

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

Lecture - 39
Hard Margin Classifier

Hello friends. Welcome to lecture series on Essential Mathematics for Machine Learning. In the last lecture we have seen that if two classes are linearly separable then we can construct a error minimizing LPP whose objective value is 0. If the objective value is does not come out to be 0 that implies that the two classes are not linearly separable. But, if we are interested to find out the optimal separating hyper plane, then the method of finding error minimizing LPP may not work.

It will give a linear classifier of course, but that may not be an optimal separating hyper plane. So, how can we find an optimal separating hyper plane? Optimal means, the margin between the two classifiers two bounding planes is maximum. How we can find out that hyper plane, where the distance between the two bounded hyper plane is maximum that is the optimal separating hyper plane. So, that comes under hard margin classifier.


So what is hard margin classifier and how we can construct an equivalent optimization problem for linear classification problem, so that we can find out an optimal separating hyper plane.

(Refer Slide Time: 01:45)

Hard Margin Classifier

Let $\{x_i \in \mathbb{R}^n, i = 1, 2, \dots, m\}$ be a set of finite patterns which is linearly separable having class of label $+1$ and -1 . $d_i \in \{-1, +1\}$ = target value of i^{th} data.
Therefore, there exists $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\begin{aligned} &w^T x_i - b > 0, \forall i \text{ having } d_i = +1 \\ &w^T x_i - b < 0, \forall i \text{ having } d_i = -1 \end{aligned}$$
$$\begin{aligned} &w^T x_i - b \geq +1, \forall i \text{ having } d_i = +1 \\ &w^T x_i - b \leq -1, \forall i \text{ having } d_i = -1 \end{aligned}$$
$$d_i (w^T x_i - b) \geq 1, \forall i = 1, 2, \dots, m$$

 2

So, let us discuss. Now, suppose we have patterns ok. How many patterns? Suppose, we are having m number of patterns; x_1, x_2, x_3 up to x_m and each pattern is element of \mathbb{R}^n ok. And we are also supposing that these patterns are linearly separable ok. Having two classes; plus 1 and minus 1. This d_i , this is belonging to minus 1 and plus 1 ok, is the target value of the i^{th} data.

Therefore, there exist w belongs to \mathbb{R}^n and b belongs to \mathbb{R} such that so, this $w^T x_i$, $w^T x_i - b$ is greater than 0, for all i which are in plus 1 label class and it is less than 0 for all i which are in minus 1 label class. We have already seen that by suitable scaling this can be transformed into greater than equal to 1, for all i which are in plus 1 label class and less than equal to minus 1, for all i which are in minus 1 label class.

So, if you want to combine these two constraints, so, these two constraints can be combined like this. $d_i w^T x_i - b \geq 1$, for all i from 1 to m because there are m number of patterns. See, if you put d_i equal to 1, d_i is either plus 1 or minus 1. These are target these are labeling basically.

So, if it is plus 1 then this will converge to the first constraint this constraint, if d_i is plus 1. If d_i is minus 1 then you multiply both the side of the inequality by minus 1 and this will convert to this inequality that is $w^T x_i - b \leq -1$.

So, I want to say that the constraint, the constraint can be convert into a single constraint of this, where d_i belongs to plus 1 or minus 1 ok. Now next is; what is our aim? Our main aim is to maximize the distance between the two hyper planes ok.


(Refer Slide Time: 04:00)

Continued...

- Among all the separable hyperplanes, we choose the hyperplane $w^T x = b$ for which the margin $\frac{2}{\|w\|}$ is maximum or $\frac{\|w\|_2^2}{2}$ is minimum.
- To find the maximum margin classifier, the optimization problem can be modelled as

$$(P1) \min_{w,b} \frac{1}{2} w^T w \quad \text{s.t.} \quad d_i(w^T x_i - b) \geq 1, \quad i = 1, 2, \dots, m.$$

- This is a **convex quadratic programming problem (QPP)**.
- We may also use Lagrangian duality to solve the above problem.


3

So, now what is the distance, what is the margin? The margin is basically 2 upon norm of w , ok. This is to maximize. Now it is same to say that this quantity is to minimize ok. Whether we are minimizing norm of w or we are minimizing norm square of w both are same because norm is a non negative quantity ok. So, it is equivalent to say that norm square w upon 2 is to minimize. Now, how we can write norm of w square? This can be written as $w^T w$; I mean inner product of w with itself. And under usual inner product it is nothing but $w^T w$.

So, what is the problem now? The problem is now converted into minimization of $\frac{1}{2} w^T w$, this is to minimize, subject to this constraint, this constraint ok. So, this constraint will come here, i from 1 to m . So, now, it is a very very simple quadratic programming problem. In fact, it is a convex quadratic programming problem because, diagonal elements are all 1 by 2 in this case and which is a positive definite and hence convex.

So, it is a and constraints are linear of course. So, it is a convex quadratic programming problem and we can have different algorithms to solve such type of problems. So, once we find w and b from these two, so, then we can find the hyper plane which is $w^T x$ equal to b and that is a hyper plane which is an optimal separating hyper plane. Now, here how many constraints we are having? Here we are having m number of constraints, if we are having m number of patterns. This m maybe 100 , maybe 1000 , maybe 500 . So, depending on the number of patterns there we are having so many constraints. But, using Lagrange duality method, we can reduce this number of constraints, so, which is computationally easy.

So, what is that Lagrange duality method let us see. See, what problem we are having here?

(Refer Slide Time: 06:34)

$$\begin{aligned}
 & \text{Hard margin classifier} \left\{ \begin{aligned} (P1) \quad & \text{Min} \quad \frac{1}{2} w^T w \\ & \text{s.t.} \quad d_i (w^T x_i - b) \geq 1, \quad i=1, 2, \dots, m \end{aligned} \right. \\
 & L(w, b, \alpha_i) = \frac{1}{2} w^T w + \sum_{i=1}^m \alpha_i (1 - d_i (w^T x_i - b)) \\
 & \nabla_w L = 0 \Rightarrow w - \sum_{i=1}^m \alpha_i d_i x_i = 0 \\
 & \Rightarrow w = \sum_{i=1}^m \alpha_i d_i x_i \\
 & \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^m \alpha_i d_i = 0 \\
 & \alpha_i \geq 0 \\
 & d_i (w^T x_i - b) \geq 1, \quad \alpha_i [1 - d_i (w^T x_i - b)] = 0
 \end{aligned}$$

The problem which we are having here that is the hard margin classifier which I am calling is P1 problem and P1 problem is basically minimization of $\frac{1}{2} w^T w$, subject to what are the constraints? Constraint is $d_i w^T x_i - b \geq 1$ and i is from 1 to m . This is this problem is basically hard margin classifier.

Now, let us write a Lagrange function of this. Lagrange will be a function of w , b and α 's. So, that is nothing but $\frac{1}{2} w^T w$ plus summation i from 1 to m because there are m number of constraints ok α_i it is $1 - d_i w^T x_i - b$ because, we have to write constraint in less than equal to format. So, that is why I put this side to right hand side that is why we are having this here.

So, this α_i is here are basically Lagrange multipliers. How many Lagrange multipliers we are having here? m ; α_1 , α_2 , up to α_m . Now, if I want to write the KKT

condition of this problem, so, how can I write KKT conditions? What are the variables here? Variables are w , b and α_i this we have to find.

So, you w , b and α_i sorry, it is w , b and α_i . So, first you differentiate with respect to w put it equal to 0. So, that implies if we differentiate this with respect to w . So, this is w only ok. Now, here if we differentiate with respect to w , so, that is nothing but minus summation i from 1 to m , ok.

It is $\alpha_i d_i x_i$ which is equal to 0. So, this implies w is equals to summation over i from 1 to m $\alpha_i d_i x_i$. So, this is a first KKT condition. Next is you can take $\frac{\partial L}{\partial b}$ equal to 0. So, this implies if you are differentiating with respect to b . So, with respect to b , if you differentiate, so, you will get summation i from 1 to m ; it is α_i , you are differentiating with respect to b .

So, that gives $\alpha_i d_i$ equal to 0. So, these are second KKT conditions. And of course, α_i should be non negative, Lagrange multipliers are non negative and next comes out from the feasibility condition; the feasibility condition is $d_i w^T x_i - b$ should be greater than equal to 1.

And next is $\alpha_i (1 - d_i)$ yeah $\alpha_i (1 - d_i) w^T x_i - b$ should be 0, for all i . So, these are the different these are the different KKT conditions which we are having here for this problem ok. Now if you write the dual of this problem P 1, then the dual we can use the KKT condition to write the duals and as I already said that if I instead of using this problem if I use the dual of this problem, then it will be computationally easy to solve the dual of the given problem P 1.

(Refer Slide Time: 10:40)

Lagrangian

The Lagrangian for the problem (P1) is given by:

$$L(w, b, \alpha) = \frac{1}{2} w^T w + \sum_{i=1}^m \alpha_i (1 - d_i (w^T x_i - b)) \quad (1)$$

where $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ and $\alpha_i \in \mathbb{R}^n$ are the Lagrange's multipliers.



So, what is the dual? So, first this is a Lagrange function which we have already defined ok.

(Refer Slide Time: 10:45)

KKT conditions:

The KKT conditions of the problem (P1) are given by:

$$\nabla_w L = 0 \quad - (2)$$

$$\frac{\partial L}{\partial b} = 0 \quad - (3)$$

$$1 - d_i(w^T x_i - b) \leq 0; \quad i = 1, 2, \dots, m \quad - (4)$$

$$\alpha_i(1 - d_i(w^T x_i - b)) = 0; \quad i = 1, 2, \dots, m \quad - (5)$$

$$\alpha_i \geq 0; \quad i = 1, 2, \dots, m \quad - (6)$$

Then it is then the KKT conditions of problem P 1 are this equal to 0, del L by del b equal to 0, that we have already discussed. The problem of the dual of the problem P 1 is given by; what is the dual now?

(Refer Slide Time: 10:55)

Dual Problem

The dual of problem (P1) is given by

$$\begin{aligned} \text{(P2) } & \text{Max } L(w, b, \alpha) \\ & \text{s/t } \nabla_w L(w, b, \alpha) = 0, \\ & \quad \frac{\partial L}{\partial b} = 0, \\ & \quad \alpha \geq 0 \end{aligned}$$

where $L(w, b, \alpha)$ is function as defined in (1). From KKT conditions (2) and (3), we have $w = \sum_{i=1}^m \alpha_i d_i x_i$ and $\sum_{i=1}^m \alpha_i d_i = 0$ respectively.



So, dual is given by this format. So, maximizing the Lagrange function, subject to gradient respect to w of L is $\frac{\partial L}{\partial w} = 0$ and $\frac{\partial L}{\partial b} = 0$ and α is greater than equal to 0.

(Refer Slide Time: 11:17)

The dual of the problem (P1) is given by

$$(P2) \quad \begin{cases} \text{Max} & L(w, b, \alpha) \\ \text{s.t} & \nabla_w L = 0 \Rightarrow w = \sum_{i=1}^m \alpha_i d_i x_i \\ & \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^m \alpha_i d_i = 0 \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, m \end{cases}$$

What this means? So, dual of the problem will be the dual of the problem P 1 is given by it is maximizing Lagrange function which is a function of w , b and α , subject to the KKT conditions; $\nabla_w L = 0$, $\frac{\partial L}{\partial b} = 0$ and α_i is non negative, for all i .

So, this problem I am calling as P 2. Now this condition I have already told you that this is nothing but, as we have already seen here ∇L respect to w is equal to 0 is nothing but w equal to this. So, we can write it here. It is nothing but, $w = \sum_{i=1}^m \alpha_i d_i x_i$. And this condition is nothing but $\sum_{i=1}^m \alpha_i d_i = 0$. That is from this constraint, we are getting this thing. So, what is the objective function now?

(Refer Slide Time: 12:47)

$$\begin{aligned}
 L(w, b, \alpha) &= \frac{1}{2} w^T w + \sum_{i=1}^m \alpha_i [1 - d_i (w^T x_i - b)] \\
 &= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j d_i d_j x_i^T x_j \\
 &\quad + \sum_i \alpha_i - \sum_i \sum_j \alpha_i \alpha_j d_i d_j x_i^T x_j \\
 &\quad + b \sum_{i=1}^m \alpha_i d_i \\
 &= -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j d_i d_j x_i^T x_j \\
 &\quad + \sum_i \alpha_i
 \end{aligned}$$

So, let us see the objective function. So, objective function is L, L which is w, b and alpha. It is nothing but, 1 by 2 w transpose w plus summation i from 1 to m ok. It is i from 1 to m alpha i 1 minus d i w transpose x i minus 1 ok. Now it is equal to 1 by 2. So, what is L? L we have defined here. It is alpha i 1 minus it is b ok.

So, what is w? w we have seen here. w is nothing but sum of sum from i from 1 to m alpha i d i x i. So, if it is w transpose w. So, that comes under double summation, summation i from 1 to m, summation j from 1 to m it is alpha i. So, for one w I am representing by index I, for another w I am representing by index j. So, it is alpha i alpha j as it is d i x i. So, it will be d i d j x i transpose x j. So, it looks little bit complicated, but it is not if you open the double summation.

So it looks it comes out to be a very simple expression plus this is summation α_i which comes here summation over i ok. Next term is minus it is summation again this w . You will replace w by this term and i is already running here. So, let us suppose that for that w it is index j . So, it is i then it is j it is α_i again α_j then it is $d_i d_j x_i$ transpose and this is x_j . And the last term is negative negative positive this is b will come out, this is summation i from 1 to m , this is $\alpha_i d_i$.

Now, this $\alpha_i d_i$ is the sum of $\alpha_i d_i$ from the second this constraint is 0. So, this will go to 0 this term will vanishes then this term plus half and minus 1 will be minus half. So, this is nothing but, minus 1 by 2 double summation over i summation over j $\alpha_i \alpha_j d_i d_j x_i$ transpose x_j and this is plus summation α_i over i .

So, this will be the objective function of the dual problem, subject to what are the conditions, these conditions we are having here. These are the these are the conditions ok.

(Refer Slide Time: 16:08)

Continued...

$$\begin{aligned}
 L(w, b, \alpha) &= \frac{1}{2} w^T w + \sum_{i=1}^m \alpha_i (1 - d_i (w^T x_i - b)) \\
 &= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j d_i d_j x_i^T x_j - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j d_i d_j x_i^T x_j + b \sum_{i=1}^m \alpha_i d_i + \sum_{i=1}^m \alpha_i \\
 &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j d_i d_j x_i^T x_j.
 \end{aligned}$$



So, here also we did the same thing and we obtain after simplification which we have discussed.

(Refer Slide Time: 16:13)

Thus, the dual problem (P2) becomes

$$\begin{aligned} \text{(P3) } \quad & \text{Max } \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j d_i d_j x_i^T x_j \\ & \text{s/t } \sum_{i=1}^m \alpha_i d_i = 0. \\ & \alpha_i \geq 0; i = 1, 2, \dots, m. \end{aligned}$$

The above problem has only one constraint apart from the non-negativity constraint. Also, the maximizing objective function is concave, thus, (P3) is computationally easier to solve. The optimal solution $\tilde{\alpha}$ of (P3) will give the values for \tilde{w} and \tilde{b} and thereafter the separating hyperplane $\tilde{w}^T x = \tilde{b}$ can be determined.



So, this will be the dual will be given by this problem. So, this constraint we have already used, you see here this constraint we have already used. So, this will be this is the only constraint which is remaining.

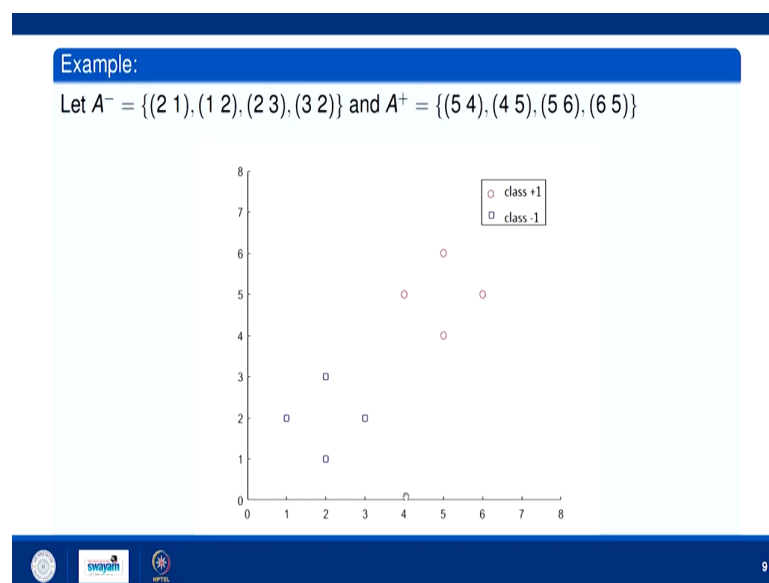
So, maximizing this function subject to this constraint and alpha i non negative. So, now, if we leave non negativity restriction; then instead of m number of constraint what we are having in the primal quadratic programming problem. Here we are having only 1 constraint. So, that is a main important application of duality theory here in this machine learning, here in this support vector machines ok.

So, computationally it reduces so much time and of course, memory also. So, the above problem has only one constraint apart from the non negativity constraint also the maximizing, here we are maximizing this function. So, maximizing a concave function and thus P 3 is

computationally easier to solve. The optimal solution α bar of P 3 will give the values of w bar and b bar and thereafter separating hyper plane w transpose x equal to b bar can be determined.

So, in this way, if you are interested to find out the optimal separating hyperplane which is w transpose x equal to b , so after solving this dual problem, we can find α bar and using α bar from this we can find w bar ok. And then from the other conditions, from the other conditions we can find b bar from which we can find w bar transpose x equal to b bar which is the optimal separating hyper plane. Now, let us discuss one example. So, here we are having two classes again; minus 1 and plus 1.

(Refer Slide Time: 18:08)



So, these are the points 2 1, see 2 1, this is 1 2. So, these square boxes are the patterns in minus 1 class and these circular boxes circular patterns are the patterns from plus 1 class. And

these of course, these are linearly separable. Now, if I want to find out the optimal separating hyper plane, so, we have to construct a convex quadratic programming problem. So, how can I construct a equivalent convex quadratic programming problem?

(Refer Slide Time: 18:48)

The slide shows the following handwritten mathematical formulation:

$$A^- = \{ (1, 2), (2, 1), (3, 2), (2, 3) \}$$

$$A^+ = \{ (5, 4), (4, 5), (5, 6), (6, 5) \}$$

$$\text{Min } \frac{1}{2} (w_1^2 + w_2^2)$$

$$\text{s.t. } \begin{aligned} &-(w_1 + 2w_2 - b) \geq 1, \\ &-(2w_1 + w_2 - b) \geq 1, \\ &-(3w_1 + 2w_2 - b) \geq 1, \\ &-(2w_1 + 3w_2 - b) \geq 1, \\ &(5w_1 + 4w_2 - b) \geq 1, \\ &(4w_1 + 5w_2 - b) \geq 1, \\ &(5w_1 + 6w_2 - b) \geq 1, \\ &(6w_1 + 5w_2 - b) \geq 1, \end{aligned}$$

w_1, w_2, b are unrestricted.

On the right side, a more compact notation is shown:

$$\text{Min } \frac{1}{2} w^T w$$

$$\text{s.t. } d_i (w^T x_i - b) \geq 1, \quad i = 1, 2, \dots, m$$

So, let us see. What are the patterns we are having here? So, patterns for A minus are patterns for A minus are 1 2, 2 1, 3 2 and 2 3. And patterns for A plus are; patterns for A plus are 5 4, 4 5 then it is 5 6, 6 5. So, these are two classes we are having here.

So, what is equivalent qpp that will be minimizing 1 by 2? So, all the patterns are in R 2. So, w will belongs to R 2. So, that will be w transpose w that means, w 1 square plus w 2 square, subject to what are the conditions. Conditions will be; so, condition we already know. See what hard margin classification problem we are having? 1 by 2 w transpose w subject to d i w transpose x i minus 1 minus b greater than equal to 1, i from 1 to m ok.

So if d_i is minus 1 because for this it is minus 1. So, it is negative of w_1 plus 2 w_2 , because first pattern first x_i ; x_1 is 1 and 2 and this w is w_1 w_2 . So, w_1 w_2 will multiply with 1 and 2. So, it will be w_1 plus 2 w_2 , it is a simple matrix multiplication and this is minus b greater than equal to 1. The second constraint will be minus of second pattern is 2 1 you substitute 2 1 here.

So, this will give 2 w_1 plus w_2 minus b greater than equal to 1, this is minus 3 w_1 plus 2 w_2 minus b greater than equal to 1, this is minus 2 w_1 plus 3 w_2 minus b greater than equal to 1. Now, come to the plus patterns, for plus 1 class. So, it will be d_i will be plus 1. So, that will be nothing but, 5 w_1 plus 4 w_2 minus b greater than equal to 1, 4 w_1 plus 5 w_2 minus b greater than equal to 1, then it is 5 w_1 plus 6 w_2 minus b greater than equal to 1, then it is 6 w_1 plus 5 w_2 minus b greater than equal to 1. And of course, w_1 , w_2 and b are unrestricted.

So, this is the this is the basically problem P 2, P 2 problem we are having now. So, we can solve either this problem or we can write the dual of this problem using this formulation; we can have the dual and we can solve either of the problem either this problem or the dual ok.

(Refer Slide Time: 22:16)

Optimization problem :

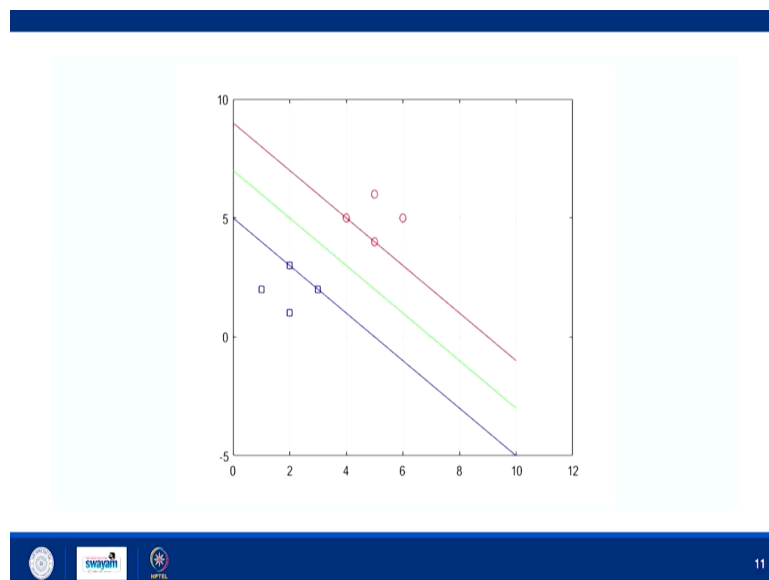
$$\begin{aligned} \min & \frac{1}{2}(w_1^2 + w_2^2) \\ \text{s.t. } & -2w_1 - w_2 + b \geq 1, \\ & -w_1 - 2w_2 + b \geq 1, \\ & -2w_1 - 3w_2 + b \geq 1, \\ & -3w_1 - 2w_2 + b \geq 1, \\ & 5w_1 + 4w_2 - b \geq 1, \\ & 4w_1 + 5w_2 - b \geq 1, \\ & 5w_1 + 6w_2 - b \geq 1, \\ & 6w_1 + 5w_2 - b \geq 1, \\ & w_1, w_2 \text{ and } b \text{ are unrestricted in sign.} \end{aligned}$$

After solving, we get $w_1 = w_2 = 0.5$ and $b = 3.5$. Hence, the equation of the plane will be $0.5x_1 + 0.5x_2 = 3.5$.

So, if you solve this problem; so, we will get these are the constraints. So, we will get you can use any solver to solve this problem. So, we will get w_1 equal to w_2 equal to 0.5 and b equal to 3.5. Hence, the equation of the plane will be $w^T x = b$, which is 0.5×1 plus 0.5×2 equal to 3.5. So, this is the this is basically the optimal separating hyper plane, where margin is maximum.



(Refer Slide Time: 22:42)



So, this plane is basically this plane is basically $x_1 + x_2 = 7$ ok which is the optimal separating hyper plane. So, we have seen that if you are having; if you are if you are interested to find out an optimal separating hyper plane, so, that the error minimizing lpp may not give the optimal separating hyper plane. If the two classes are linearly separable, then we can use hard margin classifier.

So, what this problem is? In this problem, we will try to maximize the margin by formulating an equivalent quadratic programming problem. That quadratic problem is also convex.

So, using KKT conditions and duality theory, we can also construct an equivalent dual of the given problem, given quadratic programming problem. So, the problem can be solved either by the formulating the qpp or by formulating its dual. It is always computationally easier to use duality approach because it reduces number of constraints significantly ok. So, in the next

lecture we will see that if they are not linearly separable then how can we find an optimization problem and how can we solve such problems; so.

Thank you.