

Essential Mathematics for Machine Learning
Prof. S. K. Gupta
Department of Mathematics
Indian Institute of Technology, Roorkee




Lecture - 36
Introduction to Support Vector Machines

Hello friends. Welcome to lecture series on Essential Mathematics for Machine Learning. In the previous lectures, we have seen that what important properties of convex functions and numerical optimization and other techniques for machine learning. Now, in this lecture we will see some of the important aspects of Support Vector Machines.

(Refer Slide Time: 00:49)

Introduction

- The term 'Machine Learning' basically deals with learning from the data which could be in the form of images, measurements, observations, patterns or records.
- The aim of any machine learning algorithm is to perform well on the training data and also to ensure good results in future.
- Machine learning involves the study of three main problems: classification, clustering and regression.

2

Now, let us start. So, first of all what do you mean by machine learning? So, the term machine learning basically deals with learning from the data which could be in the form of images, measurements, observations, patterns or records.


So, suppose you have some data, which may be in the any of the form of maybe in the form of images, measurements, observation, patterns or records then learning from these data is basically called machine learning. Now, what is the aim of any machine learning? The aim of any machine learning algorithm is to perform well on the training data and also to ensure good results in future.

Whatever training data we choose, in that training data, machine learning algorithm should perform good and it also give good result good also ensure good results in future. Machine learning involves the study of three main problems what are three main problems? Classification, next is clustering and next is regression.

(Refer Slide Time: 02:08)

Continued...

- There are many approaches available in the literature to deal with these problems.
- Mangasarian [1] has introduced an approach called 'Mathematical programming approach'. The advantage of using this method is that most of the models involved in Machine learning result into linear programming problem (LPP), quadratic programming problem (QPP) or some convex programming problem (CPP).
- KKT conditions and duality theory are applicable to handle such problems (convex programming problems).

 9

So, these are the three important problems of any machine learning of machine learning algorithms. There are many approaches available in the literature to deal with these problems, these problems means classification, clustering and regression. To deal with these type of basic problems of machine learning there are many algorithms given in the literature.

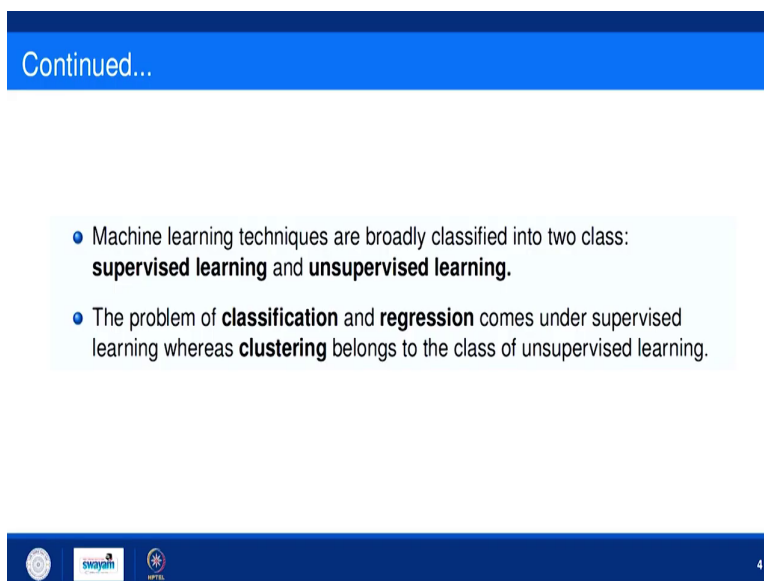
Now, Mangasarian has introduced an approach called 'Mathematical programming approach'. So, there is one approach which is called mathematical programming approach and that was given by Mangasarian. What is the advantage of this approach? Why we are using this approach? Why not some other approach?

So, the advantage of using this method is that most of the problems involved in machine learning result into linear programming problems very simple to handle, quadratic programming problems or some convex programming problems.

And we already know that if we are having some convex optimization models or convex programming problems then the KKT conditions become sufficient, sufficient means that whatever KKT conditions we are having the point, which satisfies those KKT condition will be global minimum point of the problem.

So, these are main beauty of using mathematical programming approach that, it results into either linear programming problem or quadratic programming problem which is very easy to handle.

(Refer Slide Time: 03:33)



Continued...

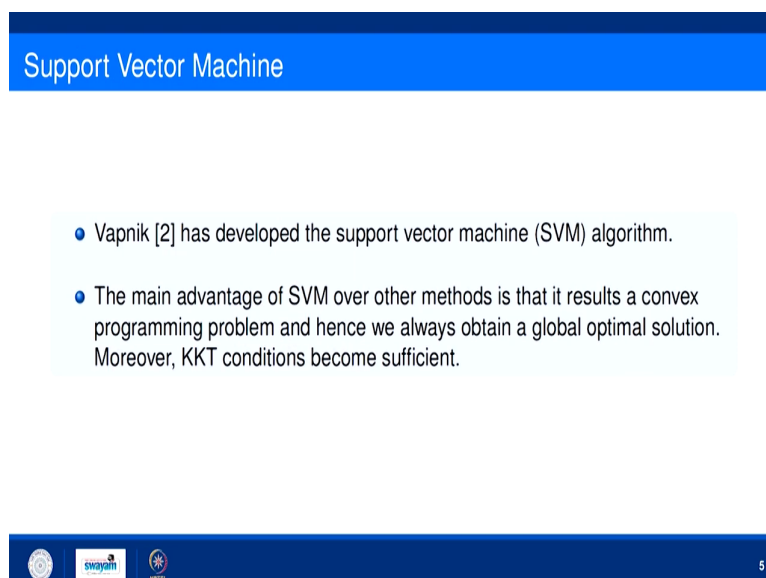
- Machine learning techniques are broadly classified into two class: **supervised learning** and **unsupervised learning**.
- The problem of **classification** and **regression** comes under supervised learning whereas **clustering** belongs to the class of unsupervised learning.

swayam
4

Now, machine learning techniques are broadly classified into two classes: number 1 supervised learning, and number 2 unsupervised learning. The problem of classification and regression comes under supervised learning while the clustering belongs to the class of unsupervised learning.




So, we are not going much detail of this topic we are may focusing on this lecture on support vector machine. So, this is a simple introduction of machine learning.

(Refer Slide Time: 04:07)

A presentation slide titled "Support Vector Machine" in a blue header. The main content area is white and contains two bullet points. The first bullet point states that Vapnik [2] has developed the support vector machine (SVM) algorithm. The second bullet point states that the main advantage of SVM over other methods is that it results in a convex programming problem, ensuring a global optimal solution, and that KKT conditions become sufficient. The footer of the slide is dark blue and contains three logos: a circular institutional logo, the "swayam" logo, and the "NPTEL" logo. The number "5" is displayed in the bottom right corner of the footer.

Support Vector Machine

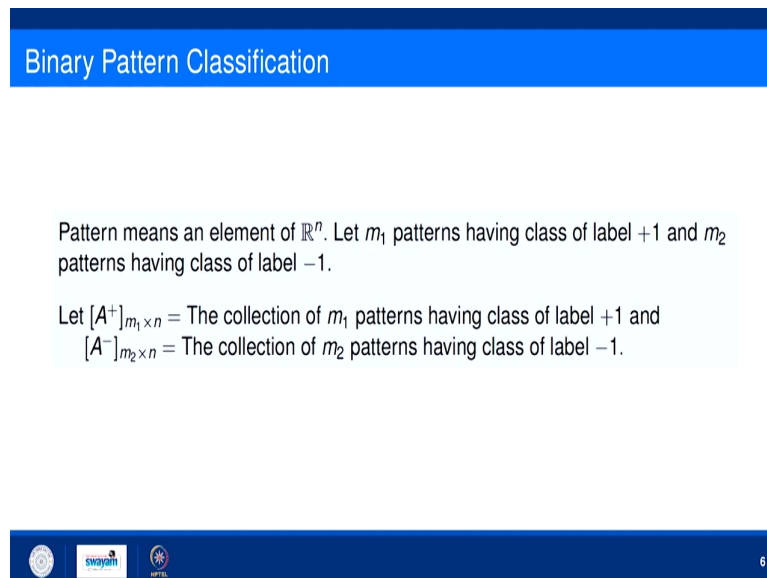
- Vapnik [2] has developed the support vector machine (SVM) algorithm.
- The main advantage of SVM over other methods is that it results in a convex programming problem and hence we always obtain a global optimal solution. Moreover, KKT conditions become sufficient.

   5

Now, what is support vector machine? Vapnik has developed this theory. So, it has developed the support vector machine algorithms. The main advantage of SVM over other methods is that it results in a convex programming problem and hence we always obtain a global optimal solution because it is a convex optimization model.

Moreover, the KKT conditions become sufficient. So, this is again the same advantage of as we are having for mathematical programming approach the same advantage here for using SVM.

(Refer Slide Time: 04:43)



Binary Pattern Classification

Pattern means an element of \mathbb{R}^n . Let m_1 patterns having class of label +1 and m_2 patterns having class of label -1.

Let $[A^+]_{m_1 \times n}$ = The collection of m_1 patterns having class of label +1 and
 $[A^-]_{m_2 \times n}$ = The collection of m_2 patterns having class of label -1.

6

Let us discuss binary pattern classification. Classification problem we are discussing here just to introduce you what SVM is? So, first of all what do you mean by binary class? Binary class means 2 class ok. Now, if you are having 2 class and you want to classify these 2 classes then, what are different methods?.

How we can classify the 2 classes? Now, these 2 data sets are not very small, these are big class of data sets. And what are different algorithms by which this big class of data set can be separated, can be classified?

Now, first of all here if we are; if we are; if we are saying pattern. So, pattern means an element of \mathbb{R}^n , we already know what \mathbb{R}^n is? \mathbb{R}^n means is set of n tuples. So, pattern means an element of \mathbb{R}^n . Let m_1 patterns having class of label plus 1 and m_2 pattern having class of label minus 1. You are having 2 classes; one class I am denoting by plus 1 and the other class I am denoting by minus 1.

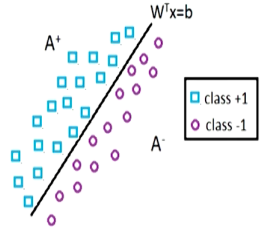
Suppose, the data points of plus 1 class are m_1 , and the data class of minus 1 points I mean minus 1 labelling are m_2 . So, total how many points? Total there will be m_1 plus m_2 points. So, we are denoting if we denote it by a matrix of this class plus 1 class ok; having m_1 patterns and each pattern is in \mathbb{R}^n .

So, it will be a matrix of order m_1 cross n ; that means, a collection of m_1 patterns having class of label plus 1. And similarly, if we say minus 1 labelling; so class of minus 1 labelling will be notify this matrix which is order m_2 cross n that is the collection of m_2 patterns having class of label minus 1.


(Refer Slide Time: 06:56)

Linearly Separable

Two sets $A^+, A^- \in \mathbb{R}^n$ are said to be **linearly separable**, if there exists a hyperplane $w^T x = b$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ such that $A^+ w > eb$ and $A^- w < eb$, where, e = vector of ones.



Linearly Separable


7

So, this is a first thing; that we you are having two data sets. One I am denoting by a plus 1, other I am denoting by a minus 1; my aim is to classify them. Now, they are these 2 classes: A plus and A minus.

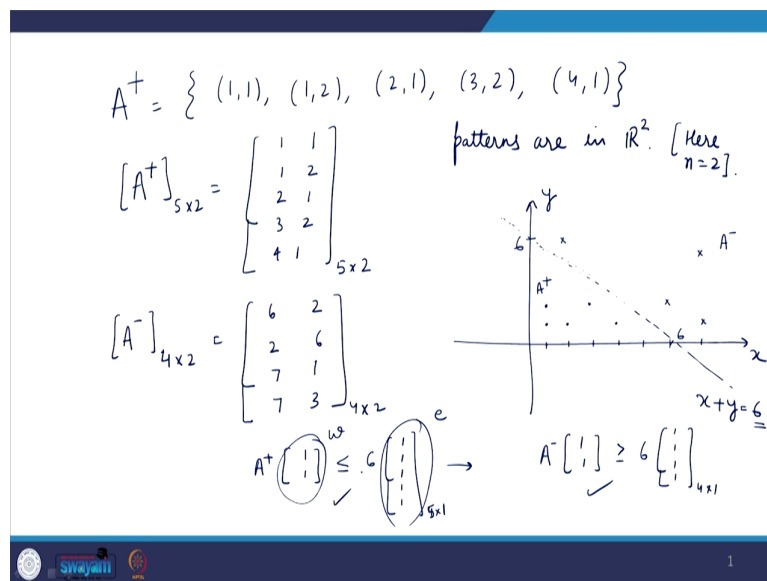
These are said to be linearly separable, if there exists a hyper plane $w^T x = b$ where, w is in \mathbb{R}^n and b is in \mathbb{R} such that for A^+ this is greater than eb and for A^- is less than eb and e is a vector of ones. See here this blue squares this denote data points of plus 1 label; that means, A^+ class and this circles this denote another class which is of minus 1 label and denoting by A^- .

So, since you are able to find a hyper plane which is $w^T x = b$ such that, these 2 classes are classified; these 2 classes are classified. So, we say that these 2 classes are linearly

separable, and if it is not possible to classify them linearly then, we say that they are non-linearly separable.

So, let us discuss this by an example so that, it will be clear to you that when 2 classes have to be linearly separable.

(Refer Slide Time: 08:26)



Suppose, you are having patterns like A plus, A plus patterns are suppose 1, 1, 1, 2, 2, 1, 3, 2 and 4, 1. These patterns are in \mathbb{R}^2 4 patterns are in \mathbb{R}^2 ; that means, here n equal to 2.

So, now how many patterns in A plus class? 1, 2, 3, 4, 5. So, there will be 5 patterns in A plus class. So, A plus will be a matrix of order 5 cross 2 and what is this matrix will be? This will be 1 1, 1 2, 2 1, 3 2 and 4 1. This is 5 cross 2.

Now, if you take another class A minus, suppose this of order say 4 cross 2 which is given by 6 2, 2 6, 7 1, and 7 3. It is 4 cross 2. So, here I am considering 2 classes this by label plus 1, this by label minus 1. These of these there are 5 patterns belonging to A plus class or plus 1 class and there are 4 patterns belonging to minus 1 class.

Now, if you try to plot these points, 1 1. So, suppose it is 1 1 1 2; so, it is 1 2, 2 1; so, it is 2 1 3 2; so, it is 3 2 3 2 then 4 1; so, it is 4 1. So, these are plus 1 label ok A plus.

Now, what are A minus 6 2, 1 2, 3 4, 5 6. So, 6 2, 6 2 means this point. 2 6, 2 6 means somewhere here suppose 2 6. 7 1, 7 1 means somewhere here. 7 3, 7 3 means somewhere here. So, these are of my A minus class.

Now, clearly you can easily visualise that you can find a hyper plane here. May not be unique; but you can find a hyper plane such that, these two classes are separable. So, since we are able to find such hyper plane; so, we say that, these two classes are linearly separable.

So, suppose one I am predicting one hyper plane. So, one hyper plane you can say x plus y equal to 6; it is x , it is y , it is 6, it is 6. Now, so if you now if you take it is x plus y equal to 6; if you take A plus times 1 1 ok, this is w_1 is 1, w_2 is 1 ok; this is 1 1.

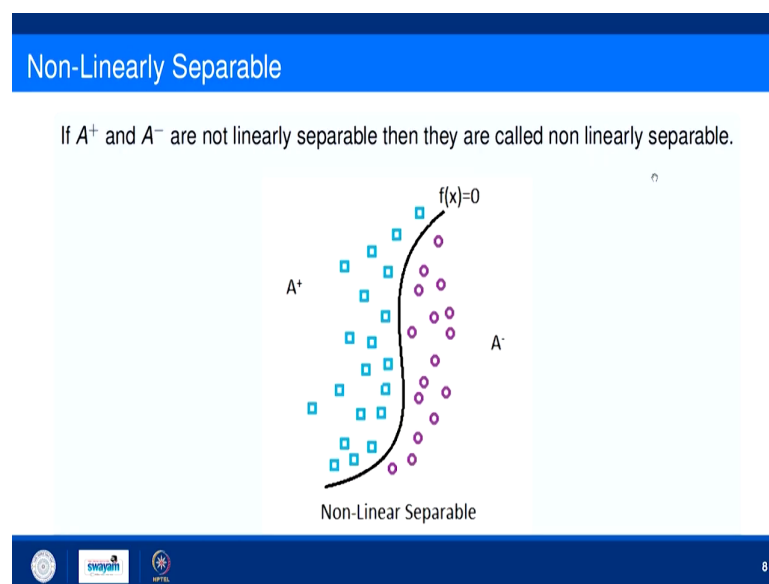
So, it is now less than equal to if you take 6 times 1 1 1; it is 1 2 3 4 5, 1 2 3 4 5. So, this you can easily verify see A plus is what? A plus is this matrix. So, this row this column the first row first column is 2; 2 is less than 6 true. Now, this row this column is 3; 3 less than equal to 6 true again, 2 plus 1 3, 3 less than equal to 6 true, 3 plus 2 5, 5 less than equal to 6 true, 4 plus 1 5, 5 less than equal to 6 true.

So, here this is your w , this is your w this is your w and this is your e vector of 1's, this is your e , this is b ; b is here this is b . So, this is satisfied. Now, again if you take A minus 1 1. So, it is greater than or equal to 6 times 1 2 3 4. So, it is 1 1 1 1 here e is this vector I mean it is 4 cross 1, here e is 4 cross 5 cross 1 ok.

Now, if you take A minus 1 1, 6 plus 2 because, first row first column is the matrix multiplication; first row first column is 8; 8 greater than equal to 6 true. It is 2 plus 6 8, 8 greater than equal to 6 true, 7 plus 1 8, 8 greater than equal to 6 again true and 7 plus 3 is 10; 10 greater than equal to 6 is again true; so this is true and this is true, so we can say that this hyper plane you can opt for this hyper plane which can which classify these two classes linearly ok.

So, hence we are having this definition A plus w greater than eb , and A minus w less than eb . Where, e is a vector of ones of appropriate dimension; it depend what A plus or A minus is.

(Refer Slide Time: 14:32)



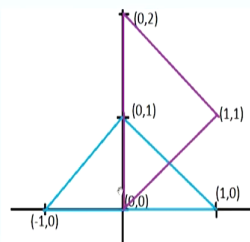
So, if it is not linearly separable; so we are having some non-linear curve here. So, we say that they are non-linearly separable.

(Refer Slide Time: 14:42)

Theorem

- Let A and B be finite sets in \mathbb{R}^n . Then A and B are linearly separable iff their convex hulls are disjoint (Mangasarian [1]).

For example : The sets $A^+ = \{(-1, 0), (0, 1), (1, 0)\}$ and $A^- = \{(0, 0), (1, 1), (0, 2)\}$ are not linearly separable since their convex hulls are not disjoint.



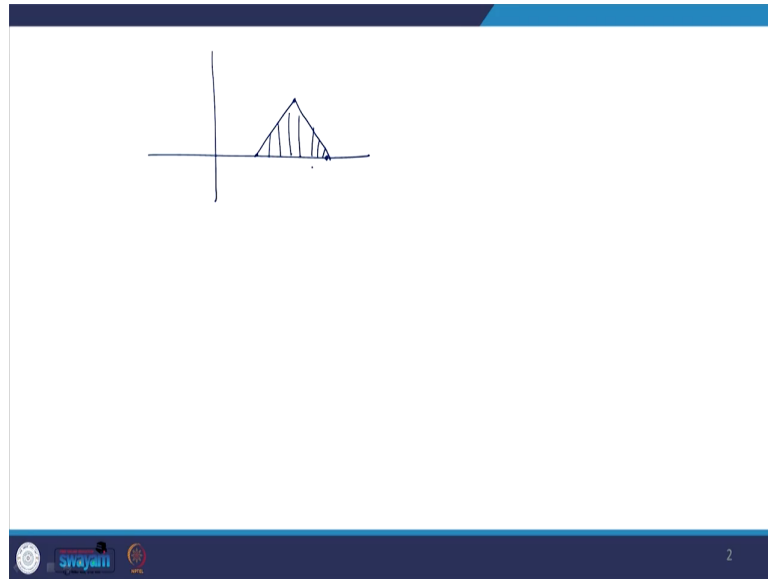
Now, the question arises geometrically if we are having a very small data data points of A plus class and A minus class we can geometrically visualise that whether they are linearly separable or not. But, in practical applications it may not be true.

In practical applications we have a large data sets. Now, to see whether these two classes are linearly separable the given classes are linearly separable or not. So, what are different methods? Because, once they are separable then only we can pay the properties of the classes.

So, the first properties given by Mangasarian; this property is Mangasarian what is the property? Properties let A and B with 2 finite sets in \mathbb{R}^n . Then A and B are linearly separable if and only if they are convex hulls are disjoint.

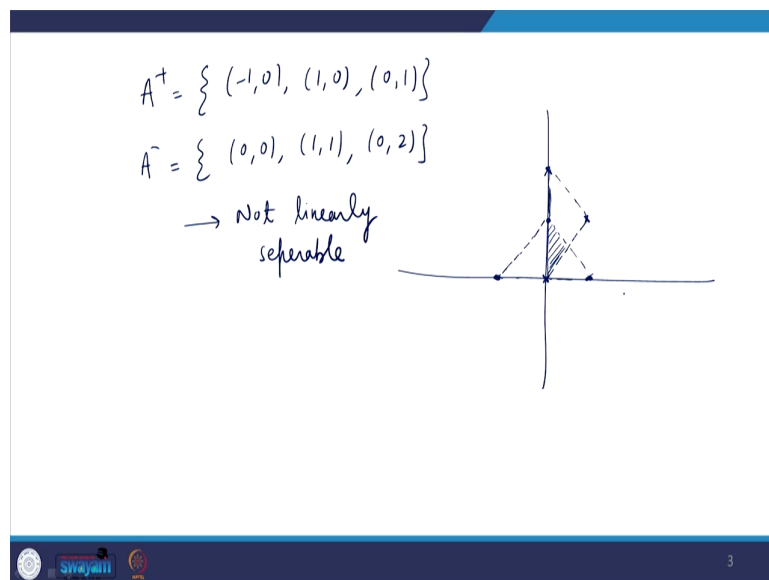
So, this is the first property that by which we can say whether two classes are linearly separable or not. So, what do you mean by a convex hull? Convex hull is a smallest convex set containing that set.

(Refer Slide Time: 15:53)



Suppose, you are having 3 points. So, what is the smallest convex set containing these 3 point and the set is convex. The smallest convex set will be this set; this set is the smallest convex set containing these 3 point and the set is convex. So, this is the convex hull of these 3 points.

(Refer Slide Time: 16:33)



So, let us understand this by an example. Let us suppose A^+ has 3 points. The 3 points of A^+ are $(-1, 0)$, $(0, 1)$, $(1, 0)$. So, A^+ the points of A^+ are $(-1, 0)$, $(0, 1)$ and $(1, 0)$. And the points of A^- are $(0, 0)$, $(1, 1)$, $(0, 2)$; it is $(0, 0)$, $(1, 1)$ and $(0, 2)$; let us see ok.

Now, what are points of A^+ minus $(-1, 0)$, minus $(-1, 0)$ is somewhat here. $(0, 1)$, $(0, 1)$ is somewhere here. $(1, 0)$, $(1, 0)$ is something here. Now, what is your convex hull of these 3 points? The convex hull of these 3 point is this triangle ok. This triangle.

Now, let us see what A^- is? A^- is $(0, 0)$ so, this point. I am denoting the points of A^- minus y cross sign, and $(1, 1)$, $(1, 1)$ is something here this is $(1, 1)$. Next is $(0, 2)$, $(0, 2)$ is something here. And what is the convex hull of these 3 points now? The convex hull of these 3 point is this triangle.

So, now are they disjoint? Because, this convex hull this triangle of A plus and this convex hull of A minus; they are not disjoint because this is common. So, since they are not disjoint. So, they are not linearly separable. So, not linearly separable.

You can also see by geometry; see we are having these 3 points here and we are having these 3 points here. So, we cannot find a plane such that we can classify these 2 data points and these 2 classes. So, this is the; this is the first most result that if the convex hull of 2 classes are disjoint then they are linearly separable, and vice versa.

But again this it is a very small set, so we can say we can find the convex hull and we can say that they are disjoint or not. But, how we can say that whether a given sets mathematically how can we show? That whether a given 2 data points are linearly separable or not.

So, we have a 1 result based on error minimising LPP which will deal with that by formulating that problem into LPP; we can see that whether a given problem with a given binary classification problem is linearly separable or not. But, before going into that result first we will see first we have to understand this inequality this result; what this result is basically?

(Refer Slide Time: 19:58)

Result


The inequalities $A^+w > eb$ and $A^-w < eb$ by suitable scaling can be re-written as:

$$A^+w \geq eb + e$$

and

$$A^-w \leq eb - e$$

respectively.



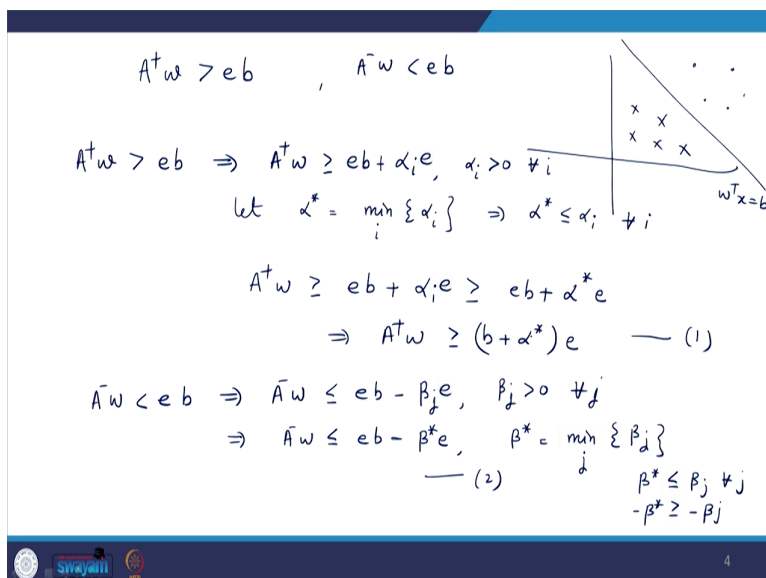
10

That the inequality $A^+w > eb$ and $A^-w < eb$ by suitable scaling can be rewritten as this why? And how? Is it true? Is it really true? So, that we have to see ok.

Because, why we are doing like this? See, if we are having a strictly greater than or a strictly less than type of constraints and we are formulating a LPP; it is not an easy to solve such problems. See, if we are having less than equal to or greater than equal to type constraint or equality type constraint in LPP; that we have different methods if a algorithms to solve such problems, but not of a strictly greater or strictly less.

So, we anyhow we want to convert these strictly the stick inequality type constraint into less than equal to or greater than equal to type constraint. So, how can we convert these 2 constraint to these types? So, let us try to understand.

(Refer Slide Time: 21:07)



$A^T w > e b$, $A^- w < e b$

$A^T w > e b \Rightarrow A^T w \geq e b + \alpha_i e, \alpha_i > 0 \forall i$
 let $\alpha^* = \min_i \{\alpha_i\} \Rightarrow \alpha^* \leq \alpha_i \forall i$

$A^T w \geq e b + \alpha_i e \geq e b + \alpha^* e$
 $\Rightarrow A^T w \geq (b + \alpha^*) e \quad \text{--- (1)}$

$A^- w < e b \Rightarrow A^- w \leq e b - \beta_j e, \beta_j > 0 \forall j$
 $\Rightarrow A^- w \leq e b - \beta^* e, \beta^* = \min_j \{\beta_j\}$
 --- (2)

$\beta^* \leq \beta_j \forall j$
 $-\beta^* \geq -\beta_j$

So, the first constraint is $A^T w > e b$ from where we get this constraint, because if you are having 2 class of data points; one denoted by a cross, other denoted by a dot suppose, and you are having a hyper plane which is $w^T x = b$ which can classify these 2 classes linearly; then for 1 class of data points this inequality will hold, and for other the class of data point another inequality will hold. Now, how can we convert this and this inequality? $A^- w < e b$ ok, 1 is greater 1 is less.

Now, let us try to understand. Now let so, this $A^T w > e b$ implies, $A^T w$ is greater than $e b$ implies, $A^T w$ is greater than equal to $e b$ minus α_i where α_i are greater than 0 for all i , it is greater, greater means plus because if it is greater than $e b$ then it is further greater than $e b$ plus α_i because, α_i is strictly greater than 0 of course. Because, $e b$ is greater than $e b$ plus α_i .

Now, let us take let α^* is equal to minimum of α_i minimum over i . It is minimum of α_i it is minimum α_i suppose, α^* . So, that implies α^* is less than equals to α_i for all i .

So, that means, $A + w$ will be greater than equals to eb which is α_i that will be again greater than equal to $eb + \alpha^* e$ because, α_i is greater than equal to α^* for all i . So, this implies $A + w$ is greater than equals to $b + \alpha^* e$. So, this is suppose 1.

Now, let us take $A - w$ less than eb . So, this further implies $A - w$ is less than or equal to $eb - \beta_i$ where, β_i is greater than 0 for all i again; β_j you can take β_j . Because eb is will be less than or equal to $eb - \beta_j$.

So, this further implies $A - w$ will be less than equals to $eb + \beta^* e$ where, β^* is equal to minimum of β_j because, here from here in the same way as we did here will be less than equals to β_j .

So, it is then minus so, here it is here it is minus sorry. So, negative of β^* will be greater than or equal to minus of β_j . So, that means, this will be minus of β_j . So, that means, minus of β_j will be less than or equal to β^* . So, this is second expression suppose.

(Refer Slide Time: 25:05)

$$\text{let } \gamma = \min \{ \alpha^*, \beta^* \} \Rightarrow \left. \begin{array}{l} \gamma \leq \alpha^* \\ \gamma \leq \beta^* \end{array} \right\}$$

$$\begin{array}{l} (1) \Rightarrow A^+ w \geq b e + \gamma e \\ (2) \Rightarrow A^- w \leq b e - \gamma e \end{array} \left. \vphantom{\begin{array}{l} (1) \Rightarrow A^+ w \geq b e + \gamma e \\ (2) \Rightarrow A^- w \leq b e - \gamma e \end{array}} \right\}$$

The above inequalities can be re-written as:

$$A^+ w^* \geq b^* e + e$$

and $A^- w^* \leq b^* e - e$

where $\left. \begin{array}{l} w^* = w/\gamma \\ b^* = b/\gamma \end{array} \right\}$

So, now if you choose now, let gamma equal to minimum of alpha star beta star. So, this implies gamma will be less than equals to alpha star and gamma is again less than equal to beta star. So, what does 1 implies then? 1 implies see alpha star from here alpha star is greater than equals to gamma ok.

So, so from here from this inequality 1 what we have concluded? So, this will be A plus w greater than equals to be plus gamma e. And from 2 what we obtain A minus w will be less than equals to be minus gamma e this you can understand from here because, minus beta star minus beta star will be further less than or equal to gamma from here ok.

So, now these 2 inequalities can be further written as this the above inequalities can be re-written as, now you can divide by γ throughout and again you are here also you can divide by γ throughout.

So, you will obtain $A + \gamma w^* \geq b^* + \gamma e$ and $A - \gamma w^* \leq b^* - \gamma e$. Where, w^* is nothing but, w upon γ and b^* is nothing but, b upon γ .

So, in this way we have converted the strict type of inequalities into less than equal to or greater than equal to types. So, that is basically suitable scaling. So, suitable scaling will not affect see, we have only changed w or b by some by applying some constant γ ok. If they are linearly separable, still they will be linearly separable and if they are nonlinearly separable then still they will nonlinearly separable.

So, we have seen that if we have 2 classes 2 classes of we are labelling with plus 1 or minus 1 and we try to see that whether they are linearly separable or not, the first way out is if you are able to find the plane which then classify them linearly; that means, they are linearly separable.

Next is if you find the convex hulls of the 2 classes and they are coming out to be disjoint then we can say that they are linearly separable. In the next lecture we will see that how can we mathematically formulate a problem by which we can see there are 2 given class of data points are linearly separable or not. So, these are references which we have used here.

Thank you.