NPTEL

NPTEL ONLINE CERTIFICATION COURSE

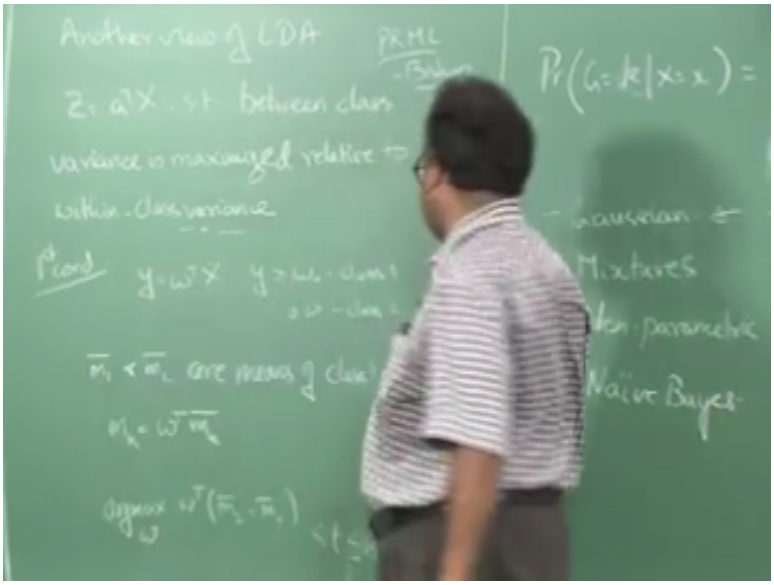
Introduction of Machine Learning

Lecture 23

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

Linear Discriminant Analysis III
- Another view of LDA

(Refer Slide Time: 00:17)



Okay so when I say between class variants I say it is the variance after class means so I will take the classes okay look at the means of those classes and look at the projected means of those classes and compute the variance among the projected means okay suppose I have K classes I can compute the variance among those if I have two classes what will this occur I want to maximizing the distance between the projected right fits two classes it will be maximizing the distance between them if it is K classes it will be maximizing the variance among the case centers right relate to do that within class variance and what would be the within class variance?

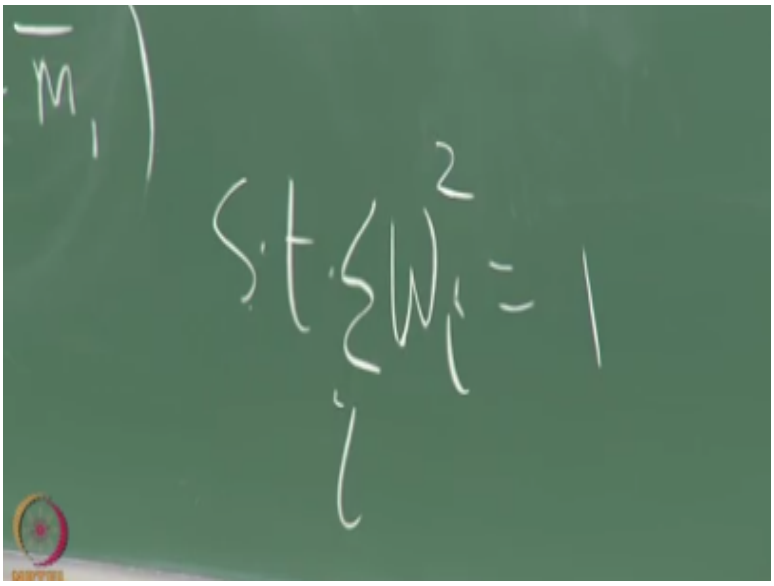
For each class the variance with respect to the class mean so that is what we already computed that right but for each class right so within class variance that essentially what I'm looking at here right so let us just treat the first condition alone all right so I will just simplicity sake start off with a two class case and then we can think of the generalization to multiple classes, so I am going to have a surface defined by $W^T X$ right so $y = W^T X$ if it's greater than some W_0 I am going to classify it as class one just less than some W_0 or less than or equal to Al classify it as class 2.

Sorry my font went too small, I am going to save one bar and M 2¯ are that means of C_1 and C_2 right and well we know how to compute M 1¯ just like you do a mu hat there and I am going to assume that when I write the MK without the bar okay this the projected one okay. So W^T M Kˆ I should see the projection of the mean okay in the direction W^T okay so that is essentially what this is so the reason I am using this funny notation is in the textbook if this is bold it is M 1 if it is unfolded it is a projection but I cannot write bold every time on the board.

So I am just using the bar right then when you read the book you can translate back and for this you read this part alone so till that part it is from hast tip scenario Friedman the ESL okay this part alone you do pyramid pattern recognition and machine learning by Chris mission the textbook reference is there on the so what is my goal when I say I want to maximize between class variants it is essentially to essentially to maximize that quantity it is a W^T M 2 is the projection of M 2 on w, w^T M 1 is a projection of M 1 w I'm trying to maximize this quantity so that is essentially my first criterion right.

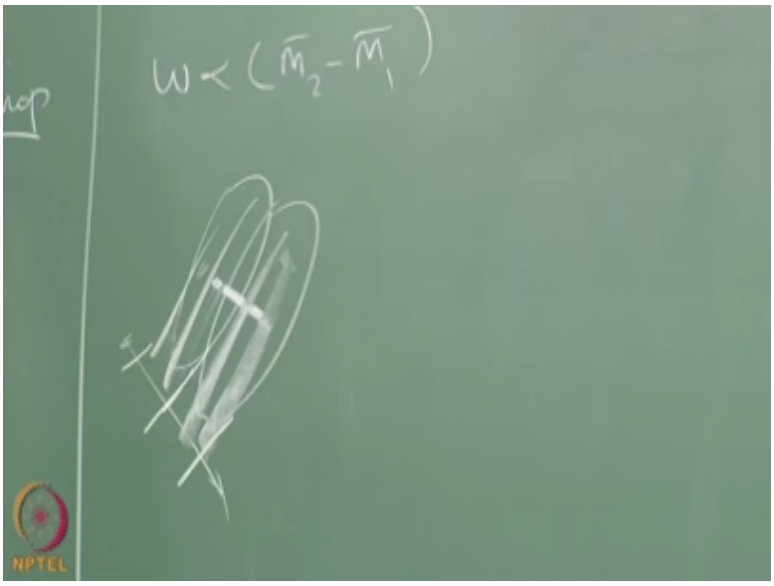
The Direction W that maximizes this right so there should be some alarm bells ringing for you what is the problem if I do not have any bounds on WI can just arbitrarily scale my W and get larger and larger values right, so I will have to have some constraints assuming summation over.

(Refer Slide Time: 05:58)



So since the norm W is one okay that is an assumption will make frequently to make sure that we do not get unbounded solutions right. So this is numerically unbounded. Yeah good question so you could impose a inequality constraint saying that summation W square is less than 1 but what we will think what do you think will happen you are maximizing the value right I am sorry you can just scale it so essentially what will happen is you will scale it says that WI hits 1 anyway so even if you are having a even if you have the lesser than or equal to constraint because you are maximizing over W you will hit it you will essentially scale W till you hit 1. So you might as well leave it as equal to 1 right.

(Refer Slide Time: 07:39)



So you can solve this right but the take-home message is that your W is going to be right, so W will be proportional okay you add that here right you take the derivative W will go and that will become W, so there will be some constants here right but essentially you are going to get W will be in the direction of m2 - m1 right, so what does this mean take the means right and if again you can go back and show that if it is spherical then the constant will be half right so it will be the midpoint of the line dividing that to means okay right.

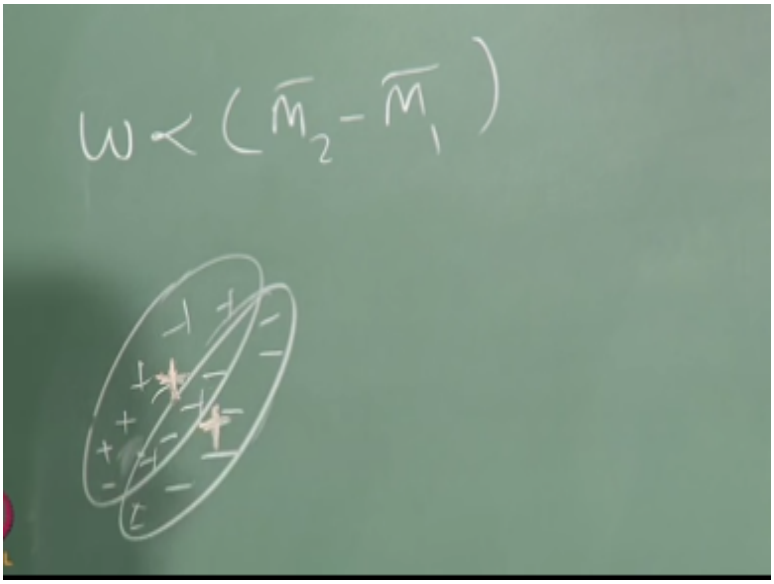
So let us do it again so I have two classes right I take the means right, so this will be the direction of the projection let us I allow to predict everything on to this right this way this will become class one that will become class 2 okay let it make sense right yeah so in this line and this line are actually parallel to each other I know you really did not want me to repeat the drawing but I think that you helped okay, so I have class one I have class two right, so I mean if you look at the data point that comes to me so the people understand when I say class one class two like this do you know the direction what I mean.

So this is the Gaussian corresponding to class one I am drawing the 1 σ contour of that right this is this is a likewise the 1 σ contour of the second Gaussian so the data point that comes to me could be something like this right this could be the training data that I am getting it will be mixed up of + and – in this region right that could be minuses here also okay already it is a drew one that could be minuses here that could be pluses here because the Gaussian still does extend

beyond the contour I have drawn okay, the contour is only the most probable region for the data points to lie does not mean that outside this contour the probability is 0 okay.

So this is essentially what it means so I am going to get data like this and I am going to model it I am modeling the Gaussian by these contours ok now let us say that.

(Refer Slide Time: 10:35)



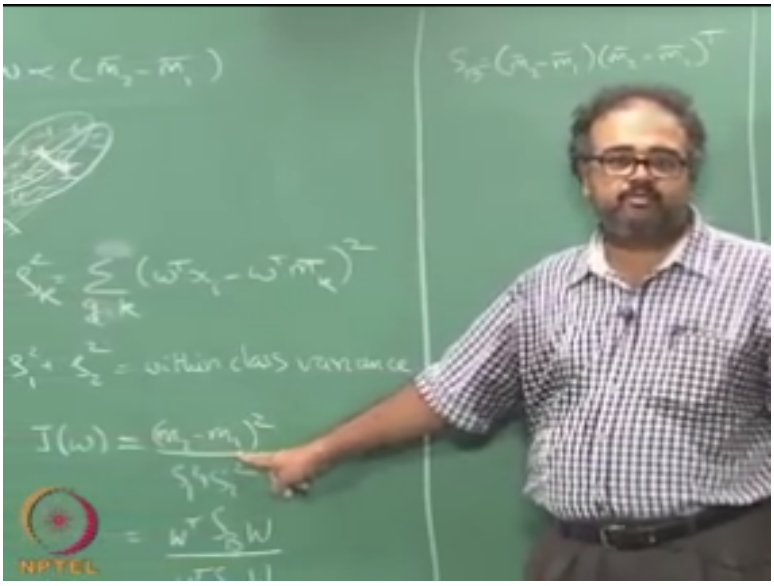
Roughly that these points are the centroids of the data that they get roughly these are the centroids of the data I get so what this tells us is that can you join these by a straight line okay and essentially you take direction that is all right like this and project all the data points to that right so you will get all the data points lying here now fix up threshold that what that is what I wrote here as w-not pick a threshold says that about that it is class 1 below then it is class 2 right.

In this case in fact if this had been spherical you can show that the threshold would light midpoint now we cannot because well you can I would guess I mean depending under special circumstances but now the point will be somewhere here and all the data points is projected above this I will say it is plus all the data points are projected below this I will say it is minus that makes sense right.

But then this is not what we are looking for right we are missing something important what is that the inter-class away this am sorry the within class variance right so this is the inter class

variance within class variance is what we are missing. So what we will do now start looking at that right.

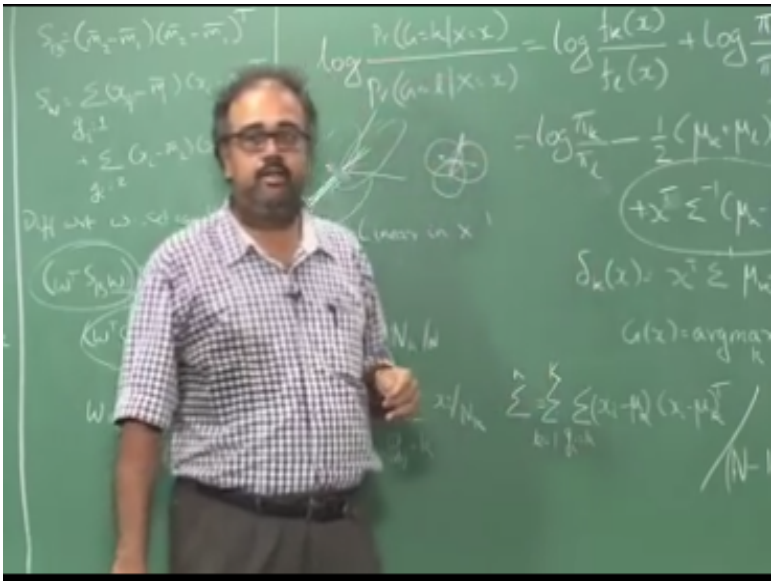
(Refer Slide Time: 12:34)



So that is a projected mean these are the projected data points belonging to class one okay keeping in with the terminology are using there so I'm picking on all the data points training data points which had class K right and looking at the projected distance from the projected mean this gives me the total within class variance yeah where I'm going to maximize everything at the end. So I am just ignoring the things that do not affect the maximization of thing okay, which squared term this way so that is essentially this is a projected data and that is a projected mean and just taking the variance of that right is exactly what we did that except that I have naught divided by the number of data points okay right.

So this criterion is called the official criterion it is called the fisher criterion after fisher was a very famous statistician who came up with LDA okay, several decades ago so here I am going to do something it is so I am going to be right it right so this is the between class covariance matrix right. So if you think about it so what I wanted was $m_2 - m_1$ what is m_2 the projected right so the projected one, so m_2 will actually be w m_2^- right W^T M 2^- right. So since neither W^T M $_2^-$ W^T M_1^- so I can take out the W^T and just have the square of the $m_2^- - M_1^-$ and I am adding the W^2 back in okay by doing W^T spw okay. Now what about SW?

(Refer Slide Time: 16:29)



So likewise so I have this as my SK^2 right so an $s_1^2 + s_2^2$ is essentially this is this S_1 right I had took take out the W from there and this is s_2 I take out the W from there so that gives me the W^T SW W okay. So now what we want to do we want to maximize this right we want to maximize the between class variants relative to the within class variance that is what we said right between class variance is maximized relative to the within class variance so that is between class variance is within class variance I have to take the ratio now I am maximizing this ratio.

So differentiate with respect to W differentiate with respect to W and set it equal to zero all right so this is what you buy V right so people want to tell me what the differential will be okay I will write it but you should recall all of this childhood memories okay you should not forget whatever you studied to get in here like so the denominator in the thing will become zero because I equated please eat already so when you take the derivative of this you're going to get some term in the denominator right.

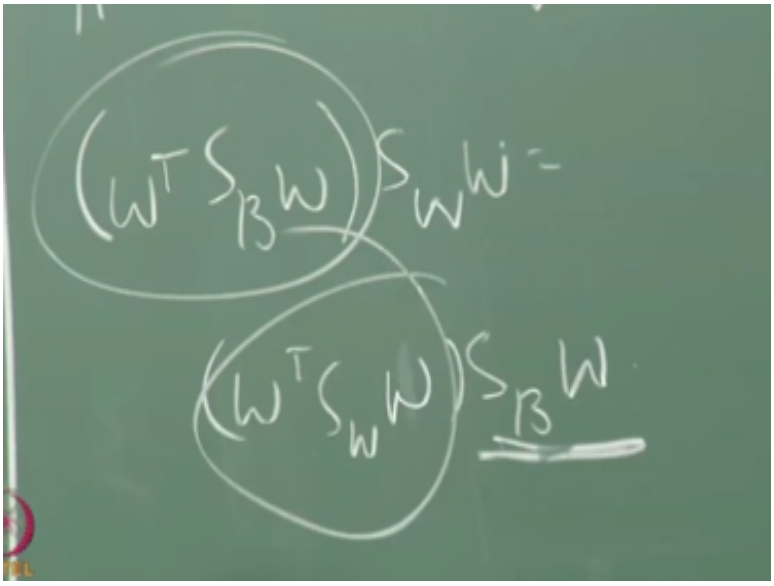
So that will go to zero so I will just have to equate the two half's in the numerator and I will get this right so just refresher derivatives the only thing that I am pretty sure putting everybody off is the fact that we are doing all of this in the matrix notation right just practice it makes life a lot easier do it a couple of times right the best way to do it is try and write it out in matrix form in gory detail okay do the term by term the derivative of it and then look at how it simplifies after you do the derivative right then you will see the pattern and then you will know exactly what we are writing it it's a very simple things like there are quadratics so you should know how to

differentiate quadratics that is the only thing that is throwing you off way W^T W is actually a quadratic in W right.

So that is the only thing so it becomes a linear in W so that is all nothing more to it actually if you think about it SBW okay, will always be in the direction of $M_2 - M_1$ right you already saw that here when we had only the constraint on SB right so here that the constraint was only on SB hey only on thee between class variants when he had the constraint only on the between class variants we ended up finding out that the solution is going to be the direction of $m_2 - m_1$ okay.

And a little bit of work you can show that always that spw will be in the direction of m_2 -m1 right so I can actually drop that and replace that with a vector proportional to $m_2 - m_1$ right, so now it makes our life a lot easier right I only have one W left so what about these guys.

(Refer Slide Time: 21:22)



They are all simplified to some kind of scalar quantities right so finally what I will get this w is not equal to but proportional to so that is essentially what I will get so if I did not have the SW constraint what I got was W as proportional time to -m1 right. But now if I am taking into

account the within class variance also then I will have to pay attention to thee within class covariance matrix.

So that is what this was rank I will have to pay attention to the within class covariance so that is basically all there is to it okay but how does this relate to this I see any relation between this and that think about it that is basically what we are doing there right. So σ inverse is SW inverse just using different notation here right, so SW inverse is just taking the variance between the in the data right in the within class variance so σ if you remember is the within class variance matrix right.

So that gives me σ inverse here and this how I got P σ inverse here and then I have m_2 - m_1 and I have mu K - μ L here, so essentially for in modulo all of these other non X related terms right so we are essentially finding the same direction right so whether you do it this way starting with data is your objective function right between class variance and within class variance or you start off by saying that your class condition density is Gaussian and then you are trying to find out the separating hyper plane right.

So in both cases you end up with the same direction modulo some scaling factors right, so you can use either motivation for deriving it but what is the nice thing about this motivation we did not make any assumption about the class conditional distribution the Gaussian assumption is missing here right the Gaussian assumption is missing and we worked only with sample means and sample variance and so on so forth right.

So it just tells you that LDA does not work only when the distributions are Gaussian right they are fine even when the underlying distribution is not Gaussian that is actually well-defined semantics to doing LDA right. People are with me on that so far okay great, so any questions let us let them move on to the next thing zfw this one DFW I told you right. So I want to look at the between class variance relative to the within class variance right so the numerator is the between class variance and the denominator is the within class variance so I'm trying to maximize the relative score.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development

Government of India

www.nptel.ac.in