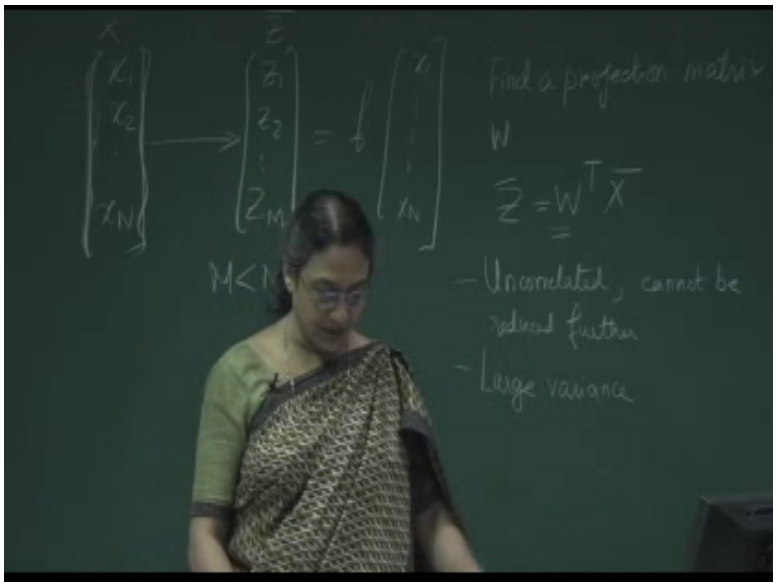
Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module - 3
Lecture - 12
Feature Extraction

Good morning. So, today we will start the third lecture in the module on Instance Based Learning and Feature Reduction.

In the last class, we have talked about feature selection. In today's class we will talk about feature extraction. So, in feature extraction what we have is that we have n dimensional features x 1, x 2, x n.

(Refer Slide Time: 00:42)



And we want to map it to a lower dimensional space which is m dimensional and we want to get the features Z 1, Z 2, Z m. So, this is the new features where m is less than n and each of these features is sum function of the original feature set x 1, x 2, x n. You note that in feature selection which we covered in the last class we said we take a subset of the features, instead what we are doing in feature extraction is that we are taking the original features and mapping it to a lower dimensional space and each feature is

obtained as a function of the feature set.

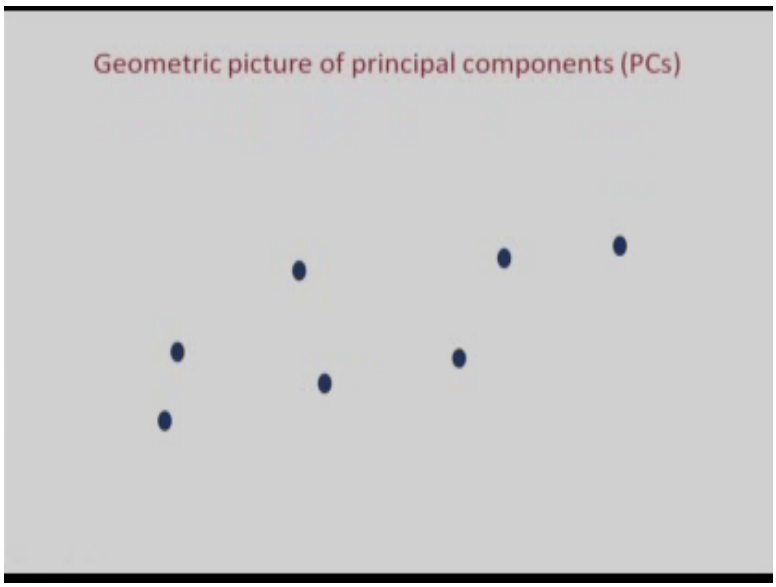
So, it is a projection of a higher dimensional feature space to a lower dimensional feature space and we want to make this projection, so that the smaller the dimensional feature set can help us to have better classification or faster classification as we have noted in the last class.

So, what we need to do is that, we need to find a projection matrix W such that Z equal to W T X, so this is Z and this is X. So, we want to find W T where Z is W T X which is the projection from n dimensional space to m dimensional space. And what we expect from such a projection is that the new features they are uncorrelated and cannot be reduced further. In the last class, we noted that features can be redundant and therefore, we can have larger number of features when we map into smaller space we want that the features are not redundant among themselves that they are uncorrelated and cannot be reduced further.

Secondly, we want features to have large variance or large variation, why? Because, if the feature takes similar value for all the instances that feature cannot be used as a discriminator, so since we want the features to be able to distinguish between the different instances we encourage larger variation or larger variance between the features because, otherwise the features would not (Refer Time: 03:51) any information.

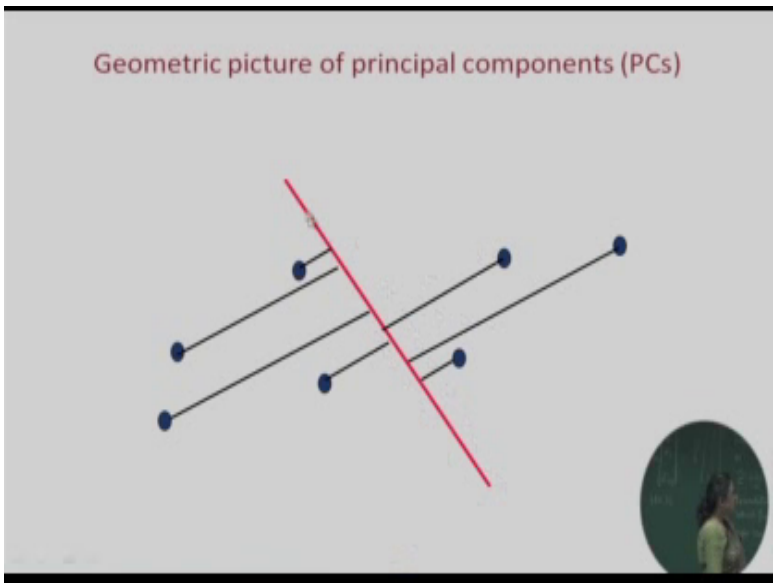
Now, let us look at the slides.

(Refer Slide Time: 03:57)



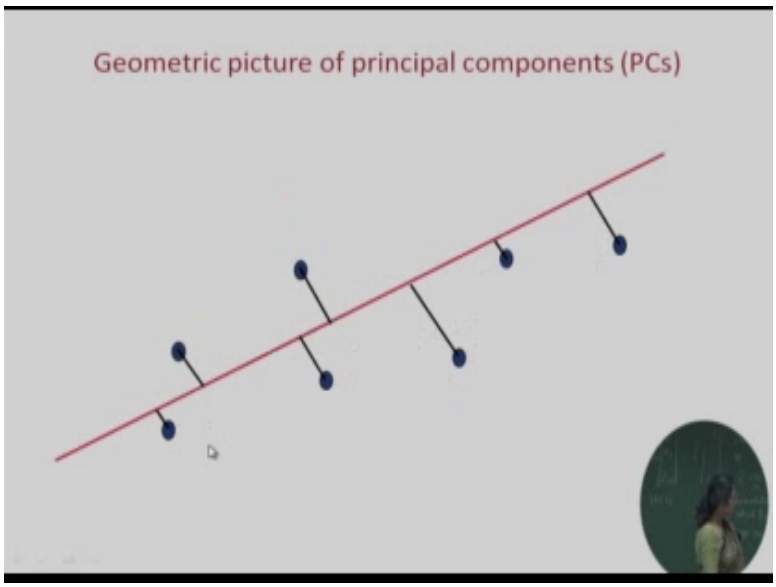
So, suppose you have the following instances and they are in the two dimensional feature space. There are two features x 1 and x 2, and in that space you have these instances. Now you have a choice of deciding suppose you want to map this two dimensional feature space to a one dimensional feature space, you have to select that axis.

(Refer Slide Time: 04:24)



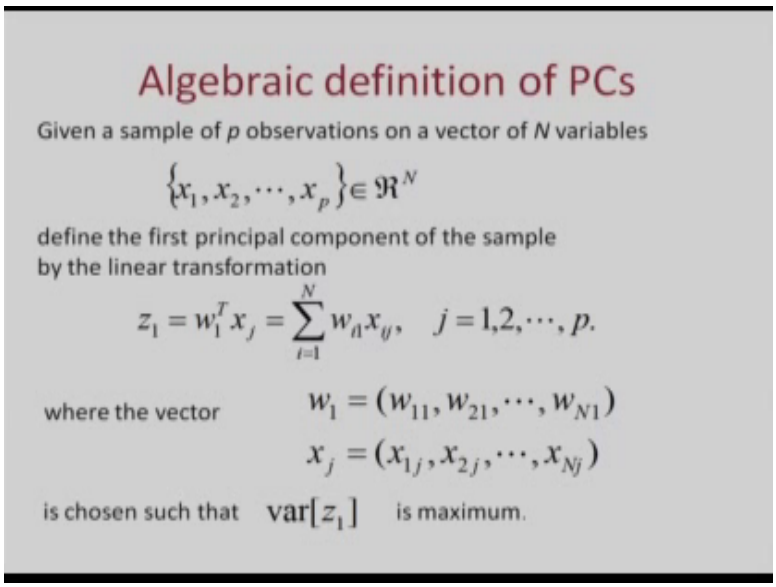
Now this is the possible axis. Now, along this axis you can, so this is the axis and you can project each of these instances on this axis. So, after projection, this point will map to this point on this axis, this point will map here and this point will map here. So, the different points will map here, here, here, here, here, here, here, right.

(Refer Slide Time: 04:49)



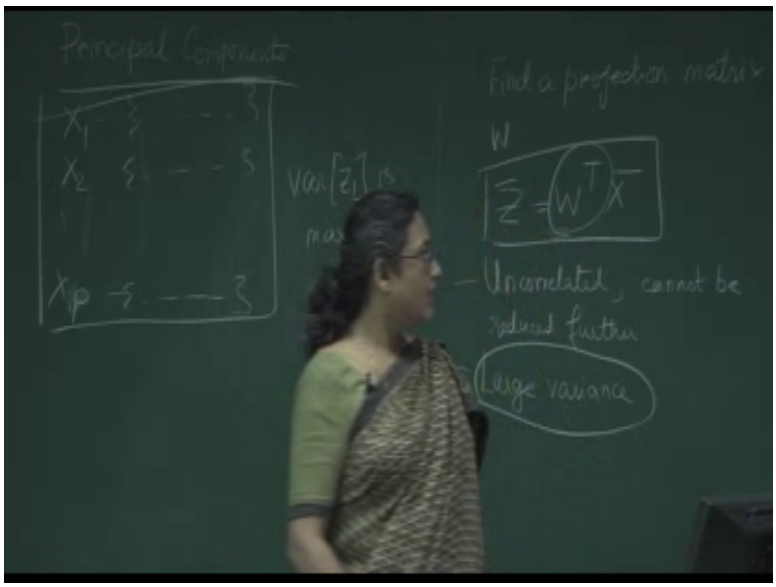
In contrast we could have taken the axis like this, and if we did that this point will map here, here, here, here, here. So, can you tell me out of these two possible projections, so this is let us say 1 and this is 2, which one would you prefer? So, you notice that in 2 there is a larger variation, this is a larger variance among the features. So, we would by the principal that we are going by prefer this (Refer Time: 05:26) this axis to the previous axis.

(Refer Slide Time: 05:28)



So, based on this what we do is that some of you may have study mathematics and are familiar with principal components.

(Refer Slide Time: 05:44)



For feature extraction this principal components play a very important role. So, we can take the principal components of the data points that we have and take the principal

components the first principal component, the second principal component, etcetera the top principal components as the new axis and that will give us certain properties which satisfy our objectives of getting uncorrelated features and features with large variance.

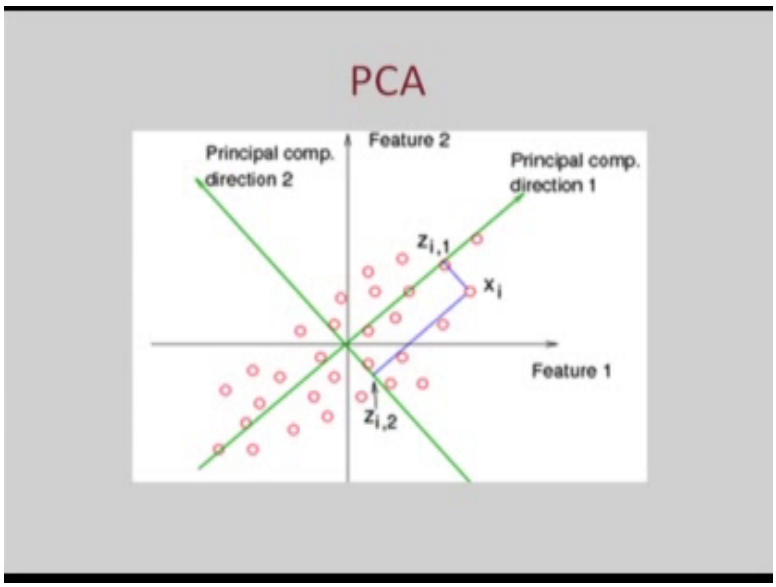
So, first let us see that if we want to give our instances in n dimensional space. If we want to select one new feature, the first feature how we should select the first feature. So, we are given a sample of p observations X 1, X 2, X p observations and for each observation we have n dimensional vector representing the n features. Now, what we want to do is that we want to find this mapping from x to Z, where Z is m dimension and let us say initially we consider Z to be one dimension. So, we want to find out the feature, the one feature which will be best for our (Refer Time: 07:28).

Now, one criteria that we can use you know because we are choosing only one feature, the question of being correlated and uncorrelated does not arise. So, we can go by this principal we can choose that feature which has the largest variance. So, we choose the feature such that the variance of Z 1 is maximum that is, we find that value of the weights for which the projection that we get correspondence to the largest variance of Z 1. So, this is what we want to do and this is where the principal components come handy, because mathematically the principal component is that vector which exactly does this.

There are two ways of considering the properties of principal components, one is it is that projection for which the variance is maximum. Secondly, principal component can be looked upon as if you map the original vectors to this new low dimensional vector space given a fixed size of that dimension the principal components is such that, if you want to recover the original instances from this reduced representation the principal components are such that the construction error is minimum. So, for more details you should refer to an algebra book.

Let us look at this picture in the slide. So, if you look at the slide we see that for these points in red, these are our instances.

(Refer Slide Time: 09:21)



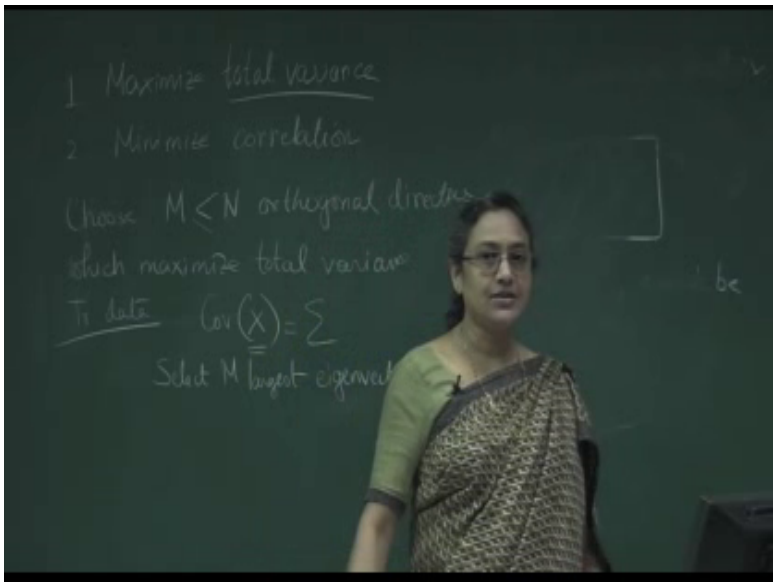
The first principal component is given by this green line. This green line gives the principal component to direction one whereas, this is the second principal component. So, the first principal component is chosen based on largest variance. How do we choose the second principal component? First of all when we choose the second feature we want to make sure that this feature is uncorrelated with the first feature. So, we want to select the feature which is orthogonal to the first feature that is, they do not share any information they are uncorrelated for that we choose an orthogonal direction.

Among the possible vectors in the orthogonal direction this is the direction for which the variance is Maximum. So, the first principal component is the one which maximizes the variance. The second is orthogonal to the first and with that constraint, the variance is Maximum. The third will be orthogonal to both the first and second, and the variance will be Maximum of course, if the original feature space is two dimensional the new feature space cannot be more than two dimensional, we are constraint by that. But you can take even the two dimensional space to a two dimensional space by making the additional property, having the additional property that then the two new feature dimensions are orthogonal.

So, there are two main principals for choosing the feature directions. So, we want to

choose the directions such that the total variation of the data is maximum that is we maximize total variance.

(Refer Slide Time: 11:29)

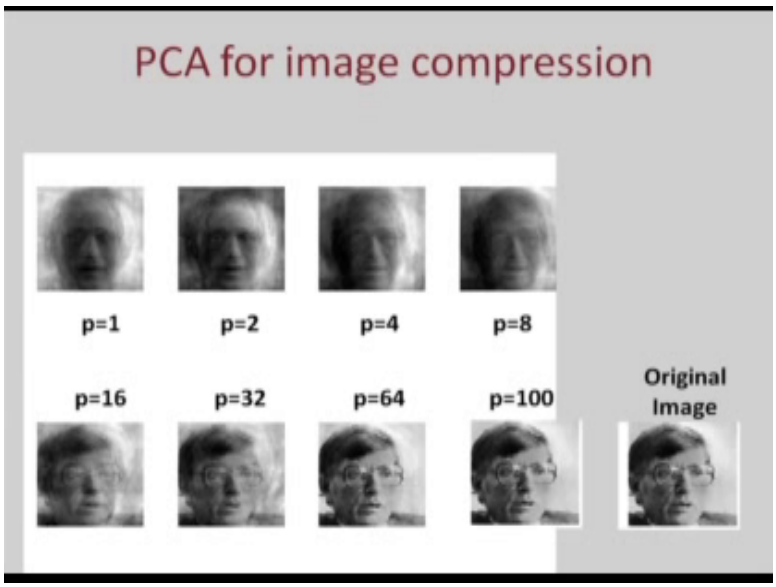


Secondly, we want to choose directions that are orthogonal, so that we minimize correlation. So, for this we choose m orthogonal direction, actually m can be less than equal to n if you wish, for feature reduction m is strictly less than n. So, we choose m orthogonal directions which maximize total variance. So, how do we do this we have an n dimensional feature space and we have n by n symmetric covariance matrix.

So, we have this training data in n dimensional feature space. So, we can have this covariance matrix of the; these are ever changing data points, this is the covariance matrix. From this covariance matrix we select the m top principal component, which are also called Eigen vectors. So, we select m largest Eigen vector of this covariance matrix. These Eigen vectors, the Eigen vectors are the one for which the Eigen values are maximum. The first Eigen vector will be the direction to the largest variance, second with the second largest variance and so on.

Now, let us look at the slides.

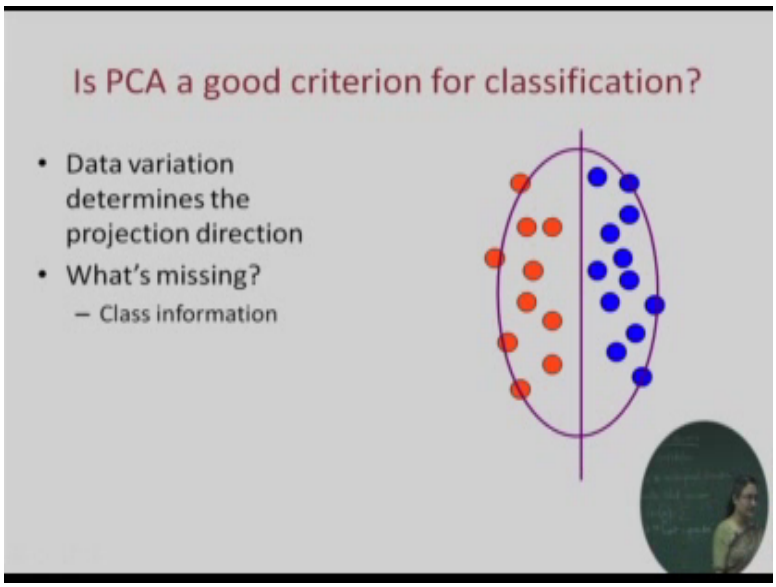
(Refer Slide Time: 13:55)



This is an application of principal component analysis for image compression. Initially we have a large number of features and what we have done is that, if we represent these pictures choosing only one principal component, the first principal component after reconstruction this is the picture that we get. If we use 2-dimensions, 4-dimensions, 8-dimensions, 16-dimension, 32, 64 and 100, this shows the reconstruction that we get, whereas this is the original image.

So, as I said that a two ways of interpreting principal components. If you take the top principal components they are selected such that they are orthogonal, and their variance is maximum, and also they can be interpreted as that selection of vectors which are orthogonal for which the reconstruction error is minimum and this can be seen from the image compression how well it works.

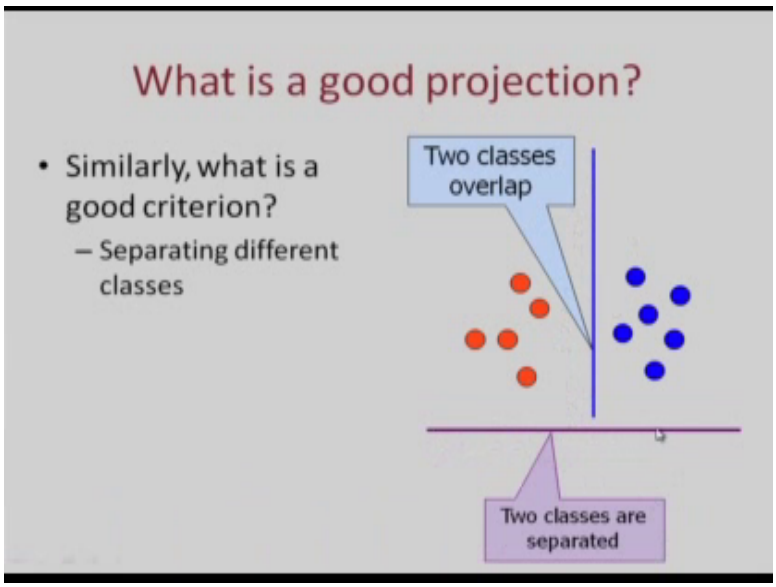
(Refer Slide Time: 15:06)



Now, let us try to see or think about is principal component analysis are good criteria for classification. So, principal component analysis if we just have the data points not considering the label, the principal component gives high variation. But, if we now look at a classification problem where we do not have just the input values of the instances, we also have label is principal component still a good way of doing that.

To illustrate this let us look at this slide, we have these points and they belong to two classes orange class and blue class. Now the principal component direction is given by this pink line, and this pink line is the one which achieves largest variance among the instances. But you see, this pink line is not very good in separating the orange points and blue points because if you project the orange points and the blue points on this axis you will see they are coming together. So, even through the total spread is high this particular axis is not able to discriminate between the orange and blue class this means this is not a very good idea when you look at a classification problem, because when we do principal components we are not taking care of class information which is very important for supervise learning.

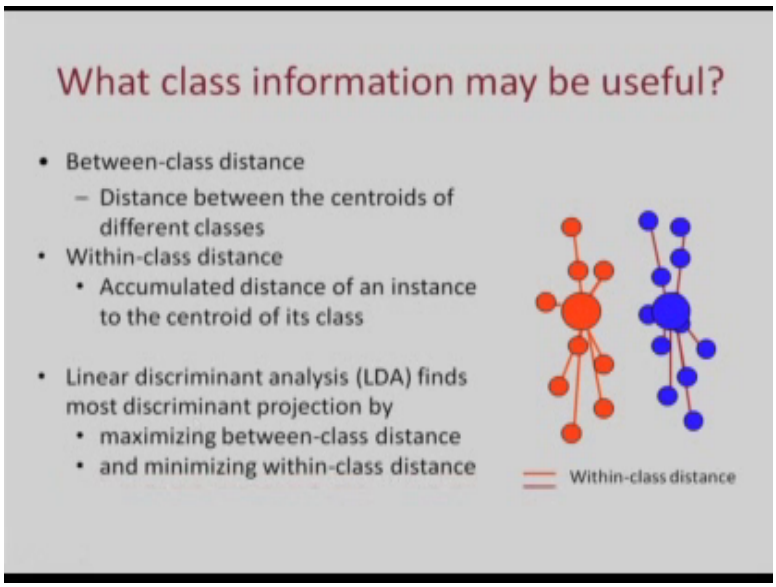
(Refer Slide Time: 15:36)



So, if you have these points what would be good criteria? So, what we really want is a feature that separates the classes, rather than the feature which simply has high variance. So, we want to make sure that the particular feature that we choose thus separates the classes. If we choose this axis, it does not separate the classes because when you project orange points and blue points on this axis they will overlap. So, this is not very good.

If we choose the pink axis, now if you project the points on these axis what you notice is that the orange and blue points are separated. So, this pink axis acts as a separator for the different classes rather than the blue line here. So, if we use this pink line the two classes are separated. So, that is to be preferred.

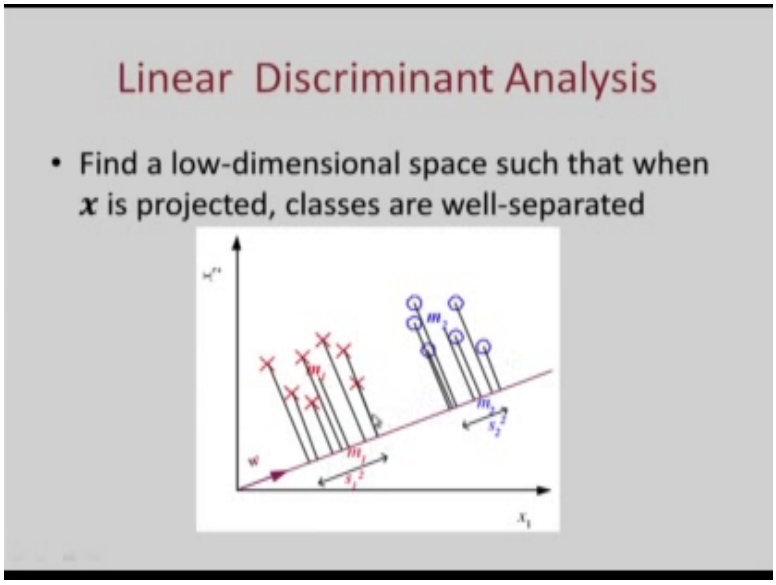
(Refer Slide Time: 17:38)



Now, in order to achieve this we have to see what is the information that we will use. First of all, we want to separate the classes, so we want to maximize between class distances. So, for this we can think of that given the points we take the centroid of the orange points and the centroid of the blue points and these two should be far away. So, we wants the between class distance to be high.

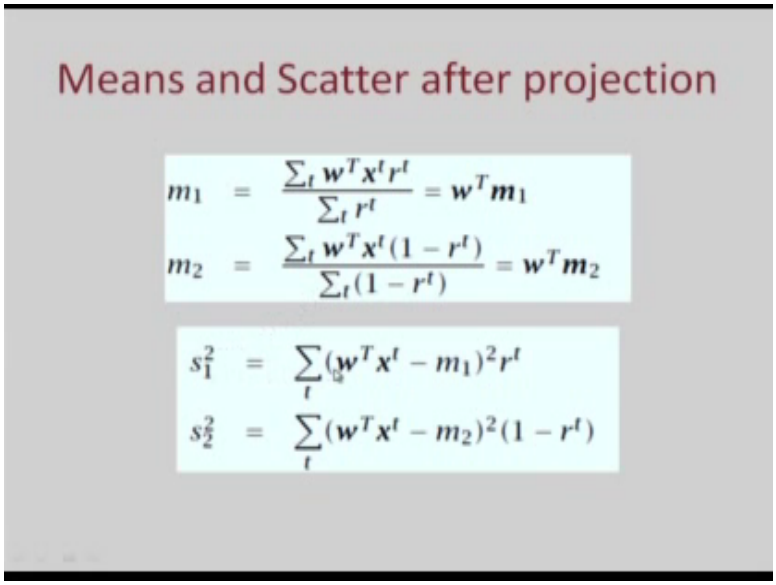
Secondly, so this is the centroid of these points, this is the centroid of these points; we want the distance between them to be high. Secondly, we want to, if we look at within class distance we want within class distance to be small. So, within class distance is the accumulated distance of an instance to the centroid of its class. So, with in the class we want to select a feature. So, that upon projecting on that feature though within class distance is small, the between class distance is high and linear discriminant analysis can be used to find the most discriminant projection which maximizes between class distance and minimizes within class distance. So, this is the representation here.

(Refer Slide Time: 19:15)



In Linear Discriminant Analysis we define the problem as this. Given the points belonging to, let us say two classes we want to find a low dimensional space such that when we project the instances on this axis the classes are well separated, they are well separated and we can use this following formula.

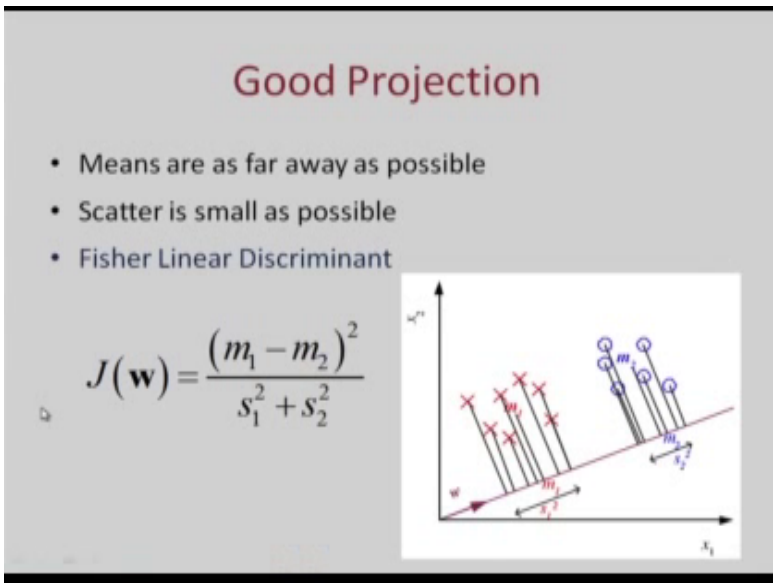
(Refer Slide Time: 19:39)



So, m 1 is the mean of the first class, so the centroid of class 1; m 2 is the centroid of the instances of class 2. S 1 square is the standard deviation of class 1 and S 2 square is the

standard deviation of class 2.

(Refer Slide Time: 20:02)



And we want the means to be as far away as possible. So, the objective function that we use looks at the difference between the means and we want the scatter between the points to be as small as possible. So, we put scatter in the denominator.

One particular method for doing this is Fisher Linear Discriminant. Fisher Linear Discriminant uses as objective function, when you have two classes its objective function is m 1 minus m 2 whole square divided by s 1 square plus s 2 square and it tries to maximize this objective function. It tries to find W, that is the projection of the feature space to the new feature space using the parameters W such that, this criteria is maximized.

This is for a two class problem. These criteria can be suitably modified when you want to work with 3 class, 4 class, or in general multi class problems, but we will not talk about it today. So, that brings us to the end of feature reduction.

Thank you very much.