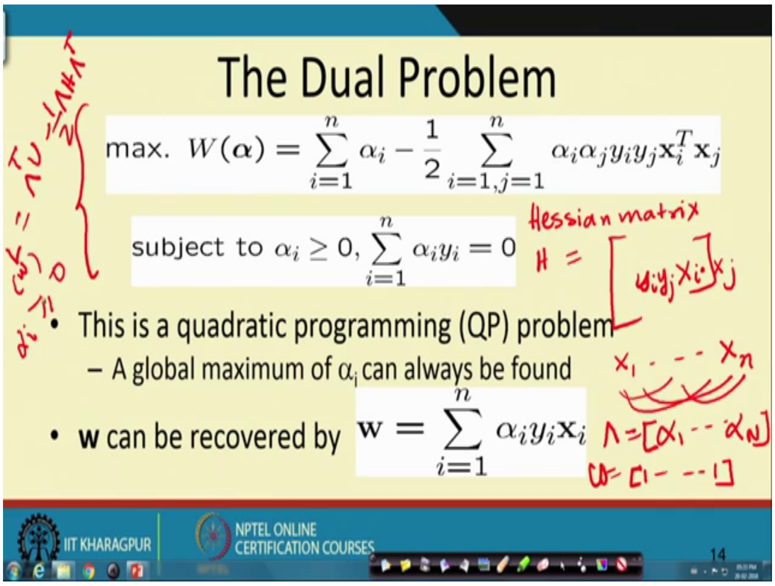
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 26
Support Vector Machine – V

Let us continue with support vector machines, if you remember we wanted to find a w and v b the slope and b value of a plane, which will correctly classify all the points and give the highest margin.

(Refer Slide Time: 00:37)



The Dual Problem

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

- This is a quadratic programming (QP) problem
  - A global maximum of α_i can always be found
- w can be recovered by $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$

Hessian matrix

$H = [y_i y_j x_i x_j]$

$x_1 \cdots x_n$

$\Lambda = [\alpha_1 \cdots \alpha_N]$

$\omega = [1 \cdots 1]$

$\frac{1}{2}\Lambda^T H \Lambda$

$\alpha_i \geq 0$

So, we formulated it as an optimization problem we converted it into a dual form; where the dual problem can be written like this, the W alpha which is the Lagrangian is summation of all the alphas the Lagrangian multipliers, and half times double summation. So, this is double summation i equal to 1, j equal to 1 alpha i alpha j, y i, y j x i dot x j which is same as xi transpose xj subject to all the alpha is being either 0 or positive and the summation alpha i yi being 0.

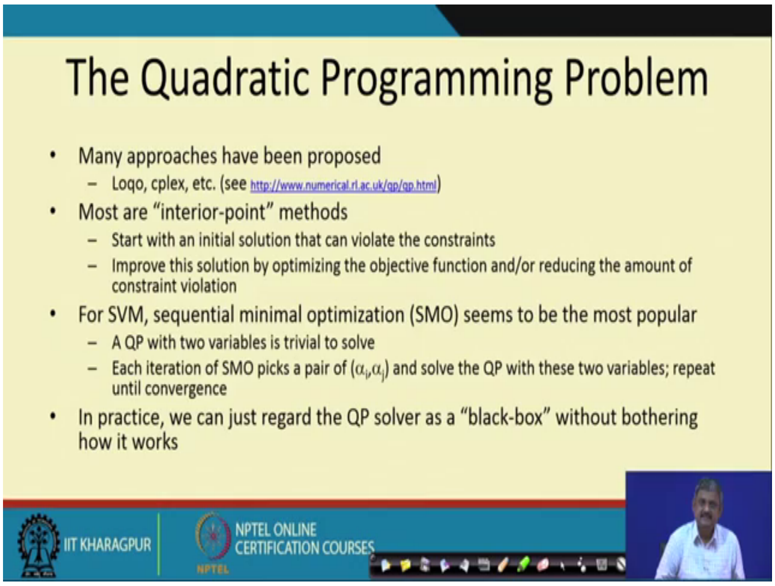
So, this is an example of a quadratic programming problem and we can solve it always and get a set of values of alpha i is equal to 1 to n, plug in those values in this equation which you already got as the value of w, when we minimize with respect to Lagrangian with respect to w and we can find the value of w.

We have also seen that you can actually write it as a matrix form, where if you consider the Hessian matrix and consider the Hessian matrix H which we construct as follows, we have X 1 2 X n we take all pairs of xi xj and the i j th entry of the Hessian is nothing, but the class level of i th point, class level of j th point the vector of i th point dot product the vector of the j th point; which is same as xi transpose x Eigen.

This y constitute the Hessian and we similarly consider another vector of the Lagrangian multipliers and another vector u which is all 1 vector and if we do this equations can be I am writing down in a simpler form is Lagrangian which is w also I write here as w is the Lagrangian vector u minus half, Lagrangian vector a Hessian matrix Lagrangian transport subject to alpha i greater than equal to 0 ok.

So, this is a typical quadratic programming problem and we will if we solve this there are numerical methods to solve this, if you solve this what we will get is the, this is how we will solve I am telling you about the this how we will solve.

(Refer Slide Time: 03:25)

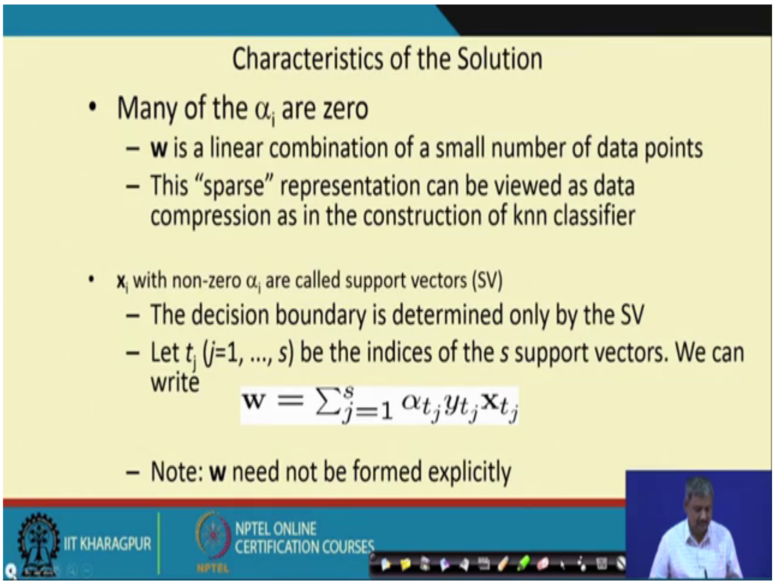


The Quadratic Programming Problem

- Many approaches have been proposed
  - Loqo, cplex, etc. (see http://www.numerical.rl.ac.uk/qp/qp.html)
- Most are "interior-point" methods
  - Start with an initial solution that can violate the constraints
  - Improve this solution by optimizing the objective function and/or reducing the amount of constraint violation
- For SVM, sequential minimal optimization (SMO) seems to be the most popular
  - A QP with two variables is trivial to solve
  - Each iteration of SMO picks a pair of (α_i,α_j) and solve the QP with these two variables; repeat until convergence
- In practice, we can just regard the QP solver as a "black-box" without bothering how it works

You get many numerical techniques like sequential, minimal optimization interior point method, which will solve this and give you a set of values of alpha i and as I told you can get alpha i to reconstruct your w. But the most interesting part of this is that if you solve most of the times if you note that see your n can be large, n is your size of the training set it can be large.

(Refer Slide Time: 03:58)

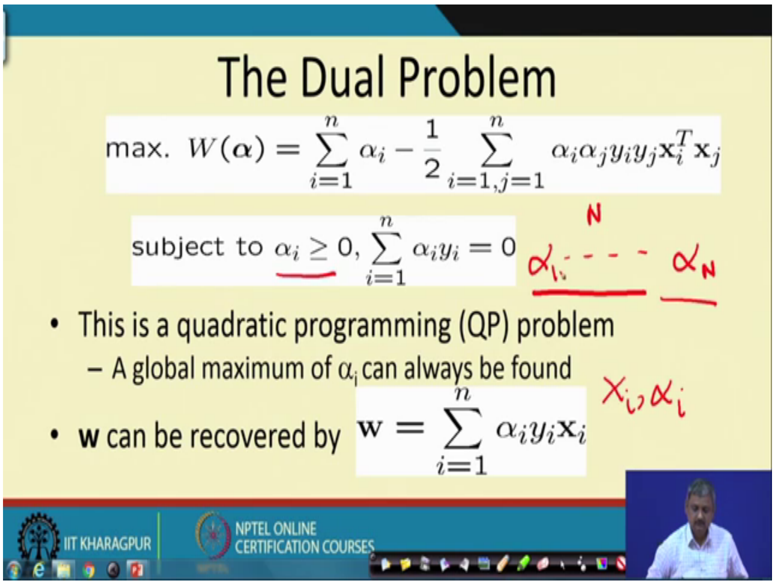


Characteristics of the Solution

- Many of the α_i are zero
  - **w** is a linear combination of a small number of data points
  - This "sparse" representation can be viewed as data compression as in the construction of knn classifier
- $\mathbf{x}_i$ with non-zero α_i are called support vectors (SV)
  - The decision boundary is determined only by the SV
  - Let t_j (j=1, ..., s) be the indices of the s support vectors. We can write $\mathbf{w} = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} \mathbf{x}_{t_j}$
  - Note: **w** need not be formed explicitly

So, you can have large values of n say 1 million 1 lakh or something.

(Refer Slide Time: 04:12)



The Dual Problem

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \underline{\alpha_i \geq 0}, \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \underline{\alpha_1 \cdots \overset{N}{-}} \; \underline{\alpha_N}$$

- This is a quadratic programming (QP) problem
  - A global maximum of α_i can always be found
- **w** can be recovered by $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ $\quad x_i, \alpha_i$

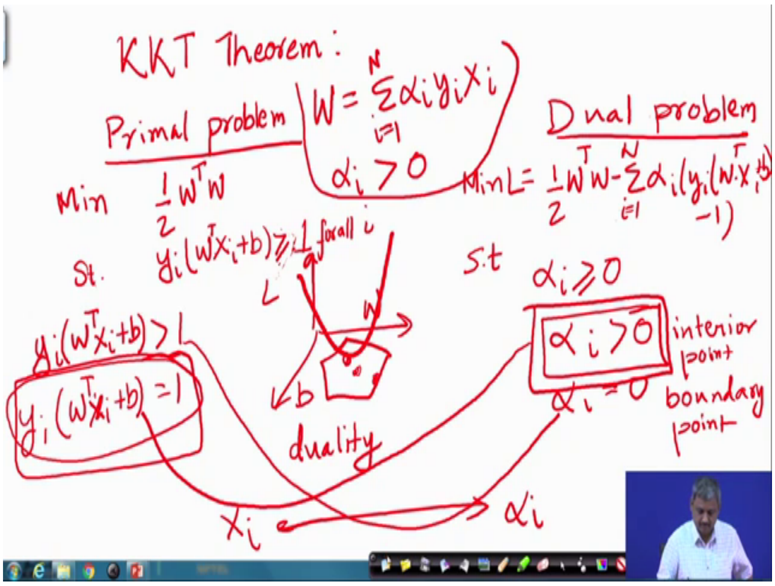
So, your Hessian matrix is really big n by n really big matrix. So, actually computational complexity of solving this QP problem is little high there are ways to reduce it. So, if we solve this and find out the values of alpha 1 to alpha N, you will find most maybe 95 percent values we have a restriction that alpha i should be greater than equal to 0, most of the values actually turn out to be exactly 0.

So, if you have n equal to 1000 maybe 99 examples note that, 1 thing you also note in this regard is that see I have as many Lagrangian multipliers alpha is as many training points as the number of training points. So, there is a kind of correspondence for every training point X i I have a corresponding Lagrangian multiplier alpha i. So, every large that comes here also when you write down the constant, it comes that for every xi there is a alpha and most of this alpha is will be 0 as I have told you, I mean if we actually solve bit will observe it also and few of the alpha is i greater than 0.

Now, we are naturally curious since every alpha e corresponds to x I, which of the xi is for which alpha is a 0 and which are the x i is the trading points for which alpha is a greater than 0. So, let us try to answer that question. The answer lies in this theorem called the Karush kuhn tucker theorem, also another thing you quickly note is that and see this w is a summation of alpha i y i. So, if an alpha is 0, the corresponding y i x i does not contribute to the value of w, corresponding x i y i do not contribute only the few alpha as I said few alpha is are greater than 0, only those few x i y i contribute to the value of w in computing the value of w ok.

So, let us see which are these ok.

(Refer Slide Time: 07:23)



Let me state you a theorem known as the Karush Kuhn Tucker theorem, 1 minute take it KKT theorem. What the Kuhn Tucker theorem says is that so you see you have 2 problems a primal problem and a dual problem, what we optimized in the primal problem is minimize half w transpose w, such that y i for all i if you remember the definition and what we did in the dual problem is that introduce Lagrangain multipliers and simplified the constant made the objective function more complex.

So, we had a objective function which are slightly more complex sorry 1. So, the constant basically the constant moves to the objective function and this is i equal to 1 to n and the constant was; so that is why the duality see you had see the constant goes to objective function whereas a simple constantly.

So, this is also actually greater than equal to 1 because for 1 point it will actually be the minimum, 1 point it will actually be the minimum so let me for sake of clarity. So now, this constant so what this optimization business basically is about, so if you look at say for example, w b space hypothetically and this is the objective function L that you are minimizing the constant sort of defines a physical area within which we should search for the solution and what we look is that the objective function gives a takes on a value, within the feasible area wherever objective function is minimum that is our solution.

Now, this constant can be satisfied in 2 way 1 is greater than 0 1 is exactly equal to 0, similarly here greater than 1 equal to 1. So, the idea is like this when the constant is satisfied as an inequality basically your solution is in the interior point and your solution constant is equality your solution is a boundary point.

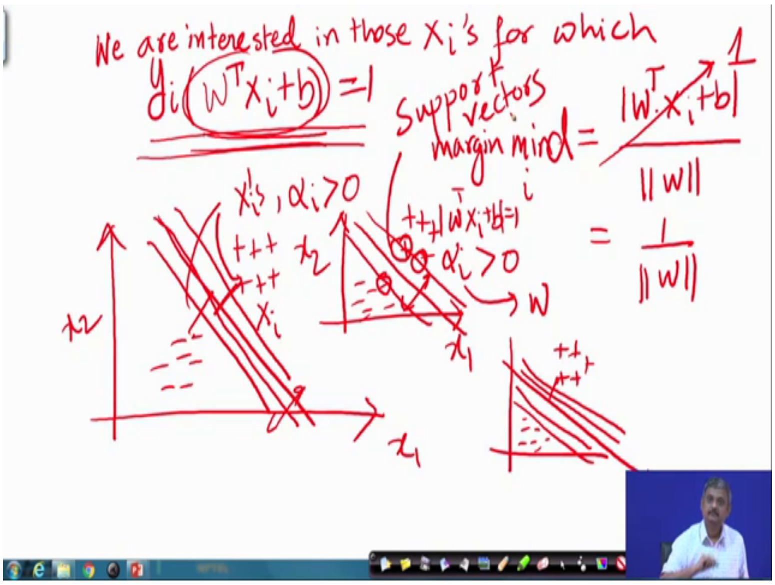
So, this is an interior point interior solution and this is a boundary point similarly here, what the KKT theorem says of course, i am not proving it you have to accept it if you open any optimization book will find it. So, what the KKT condition says is that the interior solutions in the dual problem correspond to boundary solutions in the primal problem. And the boundary solutions in the primal problem correspond sorry, the interior solution in the primal problem correspond to the boundary solutions in the dual problem. So, that is the duality which is the duality ok.

So, but if you remember from our previous slide is that, we said that we basically the weight is nothing but this. So, we are interested only those alpha is which are greater than 0 basically not equal to 0 because, it is only those alpha is which will contribute to my w only those alpha is will contribute. So, we are interested in these alpha is, and he also mentioned before that they are a very small minority, among all the alpha is the onewhich are not 0 there is very small fraction of all the alpha.

So, if n is thousand they will be only 10 maybe. So, it is this which is really important for us these alpha is so excuse me. So, by duality this interior alpha is correspond to the boundary x is note that every alpha i as a corresponding x i. So, basically we are doing a alpha i to xi mapping ok.

So, we are actually interested in this particular xi, which are which give y i into W T X i plus bas 1. So, if you I have used up all the board I have to erase it keep this picture in mind.

(Refer Slide Time: 15:03)



We are interested in those x_i's for which $y_i(W^Tx_i+b) = 1$

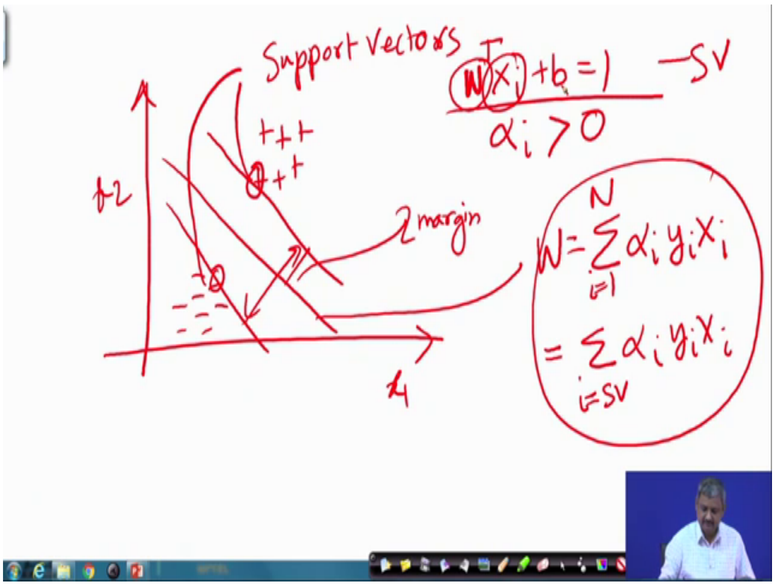
$$\text{margin } \min_i = \frac{\cancelto{1}{|W^Tx_i+b|}}{\|w\|} = \frac{1}{\|w\|}$$

Now, y i is either plus 1 or minus 1. So, if it is equal to 1 these quantities should also take on a value 1, see magnitude of this quantity should be 1; if you remember what we did we define when you define margin is we said distance from x i to the line, and we said that minimum of this over i is we normalized w and b to make it 1 that is why we get a margin of 1 by norm of w. If we remember tally I think we scale w and b each multiply both w and bi a quantity note that the distance does not change because, both numerator denominator I am multiplying by the same factor. So, distance is same margin value remains the same the numerator of the margin by scaling w and b can be normalized to having a smallest value of 1.

So, basically this point corresponds to the smallest value of this quantity. So, when does the smallest value appear? The smallest value corresponding to the x i which is closest, w is same for all points. So, only these x i and there may actually multiple x i which are closest. So, only these xi will have alpha i greater than 0 the closest 1, how to again find this closer, as if we take the line and parallel spread out 2 lines on both side increase.

So, you sort of take this line then do these spread out and when you touch 1 of these points stop. So, if I redraw I will get a picture like this I stop. So, there will be no other point in this gap and these gap would be called the margin of the lines, and the points which lie on this boundary would be called would have a alpha i; as I have said will they will have w x i, w transpose x i plus b equal to 1 absolute value equal to 1 and they will have a alpha i greater than 0 and they will only determine w.

(Refer Slide Time: 19:47)

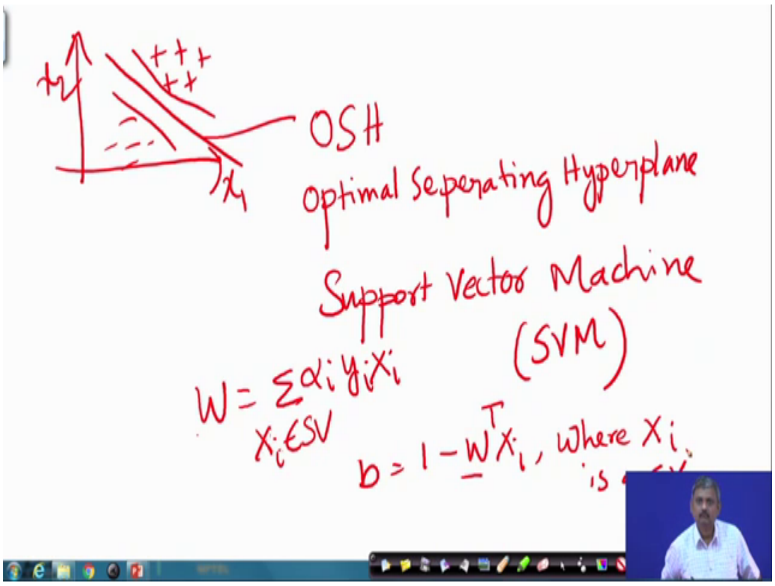

Support Vectors
$W^T X_i + b = 1$ —SV
$\alpha_i > 0$
f2
f1
2 margin
$W = \sum_{i=1}^{N} \alpha_i y_i X_i$
$= \sum_{i \in SV} \alpha_i y_i X_i$

This critical and they are a small fraction expectedly, these critical values would be called the support vectors, critical points would be called the support vectors let me draw.

In fact, this is twice the margin is only this much these and these are the support vectors, the points you just touched the margin, they are the support vectors and they have the property that b equal to 1 for them and all they also have their property that corresponding Lagrangian multiplier a greater than 0 and this slope of this line rather I can express as X i.

In fact, since except support vectors all other alpha i is at 0, I can rewrite this summation as i is a support vector, only the other I can drop up the other alpha is I can consider only the support vectors only this point and some of their a Lagrange multiplier y i x i get. So, how do I get my b? It is very easy to get b because, I know for the support vectors this holds I know my w I have computed w this way, I pick up a support vector, I know this equation will hold I know w, I know x i I will find out b from this equation. So, I liquid to 1 from 1 minus wxi will be the b, what is any support vector I can pick up and find out b.

(Refer Slide Time: 22:26)



OSH

Optimal Separating Hyperplane

Support Vector Machine

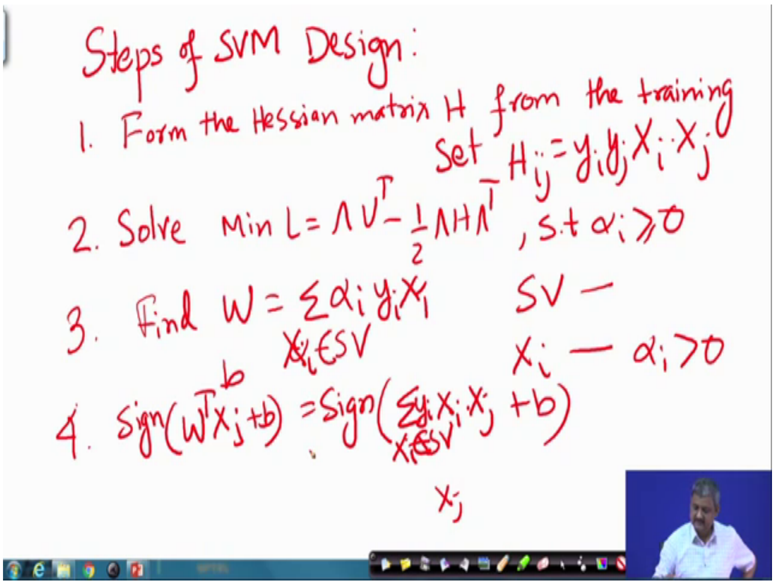
(SVM)

$W = \sum_{X_i \in SV} \alpha_i y_i X_i$

$b = 1 - \underline{W}^T X_i$, where X_i is

So, let me summarize the steps oh by the way the this line which is actually the main thing of our interest, which has this margin is called a optimal separating hyper plane or a support vector machine. We already know how to get w of that point it is take all the support vectors sorry take X i to be a support vector sum up these values, how do I find b? I know b equal to 1 minus w transpose X i, I know w Ii have already computed, where pick up any support vector to summarize.

(Refer Slide Time: 24:18)



Steps of SVM Design:

1. Form the Hessian matrix H from the training
   Set $H_{ij} = y_i y_j X_i \cdot X_j$

2. Solve $\text{Min } L = \Lambda V^T - \frac{1}{2} \Lambda H \Lambda^T$, s.t $\alpha_i \geq 0$

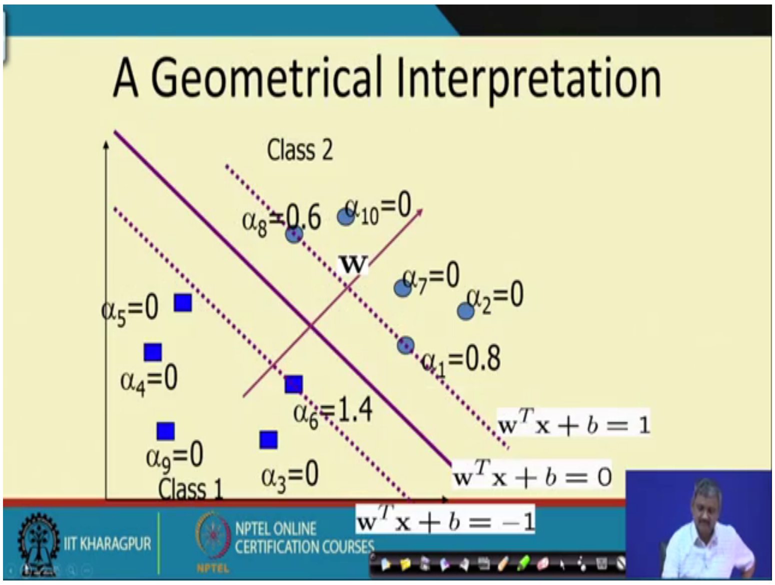
3. Find $W = \sum_{X_i \in SV} \alpha_i y_i X_i$ SV — X_i — $\alpha_i > 0$
   b

4. $\text{Sign}(W^T X_j + b) = \text{Sign}\left(\sum_{X_i \in SV} y_i X_i \cdot X_j + b\right)$
   x_j

So, Hessian matrix actually contains all the information you need to get from the data point no other place it is used. Find for all the support vectors. So, support vector is all the x i for which alpha i is greater than 0, for all the support vectors add up this. Can you tell how do I classify a new point let us say call it x j, how do I classify in it point? Very simple what I will actually do is that, I will find the sign off and find b accordingly. I will find the sign of w T transpose x j plus b it is positive plus class negative minus class.

So, this is equivalent to sign of if we substitute w value here, what you need to do is to just take the support vectors and take the dot product of each of the support vector with the new point to be classified into y i actually y i x i x. So, yi xi is this support vector x is the new point compute this summation add b look at this summation, look at the sign of that you get here classification of a new point. So, that is how you do it all right. So, this is the simple thing so what I want to do is that so there are actually many toolboxes.

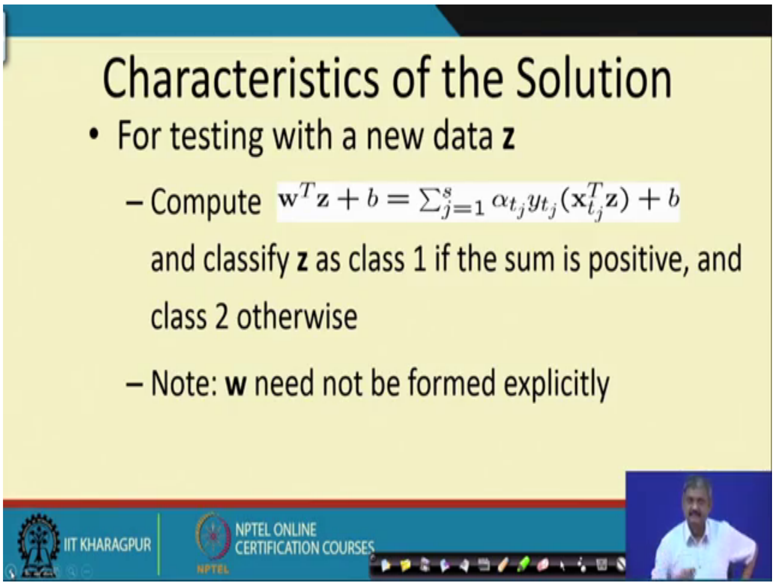
(Refer Slide Time: 27:06)



A Geometrical Interpretation

So, this is the pictorial representation of whole I wanted to say, this will depict the values of alpha you can go through the slides, so I will go into that later. So, this is the characteristics.

(Refer Slide Time: 27:23)



Characteristics of the Solution

- For testing with a new data **z**
  - Compute $w^T z + b = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} (x_{t_j}^T z) + b$ and classify **z** as class 1 if the sum is positive, and class 2 otherwise
  - Note: **w** need not be formed explicitly

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, there are many software's which can solve this problem solve the QP and give you alpha is which you can use and do the training set. There are 2 more extensions to this problem all my theory so far has been for 2 classes, can I extend it to multi class and also I have affirmed that the 2 classes are separable linearly; what happens if they are overlapped, what happens if they are non-linearly separable. So, we will discuss that in our next lecture on when we extend in 2 overlapping support vector machines, soft margin support vector machines, and Cornell machines which are a non-linear version of the support vector machine.

Thank you for today.