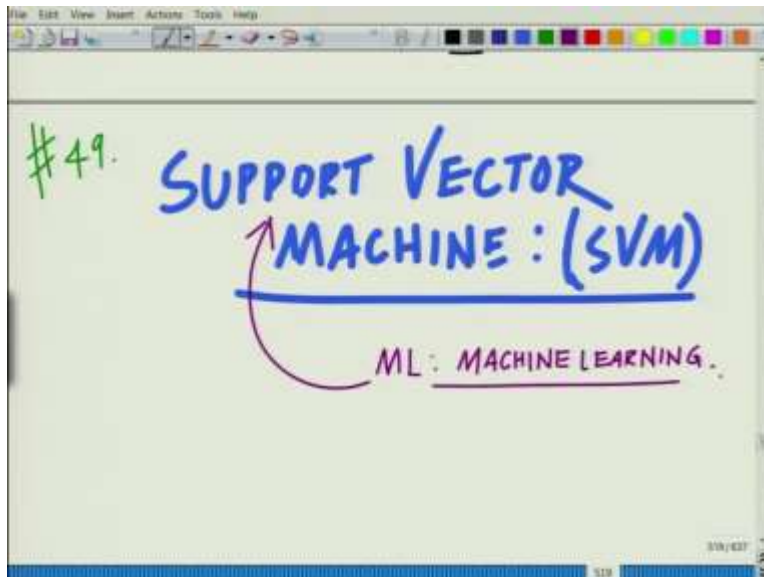


Applied Linear Algebra for Signal Processing, Data Analytics and Machine Learning
Professor Aaditya K Jagannatham
Department of Electrical Engineering
Indian Institute of Technology Kanpur
Lecture 49

Machine Learning Application: Support Vector Machines (SVM)

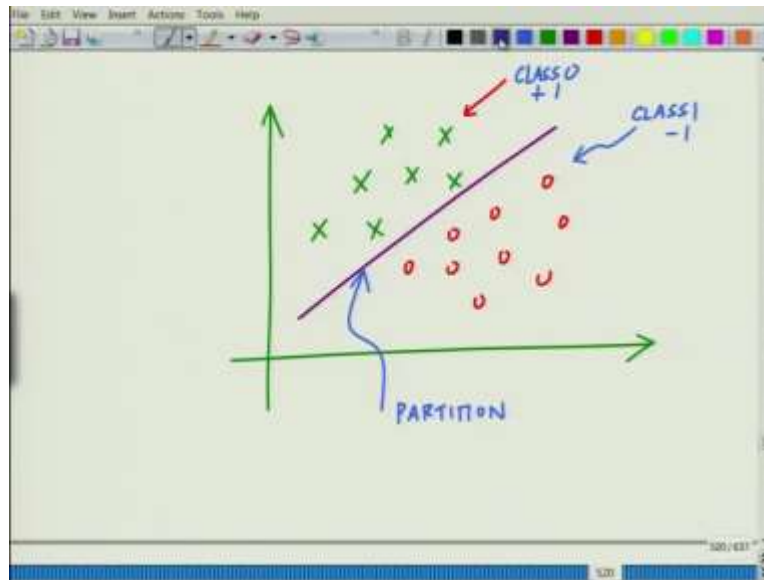
Hello, welcome to another module in this massive open online course. So in this module let us look at an interesting application of the principles of linear algebra in the context of machine learning and this is a very important application termed as SVM, it stands for Support Vector Machines which is in, which you can is a very important classifier or a concept in the context of machine learning. So let us understand this.

(Refer Slide Time: 00:43)



So what we want to explore today is this very important classifier termed as a support vector machine. Very important, so naturally this is a very important, so the abbreviation is SVM. This is a very important application in the context of ML that is as you are already familiar, that is machine learning. Now what is a support vector machine? What is the principle of support vector machine? Let us try to understand this.

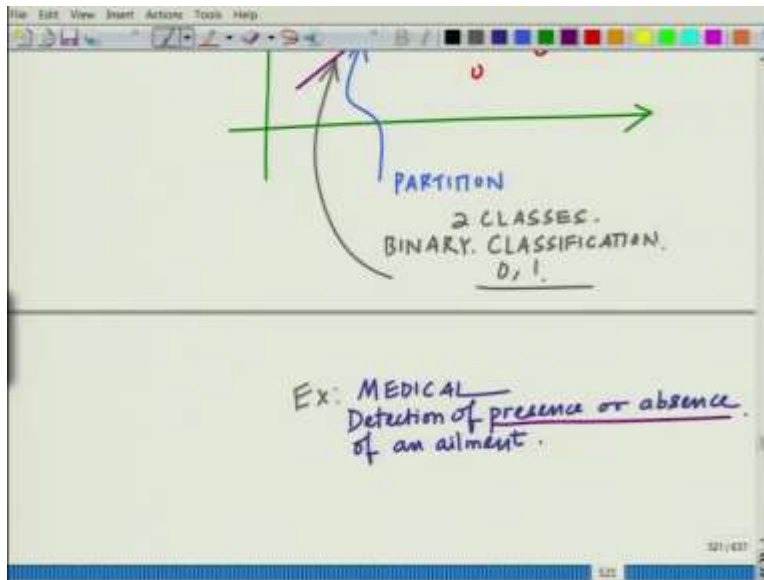
(Refer Slide Time: 01:39)



So support vector machine does the following job. It is a binary classifier, so you have... if you recall what is a classifier? You have two classes of data. You are partitioning it in to two classes so you have data which is to be partitioned in to, so this is your class 0 let us say and this is your class one. I can denote this by plus 1, by minus 1 we will look at more at this later, and so this is basically a partition, right?

So you are partitioning your data into two classes and this is also, or you are basically classifying, this is also basically a binary classification right? You are classifying your data in to two types class 0 and class 1. So there is a large amount of data that is available and you are performing classification in particular binary because you are classifying it in to two classes 0 and 1.

(Refer Slide Time: 02:59)



So this is essentially, you can also turn this as binary classification problem, this is 0 and 1. Binary because there are essentially two classes. So this is a binary classification problem. For instance example let us take this for instance. Where does this binary classification arise? For instance let us say you have a medical detection.

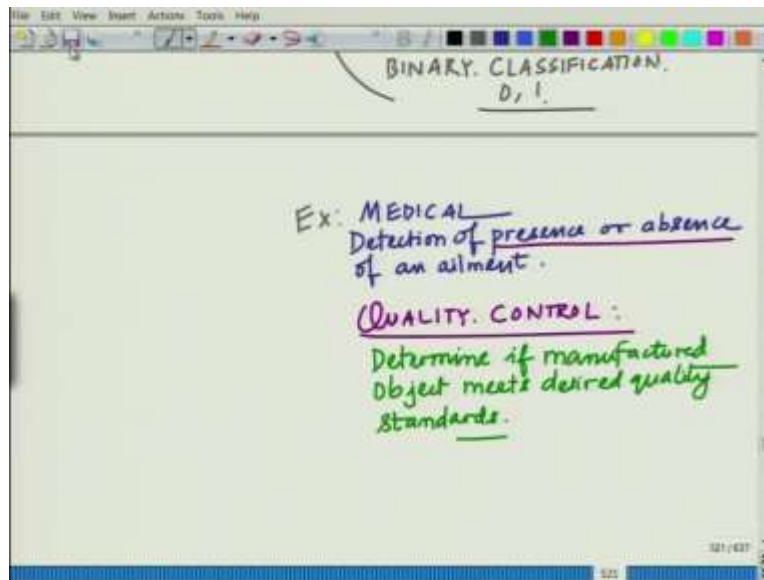
For instance let us take a look at a medical example in the medical profession. You have for instance the detection of presence of absence of a particular ailment. So you have a large amount of data let us say in terms of blood test or x-rays or CT scans or so on and from this data you want to classify it into two types. Essentially as either any ailment is present or an underlying disease is present or an ailment or essentially an underlying disease is absent.

So that is a very, very important problem, a classification problem in the context and problem in the context of medicine, and as you can see that is a very important problem because it has a lot of applications and the answer, the response to this problem will be yes or no. If the ailment is present or not. So it is binary, 0 or 1.

It is not, what I mean to say is you have to distinguish it from a continuous regression problem where the output value is continuous that is you can take any particular value on the set of real numbers or complex and so on.

In a binary classification problem the output has to belong to a particular set of discrete values. In particular for binary classification it has to be either 0 or 1. That is the output response can either be or, it cannot be for instance 1 point 5 or minus 2 point 3, right? You simply want to know if the ailment is present or absent.

(Refer Slide Time: 05:19)



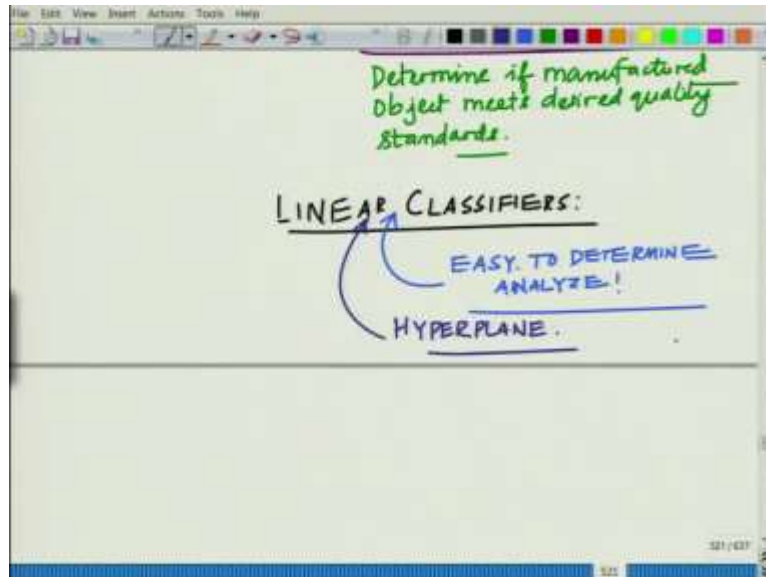
A similar another interesting example for instance in the context of quality control and this is very important in manufacturing again for instance another example in the context of quality control for instance if the manufactured object, right? If manufactured object to check to, or to determine manufactured object, needs the desired quality, to check if the manufactured objects needs the desired quality standards for instance, you might fabricate a new car or manufacture a new car.

How do you determine if this car for passes all the, this car is, meets the particular standard that has to be ensured by that company, who probably run a rod of tests, acquire a lot of data, put this data through an algorithm and the algorithm outputs a value 0 or 1. 0 means probably acceptable or 1 means it is not acceptable.

Now you can see the 0 and 1 itself, by themselves do not mean anything, these are basically what you assign. So 0 either, so 0 and 1 can either mean depending on what you assign can either mean that we decide it meets a certain set of criteria or does not. So one will automatically be, so if 0 denotes certain aspect 1 obviously means the opposite of that. The absence of that aspect.

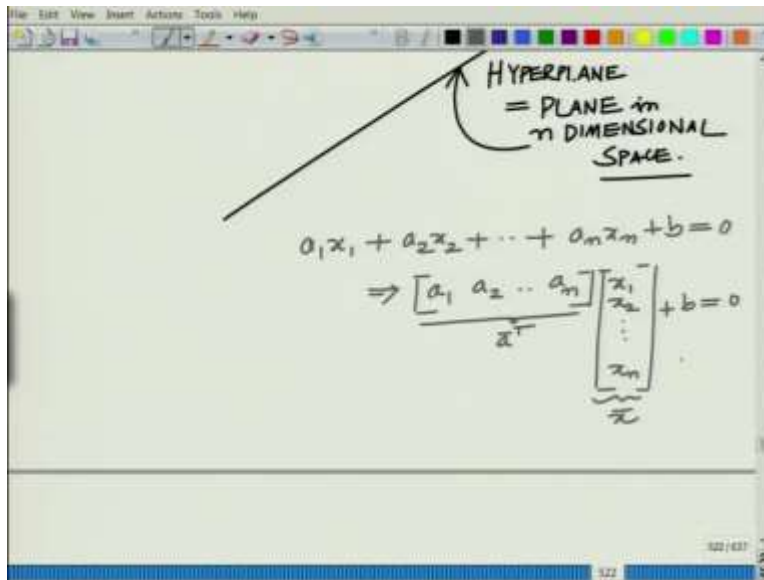
For instance if 0 is the presence of disease 1 would mean then the absence of disease and so on and so on.

(Refer Slide Time: 07:13)



The point is now there are many classifiers that can be designed. In particular we are going to be interested in the linear class of classifiers because these are interested in linear classifiers, because these are easy to determine and easy to analyze. So non-linear are typically much more interactive or easy to determine and more importantly also easy to analyze. So this is the linear classifier. And the linear classifier, the linear classification can be achieved by what is known as a hyper-plane what do you mean by a hyper-plane?

(Refer Slide Time: 08:16)



A hyper-plane is essentially a plane in n dimension that is if I look in n dimensional space, right? If I ask the question what is a general that is extension of this, generalization of the notion of a plane to n dimensional space that is a hyper-plane and the expression of the hyper plane is given as it is characterized by, if you have any n dimensional hyper-plane a 1 x 1 plus a 2 x 2 plus so on a n x n plus b equal to 0 which implies, now writing it using our principles of linear algebra, writing in terms of vectors.

Remember that is what linear algebra is about. Developing compact notation, representing it efficiently and analyzing is to I can write it as a 1, a 2, a n; x 1, x 2, x n plus b equal to 0 this you can call this a transpose. This is the vector x bar.

(Refer Slide Time: 10:16)

$$\Rightarrow \underline{\bar{a}}^T \underline{\bar{x}} + b = 0$$

$$\bar{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Weight Vector

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Variable Vector

So you can write this as a bar transpose. So this can be written as a bar transpose x bar plus b equal to 0. so the expression for the hyper-plane in n dimensions is given by a vector a bar. This is essentially a weight vector comprising of the weight a 1, a 2, a n, then the vector of variable x 1, x 2, x n. a bar transpose x bar, plus b equal to 0. So I can write this as you have a bar which is essentially your, weight vector. I can call this as my weight vector or co-efficient vector and I have x bar which is equal to x 1, x 2, x n. This is my variable vector, right? And therefore, and b is nothing but a constant.

(Refer Slide Time: 11:31)

$$\bar{a}^T \bar{x} + b = 0$$

Weight Vector $n \times 1$ a_n x_n Variable Vector $n \times 1$ constant

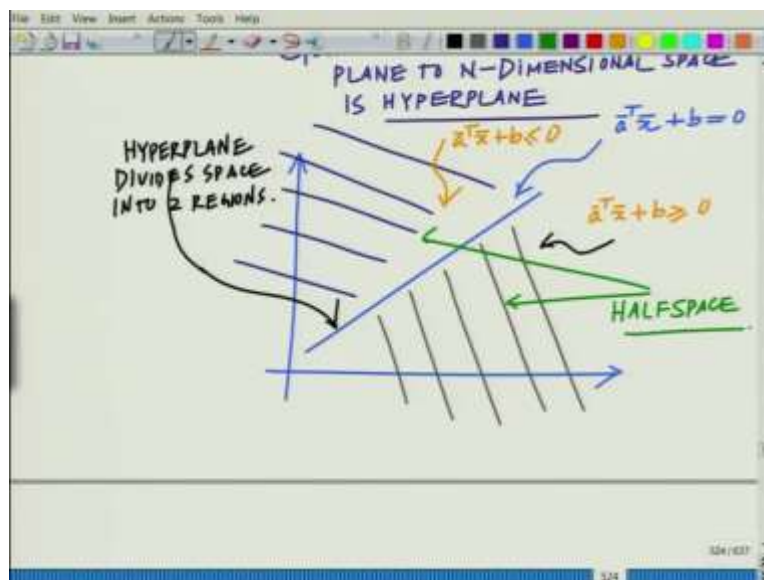
$n = 2 \Rightarrow$ LINE in 2D. Line in 2D.
 $2x_1 + 3x_2 + 8 = 0$

$n = 3 \Rightarrow$ PLANE in 3D. Plane in 3D.
 $4x_1 - 5x_2 - 3x_3 + 9 = 0$

So I can write this as a bar transpose x bar plus b equal to 0, this is a constant, this is your n cross n 1 variable vector, n cross 1, co-efficient vector or weight vector and so on. And for instance now if n is equal to 2 implies this becomes a line. For n equal to 2 this becomes a line in 2D space. In 2D we are well aware that so you have for instance let us take a simple example $2x_1 + 3x_2 + 8 = 0$. This is a line in 2D.

Now n equal to 3 implies this is our regular plane what we know as simply a plane in 3D, you might have learned this in your high school for instance. You have $4x_1 - 5x_2 - 3x_3 + 9 = 0$. so this is simply what you have, very much learned earlier this is a plane in the 3D, in the regular 3 dimensional space and so on and so forth. And the generalization of this to n dimensional space is a hyper-plane.

(Refer Slide Time: 13:15)



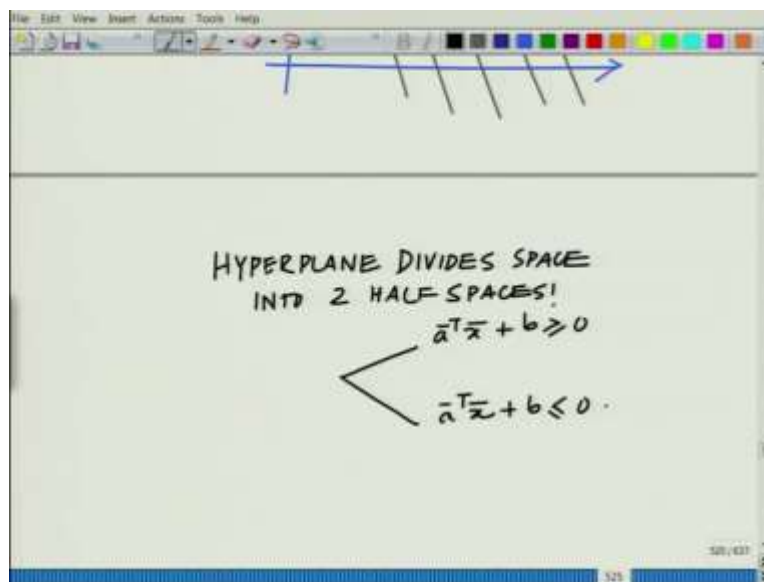
So the generalization of this. So the generalization of plane or concept of plane to n dimensional space is nothing but your hyper-plane. This is basically the generalization of the concept plane to n dimensional space that is essentially to a hyper-plane and what you will see very interestingly. It is not very difficult to see is that whenever you have this hyper-plane that is arbitrary n dimensional space, you can visualize this for now in 3D, in 2D, a bar transpose x bar plus b equal to 0.

Now it divides this into two regions. Each hyper-plane divides the space, so this hyper-plane, so not just this hyper-plane any hyper-plane. So hyper-plane divides space into two regions. So this

corresponds to the region let us say $\bar{x}^T a + b \geq 0$ then immediately the opposite one will be $\bar{x}^T a + b \leq 0$ and this is of course your hyper-plane that is $\bar{x}^T a + b = 0$.

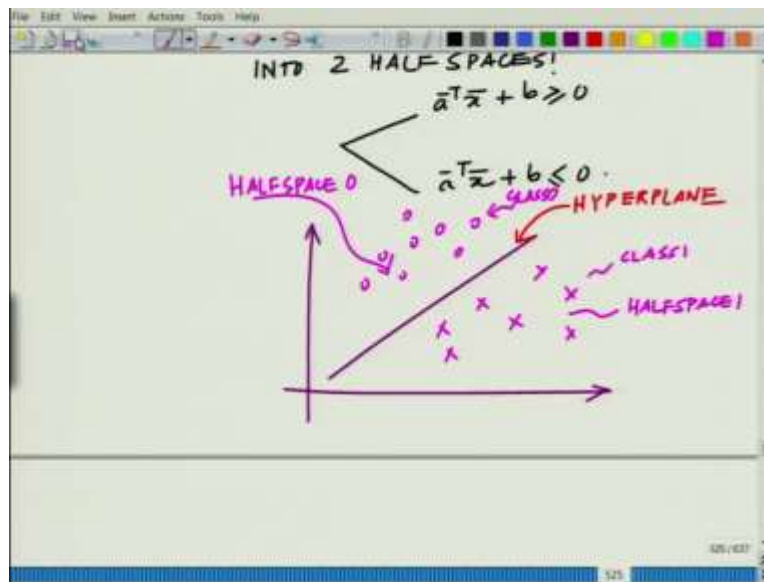
So these two, so the hyper-plane divides the n dimensional space into two regions. Each of these regions is known as a half space. So it is like taking a knife and slicing it, the object into two parts. One is which lies to one side and other which lies to another side. One is basically characterized by $\bar{x}^T a + b \geq 0$, the other one is $\bar{x}^T a + b \leq 0$. Each of these regions I known as a half-space. So these two regions these are known as basically, these regions, each is known as a half-space. So hyper-plane divides the nD space in two two half-spaces.

(Refer Slide Time: 16:35)



So hyper-plane, so we have the hyper-plane divides the space. What is the first half space? That is your $\bar{x}^T a + b \geq 0$, and the other one is $\bar{x}^T a + b \leq 0$, and now you can see I can use this for classification, I can use this hyper-plane now to separate my data, such that the data belonging to class 0 lies in one half-space, data belonging to class 1 lies in the other half-space. That is the idea of the support vector machine.

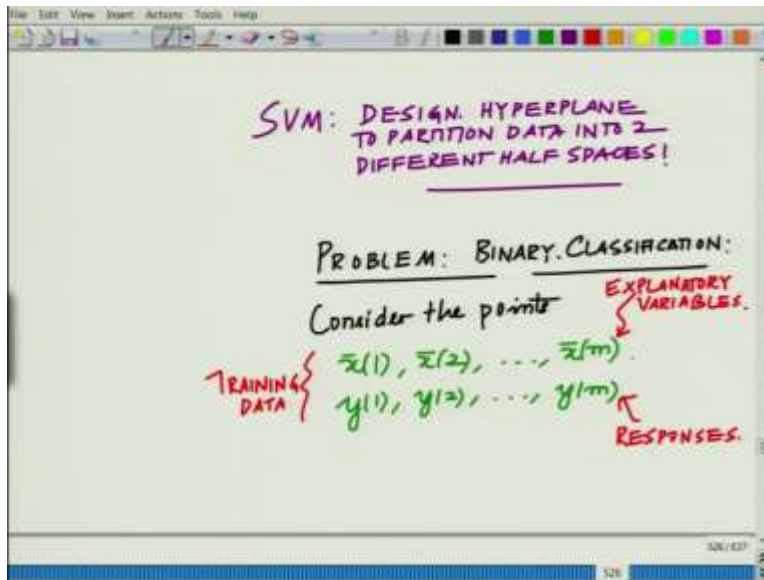
(Refer Slide Time: 17:51)



So what is the idea of a support vector machine? The idea of the support vector machine is the following thing that is if you look at it you have the hyper-plane that is how do you find the hyper-plane such that this is your data. So each of these lies in its own half-space. So we design a hyper-plane such that it partition the data into two different half-spaces. That is essentially the idea behind a support vector machine.

Support vector machine is to design essentially a hyper-plane that optimally partitions and we are going to look more at that. How to optimally partition the data into two different half-spaces. That is essentially the idea behind a support vector machine. Support vector machine is to design essentially a hyper-plane that optimally partitions and we are going to look more at that. How to optimally partition the data into two different half-spaces?

(Refer Slide Time: 19:07)



So hyper-plane, so what is SVM doing, support vector machine design hyper-plane to partition data into two different half-spaces. That is essentially what the hyper-plane is doing. The half-spaces corresponding to $\bar{x}^T \bar{x} + b \leq 0$, $\bar{x}^T \bar{x} + b \geq 0$.

Now let us formulate the problem for this. what is the specific problem? How do we do this? Problem which means problem for binary classification. Let us say we have considered the points, we consider the points we have m points and this is our training set and we have the corresponding classes for these m points, sorry! This is not y bar 1 but these are just 0's and 1's so these will be the scalar quantities.

So each of these are your, this is your training data, again you can think of this as the response. So similar to regression problem we can think of this as the responses, output variables, etc. these are essentially or explanatory data, explanatory variables, regression variables so on or regressors.

(Refer Slide Time: 21:46)

DIFFERENT HALF SPACES?

PROBLEM: BINARY CLASSIFICATION:

Consider the points

EXPLANATORY VARIABLES.

TRAINING DATA { $x(1), x(2), \dots, x(m)$
 $y(1), y(2), \dots, y(m)$

RESPONSES.

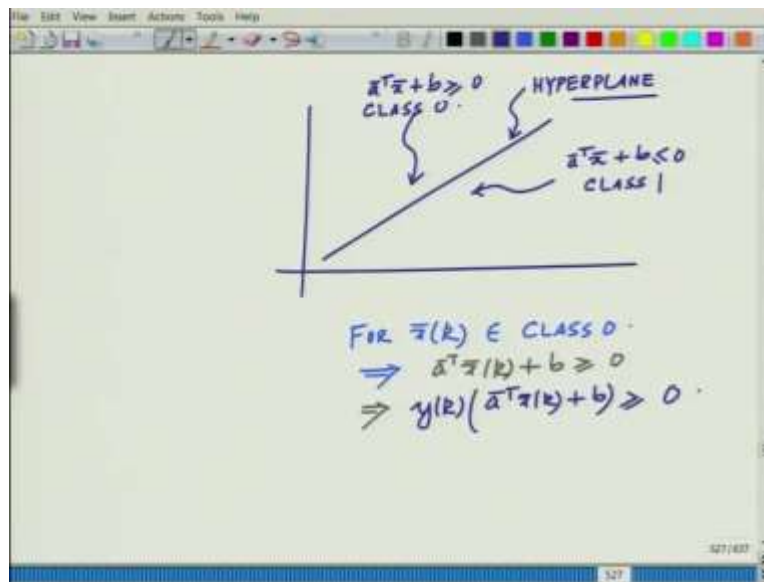
EACH $y(i) \in \{-1, +1\}$

class 0: $y(i) = +1$
class 1: $y(i) = -1$

These are essentially your explanatory variables and each y_i , you can see right? Each y_i belongs to the set either, we can call this as not exactly 0 or 1 we will call this as minus 1 or plus 1 that is we denote the classes as let us say class 0, we are going to see, this is going to have a very big use. Class 0 y_i equal to minus 1 plus 1 y_i equal to, or let us say y_i equal to plus 1 for class 0.

It does not really matter as long as you are consistent with the notation. So class 0 we are saying y_i equal to plus 1 that is the response, corresponding to class 0, if the class is 0 response has to be plus 1. If the class is 1 the response has to be responses minus 1.

(Refer Slide Time: 22:48)



Now let us take a look at it. Therefore now if we look at this, right? Let us say we have our hyper-plane, this is our hyper-plane and this is our $\bar{a}^T \bar{x} + b$, this is us say $\bar{a}^T \bar{x} + b \geq 0$ and this is let us say $\bar{a}^T \bar{x} + b \leq 0$ so this is essentially your class.

Let us say this is your class 1 and this is your, let us say your class 0, now what you can say is that for all points that is $\bar{x}(k)$, for $\bar{x}(k)$ belonging to class 0 we must have $\bar{a}^T \bar{x}(k) + b \geq 0$, which also implies. I can now multiply this by $y(k)$. $y(k)$ times $\bar{a}^T \bar{x}(k) + b \geq 0$.

(Refer Slide Time: 24:27)

$\Rightarrow y(k)(\bar{a}^T \bar{x}(k) + b) \geq 0$
CLASS 0
CLASS LABEL $y(k) = +1$

FOR $\bar{x}(k) \in \text{CLASS 1}$:
 $\bar{a}^T \bar{x}(k) + b \leq 0$
 $\Rightarrow y(k)(\bar{a}^T \bar{x}(k) + b) \geq 0$

So this is for all points of class 0. So $y(k)$ you can also call this as the class label or class response or whatever it is. Now similarly for all $\bar{x}(k)$ element of class 1 we have $\bar{a}^T \bar{x}(k) + b \leq 0$. now again once you bring $y(k)$ into the picture, remember here $y(k)$ equal to 1 for class, $y(k)$ equal to plus 1 for class 0 here $y(k)$ equal to minus 1.

So once again this means since you are multiplying by a negative quantity you will have $y(k)$ into $\bar{a}^T \bar{x}(k) + b$ greater than equal to 0 why? Inequality reverses, why? And this is important.

(Refer Slide Time: 25:49)

FOR $x(k) \in \text{CLASS 1}$:

$$a^T x(k) + b \leq 0$$
$$\Rightarrow y(k)(a^T x(k) + b) \geq 0$$

BECAUSE $y(k) = -1$

INEQUALITY REVERSES!

Look here, inequality reverses, why? Because y_k equals minus 1 so that is the thing, so therefore the same equation y_k into $a^T x(k) + b$, greater than equal to 0 can be used to represent both the classes, right? Because y_k equal to plus 1 for class 0, y_k equal to minus 1 for class one and that essentially captures both the half spaces.

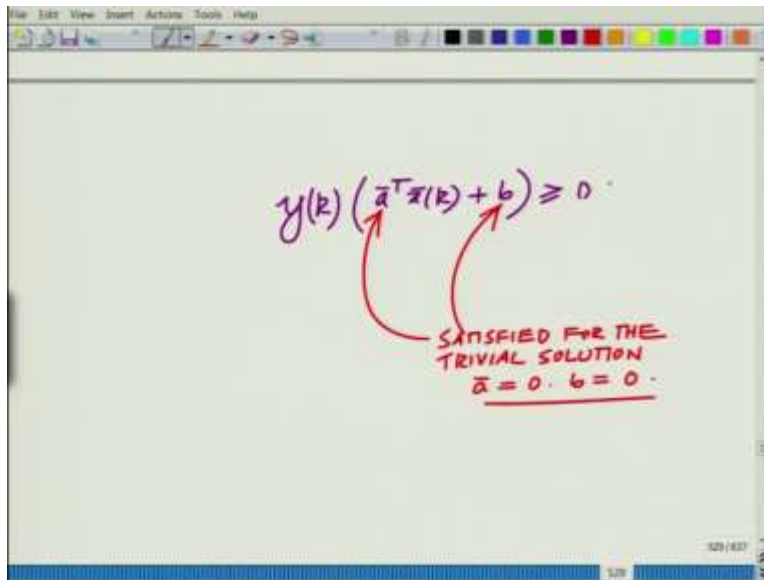
(Refer Slide Time: 26:48)

FOR BOTH CLASSES
COMBINED CONSTRAINT.

$$y(k)(a^T x(k) + b) \geq 0$$

So for both classes the combined equation, both classes combined constraint is y_k into $a^T x(k) + b$ greater than or equal to 0, right? Where we have seen y_k equal to minus 1 for class 0 or y_k equal to plus 1 for class 0 or y_k equal to minus 1 for class 1.

(Refer Slide Time: 27:32)

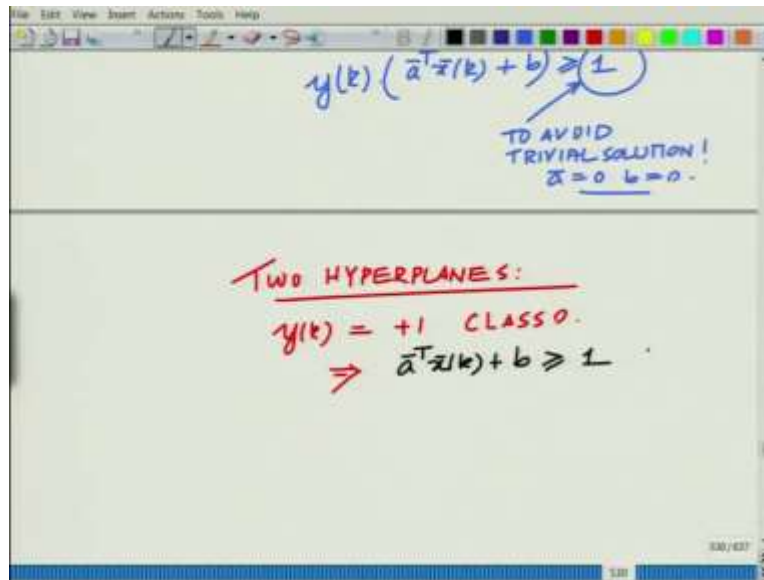


The image shows a whiteboard with a handwritten equation in red ink: $y(k) (\bar{a}^T x(k) + b) \geq 0$. Two red arrows point from the text below to the \bar{a} and b terms in the equation. The text below reads: "SATISFIED FOR THE TRIVIAL SOLUTION $\bar{a} = 0, b = 0$ ". The whiteboard has a standard toolbar at the top and a status bar at the bottom.

Now the point is, now if you look at even this problem, there is a problem, what is the problem here? The problem is the following, please look at $y(k)$ is, $\bar{a}^T x(k) + b$ greater than equal to 0. Now look at this, this is satisfied for the trivial solution $\bar{a} = 0, b = 0$ to 0.

Therefore we cannot have this. we will have to twig this a little bit, right? Because otherwise if you simply say greater than equal to 0 I can simply say $\bar{a} = 0, b = 0$ then this is always equal to 0 so it trivial satisfies this.

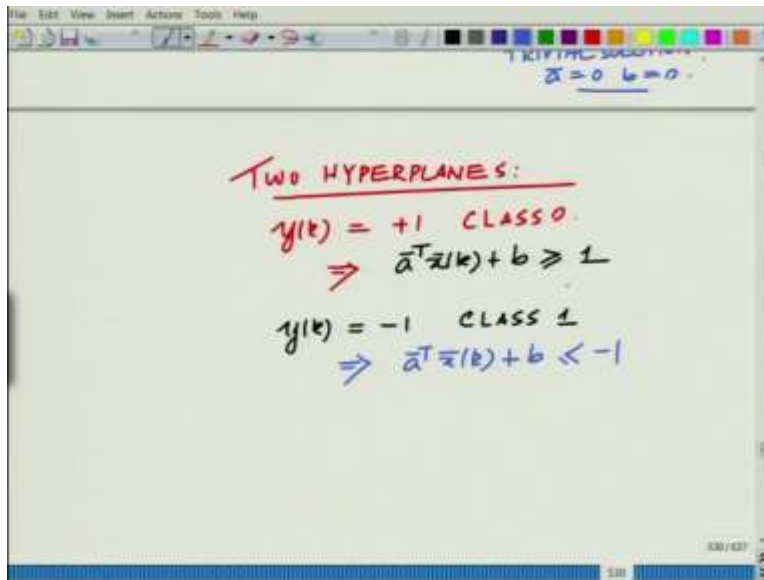
(Refer Slide Time: 28:48)



So to avoid that we here just have to tweak it a little bit and what we are going to say is that $y(k) (\bar{a}^T \bar{x}(k) + b) \geq 1$. So we will introduce this constant 1 rather than 0. This is to avoid the trivial solution. What is the trivial solution?

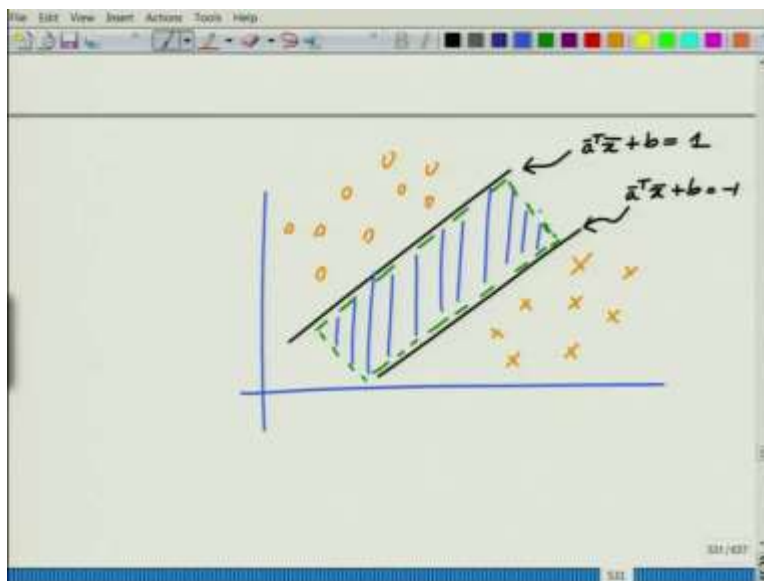
$\bar{a} = 0$ $b = 0$. So to avoid the trivial solution we will tweak this a little bit so now we have the two hyper-planes. What are the two hyper-planes? We have the two hyper-planes, so for $y(k) = +1$, that is class 0, this implies, what does this imply? $\bar{a}^T \bar{x}(k) + b \geq 1$.

(Refer Slide Time: 30:03)



Now for y_k equal to minus 1 class 1, this implies $\vec{a}^T \vec{x}(k) + b < -1$.

(Refer Slide Time: 30:22)



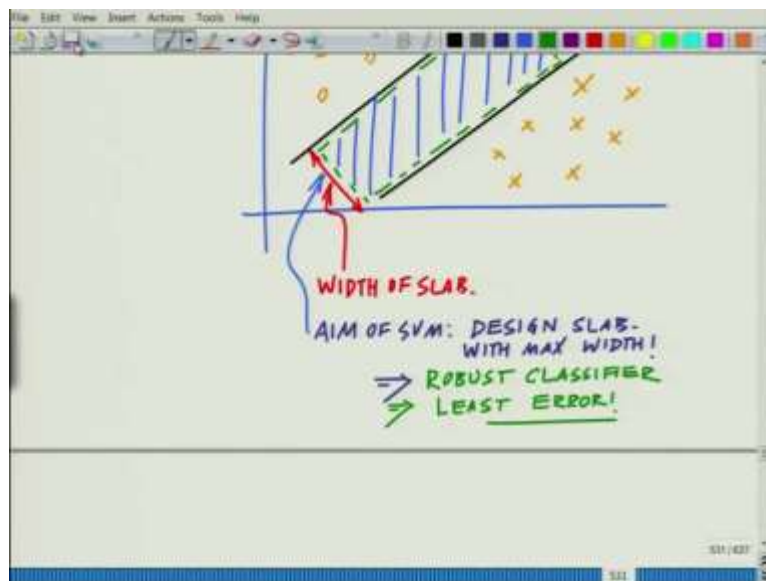
So if you see now what happens to this is you essentially have, now these two hyper-planes. So now you do not have one but you essentially have two hyper-planes so you can say this is your $\vec{a}^T \vec{x} + b = 1$ and this is your $\vec{a}^T \vec{x} + b = -1$. So you have these two hyper-planes and now what you are doing is you are, so you have these different classes now.

Let us say you have this class 0 and you have the class 1 and now you are not just fitting a hyper-plane but you are separating so this is the separation. So you are maximizing this separation or in other words you are fitting a slab. This is essentially known as a, so you are essentially fitting not just hyper-plane but you are inserting a slab, a thick partition, like hyper-plane you can think of it as a thin partition between these two classes.

And that such a classifier or a machine always going to be a very fragile, you can see just like a thin cloth partition between the two classes, easily broken. Now what you are doing is not inserting a thin partition but a thick slab, like think of this as a concrete slab so as that you want to very rigidly partition these two into two different classes, so that there is no chance of confusion or overlap.

And therefore the thicker the slab, the more robust is your classifier, the less is your chance of error and therefore you want to design your classifier in such a way such that the width of the slab, width of this partition is maximized. That is the idea essentially in your support vector machine. So what is the central idea in your support vector machine?

(Refer Slide Time: 32:41)



So this, I call this as the width of the slab. So, what is the aim of SVM? Now let us look at the aim of SVM design the slab, right? Design slab with maximum width. This implies robust classification or this implies you have a robust classifier, implies least error. We are talking

about classification error. So the most thicker your slab, the better is your classifier, the more robust, is your classifier, the lower is your classification.

All right so that is essentially the idea behind the support vector machine and how do we actually design the support vector machine, that is something that I would like to, in next module so please go through this and try to understand this concept at this level and then we will look at it in detail at the working of a support vector machine in the next module. Thank you very much.