NPTEL

NPTEL ONLINE CERTIFICATION COURSE

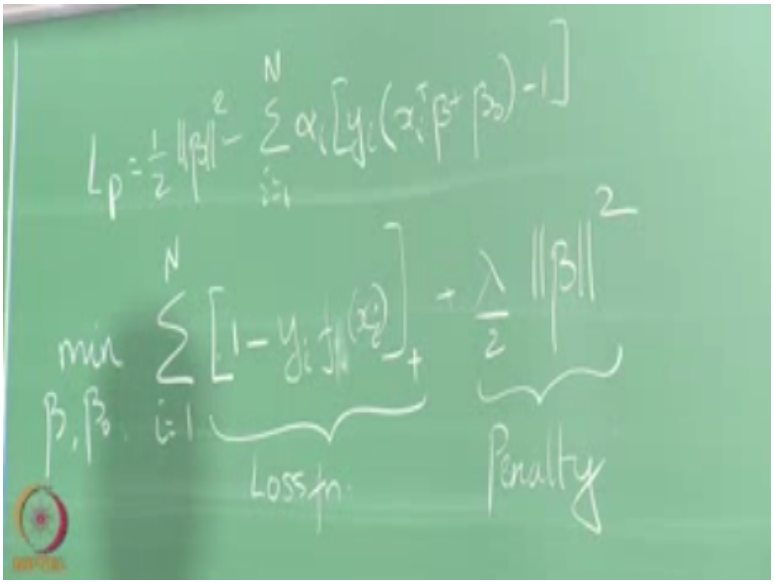
Introduction to Machine Learning

Lecture 31

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

Hinge Loss Formulation Of the SVM Objective Function

(Refer Slide Time: 00:19)



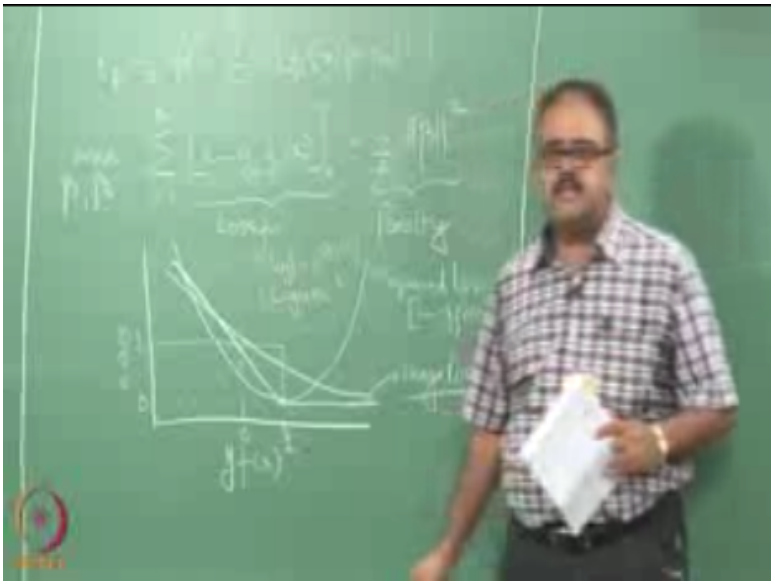
Okay so people remember the primal objective function that we had for SVM's so this is a primal objective function we had for SVM's so one way of thinking about it is to say that I am going to write it the following way, maybe some jugglery so the α I have replaced it with a λ here okay and well you know xi^T $\beta+$ β_0 is actually f at f of xi right, so I have written f of f -4 s should be f of xi sorry let us say f of xi and then essentially the same objective function except for this plus thing here.

So what is a plus thing you mean so it means that I will call this okay only whenever this is positive right whenever it is negative I will read it as 0, does it make sense I will count this only

whenever it is positive whenever it is negative I will make it I will consider it as 0, so that is what the plus term here indicates they went into λ I mean I am kind of redid this thing right so I divided everything by some factor of α and moody to λ okay right, so if you stop a minute this should look familiar to you what does it look like Richie regression.

Right this regression so you have a loss function okay and you have a penalty term right listen it look like that, so far we have been talking about non β^2 as being the objective function that you are trying to minimize and the other thing is constraints right and then we then wrote the Lagrangian and then we got the constraints into the objective function, so now I am saying you can think of another way of writing the objective function which is to say that there is this loss function right which is accounted whenever it is negative right, so now your goal is to minimize this right.

(Refer Slide Time: 04:03)



So how will this loss function look like right, so when y fx is one after that it will be 0 right that loss function talk about the loss function not about the pellet eater right but till yfx becomes one it is going to be a linear function right you can see that I it is just 1 - y effects so it is going to be a linear function of y fx right is it, clear so this kind of a loss function where this is like a door or a book opening on a hinge right if you think about it this is like two flaps of a book or a door right and it is opening on the hinge which is here right.

So it is also called right so sometimes if you have read about SVM's elsewhere you might have heard that the SVM's minimize hinge loss right so this is exactly what we are doing here so the hinge loss actually arises from the constraints, that we are imposing on the SPF right but if you think about it whether the constraints come from why were the constraints imposed what is the semantics of the constraint I do not want you to get yeah well what was the what is that we wanted to make sure that they are correct and a certain distance away right that is the reason for this.

So in effect the constraints are enforcing the correctness of the solution right and what the objective function originally was enforcing was essentially the robustness of the solution how far away are you from the hyperplane right, the constraints were making sure that you are on the right side of the hyperplane and if you think about it so in effect the constraints are an important part of what you are trying to optimize it is just not the distance from the hyperplane that matters but it is also matters that you should be on the right side of the hyperplane right.

So the putting area is a hinge loss makes it explicit and I am saying okay this is the class function I am interested in right so that essentially tells me I am interested in the correctness I want to make sure that all my data points are correctly classified okay, and the penalty tells me okay make sure it is a small norm solution it essentially becomes like Richie regression you make sure that the squared loss is as little as possible at the same time make sure that the norm of the solution is also small right.

So that is what we did right we did we enforce the l2 norm in the rigid aggression case and we are doing the same thing in the extreme case okay does it make sense now we can ask interesting questions like, okay if I replace this with some other non penalty what will happen can you do l1 regularized SVM's no that was errors regression so l1 regular is regression was last so can you do like loss so like regularization for SVM's since the β^2 if you put β what happens what do you think will happen.

Hello much harder optimization problem on your hand but it is actually a valid thing right, so what it will try to do if you remember we talked about this in last some an I did in a admittedly a little hand wavy fashion but we talked about how it will enforce sparsity right, we said it will try to make as many coefficient 0 as possible right, so in this case what do you think will happen if I put norm can attend for sparsity will it reduce the number of support vectors the statin for sparsity think about it what is that.

Now the squared loss is actually like this okay if you think about it is little where right, so if you are to this side you are actually correct right but, the further away you are from the hyperplane on the right-hand side also you still contribute to the loss because of this quiet error function whether you are on the right side of the wrong side of the hyperplane you still contribute to the loss okay so, so that is why sometimes the squared error function is not the ideal thing to minimize right.

So the hinge loss more often than not gives you a much better solution than optimizing squared error right, so what will the squared error be right that is what the square loss function is so normally you are used to seeing this as y - f of x the whole square but I have written it as 1 - yf of x that is also fine because if it is correct yf of x will be one all the time right so what is the actual loss function that you want this is fine that is the actual loss function you want right what

is the loss function call 0 1 0 1 is what you really want it should be 0 if it is correct and it should be one if it is incorrect at 0 1is what you really want and a lot of this just like a cig.

You right there is not really not going to text you or anything on it just for your interest a lot of work in theory in machine learning goes into showing that a if you optimize some other loss function will end up with the same solution as if you optimize the 01 loss right, so if you take the 01 loss I try to find a solution for it right I am trying to find the β that gives me the smallest possible 01 loss right it is a small as possible 01 loss 0 depends on the and yeah, so there is two points later depends on the data and you say linearly separable.

But why because you chose to use a linear classifier right so depending on what family of classifiers you choose and the and the data okay the minimum 01 loss could be 0 or it could be something higher right, so you say minimizes 01 loss I mean whatever is the minimum possible achievable given the data distribution and the class of I mean the class of classifiers of the family of classifiers your chosen given that what is the minimum achievable will you be somewhere close to that.

If I minimize a different loss function right so that is interesting question to ask right so I can arbitrarily come up with other loss functions I can come up with hinge loss height a squared loss so if you minimize hinge loss of squared loss will I get the same solution as I would have gotten if you had minimized 01 loss all right, so that is something people do think about right so we did look at one other loss function which is the I guess it goes something like that so that is what we minimize actually in the logistic regression case.

Even though we did not write it out explicitly as a loss function right so if you think about it this is what we actually minimize in the logistic regression case also you are trying to what were we trying to do what we do in the logistic regression, case what will be due to estimate parameters it is maximum likelihood right, so we made some assumptions about the distribution and then we try to maximize the likelihood and so on so forth right so if you work through that you can write it out as a loss function.

It turns out that this is what you are trying to so you can see that this never goes to 0 right this is going to go like this okay but then you can still think of minimizing that, so we will just an aside you do not have to worry about the logistic loss function right now will come back to that later.

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

www.nptel.ac.in