NPTEL

NPTEL ONLINE CERTIFICATION COURSE

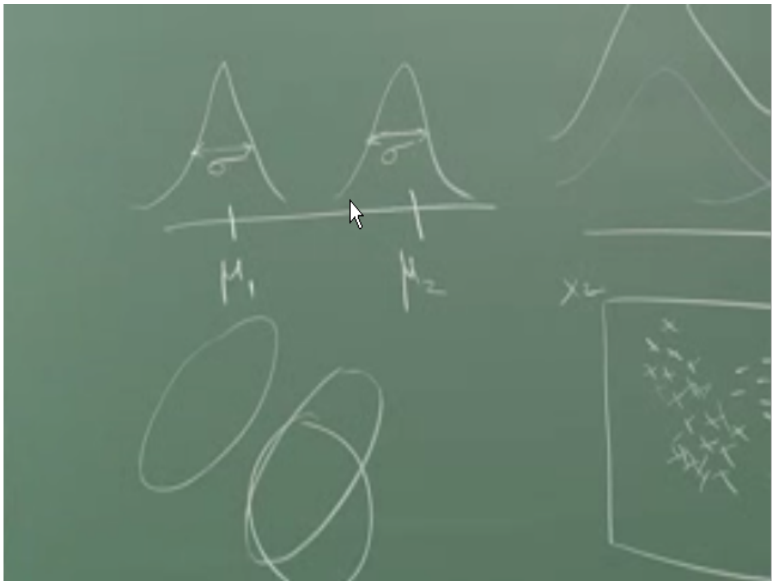
Introduction to Machine Learning

Lecture 22

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

Linear Disciminant Analysis II

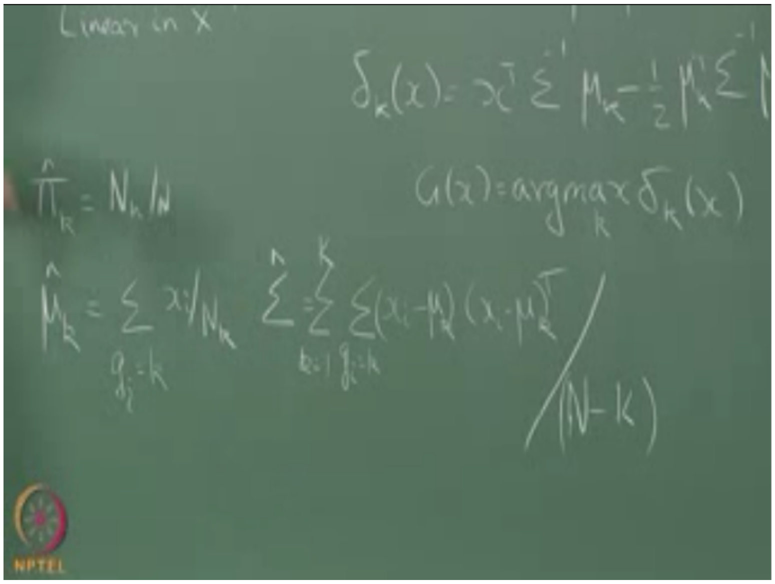
(Refer Slide Time: 00:17)



Okay so in LDA we make an further assumption than the fact that the class conditional density is Gaussian so what else can you assume so I am going to assume that $\sum_K$ it is the same for all the classes K so what does that mean it means that I will do that in 1d if it means that so class one could have the mean of my Gaussian here class two could have the mean of the Gaussian somewhere else right but when I look at this right so that is the same.

And all I can do is shift the Gaussian around but I cannot change the shape of the Gaussian okay is it clear when I say $\sum_K$ it is the same for all the classes it essentially means that I can just shift the Gaussian around I cannot change it so in 2d it will be like okay let us, let us assume that that is the equivalent of one ∑ contour right if that is the case for one class right for other class also it has to be similar okay.

(Refer Slide Time: 02:53)


Linear in X
$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu$
$\hat{\pi}_k = N_k / N$
$G(x) = \arg\max_k \delta_k(x)$
$\hat{\mu}_k = \sum_{g_i = k} x_i / N_k$
$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \mu_k)(x_i - \mu_k)^T / (N-K)$

So I cannot have one class looking like this and the other class looking like that does it make sense people are able to visualize what I mean can you can I see it that at the back or at the on the TV there right so it has to be that both the classes look similar in terms of the variance okay that is essentially the assumption we make here so in looking at logistic regression we saw that we could look at the log of so likewise I am going to look at log off so when I am at the class boundary what will this ratio be one rate and log of that is going to be zero right.

So I can actually solve for this ratio equal to zero and I can get my boundary right so now we know what is the form of probability of Z equal to K given X right now I can route it in here and I can solve for what yeah but what, what are the parameters I should be solving for and μ and ∑ and anything else we are talking about this we had it is all for ρ as well right it is solving for ρ is rather straightforward right.

I mean you just count the number of data points that belong to a class they divided by the total number of data punch you have that gives you ρ okay so it is not like complex but you still have to solve for it right it is not that it is given to a prior I so you have to estimate from data so all these three parameters your estimate right yeah, yeah so this gives you the boundary right so when the probability of it belonging to K is higher than the probability of it belonging to L then you will put it in K right.

So you will have to do this for every pair of classes to make sure that you have the right class so assuming that only two classes just you have to make one comparison but the verse there are really K classes he will have to make K minus 1 comparisons for to figuring out which class it belongs to so this essentially will give you the boundary so when the probabilities are equal then you know that well it could go either way so this is going to be 0 right.

So when the problem it actually belongs to class K the numerator will be higher when it belongs to class L the denominator will be higher so based on that you can decide which side it is going to go okay yeah so what about now solving for this, this is essentially log off right so the denominator will get canceled out like you only worry about the numerators in the-- okay and the fact that we assume that the variances are the same is also going to allow us to cancel out a whole bunch of other terms.

So what other terms can cancel out that can go right so this is $\sum_k$ it will become $\sum$ so when I take the ratio of the two things this thing can go right so I do not have to worry about that and it is log and there say Yi right so all of that will go away right so roughly the way to think of it is if I had taken out the product here right if you think of the term by term product here right so I will have some terms that will have an x^2 okay sometimes that is going to have Xμ okay and some terms that will have μ^2 right.

So I am taking the ratio so I am going to get μK^2- μl^2 right so that is essentially what I am writing out here the first term here corresponds to μK^2-μL^2 right so you have to get familiar with doing this in the vector notation it makes life a lot easier I am just giving you the intuition here if you can write it out and see that this is the right way to simplify it right so if you think of this essentially you are going to have a x^2 term μx term under μ^2 squared term.

And we take the ratio so you will have a power this divided by E power something else so that is going to become minus μ in the numerator so you are going to have new K- μL^2 right so that is essentially what I am writing out here and it is heavier what about x^2 terms x^2 will get cancelled out because $\sum$ are the same right so $\sum_K$ and $\sum_L$ is the same so x^2 will get cancelled out I only have the X terms left right that is basically right.

So I will have $X\mu K - X\mu L$ so I am going to get that term as well okay yeah and is actually plus so turns out that this separating hyper plane that we have this is essentially the solving this for zero gives us the separating hyper plane the separating hyper plane turns out to be linear in X right it is a separating surface shirts out to be a hyper plane right separating surface turns out to be a hyper plane.

So I should be saying it yes right so to get the linearity we needed to make this assumption if you do not make this assumption so what will happen the x^2 term will stay there right and what we get this QDA I told you about QDA right so if I do not make this assumption I will get you QDA I said in discriminating function case we are we always have some function like this ΔK of X right and if ΔK is greater than any other ΔL then we will classify the X into K right.

This is what we said was the idea we had disciminant functions in the very beginning so what would be the disciminant function version of LDA Hey the police note that for most of this bar anyone had him $\sum$ here on this side okay this covariance okay and I will make sure I write limits whenever I write this summation $\sum$ this work whenever we use multivariate Gaussians okay so I knew I left out one supply yeah okay.

So this is essentially the discriminant function right so you can just compare this and this is whichever has the highest discriminant value will become the class so ρ hat like I said earlier you count the number of data points in the training data there belongs to class K divided by the total number of data points you get by k and you had K you pick out all those data points for which the class was K from the training data find the center of them that gives me μ hat they make sense right.

So what about $\sum$ what is that μ like a well presumably so because these are all data points that belong to one class right, right so when the training data comes I am assuming that I have sufficient number of data points of each class otherwise I will not be able to learn anything it will work to the extent possible with the small data set right suppose I give you only ten data points for training okay then you are anyway in a soup like most of your parameter estimation algorithms will not work if you have very few days very little data there are some class of algorithms is work with very little data right.

So one such thing which we will look at depending on time today or tomorrow it is support vector machine so it works with very little data but most of the other parameter estimation methods require you to have some amount of data yeah so what we do with the variance so they will remember that variance is not limited to the class okay it is across all the classes we really want the same variance so essentially we do what is called a pooled estimate.

So we essentially use all the data points for estimating the variance not just the data points belonging to one class right but then you know what is variance right so that is the that is very inside you add that up and divide by minus one right so remember that you always do a minus one for variance estimates it is cool stuff given, given a sample mean and sample variance you must have done all of that right sample mean is divided add up the data points event by N and sample variance is data point minus μ^2 divided by N -1 usually right to adjust for the fact that mean is said dependent variable on all the data points okay.

So the N-1 essentially gives you an unbiased estimate of the variance right but then what is the mean you plug in here the mean corresponding okay so the mean of mean is a bad idea no, no, no so now I thanks for bringing it up because that is a natural confusion when they say you are going to estimate the variance across the entire population so the natural mean you are going to plug in is the mean of the mean of the whole data right but that is not correct why because I'm I had not fortunately rubbed the picture.

But I am only worried about the variance within the class right so I should plug in the mean here of the class fx_i right remember all of this is from the training data so I know what class6a actually belongs to right so I take the class of XA so I ask you I will take all those data points belonging to class k and then I will use μk here in computing this quantity right then I will do this over all classes I will do this over all classes.

And then I will divide by k okay well divide by N-K so this is a slightly different way of doing the variance as opposed to computing the variance of each class and taking some kind of a mean right this gives you a slightly more robust estimate of the variance okay so this is called a pooled estimate and all right great so if you think about let us say I have three classes that look like that so this will become what the separating surfaces that I learn if this had been completely spherical if this had been like this then the separating hyper plane would have been perpendicular to the line joining the means okay.

This is something which I just want you to note in fact I can ask you to show that it is fairly straightforward on univariate Gaussians but if it is if it had been spherical it would have been perpendicular to the line joining the means because I have now this is slanting right so this will also be the, the separating hyper plane will also be at an angle to the line joining the means okay fine any questions so far so if you look at many pattern recognition text books right they will talk about LDA as a feature selection mechanism right.

So you remember we looked at PLS when we did regression right so we did the principal component regression where we said we are looking at the directions in the input only taking into consideration the input right so that we are looking at the direction N that maximizes the variance in the input and then when we did pay less we took into account the class labels as well right.

So the equivalent of that in classification is LDA so you can think of LDA as actually finding directions along which the variance between the classes is maximized at the same time minimizing the variance within the classes right so PCA this maximizes the variance of the data right LDA maximizes the variance between the classes how does it achieve that it tries to find a direction say set the means or a spread apart as possible the mean of the data ray of each class is a spread apart as possible right. So in this case suppose this is the mean and this is the mean I am choosing this where I was using some direction where the mean such as spread apart as possible right so this is essentially the idea behind LDA right and we will just take that as our assumption.

IIT Madras Production

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

www.nptel.ac.in