Essential Mathematics for Machine Learning

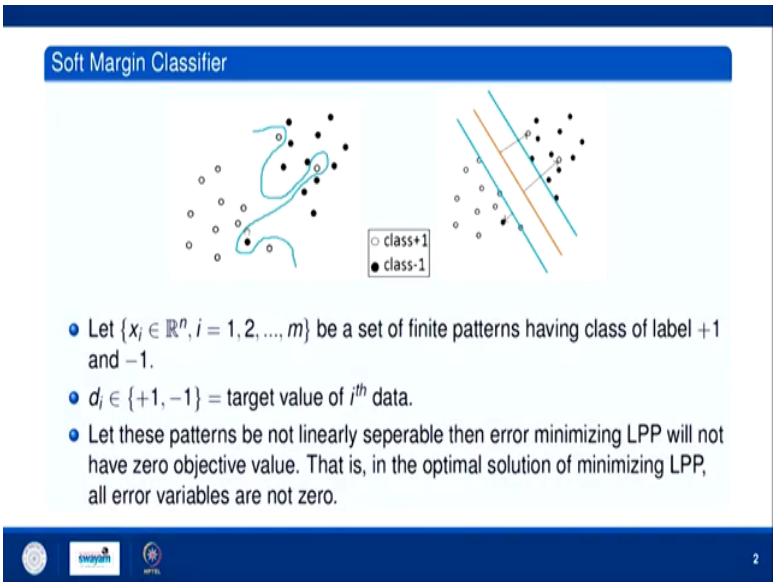
Prof. S. K. Gupta
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 40
Soft Margin Classifier

Hello friends. Welcome to lecture series on Essential Mathematics for Machine Learning. In the last lecture, we have seen that how we can find a hard class hard margin classifier to find out an optimal separating hyperplane; that can be obtained by constructing an a convex quality programming problem and that problem can also be solved using duality. Now if the two classes are not linearly separable, then how can you find out an optimal separating hyperplane, at least which is best suited.

If they are not linearly separable, at least we can find out a hyperplane which is best suited.

(Refer Slide Time: 01:11)

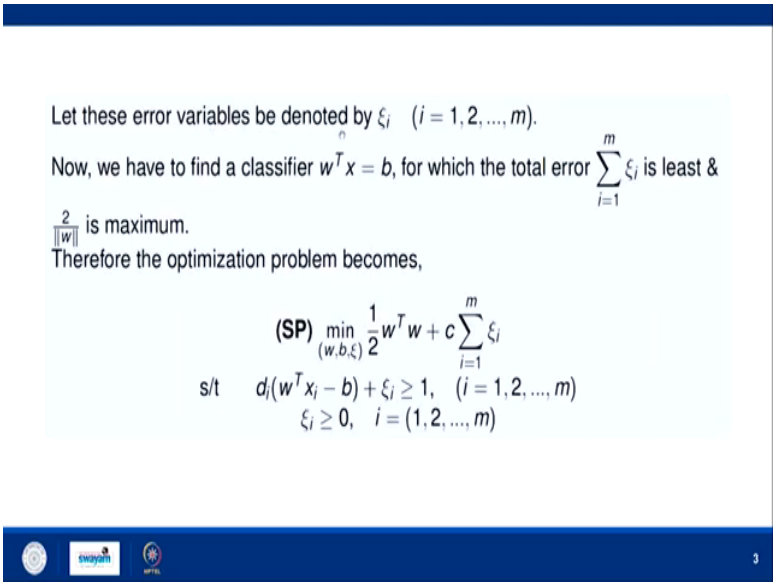


So, how can we find that? Now we have two classes. See in this figure you are having two classes; this is plus 1 class which is this hollow circles and which is a dark circles these are of minus 1 class ok. Now 1 pattern from this class is in this class and 1 pattern from this class is in this class suppose. So, of course, these two classes are not linearly separable. So, of course, they are actually non-linearly separable.

But if we are interested to find out a hyperplane the best fitted hyperplane which can separate these two. So, how can you find? So, in this case we will find misclassification error. We will find misclassification error corresponding to all the patterns ok. All the patterns of plus 1 class and minus 1 class. The patterns which are correctly classified for those patterns misclassification error will automatically be 0 and those patterns which have not correctly classified or which are misclassified for that we will try to minimize the misclassification error.

So, now our objective are 2, 2 are our objectives; the one is we have to maximize the margin which is nothing but minimization of norms norm w square by a 2 or 1 by 2 w transpose w or second is we have to minimize the sum of misclassification errors ok. So, we have two objectives, now how can we model corresponding optimization model? So, let us see. So, suppose a patterns are not linearly separable then the error minimizing LPP will not have a zero objective value that is in the optimal solution of the minimizing LPP all error variables are not zero.

(Refer Slide Time: 03:07)



Let these error variables we denoted by xi i ok. i from 1 to m for all the patterns. Now we have to find out a classifier for with a total error is least which is a sum of xi i and the margin is maximum. So, the now the equivalent optimization problem will be 1 by 2 w transpose w which is the minimization of the minimization of norm square w upon 2 plus c times summation i from 1 to m xi i subject to this constraint. We are adding xi i here in the hard

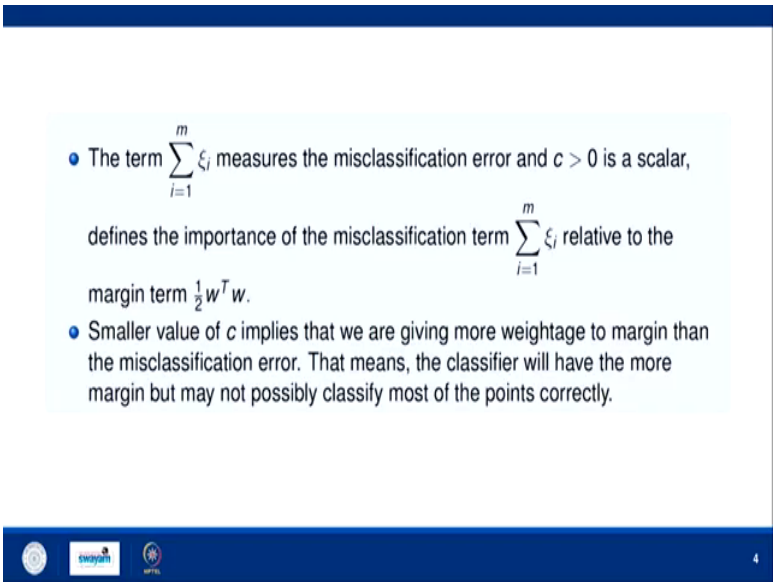
margin classifier, we are having we are not dealing with xi i. Because all misclassification errors are 0 there they are linearly separable.

So, we are adding xi i and that is greater than equal to 1. So, and xi i are non negative, i from 1 to m. Now what is the c? This c a free parameter which is greater than 0 is a scalar and. So, it is it is greater than 0. So, what is the importance of writing the c? See here we are having two objectives; the one objective is to maximize the margin, the other objective is to minimize the misclassification error the sum of misclassification error. Basically it is bi objective.

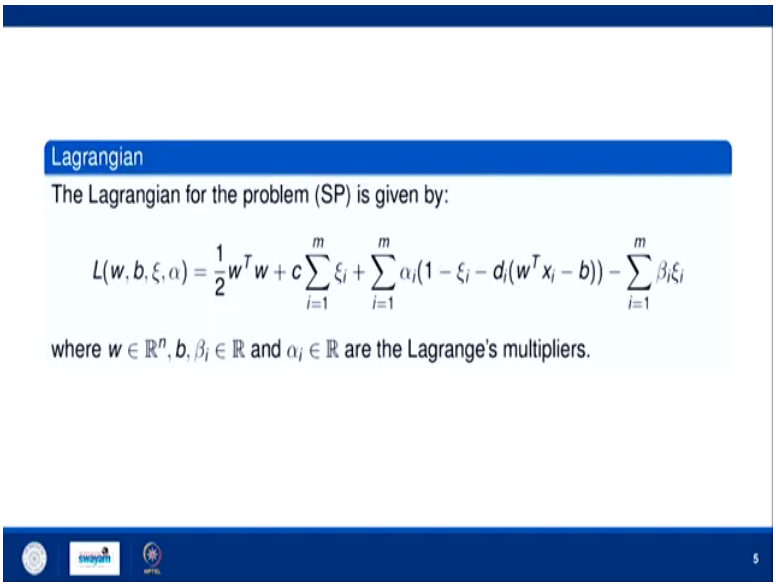
So, we are trying to make a single objective problem by adding this c. If this weightage if this c is very small is small; that means, we are giving less preference to the misclassification error and we are giving more preference to the margin; that means, the margin will be maximum, but the patterns may not be all the patterns may not be correctly classified. However, if we are giving a c a very hard if you are giving c a large value; that means, the larger value of c implies that we are giving more weightage to the misclassification error then giving weightage to margin.

So, basically it depends at which c we take ok. It depends on a decision maker. So, usually we take c as 1. So, that we give equal weightage to both the objectives margin as well as misclassification error.

(Refer Slide Time: 05:26)

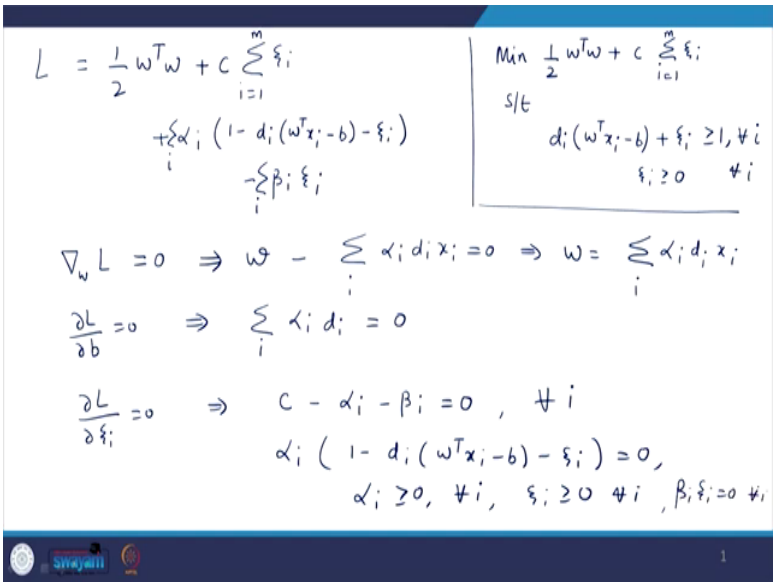


(Refer Slide Time: 05:30)



So, this I have already discussed in the importance and the implication of c. Now what is a Lagrangian of this function this problem? As we see in a hard margin classifier we can construct a problem and we can solve that problem using KKT conditions or using finding that duel of the given problem. In the same way here in soft margin classifier also we can find out the Lagrange function hence the KKT conditions and hence we can construct the duel of the problem which will make the problem computationally easy.

(Refer Slide Time: 06:04)



So, what is the Lagrange of this problem. Let us write the problem first, what is the problem soft margin classifier what is the problem? It is minimization of 1 by w 1 by 2 w transpose w plus c times summation i from 1 to m xi i ok. Subject to what are the conditions? Conditions are d i w transpose x i minus 1 plus xi i is greater than or equal to 1 and xi i non negative for all i and for all i. So, this is our problem ok.

Now first let us define the Lagrange of this problem. The Lagrange of this function will be this problem will be given by 1 by 2 w transpose w plus c times summation i from 1 to m xi i then it is a plus alpha i the sum of alpha i over i.

It is 1 minus d i w transpose x i minus 1 minus xi i. And it is minus beta i xi i it must be sum, sum over i ok. So, now, we can write it is di w transpose x i it is a minus b, it is minus b as we

have seen here it is minus b. So, minus b will come here ok. Now this is the Lagrange of this problem.

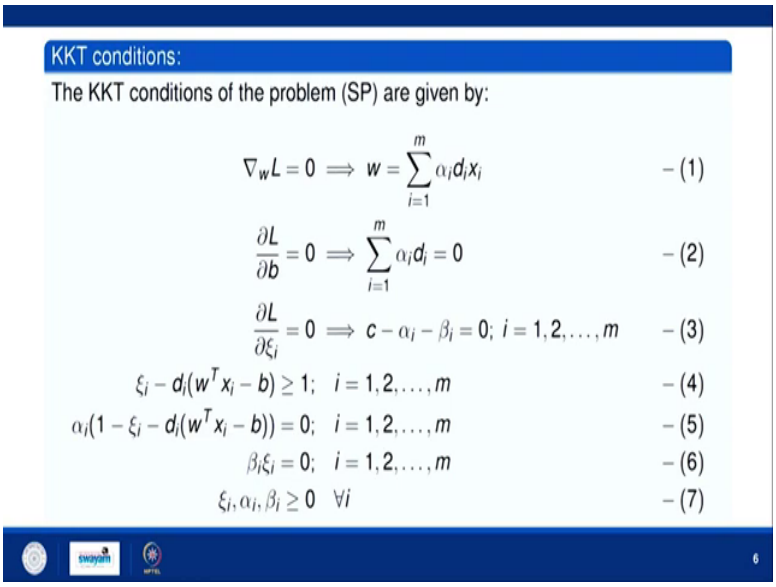
So, we will write the KKT conditions as we first take gradient of w l equal to 0 and this implies if we take the gradient respect to w. So, what this what this gives? It is w from here we get w no term of w so, it is 0. So, from here we will get minus summation over i alpha i d i w equal to 0.

So, this implies w equal to summation over i. So, it is we are differentiating with respect to w. So, it is x i. Sorry, it is a x i. So, it is alpha i d i x i. So, this is the first KKT condition the second is del l by del b equal to 0. So, this implies when you differentiate with respect to b. So, that will be summation alpha i bi alpha i di it is di over i is equal to 0. So, this is a second condition.

Now we differentiate with respect to xi i del l upon del xi i equal to 0. So, that implies c minus alpha i minus beta i equal to 0. So, these are the conditions for all i of course, and of course, this alpha i times 1 minus d i w transpose x i minus b minus xi i is equal to 0 alpha i must be non negative for all i and xi i must be non negative for all i.

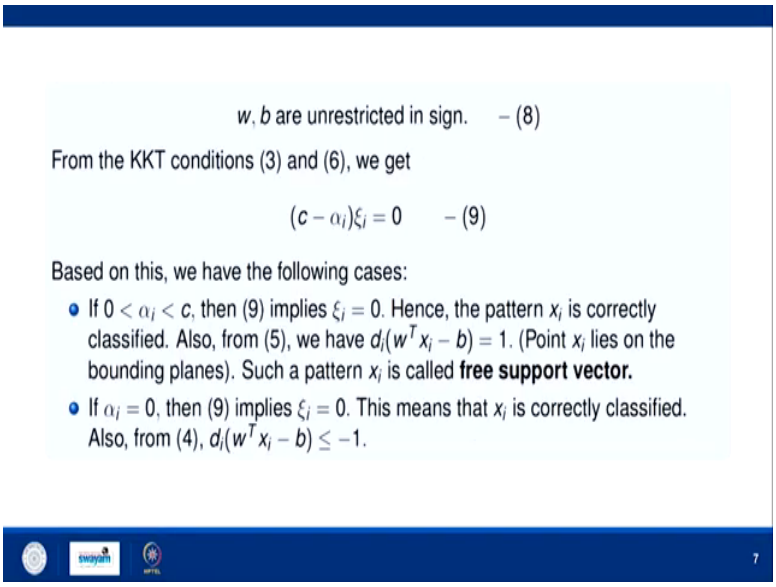
So, and of course, a feasibility conditions must be maintained. So, these are the various KKT conditions which we can obtained using the Lagrange function of the given problem.

(Refer Slide Time: 09:43)



So, these are the KKT conditions which I have just obtained. So, one more condition it is beta i xi i equal to 0 which also which can also be obtained; because beta is a Lagrange multiplier corresponding to this constraint xi i. So, beta i xi i must be 0 for all i plus feasibility conditions.

(Refer Slide Time: 10:05)



Now, now let us try to analyze these conditions c. If you see 3 and 6; it is c minus alpha i minus beta equal to 0 ok. So, from here it is ci minus alpha i times xi i equal to 0. So, let us see how we are obtaining this condition. So, from this it is ah. So, this is easy to obtain you see what is beta i from here beta i is c i, c minus alpha i.

So, you can substitute beta i here. So, it is c minus alpha i times xi i equal to 0. So, we will get this condition. Now from this condition; the following cases we may obtained what are the cases. Now first of all since alpha i is are non negative alpha i s and beta i s are non negative so; that means, this alpha i is alpha i is between 0 and c only. Because beta i is non negative alpha i is non negative so; that means, this alpha i is between 0 and c only it cannot be more than c ok. If it is more than c so, then beta i will become negative, but it is non negative ok.

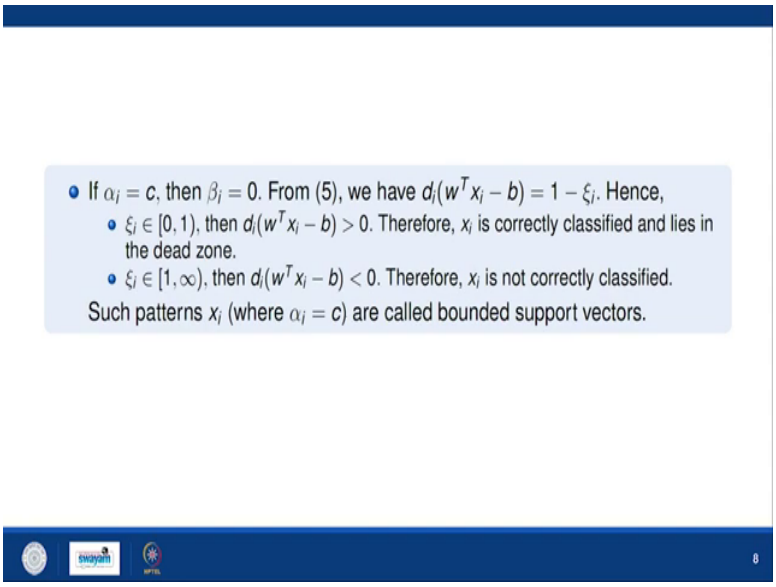
So, the first condition which may arise that this alpha i is between 0 and c. If alpha i is between 0 and c; that means, this term is nonzero this term is non zero; that means, xi i equal to 0. If xi i equal to 0; that means, that pattern is correctly classified. If you come to question number 5 from this equation number 5; this alpha i is nonzero xi i equal to 0 so; that means, that means 1 is equal to this 1 equal to di w transpose x i minus b; that means, that particular pattern belongs to the bounding that belongs to the a hyperplane. That means, point x i lies on the bounding planes ok.

Since it lies on the bounding planes that such patterns are called free support vector ok. So, what I want to say that if alpha i is between 0 and c, then the pattern is correctly classified and that pattern will lie on the bounding plane itself and such patterns are called free support vectors.

Now the second case suppose alpha is equal to 0. If alpha equal to 0 again c is non zero. So, xi i will be 0. So, xi i will be 0; that means, that means the pattern is correctly classified ok. Now from 4, now go come to equation number 4. So, from this equation 4 xi i 0 so; that means, this is less than equal to minus 1.

So, that does not lie on the bounding hyperplane, but lies on the correct side. If it is on the positive pattern lie on the positive side, if it is negative pattern lies on the negative side. So, if alpha equal to 0; that means, this condition is satisfied. Here equality is satisfied; however, here it is less than equal to ok. Now the third case left when alpha i is c itself it may be c also, but it cannot be more than c. If alpha i equal to c then we are having 2 cases based on xi i.

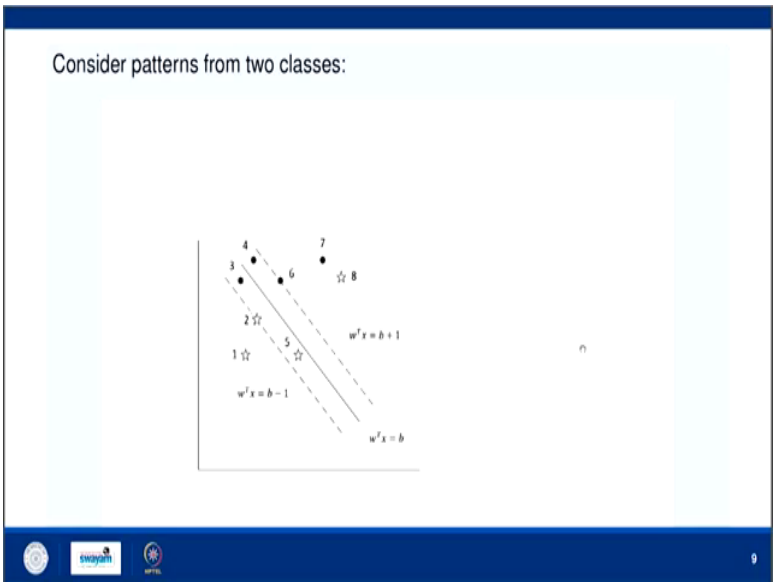
(Refer Slide Time: 13:27)



So, if alpha i equal to c then beta i will be 0 from here you can see if alpha i equal to c if alpha i equal to c then from this equation, then from this equation beta i will be 0 ok. And from 5 what we obtained from this equation 5 what we obtain. So, alpha i is c. So, which is non zero. So, from this equation we will obtain that this expression is nothing, but 1 minus xi i. So, this expression is nothing, but 1 minus xi i.

Now, if this xi i is between 0 and 1. So, what does it mean? If it is between 0 and 1 this means this is positive. This is positive means; this is positive means the pattern is correctly classified because, it satisfy this inequality so; that means, the pattern is correctly classified and lies in the dead zone ok. But if it is more than 1 so, this is negative this is negative means this inequality reversed; that means, the pattern is not correctly classified ok. So, such pattern x i where alpha equal to c these are called bounded support vectors ok.

(Refer Slide Time: 14:45)



Now, let us try to understand this by this figure we are having here two classes; one is denoted by star and other is other are denoted by a bold circles.

So, we are having 1, 2, 3, 4, 5, 6, 7, 8 patterns 4 from pattern of the star and from 4 from pattern of bold circles. Now suppose a pattern suppose they are studying this pattern 1. Now for this pattern 1; since this lies on the correct side because this we are denoting by this side we are denoting by a star. I mean; suppose it is plus 1 class and suppose or suppose it is minus 1 class and this side is suppose plus 1 class. So, the patterns of the star are basically minus 1 class and patterns of bold circles are from plus 1 class.

So, if it is if we are taken the first point so; that means, it is correctly classified ok. And correct correctly classified so; that means, xi 1 equal to 0. So, xi for this pattern will be 0 and what we can say about alpha for this pattern. See here alpha for this pattern will be 0. If it lies

on the bounding planes then alpha will be between 0 and c if it is correctly classified then alpha will be 0. So, here for this alpha will be alpha will be 0 and xi will be also be 0 and this is nothing but inequality hold ok.

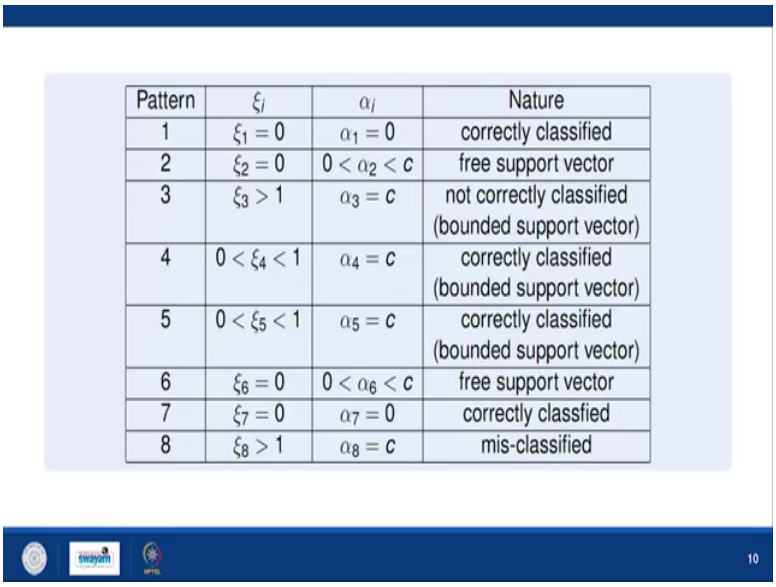
Now, if you come to pattern number 2. So, pattern number 2 of course, it is correctly classified. So, xi 2 will be 0 and what about alpha 2 alpha 2 lie between 0 and c, because it lies on the bounding hyperplane it is nothing but free support vector. Now if you go to pattern number 3 pattern number 3 is for plus 1 class and lies on other side ok. So, this is not correctly classified. For this xi will be greater than 1 and alpha will be c. If you go to pattern number 4. So, pattern number 4 this is in this side. So, for this is basically xi i will be between 0 and 1 more than equal to 0 less than 1 and alpha, alpha i is basically c for this particular pattern.

Now, if you come to pattern number 5. Now pattern number 5 is also in that zone, but it is from it is in this side of this plane this plane this side so; that means, alpha is c for this and xi is between 0 and 1 ok. Now if you come to 6; for 6 it is correctly classified lies on the bonding plane so; that means, alpha i will be 0 and c will be 0 alpha will be 0 and the xi will be 0 for this and this is nothing but free support vector.

And if you come to pattern number 7. For pattern number seven this is a correctly classified so, but it is not on the boundary. So, alpha will be c sorry alpha will be 0 alpha will be 0 and xi will be 0.

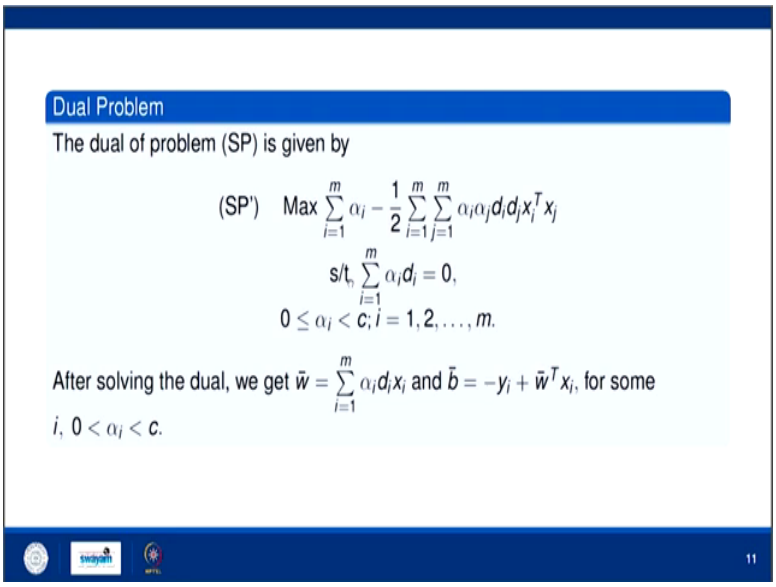
Now, for this pattern for this pattern this belongs to basically this class minus 1 class, but it is here. So, for this xi, xi will be greater than 1 and alpha will be c. So, in this way we can analyze we can analyze if you are having a 2 patterns that by a simply seeing a pattern we can see that it is a free support vector or it is correctly classified or it is misclassified.

(Refer Slide Time: 18:37)



So, this is basically the analysis of whatever I have discussed here.

(Refer Slide Time: 18:41)



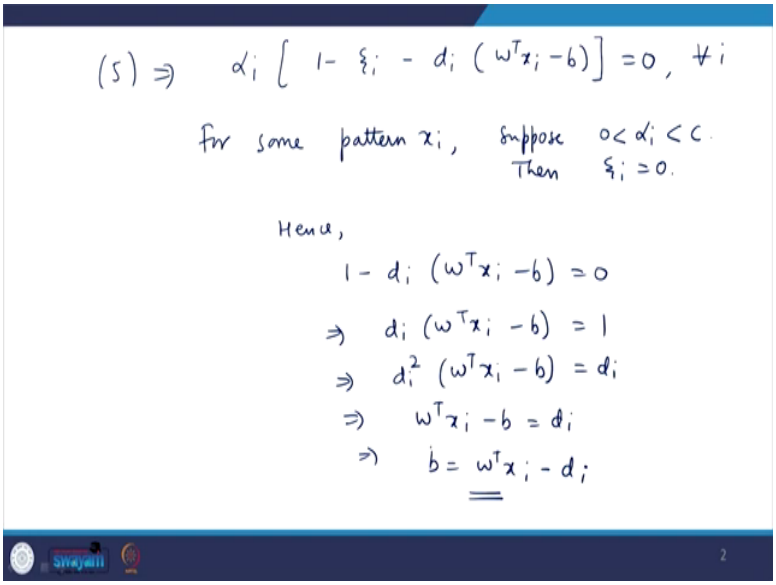
Now, the dual of the problem as I already told you the duel is simply maximizing the Lagrange function ok, subject to the constraints.

So, constraint basically come from here. This is a constraint ok. Now this is Lagrange; maximizing this function Lagrange. So, you simply substitute w of this given by this expression here ok. So, what we obtain this is ah two ws are here. So, there will be two summation; one for i other for j and here also the two summations. Now plus half minus 1 will give minus half as we have in hard margin classifier, the same concept will work here.

So, on simplifying this expression after substituting all these after using all these expressions; we get the dual as this problem. So, again it is easy to use the dual problem computationally easy to use a dual problem rather than using the primal problem as well.

So, after solving this dual problem we will obtain w bar and b bar. So, w bar can be obtained from this equation. You can see from this equation w bar can be obtained once you obtain alpha bar. So, w bar you can obtain from this equation. Now how to obtain b bar? So, b bar can be obtained if you go to this KKT conditions see KKT condition number 5.

(Refer Slide Time: 20:09)



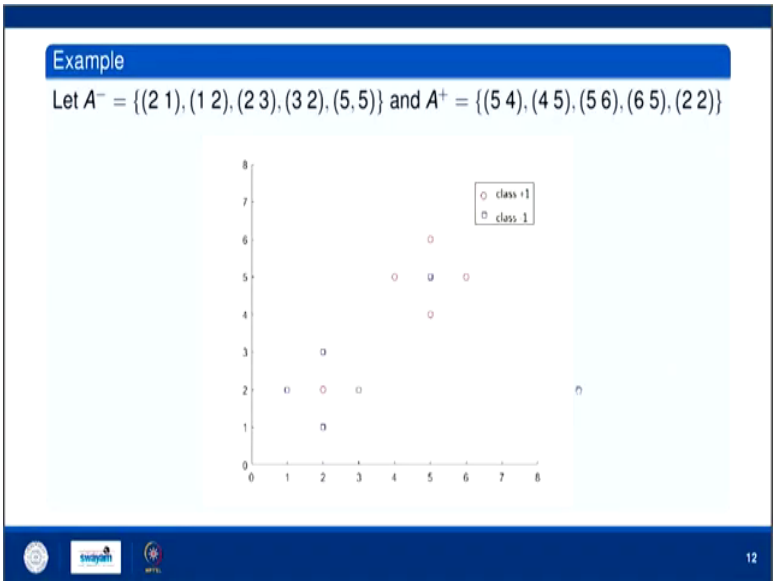
So, if you see the KKT condition number 5, from 5; what we obtain 5 implies alpha i times 1 minus xi i minus d i w transpose x i minus b is equal to 0 for all i ok. This is by this equation number 5. So, for some pattern x i for some pattern x i suppose alpha i is between 0 and c.

Then of course, xi i will be 0 as we have already discussed. So, xi i will be 0 and alpha i between 0 and c. So, this expression this expression implies. So, hence it is 1 minus d i w transpose x i minus b will be equal to 0. So, that means; di w transpose x i minus b equal to 1.

You multiply di both the side. So, it is di square w transpose x i minus b equal to di. Now di is either plus 1 or minus 1.

So, di square will be always plus 1. So, this implies w transpose x i minus b equal to di. So, this implies b is nothing but w transpose x i minus di for that pattern for which alpha i is between 0 and. So, in this way we can find b bar. So, this is an optimal b or b bar ok. So, we can find we can find b bar from here.

(Refer Slide Time: 21:54)



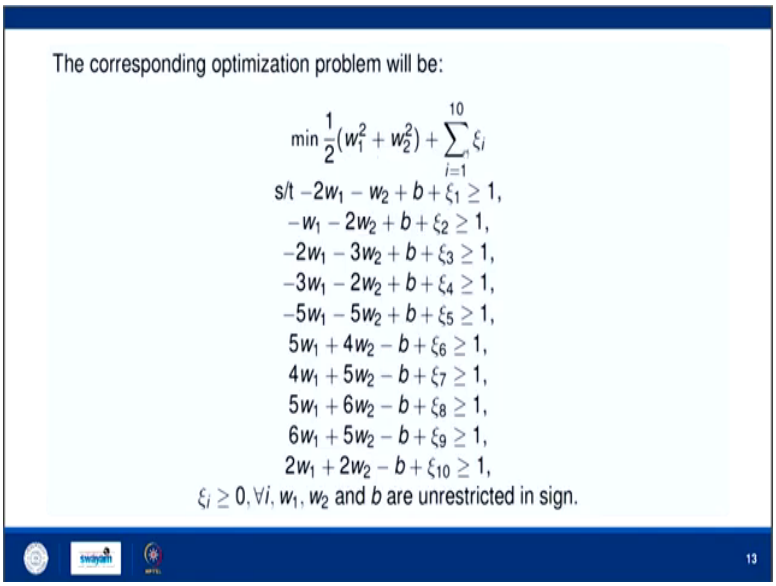
Now, let us discuss one example based on this. Now we are having two classes here; minus 1 and plus 1 ok. We are having 1, 2, 3, 4, 5 patterns of this square 5 patterns of minus 1 class and 5 patterns of plus 1 class. So, of course, these two classes are not linearly separable. So, how can we find out soft margin classifier? So, we can construct an equivalent quadratic

optimization problem which will not only maximize the margin, but also minimize the misclassification error.

So, we will use the help of the same optimization model will which we have discussed here this SP model. So, we will construct the SP model for this given example; we will solve it and find out the best possible I mean soft margin classifier ok.

So, how we can formulate? So, formulation is quite easy. What will the objective function? Patterns are in r 2. So, it will be 1 by 2; 1 by 2 times w 1 square plus 2 square plus c times sum of misclassification errors. How many patterns 1, 2, 3, 4, 5, 6, 7, 8, 9 10. So, there will be 10 number of 10 xi s; xi 1, xi 2 up to xi 10.

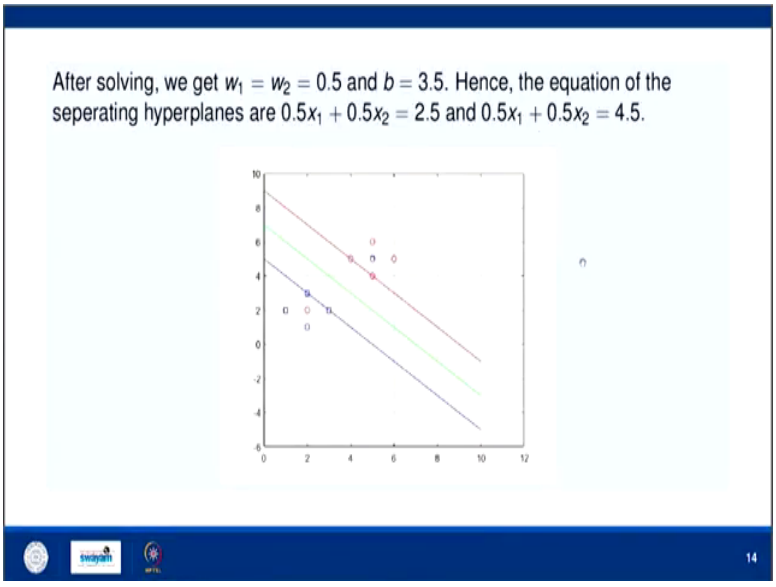
(Refer Slide Time: 23:26)



So, that will be the objective function of this problem. I have taken c equal to 1. I am giving weightage equal weightage to both the objectives ok. So, this I have taken. Now of course, these patterns are correctly classified for these patterns xi i automatically come out to be 0, not for this and pattern not for this pattern. And of course, for these 4 patterns also xi will automatically comes out to be 0, because they are correctly classified ok.

Now, what is what is the, what are the constraints? For constraints suppose we are having the first a pattern 2 and 1. So, it will be a minus of minus of 2 w 1 minus w 2 plus b plus xi 1 greater than equal to 1 from the constraint of SP model. And similarly we can construct all the 10 constraint of this problem ok. Where, xi i are non negative for all i w 1 w 2 and b are unrestricted in sign. So, this is the equivalent SP model or the soft margin classifier problem of the given numerical example.

(Refer Slide Time: 24:43)



So, after solving is this problem by using any solver. We will we get we get back to this as an 0.5 times x 1 plus 0.5 times x 2 equal to equal to 3.5; which is the which is the class soft margin classifier we are we are having.

These two are the bounding planes this is point 2.5 right hand side minus 1, b minus 1 this is b plus 1. These two are the bounding planes. So, in this way we can find a soft margin classifier of the problem. Not only this is the problem which I have used SP model as such we can also use the dual approach the dual model. This is a dual model which we have formulated which we can also use the dual model to find out the corresponding soft margin classifier.

So, in this way we have seen that if that if the patterns are not linearly separable, then also we can find out a classifier which not only maximize the margin, but also minimize the minimize the misclassification error. The model we can simply formulate a quality programming problem or we can use the KKT conditions to find out the dual of the given soft margin classifier problem.

Thank you.