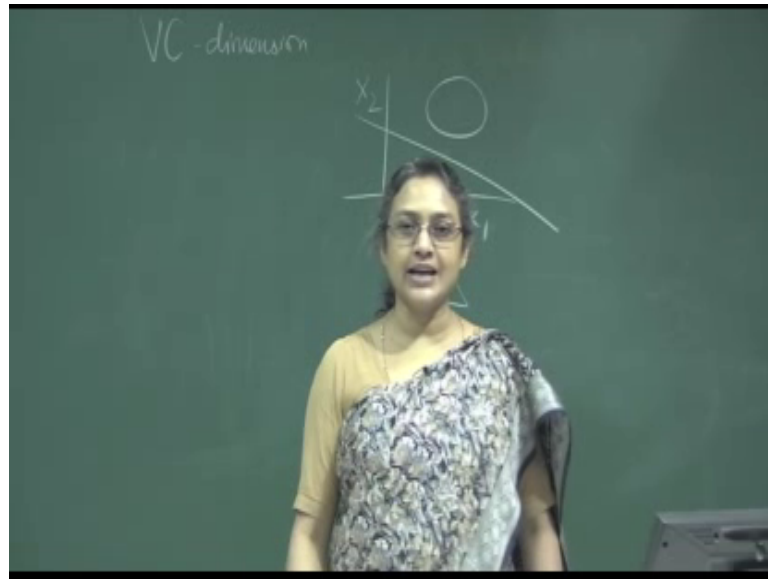
(Refer Slide Time: 00:36)



So, we will see that we want to find out, if the hypothesis spaces infinite how we can find out the required number of examples. In the last class, we looked at finite hypothesis space and we showed that the number of examples required to ensure pack learnability depends on the log of the size of the hypothesis space.

But, if the hypothesis spaces infinite in size then this will not be finite value. In that case we will not be able to come up with the required hypothesis. When can hypothesis space is infinite? Suppose your hypothesis is and you have  $X_1$  and  $X_2$  are two attributes and your hypothesis in straight line, right.

(Refer Slide Time: 01:23)



If  $X_1$  and  $X_2$  are real valued attributes, then the number of linear functions can be infinite or your hypothesis can be a circle or it could be a triangle in this particular space. In all these cases the hypothesis space is infinite and the bounds that we looked at in the last class will not apply. So, we will have to look for other measures by which we can measure the complexity of the hypothesis space and in terms of which we can specify the sample complexity required for learning.

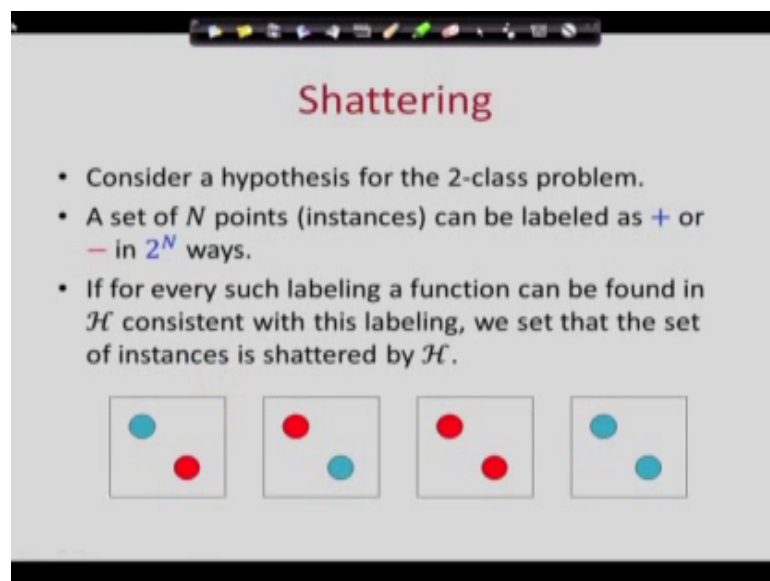
(Refer Slide Time: 03:01)

A presentation slide with a title bar at the top. The title is 'Sample Complexity: Infinite Hypothesis Spaces' in red. Below the title, there is a list of three bullet points in black text.

- Need some measure of the expressiveness of infinite hypothesis spaces.
- The *Vapnik-Chervonenkis (VC) dimension* provides such a measure, denoted  $VC(H)$ .
- Analogous to  $\ln |H|$ , there are bounds for sample complexity using  $VC(H)$ .

For this we will discuss a concept called the VC dimension which stands for Vapnik-Chernvonenkis dimension after the names of these two people. So, the VC dimension provides a measure of the complexity of the hypothesis space which is denoted by  $VC(H)$ . VC dimension can be defined for a finite hypothesis space also, but for infinite hypothesis space it is very useful because the size of the hypothesis space cannot be used. So, VC dimension, in terms of VC dimension we can find bounds for sample complexity as we will see.

(Refer Slide Time: 03:01)



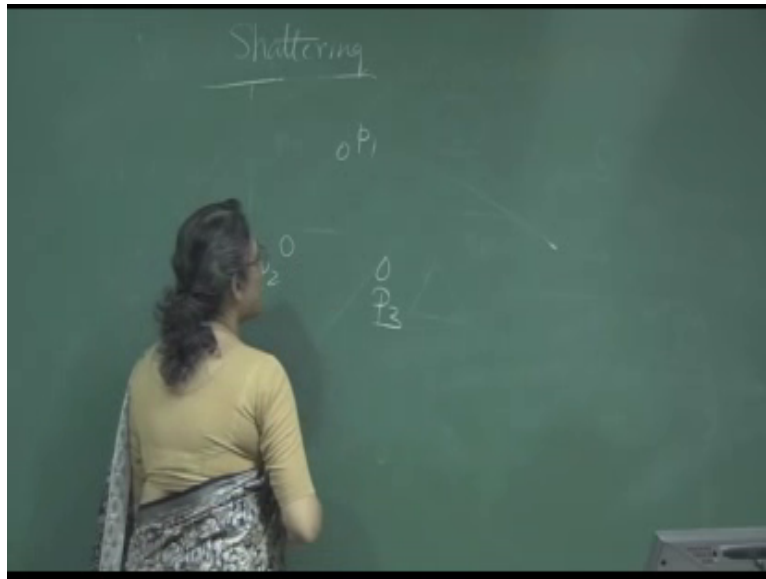
**Shattering**

- Consider a hypothesis for the 2-class problem.
- A set of  $N$  points (instances) can be labeled as  $+$  or  $-$  in  $2^N$  ways.
- If for every such labeling a function can be found in  $\mathcal{H}$  consistent with this labeling, we set that the set of instances is shattered by  $\mathcal{H}$ .

The slide contains four square boxes, each representing a set of two points. In each box, one point is cyan and the other is red, representing two different labelings. The boxes show the points in different relative positions (top-left, top-right, bottom-left, bottom-right) to illustrate that for any possible labeling of the two points, there exists a hypothesis function in  $\mathcal{H}$  that can separate them.

So, let us consider a hypothesis. So, first let me introduce the concept of Shattering. Suppose we have a hypothesis space capital  $H$  and there is two class problem and we have a set of  $n$  points. Suppose we have two points.

(Refer Slide Time: 03:27)

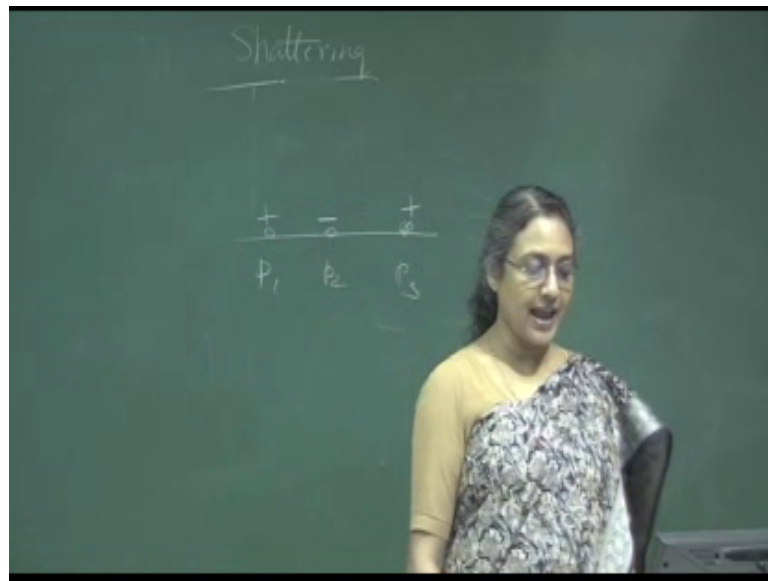


We have two points  $P_1$  and  $P_2$ . Now, if there are two classes plus and minus how many ways can we label these two points. They can be labeled in  $2$  to the power  $N$  ways for example, here we have two points we can label this as blue this as red, or red blue, red red, blue blue there are 4 ways in which we can label the set. If there are  $n$  points there will be  $2$  to the power  $N$  ways by which we can label the set.

If for every such labeling, there is a function in the hypothesis space which is consistent with that labeling then we say this set of points is shattered by the hypothesis space. If your hypothesis space is a linear function, if both of them are plus you can have a function so that this side is positive, this side is negative, which will be consistent with this labeling. If you have  $P_1$  as plus and  $P_2$  as minus, you can have this as a decision surface where this side is positive, this side is negative. If  $P_1$  is negative,  $P_2$  is positive you can have a decision surface where this side is positive, this side is negative. Where both are minus you can have a decision surface where this side is negative, this side is positive.

So, for all possible labeling of these two points you can find a hypothesis from the hypothesis space which is consistent with the labeling. Now, let us take 3 points  $P_1$ ,  $P_2$ ,  $P_3$ . How many possible ways can you partition this space? You can partition this space into two cube that is 8 ways, and we can show that for every possible such labeling we can find a line which separates those points.

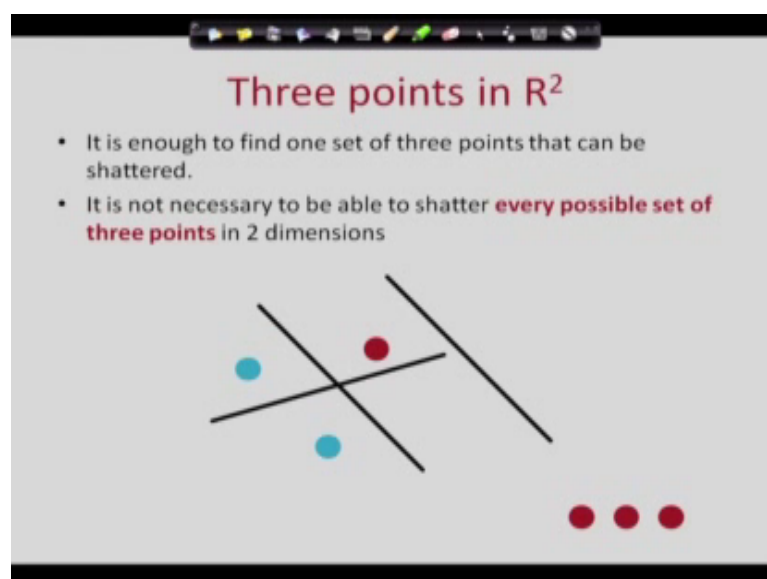
(Refer Slide Time: 05:41)



But if we have these 3 points which happen to be in a straight line, there is a labeling plus minus plus for which we cannot separate the plus and the minus points using a straight line. So, this particular set of 3 points cannot be shattered by the hypothesis space which comprises of straight lines.

So, the definition of shattering says that given a set of  $n$  points, if for every labeling of those points there is a function in the hypothesis space which is consistent with the labeling then we say that set of those points is shattered by the hypothesis space.

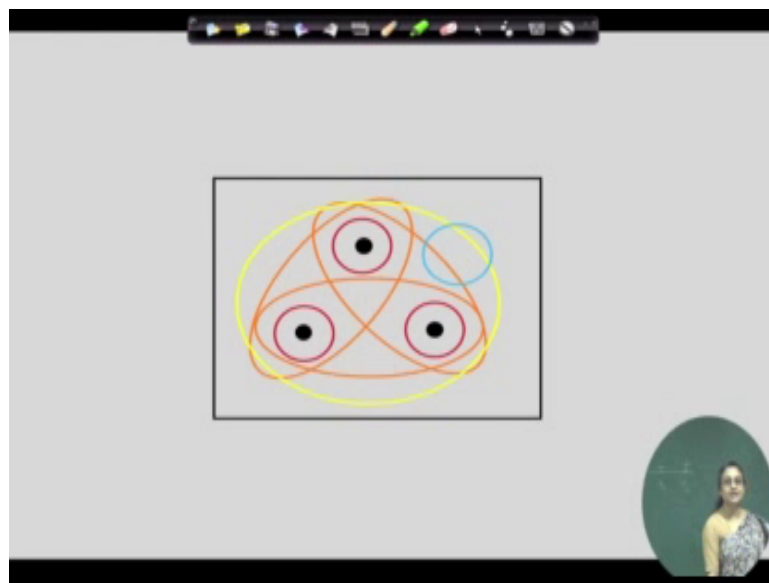
(Refer Slide Time: 06:34)



And this example shows 3 point in the real space, in the two-dimensional space and we can show that for these 3 points for all 8 possible labeling we can find a separator. But for these 3 points, there is a labeling for which there is no separator. So, it may be that there are some sets of points for which we cannot find a labeling, but what we really want is that does there exist at least one set of three points which can be shattered.

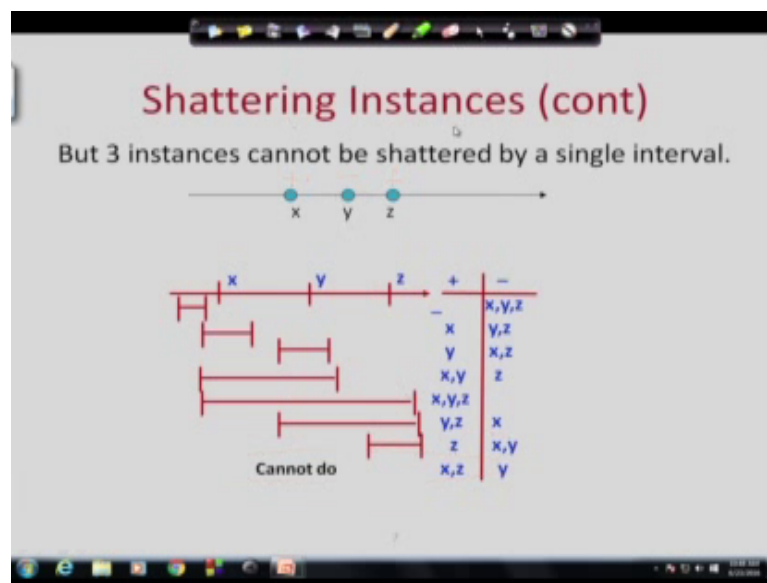
So, it is enough to find any one set of three points that can be shattered by the hypothesis space then, we say that this hypothesis space shatters 3 points in two-dimensions.

(Refer Slide Time: 07:26)



So, for this 3 points you know these blocks show the different hypothesis which are ovals corresponding to all possible labeling of these 3 points.

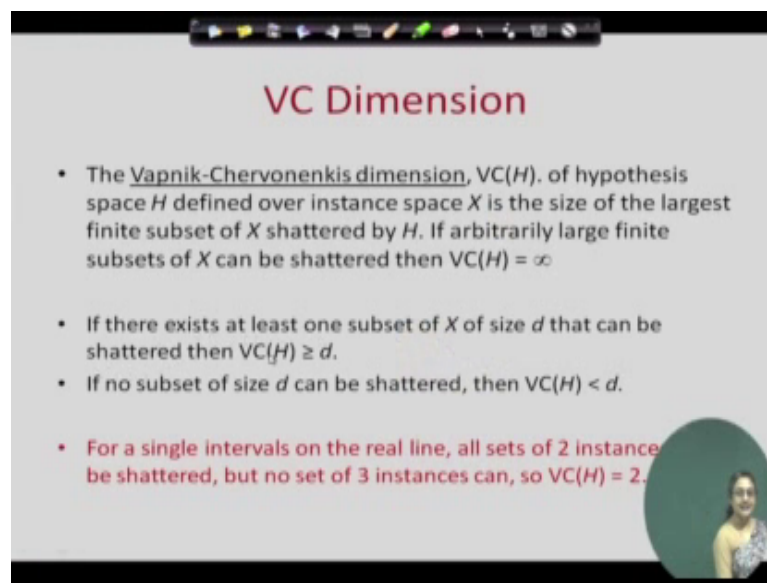
(Refer Slide Time: 07:46)



This is another example, suppose we have two points  $X$  and  $Y$ , and our hypothesis space comprises single real valued, single intervals on the line and we can see that, if there  $X$  and  $Y$  are 2 points,  $X$  and  $Y$  can be both positive; if they are both positive we can use, let us just if both are positive we can use these interval. If both are negative we can use this interval. If this is positive, this is negative we can use this interval. If his is positive, this is negative we can use this interval. So, there are 4 possible labeling and corresponding to this we can find the interval, so these two points can be shattered.

If you have 3 points on the real line, on the other hand you cannot shatter them. For example, let us look at this example here we have 3 points and corresponding to these 3 points we have 8 possible labeling. First case, all of them are negative  $xyz$  is negative, none is positive,  $X$  is positive,  $Y Z$  is negative, like this we have 8 possible situation. And for some of this situations we can find an interval which is consistent with this labeling, but we cannot find a consistent function in the case where  $X$  and  $Z$  are positive  $Y$  is negative in that case we cannot find any consistent function. So, these 3 instances on the real line cannot be shattered by a single interval. So, in this case we say it is not shattered.

(Refer Slide Time: 09:57)



### VC Dimension

- The Vapnik-Chervonenkis dimension,  $VC(H)$ , of hypothesis space  $H$  defined over instance space  $X$  is the size of the largest finite subset of  $X$  shattered by  $H$ . If arbitrarily large finite subsets of  $X$  can be shattered then  $VC(H) = \infty$
- If there exists at least one subset of  $X$  of size  $d$  that can be shattered then  $VC(H) \geq d$ .
- If no subset of size  $d$  can be shattered, then  $VC(H) < d$ .
- For a single intervals on the real line, all sets of 2 instances can be shattered, but no set of 3 instances can, so  $VC(H) = 2$ .

So, based on this discussion, we now come up with the definition of the VC dimension. The VC dimension of a hypothesis space  $H$  is defined over instance space  $X$ , it is the size of the largest finite subset of  $X$  which can be shattered by  $H$ . So, for example, we saw that when we have points on the real line there is a set of 2 points which can be shattered by a hypothesis space which consists of single interval on the real line. But, if you take any 3 points, any 3 points which lie on the line there is a labeling for those points for which you cannot find an interval consistent with that labeling.

So, for the hypothesis space which comprises of a single interval on the real line and instance space which consists of points on the real line, the VC dimension is 2, because there is a set of 2 points that can be shattered and no set of 3 points can be shattered. So, in general if there exists at least one subset of instance space  $X$  of size  $d$  that can be shattered then the VC dimension is greater than equal to  $d$ . If no subset of  $d$  can be shattered then VC dimension is less than  $d$  and as I have said for single interval on the real line all set of two instances can be shattered, but no set of three instances can, so VC dimension is 2.

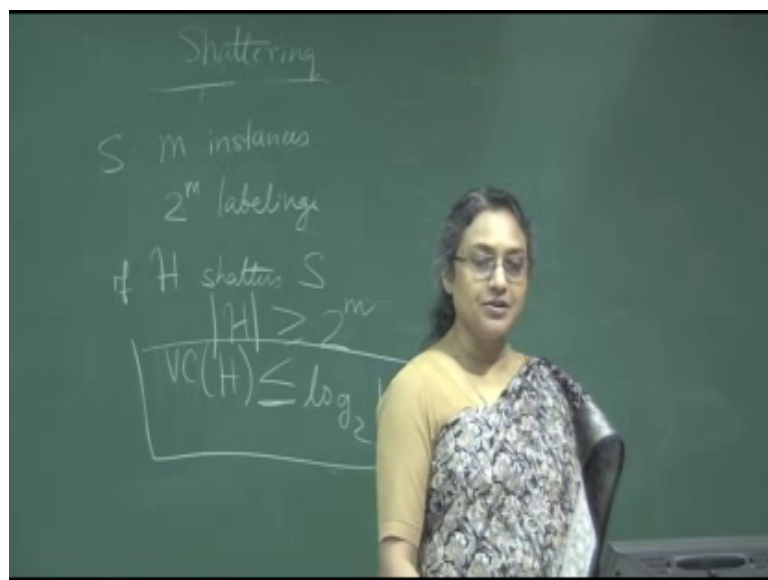
Now, if the hypothesis space is unbiased for example, if we have the set of Boolean variables,  $n$  Boolean variables and we can look at any Boolean formula then that hypothesis space shatters the entire instance space. So, there are  $2^N$  possible instances; for any partition of those instances we can come up with the Boolean



formula. So, the Boolean formula, unrestricted Boolean formula hypothesis space shatters the entire instances space and it is unbiased. But when we go for some restrictions that is when we go for only conjunctive formulas then, the hypothesis space becomes constraint and for all possible values of that attributes of instances we will not be able to find the conjunctive Boolean formula that is consistent with the labeling.

So, the larger is the (Refer Time: 13:11) subset that can be shattered, the more expressive the hypothesis space is. So, the Unrestricted Boolean formula is very expressive, Conjunctive Boolean formula is much less expressive. The VC dimension of the set of lines in two-dimension is 3, let us see how.

(Refer Slide Time: 13:38)



We have already told that if I have these 3 points there are 8 ways of labeling these points, and for each such labeling for example, this is plus, these two are minus we can find a separator or these two are plus this is minus we can find a separator, all three are positive we can find a separator and so on, for all 8 possibilities we can find a separator.

But, we can show that if we take any 4 points then in all such cases there will exist labeling for which we cannot shatter them. For example, these two are plus, these two are negative there will be no linear function existing, but then you can say maybe we can have a different arrangement, but whatever arrangement you do it can be shown that there will be a labeling for those 4 points cannot be shattered. No 4 points can be shattered by

a hypothesis space as a straight line. So, the VC dimension of the set of oriented lines in two d is 3.

And also we can look at the general relation, if there are  $m$  instances there are  $2^m$  to the power  $m$  labeling and if hypothesis space can shatter this instance space, if  $H$  shatters  $S$  comprises  $m$  instances, if  $H$  shatters  $S$  - for each of these labeling there will be one element hypothesis space, so the size of the hypothesis space will be greater than equal to  $2^m$ ; which means that, the VC dimension of  $H$  will always be less than  $\log$  of the size of the hypothesis space.

Earlier we found a bound on sample complexity based on the  $\log$  of the  $\log$  or natural logarithm of the hypothesis space, we shown now that the VC dimension of the hypothesis space is less than equal to  $\log H$ . This is one more example of VC dimension on the slide.

(Refer Slide Time: 16:21)

**VC Dimension Example**

Consider axis-parallel rectangles in the real-plane, i.e. conjunctions of intervals on two real-valued features. Some 4 instances can be shattered.


Some 4 instances cannot be shattered:

The slide contains two rows of diagrams. The top row shows 16 different axis-parallel rectangles, each enclosing a different subset of 4 points (represented by dots). The bottom row shows a single diagram of a rectangle enclosing 4 points, with a small inset showing a different configuration of 4 points that cannot be shattered.

Suppose our hypothesis space consists of rectangles, which enclose the positive points and these rectangles are axis parallel rectangles that is the two sides are parallel to the  $X_1$  and  $X_2$  axis, respectively. If we take a set of 4 points, there are 16 possible labeling possible and we show that for each of these 16 possible labeling or 16 possible partitions there is a rectangle which encloses the positive points. So, this set of 4 points can be shattered by the hypothesis space which comprises of axis parallel rectangles.

But there is another set of 4 instances that cannot be shattered; that does not matter. These set of 4 instances, this can be shattered. So, VC dimension of this hypothesis space is greater than equal to 4. It can be shown that there is no set of 5 points which can be shattered by this hypothesis space and therefore, we can say VC dimension of this hypothesis space is equal to 4.

(Refer Slide Time: 17:39)



### VC Dimension Example (cont)

- No five instances can be shattered since there can be at most 4 distinct extreme points (min and max on each of the 2 dimensions) and these 4 cannot be included without including any possible 5<sup>th</sup> point.
- Therefore  $VC(H) = 4$
- Generalizes to axis-parallel hyper-rectangles (conjunction of intervals in  $n$  dimensions):  $VC(H)=2n$ .

11

So VC dimension is equal to 4.

(Refer Slide Time: 17:44)

### Upper Bound on Sample Complexity with VC

- Using VC dimension as a measure of expressiveness, the following number of examples have been shown to be sufficient for PAC Learning (Blumer *et al.*, 1989).

$$\frac{1}{\epsilon} \left( 4 \log_2 \left( \frac{2}{\delta} \right) + 8VC(H) \log_2 \left( \frac{13}{\epsilon} \right) \right)$$

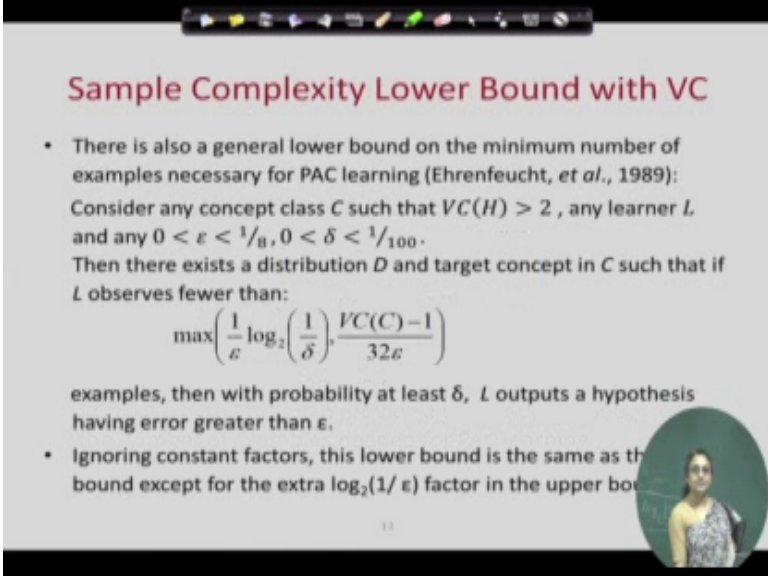
- Compared to the previous result using  $\ln |H|$ , this bound has some extra constants and an extra  $\log_2(1/\epsilon)$  factor. Since  $VC(H) \leq \log_2 |H|$ , this can provide a tighter upper bound on the number of examples needed for PAC learning.

12

Now, why we are looking at VC dimension, it has been shown by Blumer in 1989 that using VC dimension it has been found that we can find a bound of the sample complexity. That is, if we look at  $m$  greater than equal to this many examples that is,  $\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{8}{\epsilon} VC(H) \log \frac{1}{\epsilon}$ . If we look at these many examples and we output a consistent hypothesis that consistent, that hypothesis will be probably approximately correct.

Earlier we had shown a result which involved  $\log$  of  $H$ , this bound includes VC of  $H$  and we already saw that VC of  $H$  is less than equal to  $\log$  of  $H$ . Of course, there is some other factors higher constant etcetera, but we know also have a  $\log$  of  $\frac{1}{\epsilon}$  into  $\frac{1}{\epsilon}$  but this provides tighter upper bound on the number of examples needed for PAC learning. And especially for infinite hypothesis space, the earlier formula does not apply here we can apply this type formula.

(Refer Slide Time: 19:19)



**Sample Complexity Lower Bound with VC**

- There is also a general lower bound on the minimum number of examples necessary for PAC learning (Ehrenfeucht, *et al.*, 1989): Consider any concept class  $C$  such that  $VC(H) > 2$ , any learner  $L$  and any  $0 < \epsilon < 1/8$ ,  $0 < \delta < 1/100$ . Then there exists a distribution  $D$  and target concept in  $C$  such that if  $L$  observes fewer than:
 
$$\max\left(\frac{1}{\epsilon} \log_2\left(\frac{1}{\delta}\right), \frac{VC(C)-1}{32\epsilon}\right)$$
 examples, then with probability at least  $\delta$ ,  $L$  outputs a hypothesis having error greater than  $\epsilon$ .
- Ignoring constant factors, this lower bound is the same as the upper bound except for the extra  $\log_2(1/\epsilon)$  factor in the upper bound.

We will also state another result which gives a general lower bound on the minimum number of examples necessary for PAC learning this was given by Ehrenfeucht in 1989. And according to that equation that theorem that proof in that paper, if you consider any concept class  $C$  whose VC dimension is greater than 2 and you have a learner  $L$  and  $\epsilon$  is between 0 and one-eighth,  $\delta$  is between 0 and 1 by 100 then there is a distribution  $d$  and the target concept  $C$  for which, if you observe few more these of examples then you do not get an approximate hypothesis.

This shows that the type of bound that we bought by Blumer for saying the number of examples needed for PAC learning is very tight, because the lower bound shows that less than this number of examples will not give you approximately correct hypothesis and these two expressions are almost close according to certain factors.

With this brief introduction to Computational Learning Theory, we will end this topic and in the next class we will study a little bit about Ensemble Learning.

Thank you very much.