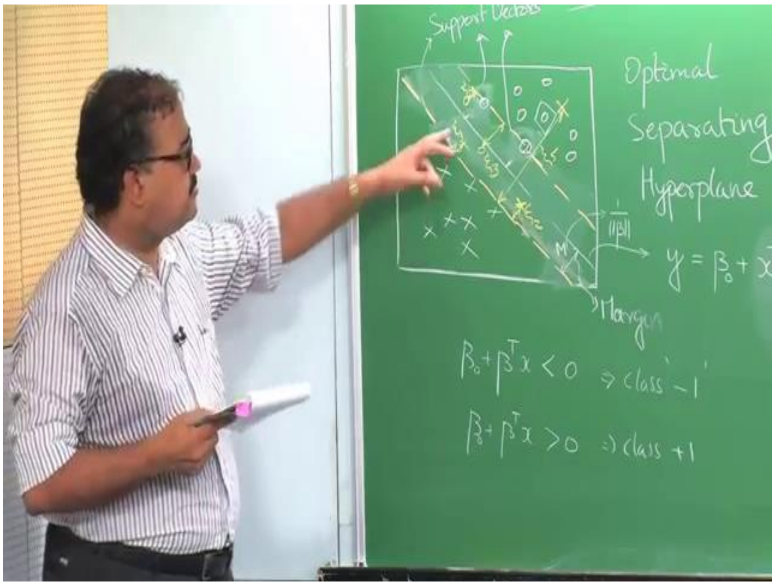
Introduction to Data Analytics
Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

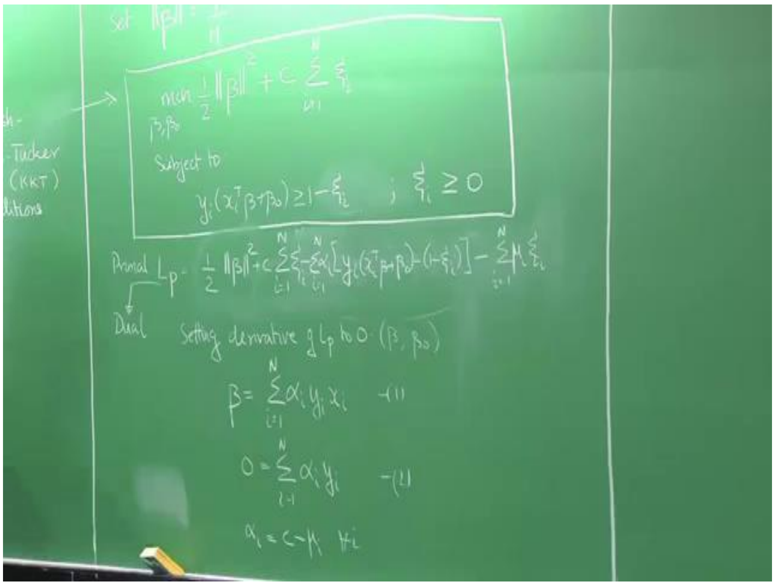
Module - 05
Lecture - 31
Support Vector Machines

(Refer Slide Time: 00:21)


Support Vectors
Optimal
Separating
Hyperplane
$\frac{1}{\|\beta\|}$
M
$y = \beta_0 + x$
Margin
$\beta_0 + \beta^T x < 0 \Rightarrow$ class -1
$\beta_0 + \beta^T x > 0 \Rightarrow$ class $+1$

Now, we look at the case where the data is not so-well behaved as you wanted to be. Specifically the data is not separable, is not linearly separable. So, you look at the non-separable case. So, I am going to introduce some additional data points whatever have looking at so far. So, this is a non-separable case, linearly non-separable case because some of my data points are really mixed up here. Now, still we will like to have large margin, but not only have I allowing data that is not separable, but I am also allowing data points to fall within the margin. So, its essentially two sides of the same coin. So, if I am allowing data to fall within the margin, so in some sense I am having the flexibility to make some kind of errors as well.

(Refer Slide Time: 02:40)


sh-
-Tucker
(KKT)
litions
Primal
Dual

Set $\|\beta\| = \frac{1}{M}$

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$

Subject to

$$y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i \quad ; \quad \xi_i \geq 0$$

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)\right] - \sum_{i=1}^{N}\mu_i\xi_i$$

Setting derivative of L_p to 0 (β, β_0)

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \quad -(1)$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i \quad -(2)$$

$$\alpha_i = C - \mu_i \quad \forall i$$

So, I essentially look at how far away am I from the margin in the long direction. So, I am going to denote these distances by which I am away from the margin by the symbol ζ, but we still have same constraints here, but now going back to our optimization problem, but I am really looking at here is, I am going to modify my constraints such that, so β_0* x_i* β + β_0, ≥ 1 - ζ_i, but ζ_i is some kind of slack variable that allows me to satisfy this constraint with some error in it. So, essentially if you look at this first data point that we drew here, if you look at the first data point I drew here. So, it has the slack of ζ_1, and the second data point as a slack of ζ_2, and this one of the really fairly large slack, but I still it is possible under the circumstances. This allows me to have a larger margin.

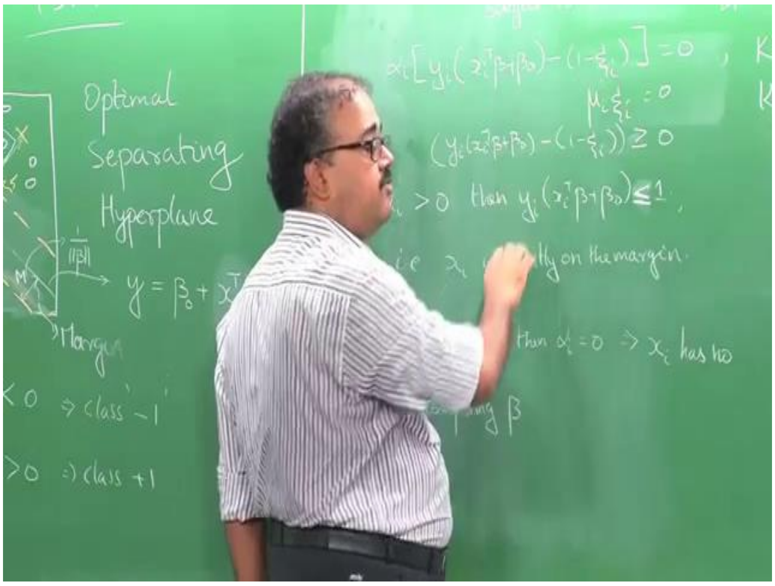
So, if you think about it if I did not, even if this data point was not there, if I did not allow these kinds of data point appear in the margin. If I dint allow these things to appear in the margin, my classifier would actually have been here. My classifier would have been here trying to separate this x_n from this o, and margin would have been very small. In all likelihood I am fitting a noise data point here which is not an optimal thing to do.

So, now I am allowing some amount of data points to fall within the margin. I am able to expand the size of the margin that I can have and I can also incorporate linearly the small number of errors that I have to make because the data is not linearly separable, but then I don’t want to this become arbitrarily large. I just cant say that hey it doesn’t matter you can have slack variable for every data point that we have and the slack variable can be as large as you wanted. I need to have a control on this as well and therefore I try to minimize that. So, are these conditions sufficient for us to define a new problem now, but still one more condition that we need. So, we are really measuring ζs in one direction and not in the other direction. So, we have to be careful. So, I also need to add another set of constraints, let us say that the ζ have to be positive. So, that is the complete constraint optimization problem for us in the case where the data is not linearly separable.

So, you have to minimize β^2, $||\beta^2||$ + the sum of the ζs that we are using subject to the condition that y_i x_i * for $\beta + \beta_0 > 1 - \zeta_i$ and $\zeta \geq 0$. So, what happens to our primal objective function now? So, what is that we need a component that corresponds to the actual objective function, and you need a component that corresponds to the constraints that we are using. So, the first part of it remains as it is, but then we have to add the newer components that we are bringing in. So, this is the second part of the objective function and this is the first constraint, and this is the second constraint and since these has to be applied to each and every data point, so you have the summation over all the data point you have, likewise here.

So, how do we go about deriving the dual in this case? So, just like we did earlier, so start differentiating the primal with respective of various parameters. So, you end up with the same condition that we had earlier, and here you end up with the same condition when you differentiate with respect to the ζs, you end up with so, one condition for each i. So, α_i = C - µi. So, I can put everything back into the primal, do some algebra and derive my dual which turns out look exactly like this, except that my α has to lie between 0 and C that you can actually see from that condition that we have there. α_i β_0 have to be equal to 0. That we already have as a condition. So, the remainder of the KKT condition that we have, I am just going to go through this very quickly because I don’t want to do the complete derivation here.

(Refer Slide Time: 09:50)


Optimal
Separating
Hyperplane
Margin
$y = \beta_0 + x^T$
$\alpha_i[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)]=0$
$\mu_i\xi_i=0$
$(y_i(x_i^T\beta+\beta_0)-(1-\xi_i))\geq 0$
>0 then $y_i(x_i^T\beta+\beta_0)\leq 1$
ie x_i ...tly on the margin.
then $\alpha_i=0 \Rightarrow x_i$ has no
$<0 \Rightarrow$ class -1
$>0 \Rightarrow$ class $+1$

So, the remainder of the KKT conditions will be, this is what we had last time except for the ζ_i part, but you also have. It is essentially or initial constraint written in a slightly different form, so the third constraint here which is the original constraint with 1 - ζ taking to the other side. So, what is that we notice from here? So, let us go back and let us do the same argument. If $\alpha_i > 0$, then $\beta_0 x_i^T \beta$, $\beta_0 < 1$ and only, then this will be zero because . So, that would mean that for the particular choice of ζ_i. So, this goes to 0. That would mean that this is on the wrong side of the margin, or on the margin. If the ζ_i is 0, then it will be on the margin and if the ζ_i is not 0, it will be within the margin. So, all the data points that are on one side of the margin or all, because again like last time, so β depends only on those vector for which α_i is greater than 0.

So, if the data happens to be on the right side of the margin, then your α_i have to be 0 as we saw earlier. So, those data points have no role in computing. So, this is essentially kind of takes us over the entire optimal separating hyper plane part, where it was linearly separable. So, we had a very easy solution, but when the data is not linearly separable, so we have to allow for the possibility of data points to lie on the wrong side of the margin. So, when we do that, we can essentially take advantage of the fact and try to push our margin away by allowing data point which are correctly classified, but are within the margin is small fraction of such data points are permitted and therefore, we need to expand the margin a little bit. Therefore, we can come up with the solution.

So, you don’t have to, at this point here don’t have to go in details of solving this optimization problem, but there are many powerful optimization techniques it have been developed that allow you to solve these kinds of problems. In fact, SVM have let to the revival of popular class of optimization algorithms called interior point methods because these are very efficient in solving these kinds of optimization problems. They are pretty robust classifiers and are very widely used for wide variety of applications, but some of you will be probably thinking now, but whenever I talk about a support vector machines, people always tell me something about kernels. I thought to support that machine all about kernels and I have not talked about kernels at all at any point here, yes.

So, the kernel idea is very crucial in support vector machines, and I will be looking at that in more detail in the next module. What you should remember is the basic optimization problem that you are trying to solve with support vector machine is the one of optimal separating hyper plane. So, in fact the kernel idea is called the kernel trick because it allows you to solve really wide variety of problems which is not easily amenable to linear classification by using a very powerful idea, but then the under lying optimization problem that you are solving is still this, the same optimization problem that on the boards so far in the last two modules.