# Pattern Recognition and Applications
**Prof. P. K. Biswas**
**Department of Electronics and Electrical Communication Engineering**
**Indian Institute of Technology, Kharagpur**

## Lecture - 17
## Multiple Discriminant Analysis

Good morning. So, in the last class we started our discussion on the problem involved with dimensionality of the feature vectors, and we have said that when the feature vector dimension is quite large then the computational complexity becomes very high. Because computation of the discriminant function for different classes, for each of the classes is proportional to d square, where d is the dimension of the feature vector. And when we have c number of such classes then obviously the amount of time that has to be spent to decide the class, we have to compute the discriminant function for each and every class.

So, the total computation time becomes of the order of c into d square and you can easily understand that in most of the real life situations the dimension of the feature vectors becomes of the order of 100 or 200 or so. And in that case what will be the amount of computation involved to classify an unknown feature vector, because for every unknown feature vector, we have to compute the discriminant function corresponding to each and every class. And then we have to decide to which class that unknown feature vector has to be classified, based on which particular discriminant function gives the maximum value.

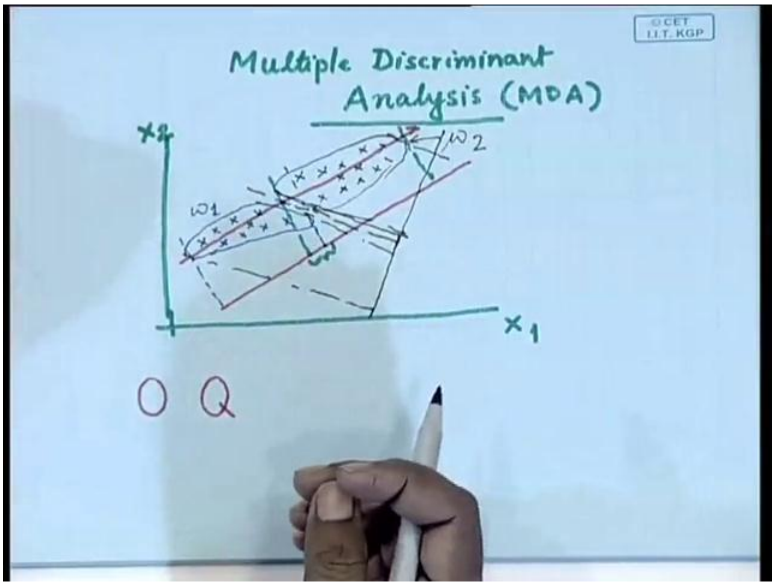
So, in the last class we have discussed about this dimensionality, reduction problem. And we have said P C A or principle component analysis is one of the techniques by which the dimension of feature vectors can be reduced. Because the principle component basically gives you those principle components which are most useful ((Refer Time: 02:04)) describe the feature vectors.

So, though the principle component analysis that gives you the feature that reduces the dimensionality of the features by taking projection on lower dimensional space either in single dimension or in multiple dimension. And we have seen that those directions are nothing but the directions of the Eigen values of the scatter matrix or co-variance matrix of the data set. And we take those Eigen values, Eigen vectors corresponding to the maximum Eigen values.

So, the error while doing the projection is minimised. Now, while doing so these feature vectors in lower dimension that we get they are the best representations of the original feature vector, in terms of the squared error sense. We ensure that the sum of squared error when we project that higher dimensional feature vectors onto the lower dimension following principle component analysis, the sum of squared error will be minimum. However, it does not necessarily mean that those projections will be the best projections for separating different classes, because when we go for classification we have to ensure that the feature vectors belonging to the different classes they are well separated.

So, while the principle component analysis tries to find out the projection directions which will give you best representation of the data set in minimum squared error sense. But M D A or multiple discriminant analysis tries to find out the projection direction so that if I take projections in those directions, I try to ensure that the data belonging to different classes or the sample vectors belonging to belonging to different classes, they are well separated in that projected space. So, today what we are going to discuss is this multiple discriminant analysis or M D A.

(Refer Slide Time: 04:12)



So, this is also a projection technique, but when we try to find out the projection directions we try to ensure that the data in the projected space they are well separated. Now, why the principle component analysis does not guarantee you that sort of separation? Let us just take a sample situation, say something like this. Suppose, I have two dimensional feature vector, one is feature component X 1, the other feature component is say X 2 and the data distribution, let us assume is something like this.

So, this is the data belonging to this space. The training sample that you get that belongs to one class. Let us say this is class omega 1, and the other set of training samples which are given they may be something like this. Suppose, this set of samples belong to class omega 2. Now, when you do principle component analysis, the principle component analysis takes the projections of data onto a direction which are aligned with the Eigen vectors. So, in this case the direction of the Eigen vector having corresponding to the maximum Eigen value will be something like this. Because this is the dimension direction in which the spread is maximum.

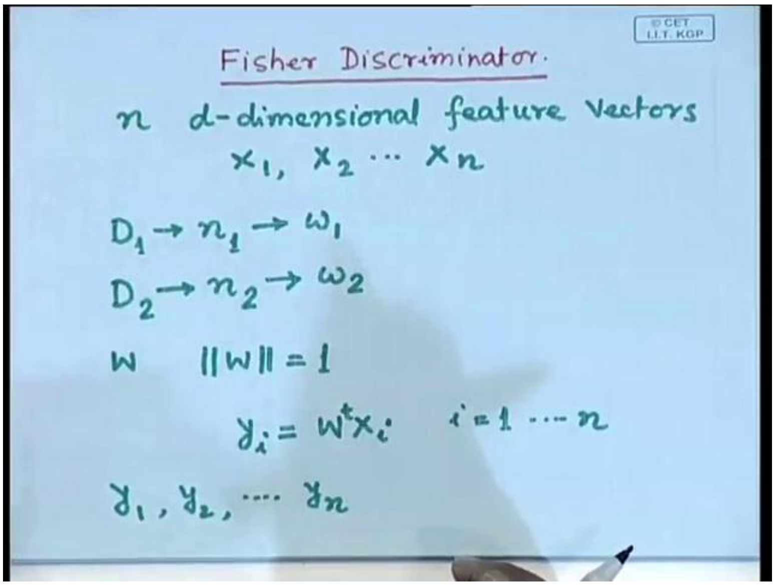
So, on this if I take the projection, let me draw the diagram somewhere over here. So, if I take the projection you find that this set of data will be projected into this space. Whereas, the data classified as levelled as belonging to omega 2 will be projected into this space. So, as a result you find that I have region where the data both from class omega 1 and class omega 2 they are intermingled. That means taking this example we can say that when I take the projections along the directions of principle, along the directions of Eigen vectors, it does not necessarily mean that the data will be well separated in the projected space.

Whereas, if I take the projection direction like this, suppose I take this as the projection direction we have discussed earlier that for best representation projection should be orthogonal projection. So, if I take this projection direction in that case we find that this set of data will be projected into this space. Whereas, this set of data will be projected into this space and now these two sets of projected data, they are well separated in the projected space. So, while principle component analysis tries to find out this direction, this projection direction, the multiple discriminant analysis tries to find out this projection direction.

So, what does it practically mean? Let us assume that we have got two printed characters, one is say printed character O, other one is printed character Q. So, when I take the principle component analysis, the principle components actually tries to capture the gross features of the patterns and while doing so it may try, it may ignore the feature. This feature corresponding this tail whereas, this tail is most important to discriminate between O and Q. Whereas, principle component may ignore the feature corresponding to this tail part and unless I have that information I cannot discriminate between O and Q.

So, that is the drawback of principle component analysis, though it gives you the best representation of data onto the projected space in minimum square sense, but it may not be the best projection. So, for us separability of the class are concerned so that is the aim of the multiple discriminant analysis. And when we discussed today, I will take a specific case of multiple discriminant analysis that I will consider that. Suppose, I have got only two classes and the approach taken for this multiple discriminant analysis to separate between two classes, and or to design a classifier which can classify between two classes after taking the projections onto lower dimensional space, that is what is called feature discriminator.

(Refer Slide Time: 09:47)



Fisher Discriminator.

$n$ d-dimensional feature vectors

$x_1, x_2 \cdots x_n$

$D_1 \rightarrow n_1 \rightarrow \omega_1$

$D_2 \rightarrow n_2 \rightarrow \omega_2$

$W \quad \|W\| = 1$

$y_i = W^t x_i \quad i = 1 \cdots n$

$y_1, y_2, \cdots y_n$

So, let us see what is this. Yes, they will pass through the mean of the samples, but what I have shown is just the projection direction. Actually, the line will pass through the mean of the samples, I have just shown the projection direction, not its exact position and then try to say see that. Because even if I pass this line to the mean of the samples, even then the same problem will occur. Either the projection will be downwards or the projection will be upwards, but the same problem will exist. I just have done it for the clarity of the figure that is all.

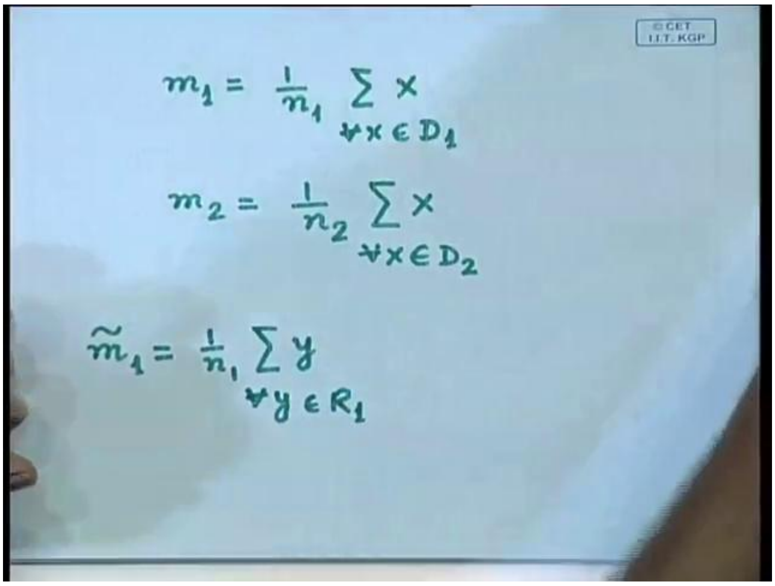
So, let us assume that we have n number of d dimensional feature vectors which are say X 1, X 2 up to X n. And this feature vectors are partitioned into two different sets, one is set D 1 consisting of n 1 number of feature vectors which are levelled as class omega 1, because we are going for supervised learning though. So, I have out of this n number of feature vectors, n 1 number of feature vectors or obviously n 1 is less than n 1 number of feature vectors. I say that this feature vectors are in the set D 1 and this vectors are levelled to belong to class omega 1. Similarly, I have another set D 2 consisting of n 2 numbers of feature vectors and these feature vectors are levelled as class omega 2.

So, obviously n 1 plus n 2 in our case will be equal to n, n is the total number of feature vectors partitioned into two subsets, one consisting of n 1 number of feature vectors coming from class omega 1. I am calling that set of feature vectors as D 1 and the other one is n 2 number of feature vectors coming from class omega 2, and I say that this set is set D 2. And suppose I take a projection direction, W direction of projection where I assume that modulus of W is equal to 1, that means this is an unit vector in the direction of projection line.

So, given this a feature vector X i when I take its orthogonal projection, because as we said that the orthogonal projections are the best projections. So, whenever we will talk about projection, we will talk about orthogonal projection. So, what I take orthogonal projection of the feature vectors onto this line. Suppose, I take a feature vector X i take the orthogonal projection onto this projection line, I get projected vector y i. So, here this y i will be nothing but W transpose X I, because W is the unit vector in the direction of line of projection and X i is the feature vector in my original space. When I take the projection of this in the direction of W, I get W transpose X i and the projected vector is y i.

So, when I have this n number of samples, X 1 to X n, the projections on this onto a projecting line will be y 1, y 2 up to y n. So, each of this I can compute following this formula, so here y varies from i varies from 1 to n. So, these are the projected vectors that I can have after taking projection in the direction W. Now, coming to the sets D 1 and D 2.

(Refer Slide Time: 14:59)



$$m_1 = \frac{1}{n_1} \sum_{\forall x \in D_1} x$$

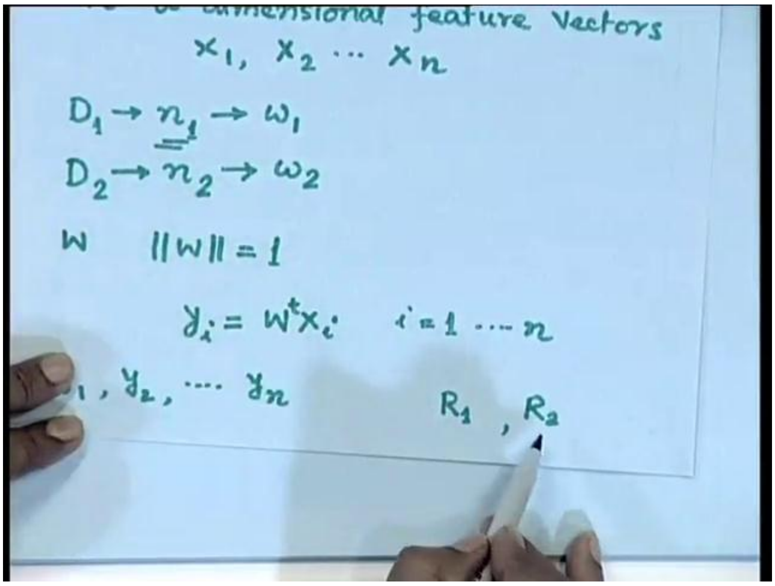
$$m_2 = \frac{1}{n_2} \sum_{\forall x \in D_2} x$$

$$\tilde{m}_1 = \frac{1}{n_1} \sum_{\forall y \in R_1} y$$

I can find out the mean m 1 which is nothing but mean of the samples in class D 1, and obviously this m 1 is equal to 1 upon n 1, because n 1 is the number of samples in class D 1. Then summation of X for all X belonging to set D 1, I take all the samples from set D 1 and average of that gives me the mean m 1. Similarly, m 2 is nothing but 1 upon n 2 summation of X, for all X in set D 2 is that. Now, when I compute the projections of this mean vectors.

Suppose, the projected mean vector m 1 I write as m 1 hat, this m 1 hat is nothing but 1 upon n 1 into summation of y, for all y belonging to say set R 1. Now, what is this R 1 when I have all these projected samples, projected vectors y 1 to y n. Now, the projected vectors which are projection of these n 1 number of samples.

(Refer Slide Time: 16:49)



dimensional feature vectors

$x_1, x_2 \cdots x_n$

$D_1 \to \underline{n_1} \to \omega_1$

$D_2 \to n_2 \to \omega_2$

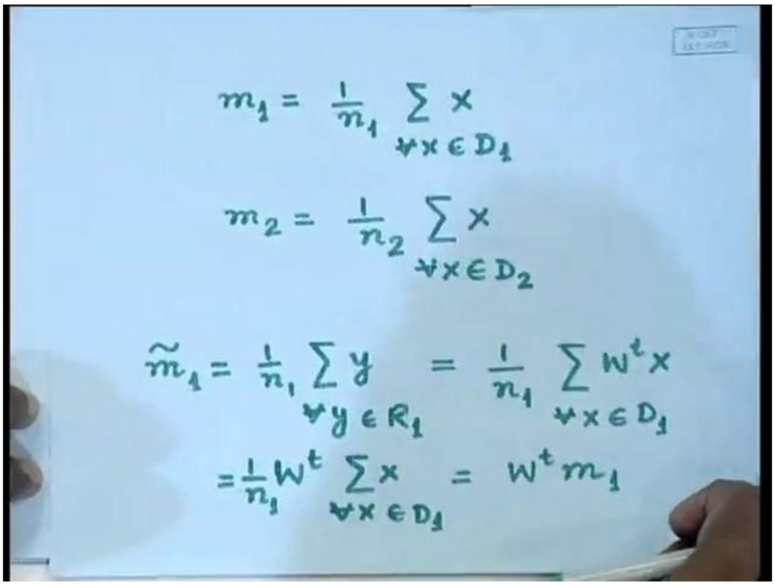
$W \quad \|W\| = 1$

$y_i = W^t x_i \quad i = 1 \cdots n$

$_1, y_2, \cdots y_n \qquad R_1 \; , R_2$

I call that set as set R 1. So, this set is the projected vectors of these n 1 number of samples. Similarly, I say set R 2 which is the set of projected vectors set of this n 2 number of projected vectors. I call this set R 2, right? So, here this m 1 tilde which is the projected mean is nothing but 1 upon n 1 sum of y, for all y belonging to class R 1.

(Refer Slide Time: 17:35)



$$m_1 = \frac{1}{n_1} \sum_{\forall x \in D_1} x$$

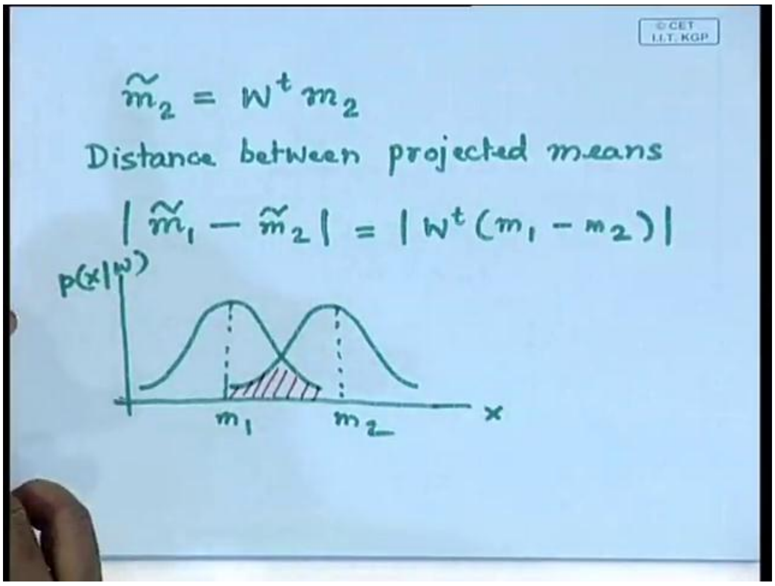
$$m_2 = \frac{1}{n_2} \sum_{\forall x \in D_2} x$$

$$\tilde{m}_1 = \frac{1}{n_1} \sum_{\forall y \in R_1} y = \frac{1}{n_1} \sum_{\forall x \in D_1} W^t x$$

$$= \frac{1}{n_1} W^t \sum_{\forall x \in D_1} x = W^t m_1$$

And this y is nothing but 1 upon n 1 summation of W transposes X, for all X belonging to set D 1. I can take this W transpose out of this summation, because it does not depend upon the samples, okay? I simply get 1 upon or W transpose. Let me put 1 upon n 1 W transpose summation of X, for all X belonging to set D 1 which is nothing but W transpose m 1. So, that is quite obvious. So, projection of mean of the vectors in set D 1 is nothing but W transpose m 1.

(Refer Slide Time: 18:48)



$$\tilde{m}_2 = W^t m_2$$

Distance between projected means

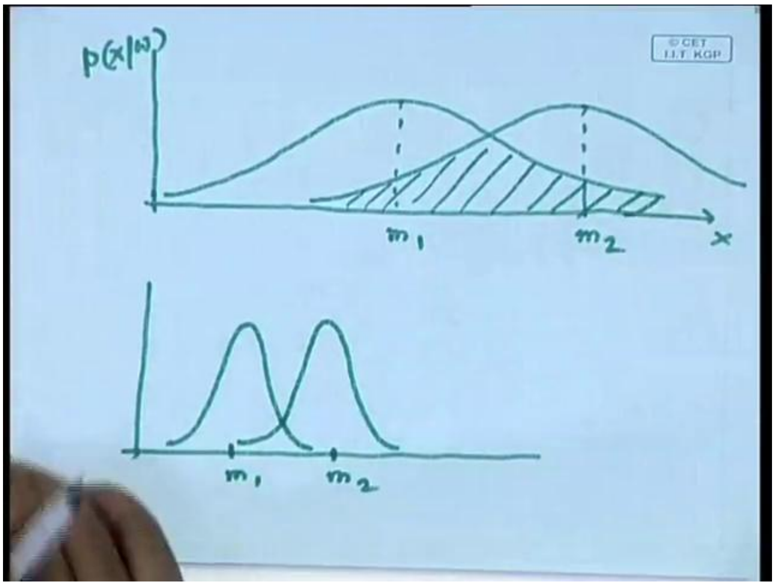
$$|\tilde{m}_1 - \tilde{m}_2| = |W^t(m_1 - m_2)|$$

In the same manner I can compute the projection of mean of samples in set D 2 which is nothing but m 2 tilde. If I follow the same procedure it will simply be W transpose m 2. So, m 1 tilde that is the projection of m 1 onto the direction of W is nothing but W transpose m 1, projection of m 2 onto the direction of W is W W m 2 tilde which is nothing but W transpose m 2. Now, what is the distance between these two projected means? That is nothing but so the distance between projected means is nothing but mod of m 1 tilde minus m 2 tilde. This is the distance between these two projected means or difference between the two projected means, which will be simply mod of W transpose m 1 minus m 2.

So, this is quite simple. Now, find that I can increase the distance between m 1 tilde and m 2 tilde by simply scaling up W. Here, I have assumed that modulus of W is equal to 1, if modulus of W is more than 1 then the distance between m 1 tilde and m 2 tilde will go on increasing. As I increase the scale factor of W, the distance between m 1 tilde and m 2 tilde will increase, and that ensures that I can increase the separability between the two classes. But the question is how much should I increase the separation between the two classes?

Should I go on increasing indefinitely? Obviously that is not justified or not very logical, because how much difference between m 1 tilde and m 2 tilde that I need for good classification depends upon, what is the variance of the samples within set D 1 and what is the variance of samples within set D 2. This variance will indicate that what should be the difference between m 1 tilde and m 2 tildes, the reason is very simple, if I take that distribution suppose I have two distributions like this. This is m 1, this is m 2, this is my x and this is say p x omega. We said the error of classification is nothing but the area bounded between these two density curves, probability density curves which is this shaded portion. Now, if the variance of the data distribution is quite large that means I have variance something like this.

(Refer Slide Time: 22:46)

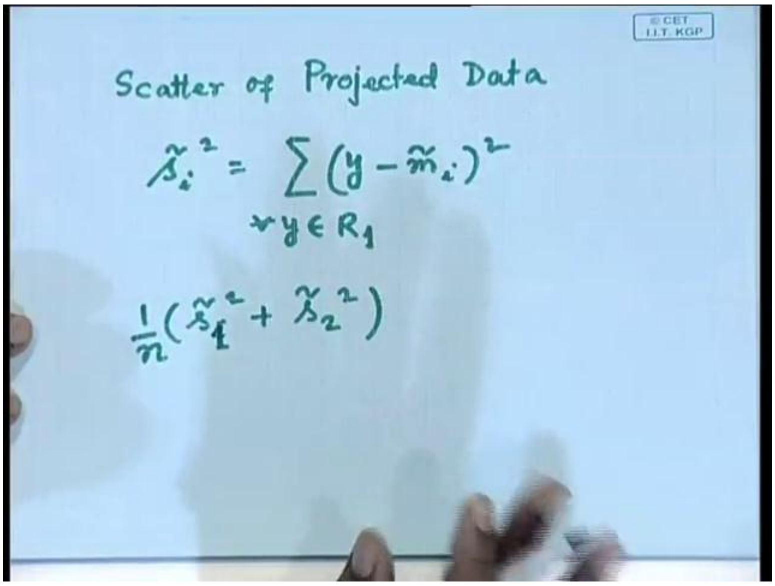


The other one is like this. So, this is my x, this is p x omega, this is m 1 and this is m 2, here the variance of the standard deviation of the data within the two different classes are quite large. So, the curve has become very flat and the error is this. So, when I have this sort of distribution, you find that I should have the difference between m 1 and m 2 to be quite large. So, that the error of classification is reduced, that means these two curves should be wide apart, right? Whereas if I have distribution of this form where the variance is very small, m 1 is somewhere over here, m 2 is somewhere over here.

So, in this situation I do not need that much separation which I need in the first case, here the separation between m 1 and m 2 can be much less than this, but even then my error the total error will be under control. So, how much should be the difference between two projected means in the reduced space that should be relative to the variance or standard deviations of data within different classes.

So, it should not be an absolute value, it should be relative to the variance of data. So, accordingly I can make use of the scatter of variance to have some criteria function, the criteria function which has to be either maximized or minimized. And when it is maximized or minimized I can say that the separation between the two means that I have got that is sufficient for classification. So, we can define the scatter of the projected data like this.

(Refer Slide Time: 25:32)



© CET
I.I.T. KGP

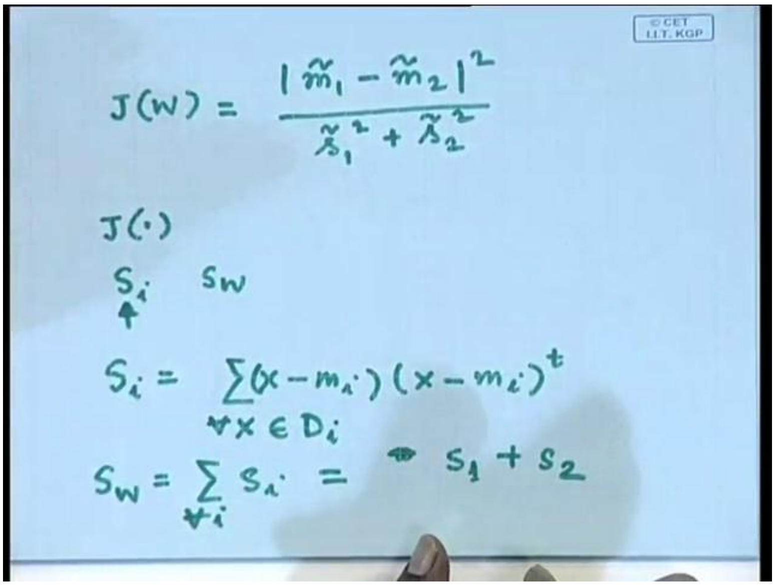
Scatter of Projected Data

$$\tilde{s}_i^2 = \sum_{\forall y \in R_1} (y - \tilde{m}_i)^2$$

$$\frac{1}{n}(\tilde{s}_1^2 + \tilde{s}_2^2)$$

So, let us see what is scatter of projected data. So, for an ith class I define the scatter of the projected data as summation y minus m i tilde square, where this summation has to be taken over all y belonging to set R 1, is that you find that this is something similar to variance. What we have not done is we have not normalized this. So, when I define such a kind of scatter for the ith for the samples belonging to set D i or the samples taken from class omega i, then I can define the total within class scatter. So, the total within class scatter will be nothing but s i tilde square or since we are considering only two classes, let me make it s 1 tilde square plus s 2 tilde square. Let us normalize this by the number of samples I have which is equal to n. So, this tells you the total within class scatter of the projected samples, right? And as I said that I want the separation between the two classes of the distance between two means should be relative to this total scatter, right?

(Refer Slide Time: 27:39)



$$J(W) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

$J(\cdot)$

$S_i \quad S_W$

$$S_i = \sum_{\forall x \in D_i} (x - m_i)(x - m_i)^t$$

$$S_W = \sum_{\forall i} S_i = S_1 + S_2$$

So, I can define now a criteria function which is something like this, J W is equal to m 1 tilde minus m 2 tilde square upon s 1 tilde square plus s 2 tilde square. So, this s 1 tilde square plus s 2 tilde square, it is the total within class scatter which takes into consideration the scatter of the samples taken from set R 1 plus the scatter of samples taken from set R 2. And what I have at the numerator is m 1 tilde minus m 2 tilde square which is nothing but the square of the distance between the projected means m 1 and m 2. So, what I should try to achieve is this m 1 tilde minus m 2 tilde square should be as large as possible, with respect to the total within class scatter. Or effectively I want to maximize this criteria function J W, which is the ratio of m 1 tilde square minus m 1 tilde minus m 2 tilde square upon the total within class scatter.

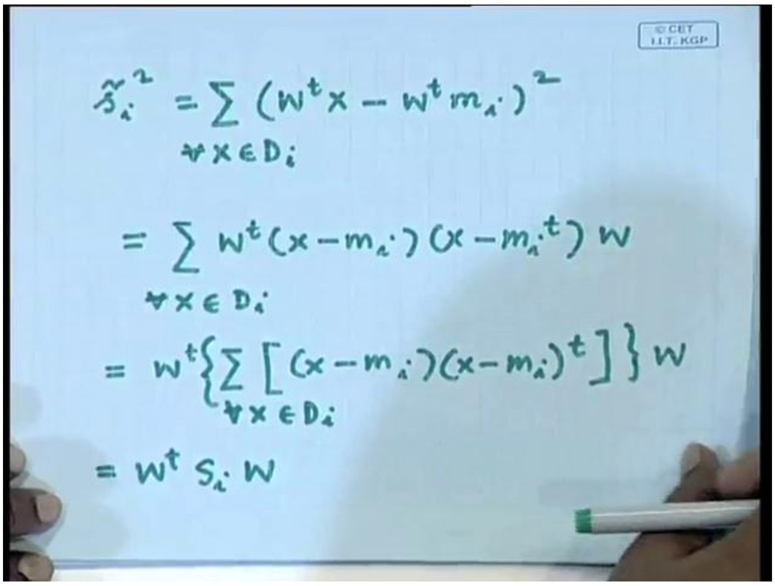
So, I want to maximize this. So, I can see effectively that this being the difference between the two classes or two means. I can say equivalently that this is something similar to within class scatter, sorry inter class scatter and this is something similar to intra class scatter. So, if intra class scatter is less that indicates that your data distribution is very compact. If it is compact then inter class scatter I need not have very large value to give the separation between two classes, right? Whereas if this intra class scatter is large that means the data are distributed sparsely or widely within every class. If it is so then this inter class scatter should be quite large so that I can have satisfactory classification performance or I can discriminate between the classes very easily.

So, effectively what I have to have is I have to maximize this criteria function J W which is nothing but something like ratio of inter class scatter and intra class scatter. And features discriminator actually maximizes this and the value of W which maximizes this gives you the projection direction. So, how do I find optimal value of W which will maximize this? If I want to do that in that case, this expression I have to write as an explicit function of W and then I take the gradient. Or I take the differentiate it with respect to W equate that to 0 and you get the solution.

So, if I want to explicitly write this criterion function J in terms of W, let us do this first. We define a scatter matrix scatter matrices, one is s i and other one is s w I, will say s i is the scatter within the ith class. And s w is the total within class scatter, right? So, as before I can define this s i that is scatter of the samples within ith class which is nothing but sum of X minus m i into X minus m i transpose, where I have to consider this summation over all X belonging to class d i.

So, this gives me the scatter of the samples in set d i or the scatter of the samples which are taken from class omega i. So, once I get the scatter for individual classes or individual sets then I can define total within class scatter, because this is within class scatter. So, I can define total within class scatter as s w is equal to sum of s i for all i and here we are considering only two classes. So, this will be s 1 plus s 2 is that total within class scatter s 1 is the scatter of the samples in set D 1, s 2 is the scatter of the samples in set D 2. If I take these two sums that gives me total within class scatter or simply it is called within class scatter. Next, let us try to see that what will be the scatter of the projected samples.

(Refer Slide Time: 33:46)



© CET I.I.T. KGP

$$\tilde{s}_i^2 = \sum_{\forall x \in D_i} (W^t x - W^t m_i)^2$$

$$= \sum_{\forall x \in D_i} W^t (x - m_i)(x - m_i^t) W$$

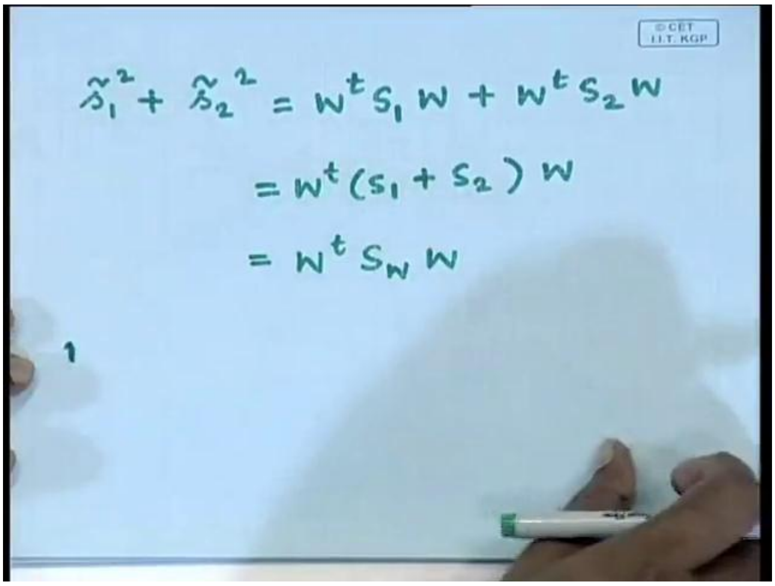
$$= W^t \left\{ \sum_{\forall x \in D_i} \left[ (x - m_i)(x - m_i)^t \right] \right\} W$$

$$= W^t S_i W$$

So, for that I define s i tilde square which is nothing but W transpose X minus W transpose m i square of this. Take the summation over all X belonging to set d i. So, earlier we have taken the scatter of the original samples. Now, we are taking the scatter of the projected samples, this expression can be written as just by some re-organisation. I can write this expression as W transpose X minus m i into X minus m i transpose into W, take the summation for all X in set d i. Now, given this you will find that this W which is independent of X i can take outside the summation, right?

So, if I take outside the summation then what I will have is W transpose, then summation X minus m i into X minus m i transpose into W, where this W is outside the summation operation. And these summations I have to take for all X in set d i. Now, what is this term within this summation? This is nothing but S i, that is the scatter of the samples in the original space. So, what I simply have is this is nothing but W transpose S i W.

(Refer Slide Time: 36:09)



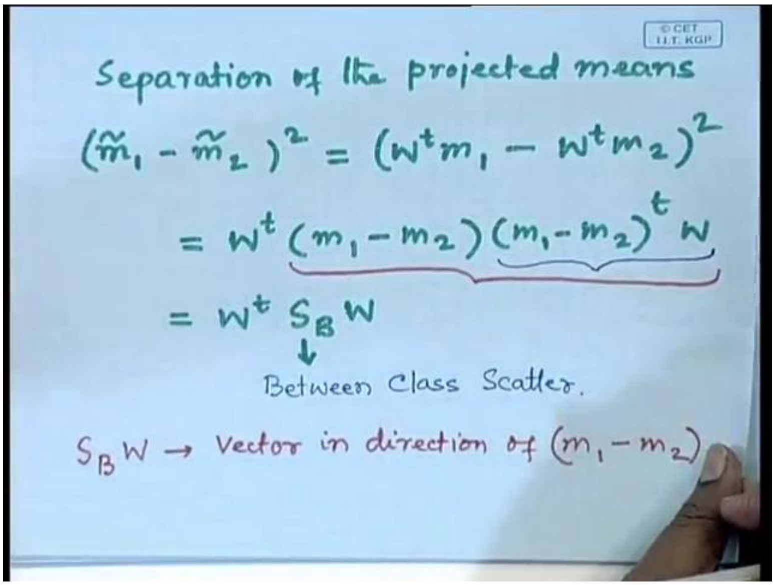
$$\tilde{s}_1^2 + \tilde{s}_2^2 = W^t S_1 W + W^t S_2 W$$

$$= W^t (S_1 + S_2) W$$

$$= W^t S_W W$$

So, when I have this then the sum of this scatter over two classes that is s 1 tilde square plus s 2 tilde square, that will be simply W transpose S 1 W plus W transpose S 2 W. This is nothing but W transpose S 1 plus S 2 W and S 1 plus S 2 as we have already defined that. This is total within class scatter S 1 plus S 2 is nothing but S W, which is total within scatter. So, this simply becomes W transpose S W W, right? So, in the same manner when I compute the difference between the two means, so that is nothing but the difference between the two means.

(Refer Slide Time: 37:30)



Separation of the projected means

$$(\tilde{m}_1 - \tilde{m}_2)^2 = (W^t m_1 - W^t m_2)^2$$

$$= W^t \underbrace{(m_1 - m_2)\underbrace{(m_1 - m_2)^t W}_{}}_{}$$

$$= W^t \underset{\substack{\downarrow \\ \text{Between Class Scatter.}}}{S_B} W$$

$S_B W \rightarrow$ Vector in direction of $(m_1 - m_2)$

Or separation of the projected means which is m 1 tilde minus m 2 tilde square of this. Where m 1 tilde is nothing but W transpose m 1. m 2 tilde is nothing but w transpose m 2 square of this. I can re-orient or re-organise this expression as W transpose m 1 minus m 2 into m 1 minus m 2 transpose W, and you find that this is something like between class scatter. So, we had this S 2 to be within class scatter, because for s omega to be within class scatter. Because for computation of this I am considering the samples belonging to individual classes computing their scatters and then adding these inter class scatters.

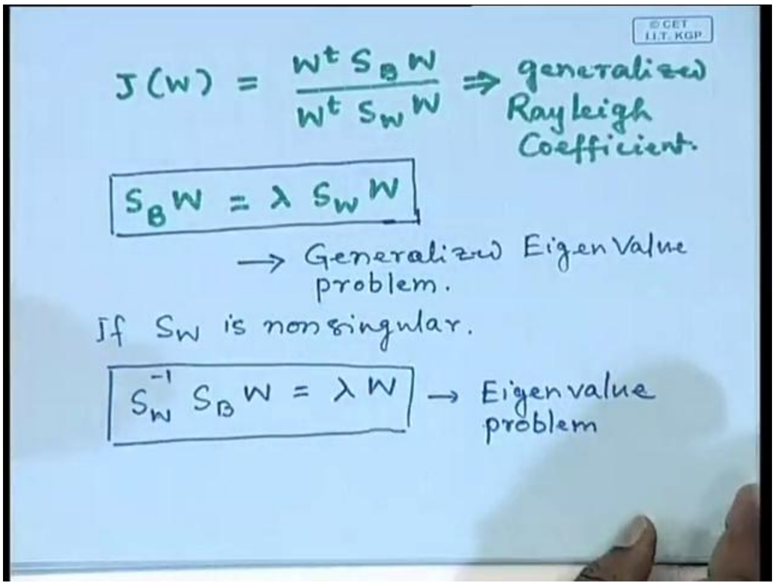
So, this is what is called within class scatter, and following the similar expression this expression m 1 minus m 2 into m 1 minus m 2 transpose. This tells you something like between class scatter. So, I can write this expression as W transpose S B W, where this S B we will call as between class scatter. Now, we will find that this is W or within class scatter. This is symmetric and positive semi definite and so is this between class scatter S B that is the property of this scatter matrices and not only that, when I am writing this expression as W transpose as P W, where this S B is m 1 minus m 2 into m 1 minus m 2 transpose m 1 minus m 2. It is nothing but a vector, row vector m 1 minus m 2, sorry this is a column vector and m 1 minus m 2 transpose is a row vector.

So, this is actually the outer product of two vectors. So, as it is outer product of two vectors so rank of this scatter matrix S B will be at the most 1, not more than 1. And not only that, I can have some more observation about S B or I can have some more observation about S B W. You will find that when I consider this part. Yes, so this portion m 1 minus m 2 transpose W, this is the inner product of two vectors. Because this is a row vector, this is a column vector of same dimension.

So, this term is inner product of two vectors. As it is inner product of two vectors so it is a scalar, and when this is scalar this total term m 1 minus m 2 into m 1 minus m 2 transpose W. This being a scalar, this whole portion will be a vector and the vector in the direction of m 1 minus m 2. So, when I am writing this as S B W you find that this S B will be a vector in the direction of m 1, sorry S B W will be a vector in the direction on m 1 minus m 2, because S B W is nothing but this, right?

So, this S B W is a vector in direction of m 1 minus m 2. So, we will make us of this while solving for or projection direction W. So, I have this different scatter matrices and the corresponding expressions. So, once I have this, I have this within class scatter, I have this between class scatter and what we said is that is should define a criteria function, which is a ratio of measure of between class scatter and measure of within class scatter, okay?

(Refer Slide Time: 43:20)



© CET I.I.T. KGP

$$J(W) = \frac{W^t S_B W}{W^t S_W W} \Rightarrow$$ generalised Rayleigh Coefficient.

$$\boxed{S_B W = \lambda S_W W}$$

→ Generalized Eigen Value problem.

If $S_W$ is nonsingular.

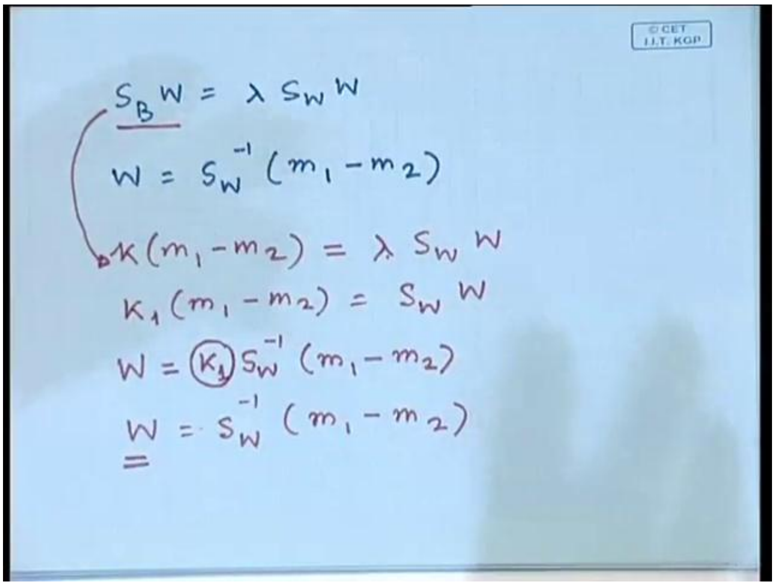
$$\boxed{S_W^{-1} S_B W = \lambda W}$$ → Eigenvalue problem

So, now if I redefine our criteria function as J W is equal to, in earlier case it was m 1 minus m 2 square upon m 1 tilde minus m 2 tilde square upon S 1 tilde square plus S 2 tilde square, right? And we have found that this m 1 tilde minus m 2 tilde square is nothing but W transpose S W W and S 1 tilde square plus S 2 tilde square is nothing but W transpose S W W. So, I get an expression of this criteria function J explicitly in terms of W is given by W transpose S B W upon W transpose S W W, where S B is between class scatter and S W is total within class scatter.

So, what I have to do is we have to maximise this ratio by varying the vector W. And W gives you the projection direction and that is the tedious mathematics. So, without going for tedious mathematics, let us see what is the solution. So, S W which will maximize this ratio must satisfy S B W is equal to lambda S W W for some constant lambda. So, the W which maximizes this ratio must satisfy this condition for sum constant lambda this expression. This ratio is known as generalized Rayleigh coefficient and this expression S B W is equal to lambda S W W. It is nothing but a generalized Eigen value problem.

So, we find that if S W is non-singular then this expression will be simply in the form S W inverse S B W is equal to lambda times W. So, this is our well known Eigen value problem, where this W is nothing but Eigen vector of S W inverse S B and lambda is the corresponding W is the Eigen vector of S W inverse S B. And lambda is the corresponding Eigen value, but we can simplify our solution instead of trying to solve the Eigen values and Eigen vectors simply, because of the fact that S B W as we said that this S B W is a vector in the direction of m 1 minus m 2.

(Refer Slide Time: 47:52)



© CET
I.I.T. KGP

$$\underline{S_B} W = \lambda S_W W$$

$$W = S_W^{-1}(m_1 - m_2)$$

$$K(m_1 - m_2) = \lambda S_W W$$

$$K_1(m_1 - m_2) = S_W W$$

$$W = \boxed{K_1} S_W^{-1}(m_1 - m_2)$$

$$\underline{\underline{W}} = S_W^{-1}(m_1 - m_2)$$

So, by using this and from the relation that S B W is equal to lambda S W W I can find out what should be the direction W, because as I said the scale factor of W is not much important. What I want to see is what is the direction of W, that is my projection direction. So, from this you can easily find that W is nothing but S W inverse m 1 minus m 2. Why is it so? Because as we said that S b w is a vector in the direction of m 1 minus m 2 so I can write this S B W as sum scale factor K times m 1 minus m 2. Because it is a vector in the direction of m 1 minus m 2, right?

So, I can write this as sum scale factor K into m 1 minus m 2 that is equal to lambda S W, lambda is a constant. So, I can put this constant under this constant K, because it simply becomes K upon lambda. So, equivalently I have sum constant K 1 times m 1 minus m 2 is equal to S W W which clearly says that this W is nothing but sum constant K 1 times S W inverse into m 1 minus m 2 is it. And as I said that this scale factor is not that important to me, because I am interested in the direction of W. The scale factors will indicate that how much will be the separation between the two classes and in the projected domain how the points will be distributed.

If I increase the scale factor, separation will be more. The points will be more spreaded if I reduce the scale factor. The separation will be less points, will be less spreaded. So, both of them the separation between the classes as well as the variance in the projected domain, both of them vary simultaneously depending upon the value of K in the same scale. So, this scale K or K 1 is not that important. What is important is what is the direction W. So, as such I can ignore this scale K 1 and if I ignore this scale K 1, I simply have W is equal to S W inverse into m 1 minus m 2. Where this S W is total within class scatter and m 1 minus m 2 is a vector in the direction of line joining m 1 and m 2, right?

So, if I use this projection direction for projecting the data into lower dimensional space, I get I can maintain the separation among the classes and features data for classification purpose. However, this particular projection may not be the best in terms of data representation as far as minimum squared error is concerned. So, what I have done is till not is a two class problem, and your classification design gets complete because what I have done is from multiple dimension. I have projected onto single dimension and in the single dimension I have to find out a threshold so that if this one dimensional vector. So, the scalars are less than the threshold, I will put them into one class. If it is more than the threshold, I have to put them into another class.

So, I have to determine the threshold value that is all, and the classifier which makes use of this concept is as I said is called features classifier or feature linear discriminator. Now, when I have seen number of classes, I have to extend from two class problem to C class problem. And it can be shown that in that case it simply becomes C minus one linear classifiers or the projection that now I am taking into one dimension. I have to take projections onto C minus 1 number of different directions.

So, the problem remains the same, simply it is extension from two dimension to C dimension or the dimensionality of the feature vectors. Here it is one dimensional feature vector and there we have to go for C minus one dimensional feature vector and obviously

I will gain if C minus 1 is less than D. Where D is the original dimensionality of the feature vectors, and C minus 1 is the reduced dimensionality of the feature vectors.

So, we have discussed about this dimensionality problem and two ways to solve to tackle the dimensionality problem. One approach tries to find out a projection which is the best in terms of data representation that is the principle component analysis approach. And the other approach that is this M D A or multiple discriminant analysis is the one which is again a projection onto a lower dimensional space. But it tries to guarantee that when you take the projection, the separability of the feature vectors between different classes that will be maintained which is not guaranteed by principle component analysis. So, we will stop here today. Next day we will start some other topic.