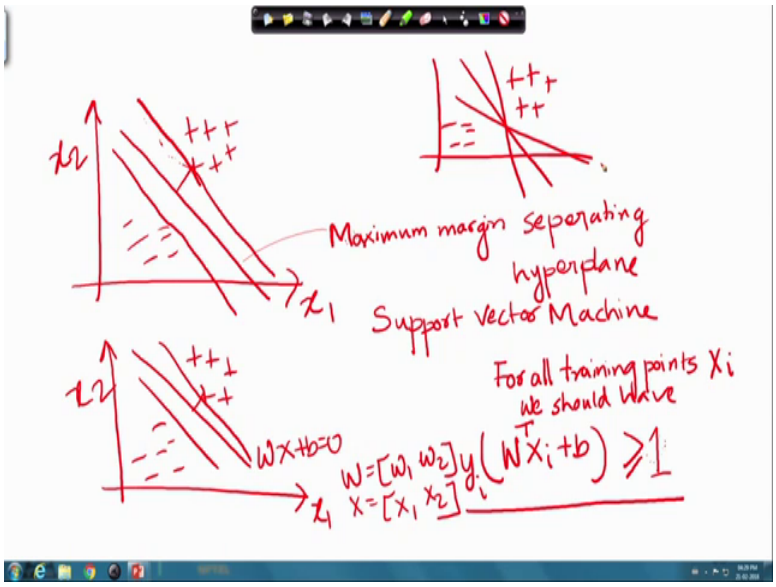
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 27
Kernel Machines

Welcome, to the final lecture on support vector machines. What we did previously was to coming back to the same figure.

(Refer Slide Time: 00:28)

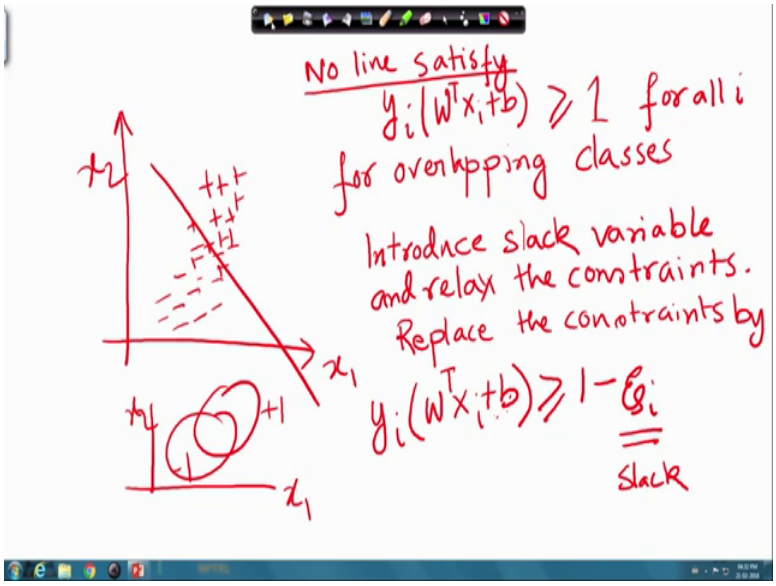


We will find a maximum margin separating hyper plane and what we did was, sorry, that is what we did was to found out a line and which has the highest margin of separation with the closest point from both classes and this boundary we called as a maximum margin hyperplane or a support vector machine.

And, the way we started was we initially sort of put the constant that for given W X plus b equals to 0, where W is a vector and X is this x 2 vector. We started was that or X i should have into y i to be greater than equal to 0. So, the plus points are in plus side minus points are in minus side. In fact, we sort of said that this is equivalent to scaling W and b, so, that we have the minimum value to be not just 0, but also 1.

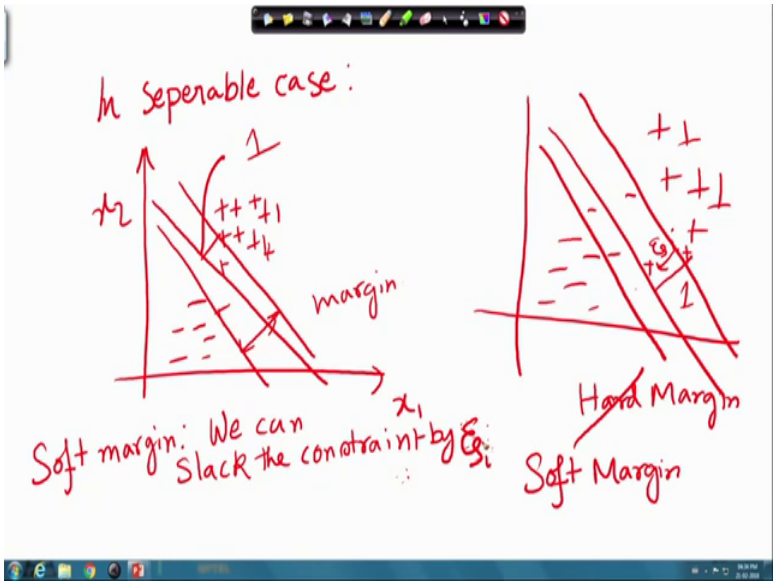
So, this was our constant and then we maximized margin given satisfying that constant. So, we actually observe that there will be if we have classes like this there will be infinite lines we pick up the one with the highest margin.

(Refer Slide Time: 04:12)



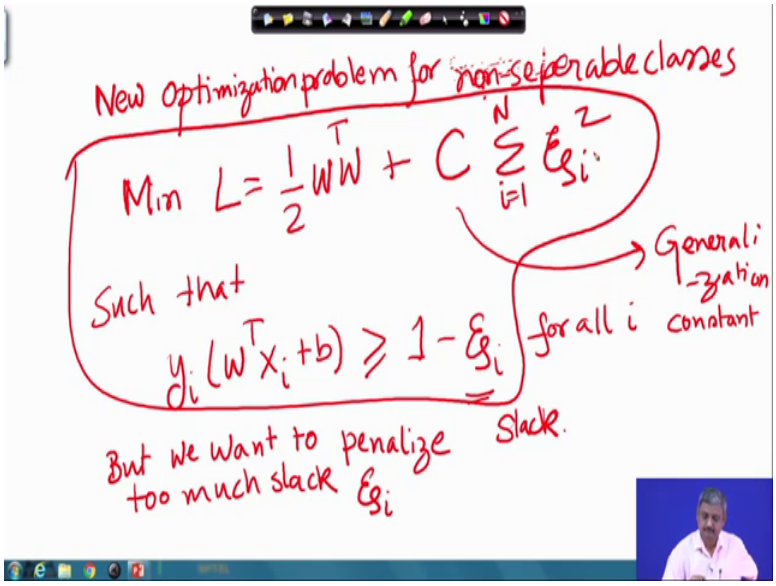
But, suppose the case is like this, instead there is an overlap and you cannot find any line which will satisfy for overlapping classes. So, you have like this and I try them. So, what to do? So, what I do is introduce some slack variable and relax the constraints replace them by some slack enjoy life some relaxation.

(Refer Slide Time: 06:51)



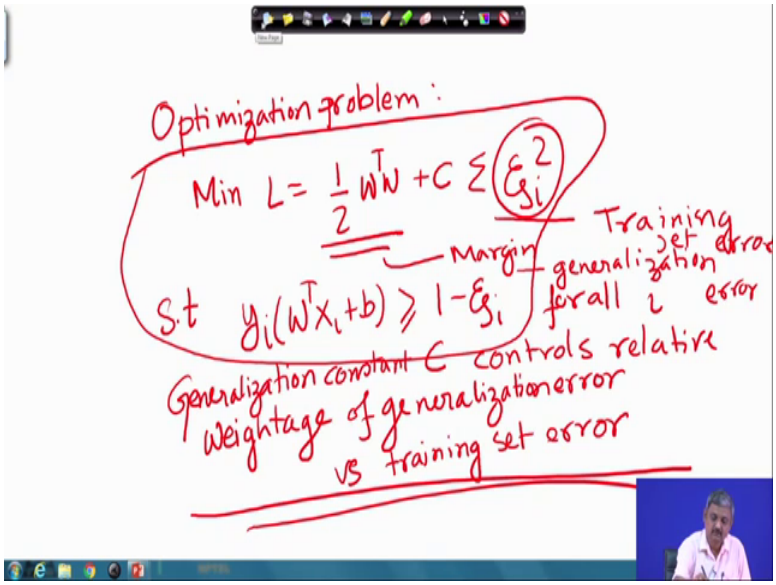
So, what is the meaning of this? These were the margin. In the closest point was at a distance of 1, not distance and a value upon. Now, inside the margin we can have points this I draw clearly, this was the hard margin. So, this point is at 1, now have 1 minus xi i, so there is a point inside also. Some points can be inside the margin it is no longer it in no man's land. So, this is called a soft margin slack by some psi I let me properly write it xi i allow slack.

(Refer Slide Time: 09:15)



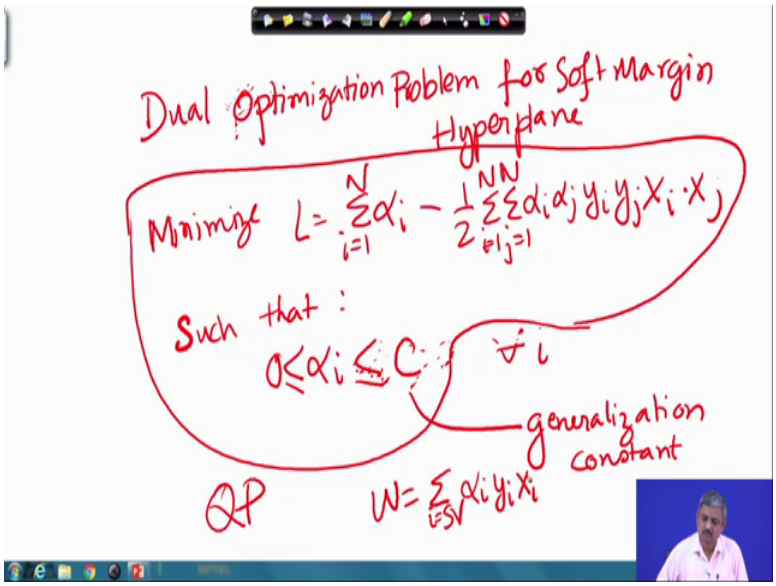
So, if I rewrite my optimization problem it becomes, sorry, such that not 1, but 1 minus xi i. So, slack, but we and do not want to have too much xi i. So, add it as a note that every point x i will have a slack xi i when you watch this and multiply these two factor by a constant C. So, C, this C I called constant it controls. So, this is clear. So, this is my new optimization problem, new constant, new objective function constant I add slack objective function I penalize slack.

(Refer Slide Time: 12:08)



So, if I look, if I write again, this is the margin 1 by this seller. So, it controls generalization error and this is the training set error. This is the error controls error it does that all right. So, let us try to solve this.

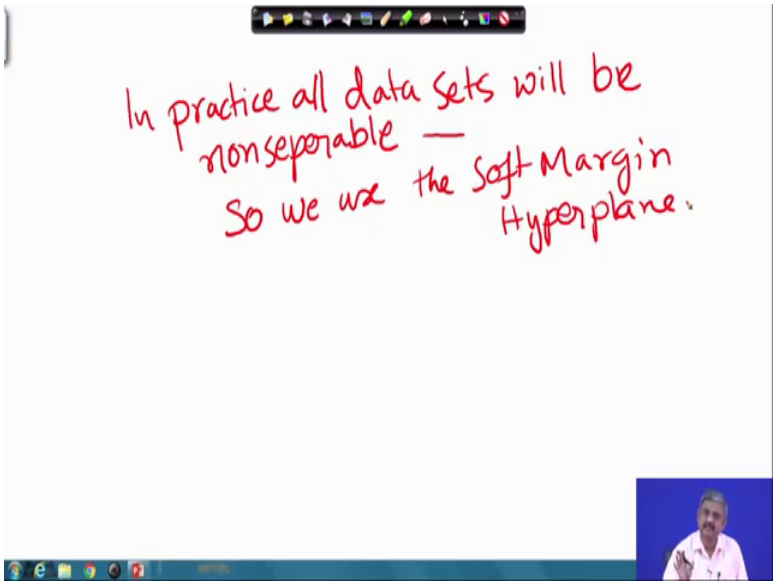
(Refer Slide Time: 14:24)



The dual problem looks like this. If you transform we will see with all the operation that we do earlier same form alpha i at the Lagrange multipliers, same only constant changes. Earlier, we had the constant alpha i greater than or equal to 0, for all i. Now, we have the constant alpha greater than or equal to 0 and less than the generalization constant. So,

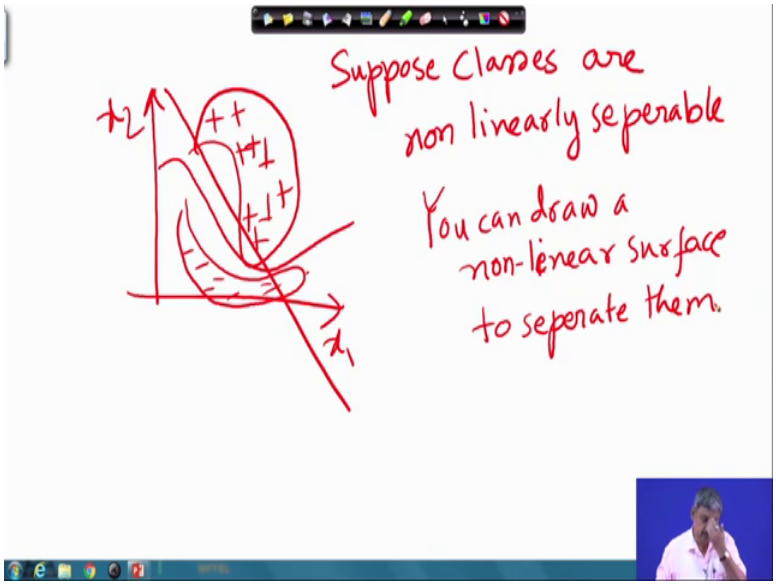
you solve it in the for all i it is, solve it in the same way as earlier quadratic programming you get your support vectors you still get X i, you still get that.

(Refer Slide Time: 16:57)



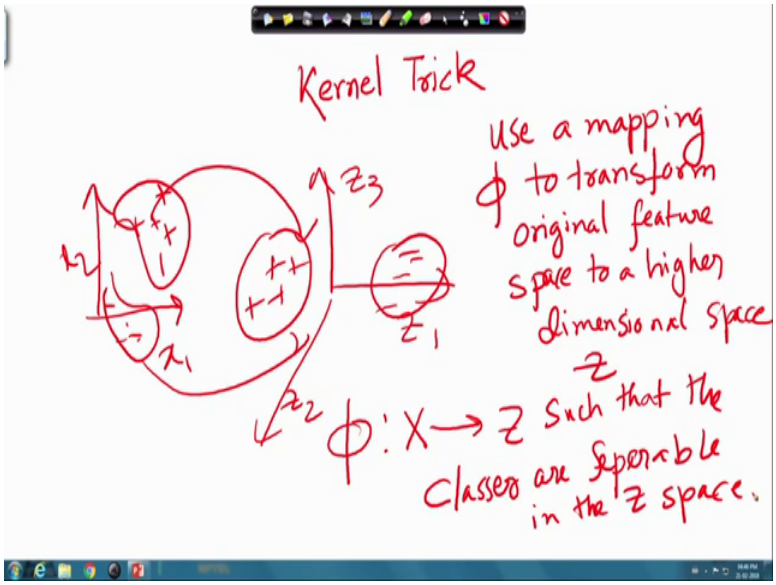
In practice, we use the soft margin hyperbola same process; see everything remains the same except for the constant of the QP changes when an upper bound on alpha is at there that is the only difference exactly same way right. So, all the softwares you use for SVM will have this soft margin actually not the hard margin. So, you have to choose a value of C.

(Refer Slide Time: 17:54)



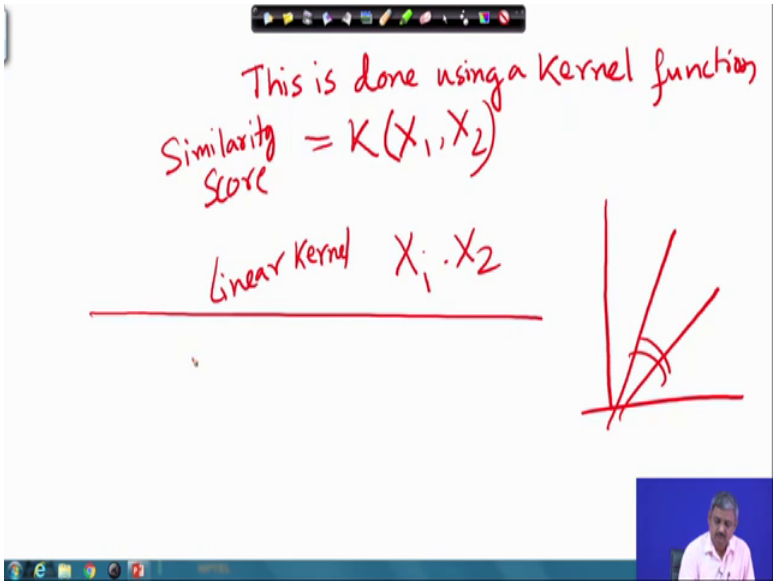
Now, let us come to another case. Now, let us come suppose, so, you cannot draw a line, but you can draw a non-linear surface to separate them. So, how to solve this problem?

(Refer Slide Time: 19:07)



The problem is solved by something known as a Kernel trick. So, what is the trick? So, phi transforms Z, such that the classes are separable in the Z space. So, you map it to high dimension and then draw a line as it is, but not on X, but on Z.

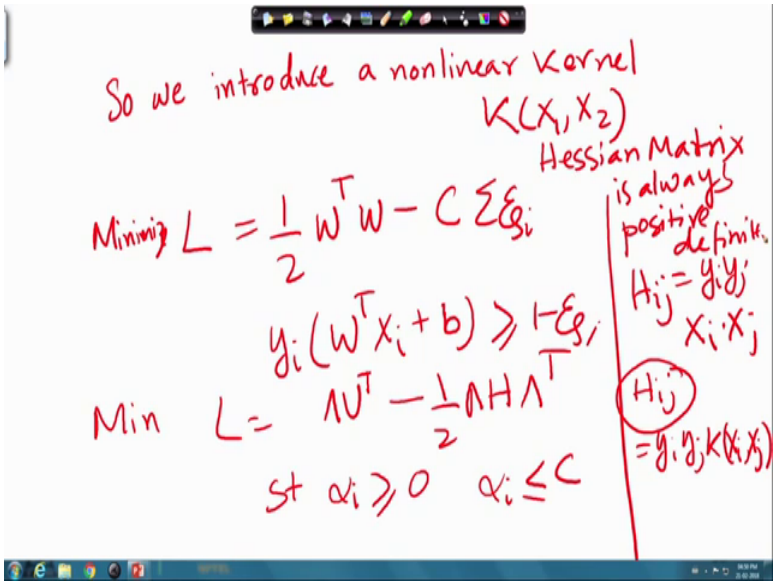
(Refer Slide Time: 21:15)



Done using a kernel function, what the kernel function does is, that it takes two vector X 1 and X 2 and returns a similarity score between them. So, the dot product it takes two is

an example takes two vectors it turns the cosine of the angle between them that is called the linear kernel.

(Refer Slide Time: 22:20)

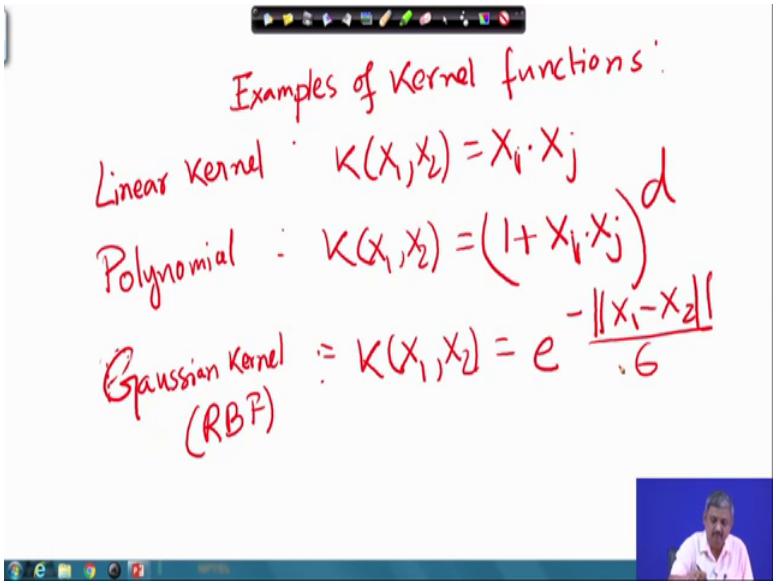


So, what happens is that, let me draw this kernel. I am actually not going into a theory there is a nice mathematical theory maybe if you ask question in the forum I will give you the details because this is I am just giving the overview. So, if you remember what we did our minimize the dual form was if you remember again minimize you took a vector of alphas minus if you remember use the Hessian matrix H and C.

So, what does the Hessian matrix H ij was y i y j times X i dot X j dot product linear kernel. We will just generalize this. Instead of X i dot X j we write y i y j; K of X i, X j, kernel of X i, X j it takes two vector returns a scalar, positive scalar, positive definite it should always be positive definite. So, one property is that this Hessian matrix should always be something called a positive definite matrix is always positive definite.

So, you just formerly say, when I was plug in that value of Hessian matrix and rest is as it is same W equal to summation alpha i X i for the support vector same thing. It does all everything.

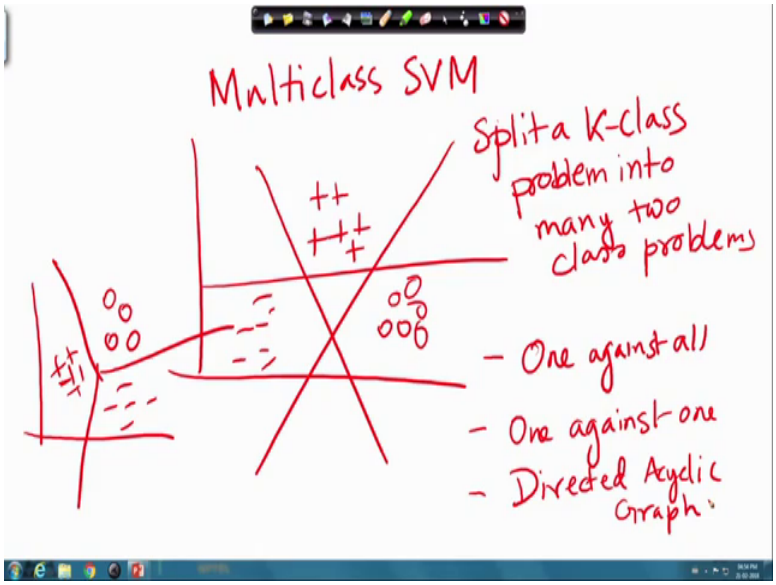
(Refer Slide Time: 24:54)



And, you have several examples of kernel function linear kernel. So, two property a kernel function should study. So, you should return a scalar and it return some value positive value X 1, X 2 are two vectors. X 1 it is X 1 dot X j. Polynomial kernel; X 1, X 2, X 1 dot d, d th order Gaussian kernel or sometimes known as the RBF kernel Radial Basis Function kernel is that sigma. Many other kernel even for discreet other cases are proposed.

So, you can choose your kernel accordingly and just use the same support. So, when you use support vector machine software you have to specify the value of generalization of constant C and what type of content you want to use, that is all. So, it will just plug in solve QP everything give you the support vectors give you W everything at the castle (Refer Time: 26:48).

(Refer Slide Time: 26:55)



Just one more point I discussed everything for two class, what to do for multi class? Suppose, there are more classes what to do what you do you do split a K – class problem into many two class problem. How? To approach one against all, so, this against rest one thing, these against rest one, this against rest, three hyperplane unit which is like this; first between these two then between these two then between these two. Both cases you need three hyperplanes actually.

There are other approach there is one more approach called a directed acyclic graph which does a hierarchical tree like splitting, which I am not describing you can see it there is a optimization problem to solve. So, this is how these also if we specify in your software how to do it will you do it, alright.

So, this concludes my talk on SVM. You can look up the lecture notes, more details are there. I did not go into some of the mathematics, you can look up and you can use try any of the software on some sample problems which I will give in the exercise.

Thank you.