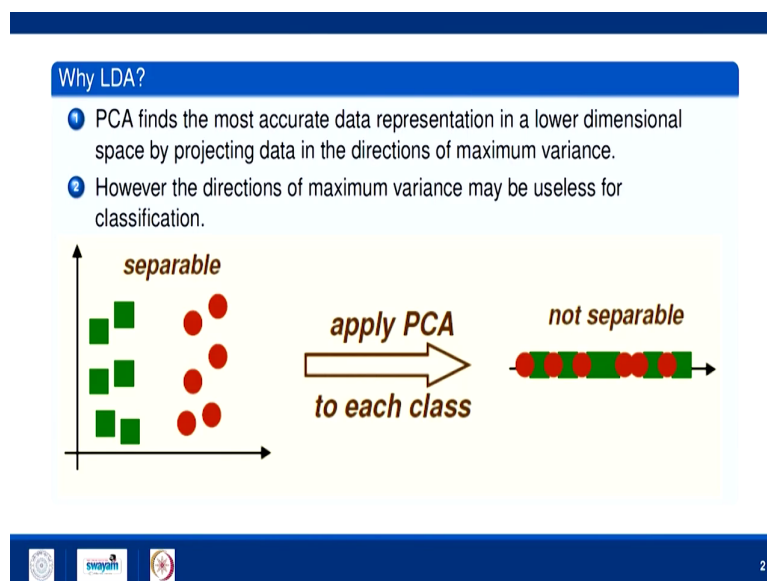


Essential Mathematics for Machine Learning
Prof. Sanjeev Kumar
Department of Mathematics
Indian Institute of Technology, Roorkee

Lecture - 18
Linear Discriminant Analysis

Hello friends. So, welcome to Module 18 of this course Essential Mathematics for Machine Learning. In first couple of lectures, we have seen about principal component analysis. In this lecture, we will talk about Linear Discriminant Analysis that is again a very popular technique to play with data in machine learning and again it is very easy mathematical concept based on eigen values and eigen vectors.

(Refer Slide Time: 00:56)



So, let us start it and why we need LDA. So if you see this data set here, it is a two dimensional data set and again two classes. One class is represented by the green square and

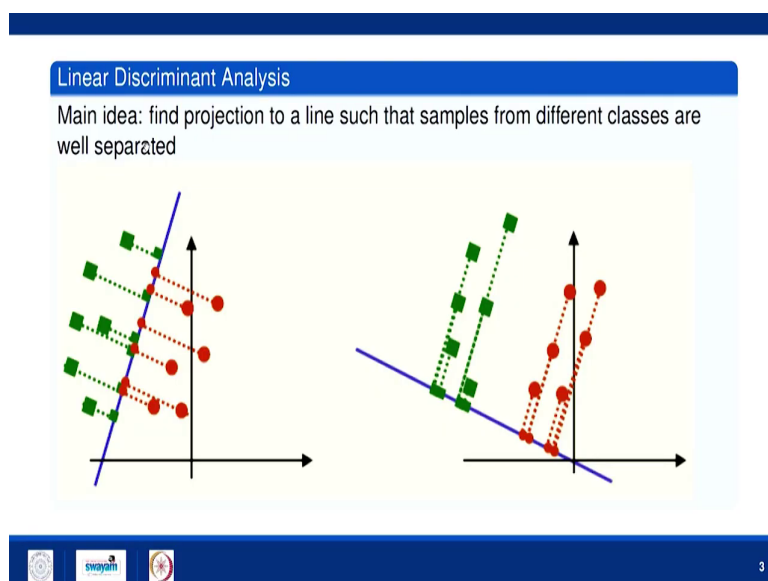
another one with red circle. So, if you see this class is easily linearly separable. If you put a line here, then the data become linearly separable. If you apply the PCA on this data PCA project it to lower dimensional space that is one dimensional space by projecting data in the direction of maximum variance.

So, if I see this is the after applying the PCA on this data, I am having this kind of data. So, now you can see I am having mixing of this data. This data is no more linearly separable like here because here you can separate this data in are two by a line, but here in one d you cannot find out any point. One side of that point you are having green patterns and another side of the point you are having the red class pattern.

So, what we are observing here the direction of maximum variance may be useless for classification. Why because my data is linearly separable here. I can use a linear classifier for classifying this data, but if I am coming to lower dimensional space, what I am having? My data is no more linearly separable.

So, why to come to lower dimensional or in other way I want to say that I can go to lower dimensional, but it should preserve the property of linear classification of the data means if the data is linearly separable in the higher dimensional space, then after projecting into lower dimensional space, the data should be still linearly separable. So, how to find out such a line, so that the property of linear separability should be preserved?

(Refer Slide Time: 03:09)



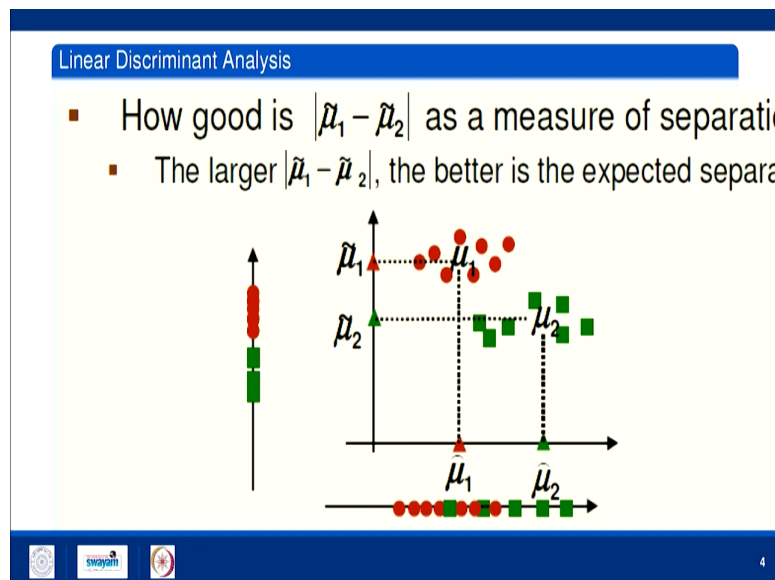
So, for example, here the main idea of linear discriminant analysis is find projection to a line such that samples from different classes are well separated or linearly separated.

So, for example, if you see again these data, so this is the example of PCA and it is not well separated, but instead of this if the same data I project onto this line, you can see this data here. You will find the green cluster here, you will find the red cluster and this data will be well separated. So, now the objective of linear discriminant analysis is to find out the direction of such a line for a given data set like the objective of PC was to find out the direction of maximum variances.

Here the objective is different. Find direction such that if we project the data on the line on those directions or on the sub-space of those direction, the data should be well separated. So, this is the idea of linear discriminant analysis. Now, in this lecture we will learn how to do it

and again I told you like PCA, it is very easy. Just we will play with eigen values and eigen vectors of the matrix.

(Refer Slide Time: 04:34)



So, let us see here if I am saying that after projection my data should be well separated or linearly separated, then one can say in that case what should we have? We should have a small clusters like here of the data or not a small cluster, but the centroid of the data after projection should be far away from each other. So, for example, if you see here this μ_1 is the centroid of the data in two dimension means before projection and μ_2 is the mean of the data before projection for class 2.

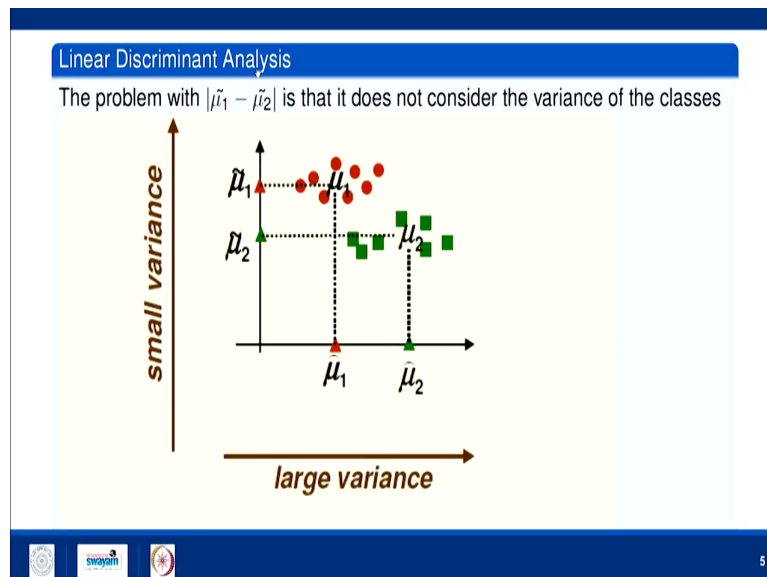
So, I am having two class data and then, if I project it on to x axis means on the on to a horizontal line this μ_1 cap and μ_2 cap are the centroid of the data after projection to the horizontal line. So, similarly μ_1 tilde and μ_2 tilde are the projection of μ_1 and μ_2 that

is the centroid of the data of different classes before projection to after projecting on to a vertical line.

So, if I project data onto a horizontal line you can see this is the distance between μ_1 and μ_2 . If I project onto vertical line, this is the distance between $\tilde{\mu}_1$ and $\tilde{\mu}_2$. So, now if I am assuming that the mean of the two classes after projection should be far away from each other means maximize the distance between the mean, then in this case this distance is bigger than this one but in this direction the data is not well separated while in this direction the data is well separated.

So, the concept or the idea which we have taken that larger the distance between the mean of the two classes, the better is the data will be well separated after projection. No, it is not true due to this example. So, we are missing something.

(Refer Slide Time: 06:48)



The mean should be far away from each other is that it does not consider the variance of the classes because in this direction if I project on to horizontal line, the data is not well separated.

Why because in this direction I am having bigger variance of the data when compared to vertical direction because in vertical direction, variance is quite small.

(Refer Slide Time: 07:25)

LDA:-
 Suppose we have two classes and a d -dimensional samples x_1, x_2, \dots, x_n , where
 * n_1 samples are coming from class-1 (c_1)
 * n_2 samples are coming from class-2 (c_2)
 → Now if x_i be a datapoint, then its projection on the line having direction given by unit vector v is given as $v^T x_i$
 → Let μ_1 and μ_2 be the means of class c_1 and c_2 , respectively before projection.
 If $\tilde{\mu}_1$ denote the mean of samples of class c_1 after projection, then

$$\tilde{\mu}_1 = \frac{1}{n_1} \sum_{x_i \in c_1} v^T x_i = v^T \left(\frac{1}{n_1} \sum_{x_i \in c_1} x_i \right) = v^T (\mu_1)$$

So, I have to normalize this idea that is the mean should be far away from each other after projection by the variance. How to do it? So, how to normalize it? So, concept of LDA is clear to you by now.

Now, suppose we have two classes and a d -dimensional samples $x_1 \times x_2 \times \dots \times x_n$ where n_1 samples are coming from class 1. So, let us say this class is c_1 and n_2 samples are coming from class 2 let us say c_2 .

Now, if x_i be a data point, so what n_1 samples are coming from class 1 and n_2 samples are coming from class 2. So, n_1 plus n_2 equals to n . Now if x_i be a data point, then its projection on the line having direction given by unit vector v is given as $v^T x_i$ that is the dot product between v and x_i .

So, we are assuming that we are having n samples from d -dimensional space, n_1 samples from class 1, n_2 samples are coming from class 2 and the projection as I told you that if I want to project a point x_i on a line having direction given by unit vector v , then this projection is given by $v^T x_i$.

Now, let μ_1 and μ_2 be the means or centroid of class c_1 and c_2 respectively. Before projection means in original dimension or for original data points then if μ_1 denote that the mean of samples of class 1 that is class c_1 after projection, then what we are having μ_1 equals to how many total points from class 1 n_1 ?

So, $\frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ all points belongs to class 1. So, x_i belongs to c_1 and these are n_1 points $v^T x_i$. So, I can take v^T out $\frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ belongs to c_1 to n_1 point x_i . This is v^T and what is this? This is μ_1 . So, μ_1 is $v^T \mu_1$.

(Refer Slide Time: 12:35)

Similarly, we have
 $\tilde{\mu}_2 = v^T \mu_2$

In LDA, we need to normalize $|\tilde{\mu}_1 - \tilde{\mu}_2|$ by scatter.
Let $y_i = v^T x_i$ be the projected samples.
then scatter for samples of class c_1 is

$$\tilde{s}_1^2 = \sum_{y_i \in c_1} (y_i - \tilde{\mu}_1)^2$$
$$\tilde{s}_2^2 = \sum_{y_i \in c_2} (y_i - \tilde{\mu}_2)^2$$

The slide features a dark blue header and footer. The footer contains a small logo on the left, the word 'swayam' in the center, and the number '3' on the right.

Similarly, we have μ_2 tilde equals to v transpose μ_2 . In LDA what we need in LDA we need to normalize the distance between the two means after projection that is absolute value of the difference of μ_1 and μ_2 by variance or I am writing scatter. So, now how to define these scatters? So, let y_i equals to v transpose x_i means be the projected sample.

So, then the scatter for samples of class c_1 is given by let us say s_1 tilde square and this is y_i belongs to class c_1 and then y_i minus μ_1 tilde whole square. So, it is variance only just we have not taking $1/n_1$ here. Similarly for class 2, it will become y_i belongs to c_2 y_i minus μ_2 tilde square.

(Refer Slide Time: 14:55)

Thus, we need to project our data onto a line having direction v such that which maximizes

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Class-1 scatter after projection should be small

Class-2

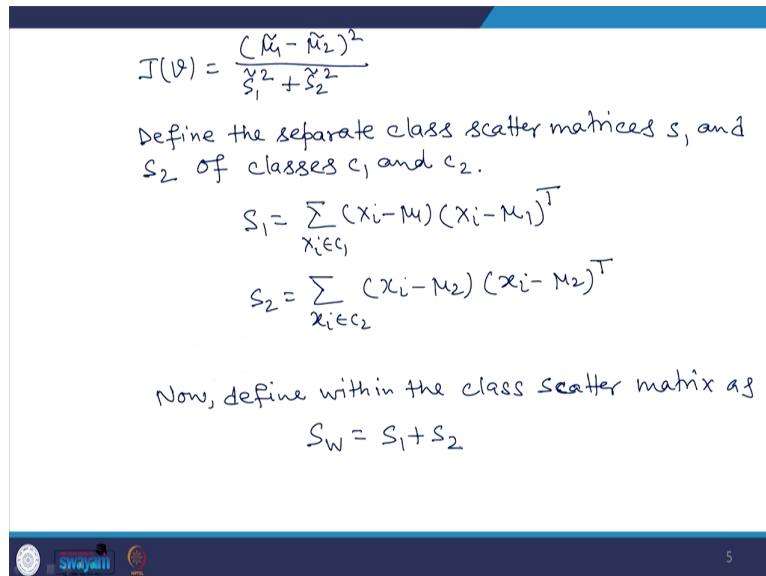
[If we find v which makes $J(v)$ large, we are guaranteed that the classes are well separated.]

So, these are these two are giving the scatters. Now, thus we need to project our data onto a line having direction v such that which maximizes. So, something like this somewhat it should maximize. So, instead of absolute I am taking the square. So, the same thing will happen. So, it will say you that mean of the two classes after projections are far away from each other.

So, distance between the two means after projection we are maximizing why we have to normalize it. So, so what it is saying? It is saying that the class 1 is scatter after projection should be small. Similarly it is saying for class 2 and it is saying you that mean of the two class should be as far away as possible. So, in that way what we are having? We are imposing both the conditions here that the mean should be far away from each other and we have normalized that by the scatter also.

So, now how to do it? So, now if we find v which makes $J v$ large we are guaranteed that the classes are well separated, ok. So, this is important point. So, first what we need to do, we need to do this objective function, we need to write in terms of v .

(Refer Slide Time: 18:09)



Handwritten mathematical derivation on a slide:

$$J(v) = \frac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

Define the separate class scatter matrices S_1 and S_2 of classes c_1 and c_2 .

$$S_1 = \sum_{x_i \in c_1} (x_i - \mu_1)(x_i - \mu_1)^T$$

$$S_2 = \sum_{x_i \in c_2} (x_i - \mu_2)(x_i - \mu_2)^T$$

Now, define within the class scatter matrix as

$$S_W = S_1 + S_2$$

The slide includes a Swayam logo and a page number 5 in the bottom right corner.

So, how to write it in terms of v ? So, now what we are having, we are having $J v$ equals to μ_1 tilde minus μ_2 tilde square upon S_1 tilde square plus S_2 tilde square.

So, we need to write this J in terms of v . So, now define the separate class scatter matrix S_1 and S_2 of classes c_1 and c_2 means before projection. So, what will be S_1 . So by the concept of covariance matrix, it will be x_i belongs to c_1 x_i minus μ_1 multiplied by x_i minus μ_1 transpose.

So, by the concept of covariance matrix only thing we are not dividing it by $1/(n-1)$, similarly S_2 will become $\sum (x_i - \mu_2)^2$ multiplied with $(x_i - \mu_2)^T$. So, what will happen using this? We will be having two matrix, S_1 and S_2 . So, once you are having these two matrix, now define within class scatter matrix as S_w equals to $S_1 + S_2$.

(Refer Slide Time: 20:20)

$$\begin{aligned} \tilde{S}_1^2 &= \mathbf{v}^T S_1 \mathbf{v} \\ \tilde{S}_2^2 &= \mathbf{v}^T S_2 \mathbf{v} \end{aligned} \quad \Rightarrow \quad \tilde{S}_1^2 + \tilde{S}_2^2 = \mathbf{v}^T (S_1 + S_2) \mathbf{v} \quad (*)$$

Define between the class scatter matrix

$$\begin{aligned} S_B &= (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T \\ (\tilde{\mathbf{M}}_1 - \tilde{\mathbf{M}}_2)^2 &= (\mathbf{v}^T \mathbf{M}_1 - \mathbf{v}^T \mathbf{M}_2)^2 \\ &= \mathbf{v}^T (\mathbf{M}_1 - \mathbf{M}_2)(\mathbf{M}_1 - \mathbf{M}_2)^T \mathbf{v} \\ &= \mathbf{v}^T S_B \mathbf{v} \quad (***) \end{aligned}$$

So, now what is \tilde{S}_1^2 ? That is the scatter after projection for class 1 patterns. So, if you do a bit calculation what you will find, it is coming out to be $\mathbf{v}^T S_1 \mathbf{v}$ and similarly you will get $\tilde{S}_2^2 = \mathbf{v}^T S_2 \mathbf{v}$. So, from here if I see the denominator of J that is $\tilde{S}_1^2 + \tilde{S}_2^2$, so it is $\mathbf{v}^T (S_1 + S_2) \mathbf{v}$. And what is $S_1 + S_2$? That is within class scatter matrix that is S_w .

So, this is the denominator of J . So, this equals to this one. So, let us say now define between the class scatter matrix that is S_B between class scatter matrix, then certainly it will be the difference of two means and then transpose of that product of that.

Now, what this S_B measures? S_B measure separation between the means of two classes before projection. So, means of two classes after projection means separation of the means of two classes after projection is just $v^T \mu_1 - v^T \mu_2$ square. This I can write $v^T \mu_1 - \mu_2^T \mu_1 - \mu_2^T \mu_2$ transpose into v , then this comes out to v^T plus S_B into v .

(Refer Slide Time: 22:56)

$$\max_v J(v), \text{ where } J(v) = \frac{(\bar{\mu}_1 - \bar{\mu}_2)^2}{s_1^2 + s_2^2} = \frac{v^T S_B v}{v^T S_W v}$$

$$\frac{d}{d(v)} J(v) = 0 \Rightarrow S_B v - \frac{v^T S_B v (S_W v)}{v^T S_W v} = 0$$

$$\Rightarrow S_B v - \lambda S_W v = 0$$

$$\Rightarrow S_B v = S_W (\lambda v)$$

$$\Rightarrow \underline{S_W^{-1} S_B v = \lambda v}$$

$$\Rightarrow \underline{M v = \lambda v}$$

v is the eigenvector of $S_W^{-1} S_B$ corresponding to largest eigenvalue.

So, this is the numerator of $J v$. So, now what $J v$ is we have to find out v which maximize $J v$ where $J v$ equals to earlier it was $\mu_1 - \mu_2^2$ upon $S_1^2 + S_2^2$ square and this we have reduce is $v^t S B v$ upon $v^t S W v$.

So, for extremizing this one what we have to do? We have to make d by $d v$ of $J v$ equals to 0 and if you do it, it will give you after certain calculation that $S B v - \lambda S W v$ equals to 0. Now, see this value what is this? It is a scalar value and that is your $J v$ means which you need to maximize.

So, let us assume that it is your λ . So, it will become $S B v - \lambda S W v$ equals to 0 or $S B v = \lambda S W v$ or I can if $S W$ is invertible $S W^{-1} S B v = \lambda v$. Now, this is a matrix let us say this I am writing $M M = \lambda v$.

So, you have to find out v which maximize the λ . And now what is λ here? By this you can see that because v should be a non-zero vector, it is a direction vector of the line. So, it is the eigen value of M by the definition of eigen values and eigen vectors. So, and what you have to maximize the λ , so which eigen value you have to take which is the largest one because you have to maximize $J v$ and what is $J v$. $J v$ is your λ only.

So, you have to maximize λ . So, you have to take λ which is the largest means largest eigen value of M . So, what is v here. So, I can write v is the eigen vector of $S W^{-1} S B$ corresponding to largest or biggest eigen value. So, what you do samples are with you. You can easily find out capital $S W$ and capital $S B$ because capital $S B$ will come from the means and capital $S W$ will come from the scatters from the covariance of two classes.

So, once you are having capital $S W$ matrix and capital $S B$ matrix, then you can easily find out $S W^{-1}$ into $S B$. So, v is the direction given by the eigen vector of $S W^{-1} S B$ corresponding to the largest eigen value. So, this is easily you can calculate.

(Refer Slide Time: 27:02)

$$\Rightarrow S_B v = \lambda S_W v$$

$$\Rightarrow S_W^{-1} S_B v = \lambda v$$

But $S_B x$ points in the same direction as $\mu_1 - \mu_2$.

$$S_B x = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T x = \alpha (\mu_1 - \mu_2)$$

$$\Rightarrow v = S_W^{-1} (\mu_1 - \mu_2)$$

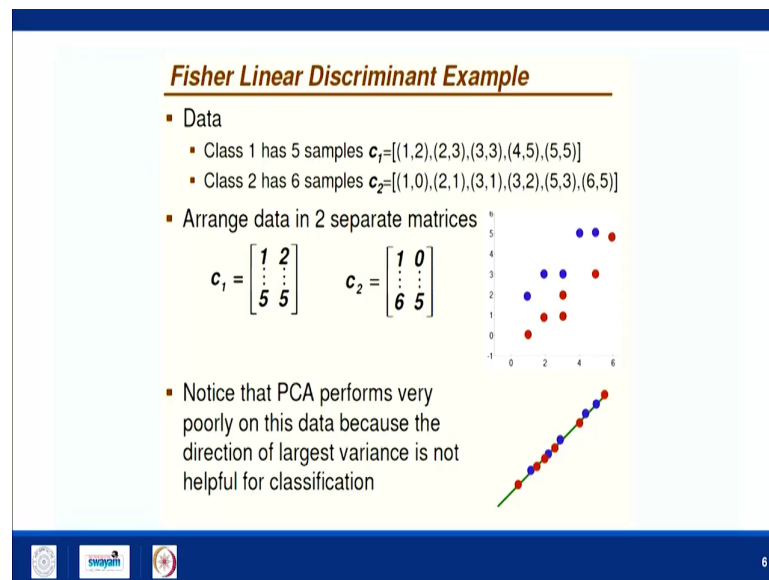
However more we can do more manipulation in this that what you are having, you come out $S_B v$ equals to $\lambda S_W v$. So, if S_W is full rank that is inverse exist $S_W^{-1} S_B v$ equals to λv , but what you are having for any vector x $S_B x$ for points in the same direction as $\mu_1 - \mu_2$.

Why? Because $S_B x$ equals to $\mu_1 - \mu_2 (\mu_1 - \mu_2)^T x$ and it is because $\mu_1 - \mu_2 (\mu_1 - \mu_2)^T x$ will be a some scalar. So, $\alpha (\mu_1 - \mu_2)$. So, $S_B x$ is some scalar times $\mu_1 - \mu_2$. So, they points in the same direction. So, what I can do? So, from here I can make that in that case v equals to $S_W^{-1} (\mu_1 - \mu_2)$.

So, even though you no need to calculate S_B because S_W is there. Once you are having S_W , find out the inverse and the direction of v is given by the product of S_W^{-1} with the

vector $\mu_1 - \mu_2$. If W is not full rank, then you can make use of some kind of pseudo inverse in this case. So, let us see an example of this.

(Refer Slide Time: 29:15)



So, suppose I am having this data. We are having total 11 points. Class 1 has 5 samples. So, these 5 points are (1,2), (2,3), (3,3), (4,5), (5,5).

So, these are denoted by these blue points. Similarly class 2 has 6 points given by these coordinates. So, this is a two-dimensional data and I want to project this data onto a line in 1D, so that it should remain linearly separable. Well, the basic separability should be guaranteed. So, what you do? First you arrange these data points into two separate matrices. So, here I am having five samples.

So, it will be a 5 by 2 matrix and it will be a 6 by 2 matrix. Let us say these 4 class 1, I am saying c 1, for this I am saying c 2. If you see in a PCA, PCA project on to this line and you can see here the data is not well separated.

(Refer Slide Time: 30:21)

Ex: $C_1 = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 5 \end{bmatrix}_{5 \times 2}$; $C_2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 6 & 5 \end{bmatrix}_{6 \times 2}$

$\mu_1 = [3 \quad 3.6]$; $\mu_2 = [3.3 \quad 2]$

$S_1 = 4 * \text{Cov}(C_1) = \begin{pmatrix} 10 & 8 \\ 8 & 7.2 \end{pmatrix}$

$S_2 = 5 * \text{Cov}(C_2) = \begin{pmatrix} 17.3 & 16 \\ 16 & 16 \end{pmatrix}$

$S_W = S_1 + S_2 = \begin{pmatrix} 27.3 & 24 \\ 24 & 23.2 \end{pmatrix}$

$S_W^{-1} = \begin{pmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{pmatrix}$

$v = S_W^{-1}(\mu_1 - \mu_2) = \underline{\underline{\begin{pmatrix} -0.79 \\ 0.89 \end{pmatrix}}}$

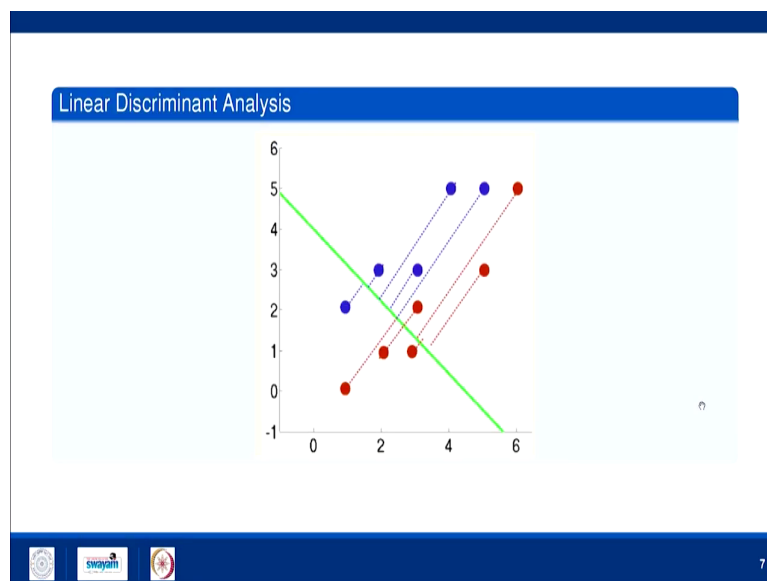
So, now you what you do? You are having c 1 which is a 5 by 2 matrix. So, I am continuing with the same example and c 2 which is again a 6 by 2 matrix. So, it will be having 1 2 and then 5 5.

So, blue points and it is 1 0, 6 5 means red points. Now you compute mu 1. Mu 1 is the mean of this class c 1 patterns. So, it comes out to be 3 which is the mean average of this column and average of second column 3.6. Similarly I calculate mu 2 mu 2 comes out to be 3.3 which is the average of this column and then 2 2 and 2.

Now, calculate S_1 . So, S_1 is the covariance matrix four times covariance matrix of c_1 and this comes out to be $\begin{bmatrix} 1 & 10 \\ 8 & 8 \end{bmatrix}$ and 7.2. Similarly S_2 will be 5 times covariance of C_2 . So, from these two columns you can easily find out covariance of c_2 and this comes out to be $\begin{bmatrix} 17.3 & 17.3 \\ 16 & 16 \end{bmatrix}$.

So, here S_W within class scatter matrix is S_1 plus S_2 and this S_1 plus S_2 becomes $\begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$. Now S_W inverse becomes $\begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$. Now, from the previous derivation what is the line v which maximize J_v that is S_W inverse into μ_1 minus μ_2 and once you calculate it, it comes out to be $\begin{bmatrix} -0.79 \\ 0.89 \end{bmatrix}$. So, this is the line and then you can project all the points there.

(Refer Slide Time: 33:10)



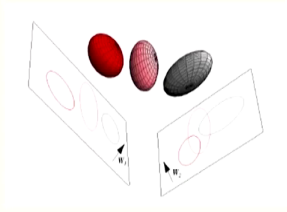
So, I have given here. So, this is the line and once you project these point here, this will be the projection and this is the best possible linear separable data after projection on to 1D. So, this

is about linear discriminant analysis. Now, you can generalize into multiple classes because this derivation we have made using only two classes.

(Refer Slide Time: 33:37)

Multiple Discriminant Analysis

- Can generalize FLD to multiple classes
- In case of c classes, can reduce dimensionality to 1, 2, 3, ..., $c-1$ dimensions
- Project sample \mathbf{x}_i to a linear subspace $\mathbf{y}_i = \mathbf{V}^t \mathbf{x}_i$
 - \mathbf{V} is called projection matrix

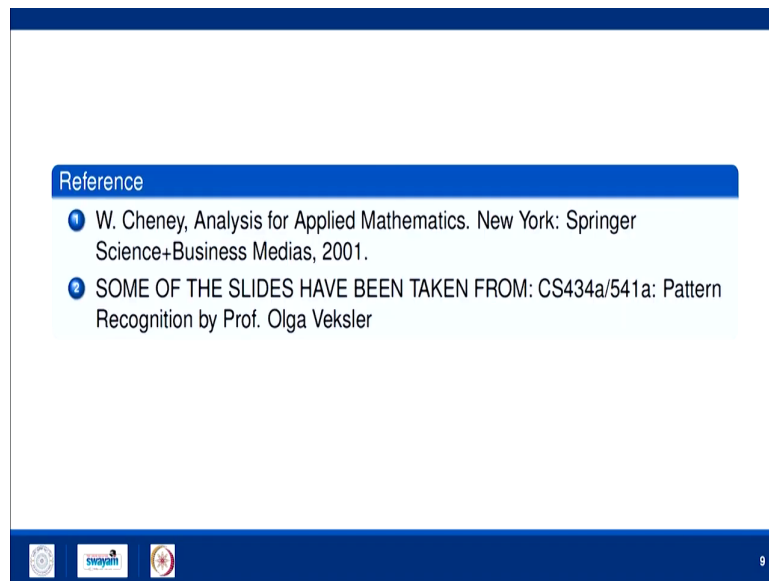


8

So, in case of c classes can reduce to dimension to up to 1 2 3 up to c minus one-dimension. So, if you are having like 10 classes, you can reduce the dimension up to 9. Project sample \mathbf{x}_i to a linear sub space \mathbf{y}_i . So, now it will be the subspace of \mathbf{x}_i and this projection will be given by the projection matrix \mathbf{v} transpose.

So, for example here I am talking about three-dimension. So, you are projecting here in w_1 , they are well separated in r_2 while here it is not well separated. So, I have to find out this projection matrix and this you can easily find out the using the concept of linear discriminant analysis.

(Refer Slide Time: 34:25)



Reference

- 1 W. Cheney, Analysis for Applied Mathematics. New York: Springer Science+Business Medias, 2001.
- 2 SOME OF THE SLIDES HAVE BEEN TAKEN FROM: CS434a/541a: Pattern Recognition by Prof. Olga Veksler

9

So, these are the references for this lecture. Some of the slides I have taken from this course notes.

Thank you very much.