NPTEL

NPTEL ONLINE CERTIFICTION COURSE

Introduction to Machine Learning

Lecture 29

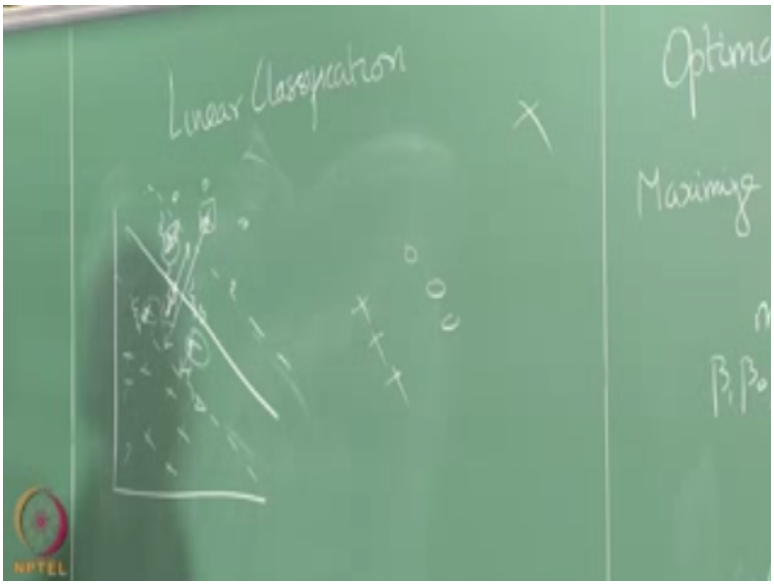
Prof. Balaraman Ravindran
Computer Science and Engineering
Indian Institute of Technology Madras

SVMs for Linearly Non Separable Data

Suppose I have some data which is not linearly separable right, so that is the problem that we are with perceptions right so what happens if the data is not linearly separable perceptions do not converge, so can we tweak our objective function that we have here to make sure that we can handle non linearly separable data is that right we are saying it is okay to say non linearly separable data was my question yeah linearly inseparable data right, so you have to be careful where you put the not the negation there right.

So what we do in this case yeah somebody had been a suggestion yeah so how will you do this right so there are many ways there are many choices you can make right let me not play around with it are many choices you could make but there is one particular choice which is seems to yield a very nice optimization formulation okay so what is a choice I am going to say that I would really like to maximize the margin right and I would like to get as many data points correct as possible right.

(Refer Slide Time: 01:42)



So if you think about it so there are a couple of things so this is the margin that I want right, so what are the problems here well these data points are within the margin right, so I have some data points that are within the margin so I would like to minimize such cases there are some data points that are within the margin and erroneous right I would like to minimize such cases as well right if you think of what if I had tried to get this correct right there is a gap here and there seems to be a gap here between the points if I try to get this correct and move my classification surface below then the margin would have been reduced even further right so it is okay to get this wrong but then what about this case is it within the margin or outside the person and within.

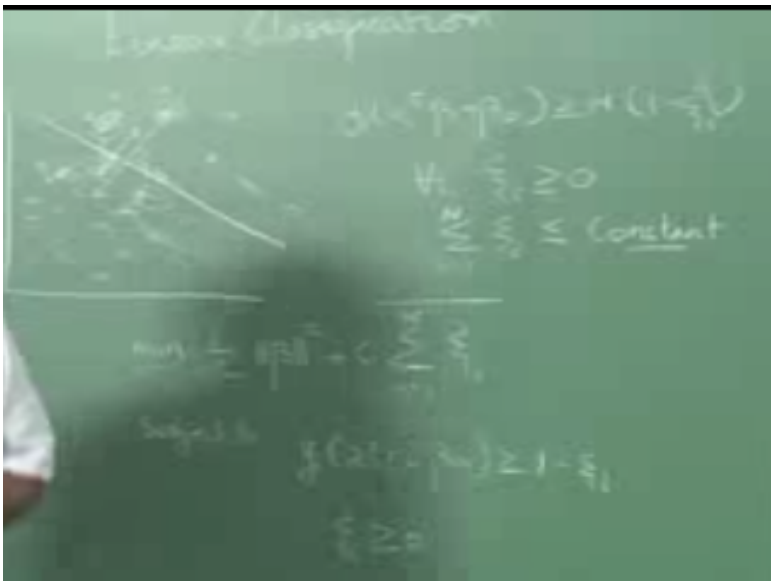
Right so the margin for that class is defined on the other side right so the margin for that classis this side so anything to this side and x is within the margin does it make sense right this will be y_i times this right, so this will actually be negative so it is within the margin if we want things to be greater than one y_i times f(x) we want it to be greater than one right ≥ 1 this is going to be negative.

So obviously this is within the margin right makes sense right so essentially what I want to do is minimize these distances so you can see the distance at their market so these distances I would like to minimize that makes sense right so this is a certain small distance inside the margin right this is a large distance inside the margin is a very large distance inside the margin like way so I can mark each one of these and I want to minimize these, so let us denote them says $\psi 1$ to $\psi 5$ and I want to minimize those right essentially.

So if I minimize the sum of these deviations I make along with my original along with my original objective function right I can handle why do not they minimize the minimum here again that could minimize the maximum would essentially mean that i will try to get as many things character as possible so in this case I do not mind getting something wrong as long as the overall deviation is not does not exceed a certain limit see that the difference between minimizing the maximum and minimizing a sum is that I might as well give up all of the sum to a single data point it might be something that is very hard to classify.

And I might have one single outlier somewhere here right let us let us let us draw so this data might be perfectly separable and I might have an outline then okay so now if I say okay minimize the sum of the things it is fine right but if I say minimize the max okay then it is going to actually give me a hyper plane somewhere there but like I said many different formulations are possible this one actually yields a very nice computation that is one of the reasons people use this.

(Refer Slide Time: 06:21)



So what I am going to do is write it here, so I am going to say that this has to be that we had already right I am going to introduce a slack variable so it does not have to be greater than M it can be some fraction lesser also M is what I would really like but I allow it to have a slack right ideally I would want most of these ξ is to be ψ as to be 0 right I want ideally I would like most of ψ this is to be 0 if I force ψ as to be zero I am back here right but I really like some lee way right.

So I am allowing myself that leeway by introducing ψi here this is a very standard technique for relaxing constraints in optimization right that is one of the reasons people adopt this is a standard constraint so another wee thing which I could have chosen is that right in fact this is a little bit more common constraint but it turns out that in this particular case if I choose M - ξ instead ofM(-1 ξi) I end up getting a non convex optimization problem.

So we do not want that right so it is to end up doing this so I drew this figure first because I wanted to get an idea of what these slack variables actually mean right, so the slack variables essentially tell you by what fraction right you are violating the margin right, so is I won this essentially what fraction of distance you are coming in here from the margin ξ2 is what fraction of the distance you are coming in from the margin.

So the margin is M so I have moved some fraction of the distance inside right, so they essentially that is what the ξ tells me right so what are the constraints we have what is your question right, so the first constraint I have is okay all ξ_i have to be ≥ 0 right I do not care about points going to that side of the margin right so all $\xi_i \geq 0$ and the second thing is whatever we have been talking

about right I do not want the ξ_i is to be very large taken in total so I want to upper bound them by a constant.

So because I am talking about going that side of the margin right if ξ_i are negative so essentially I will be imposing a tighter constraint than what I was looking for so this will be like it will larger than M right this is well I will be having a thing that is larger than him on the returns I am sorry it is a relative distance, so it is why it is a relative distance right so essentially this becomes on M- M ξ_i right so if I originate should be M, so it is now M ξ_i a away from their okay so ξ_i is essentially a relative distance right and if I make ξ_i is negative.

So this will become plus so that will essentially mean that not only do I want the data points to be aware than m actually masking it to be further away than, so it just imposes a tighter constraint, so I do not want that to happen so and here we are essentially giving it a budget that we do not want it to be greater than the budget right fine, so we saw such an constraint earlier where it we see such a constraint earlier we had a budget we did not want it to be greater than a budget.

So yeah rich regression and last so on other things we had this thing right, so wherever we are looking at this regularized regression, so we had this greater than it or lesser than a constant and what did we do in those cases we push it into the objective function right and then added a multiplier there and then we said okay it has to be right, so then there is a relationship between this constant on the multiplier that we put in the objective function right so likewise will do the same thing here I will do all the other transformations that we need to do right to normalize β and things like this.

So essentially I will end up with the same objective function I had there right you what $\geq$ to because they have gotten rid of the M right why how do we get rid of them M is 1 / β right so 1 by nom β so we got rid of that just anything else we need here right, so now that we have this objective function what should be the value of C if I want a linearly separable case we want to solve the linearly separable problem right or I want to ensure that all ξi are 0 what should the value of C be this is the simple question infinity right c should be infinity.

So the larger the value of C the more you are penalizing the violations right, so the smaller the ξi should be right so the larger the value of C the smaller the ξi should be right so this is a trade-off

so the larger you make C the smaller will the margin be but work will be getting more of the training data correct right so for large values of C you are allowing a little bit more leeway in does it make sense so C is very large see small then you are allowing lot more errors to happen if C is very large then you are forcing the classifier to classify as much of the training data is current as possible okay.

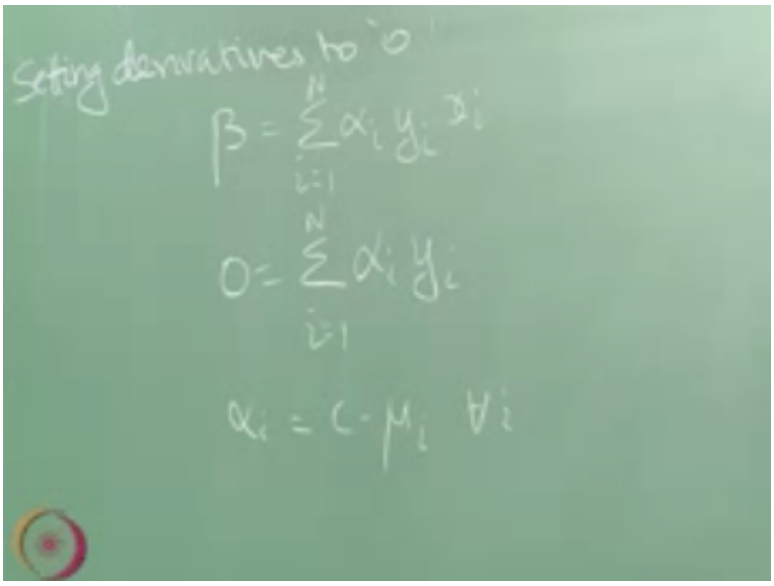
The data is truly linearly separable and you make C very large what will happen you will find the correct linear separator right but if the data is truly linearly separable but you keep C small what might happen you might trade off errors in the training data for a larger margin even if the data is linearly separable it set a desirable thing when narrator exactly, so if the data is noisy such that there are some data points that are closer to the margin it may be one or two data points that are closer to the margin.

So if you are trying to find the perfect linear separation you will pay attention to them as well right and therefore you will end up having a low margin right but then if you are willing to ignore a few noisy data points right even if the training data looks perfectly separable right you might end up making a few errors on it but you will get a more robust classifier right so can people visualize a situation I am going to try and do something here let us see that works it has looks perfectly separable right.

That is noise is it still separable yes hey let us more or less linear see the data points are point okay there you go there it is still separable right if you try to solve it as a perfectly separable problem that is the separator that you are going to get but if you allow errors right so that will probably be the separating hyper plane you get and that is probably a more appropriate hyper plane right.

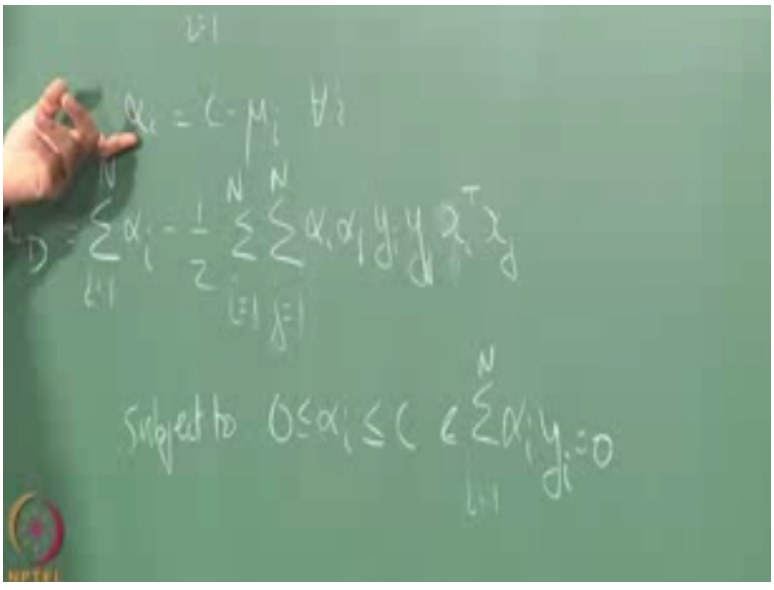
Apart from being robust it is it is also connecting an expected sense so far no questions now we will move on to the whole the primal do so I just wanted to leave this on board till I wrote this note so that you can compare it right so it is all good okay this fine.

(Refer Slide Time: 16:46)



So that is the Prim value also having have α right ξ α μ has to be ≥ 0.

(Refer Slide Time: 18:59)

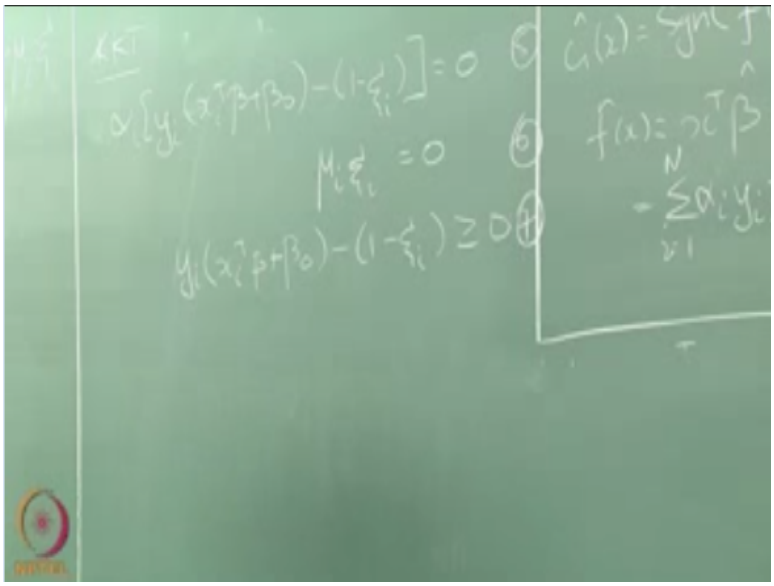


Yeah I do not have to I do not have to do this what is that why is it what is the single condition not a single condition is there for each I for we need the σ Π ≤ constant do we now we do not write, so that is why we consider constructed put that into the optimization objective function itself right, so by minimizing this right we are ensuring that σ ψi will be less than some limit right and like I was mentioning in the ridge regression discussion, so you can find a relationship between this constant and this C right.

It is also a function of the range of the objective function but you can always find so basically they are equivalent ways of writing the optimization problem except that you have to win this constant and the C will not be the same there will be different values so this constraint is gone this is no longer present here that went into the objective function okay.

So putting all of this back in and doing some algebra can be surprised at the algebra outcome of this anyone has already solved it know what it is look similar right it is essentially the same duel you will get but your constraints are different this is already there, so it is just added for completeness sake but what is important here is earlier while I had only a non negativity constraint on α now I have a upper bound on the value of α. So why is that because α is only - μ right since α C - μ so there has to be a upper bound on α okay good so what about the other KKT conditions.

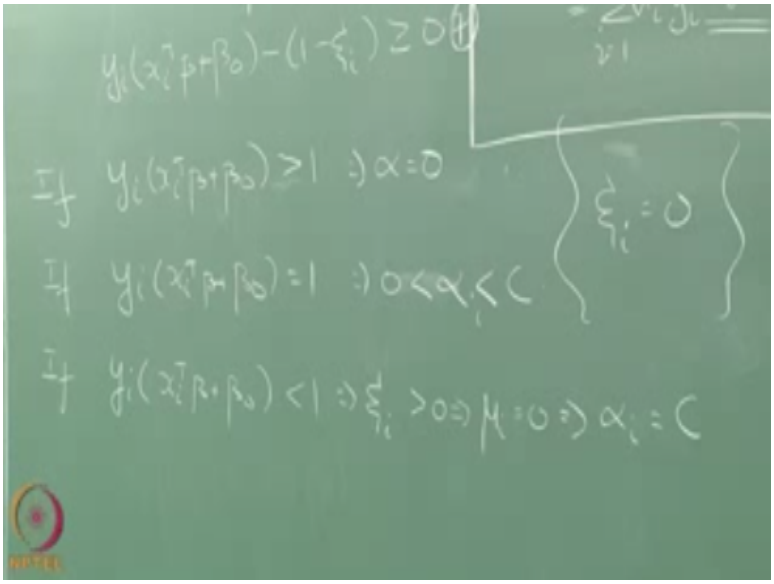
(Refer Slide Time: 22:30)



So 1 to 7 or the KKT conditions okay. So what do you notice here again well you notice again that your β is determined by your α Yi Xi just like you hand earlier right it β is given by those excise for which α will be nonzero right so like we had earlier those excise for which α is non zero or called support points okay our support vectors depending on how we want to look at it okay now let us go look at when it will be nonzero right so when will α be 0 whole thing is nonzero when will this whole thing be nonzero well it lies at a large enough distance on the right side of the margin right what about ξi is a will be 0.

Then ready somewhere here the ξi will be 0 right, so in the ξi 0 so it will be left with this term alone right -1 that just takes exactly the same condition that we had earlier right, so if this is far enough away from the margin then this will be nonzero so α is have to be 0 so we know for sure okay the same thing right things that are on the right side of the margin right means correct side okay things that are on the correct side of the margin then α will be 0 so they would not contribute anything right.

(Refer Slide Time: 26:25)



So now what about things that are on the margin right is that a third case we have to consider third case now right we have to consider the third case in which case what happened as ξi will start increasing right as is will become non zero, okay I ξi is nonzero what does it imply okay because my α is C – μi I right so if ξi is nonzero then μi will become 0 that for my α is will become C right so now how will this term go to 0 by suitably makings ξi a large enough right.

So I will make ξi a large enough, so that this term will go to 0 right because this is this will be negative hours will be less than 1 all right so I will make this I will just ξi so that this term in the square bracket goes to 0 because my α I will be C what is that in case this is because I do not really do not want to penalize this case right this case also ξi will be 0 right, so this case ξi is 0 this case also ξi will be 0 because what I really need is that is my condition ≥ 1- ξi so if is equal to 1 so I can set ξi to 0 correct in both these cases ξi 0 should that make sense everyone with me k.

So what are all the support vectors everything on the margin everything on the wrong side of the margin as well right everything for which alpha is nonzero now becomes support vectors so at the end of the day I am going to say that you are just going to use a package to solve all of these things right but it is like saying yeah anyway you are going to use Microsoft Windows or I mean Mac OS X or something why do you learn operating system right.

So you need to know what the internals are it is not the fact that you just use the tools that matters it when you can just use the tools well yeah we can do a tool course right how to use the tools right how will you start up limbos vm it is not trivial right, so many people I know actually run experiments with SVM's by just using the default parameter settings that the package will give right.

I do not laugh I mean there might be some of you in that but then it Is you need to understand what is it that you are tuning right, so now I told you about the c parameter right, so you understand have some idea of what a large C means versus what a small C means hey instead of blindly saying that okay I am going to Q and C from some number to some other number right, so to have an appreciation of what these things are doing actually helps you even use the tools better right so that is the whole idea behind doing all of this is not that I am going to expect you to come and derive a large margin classifier tomorrow when ideally.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in