Machine Learning For Soil and Crop Management
Professor Somubhra Chakraborty
Department of Agricultural and Food Engineering
Indian Institute of Technology, Kharagpur
Lecture – 16
Application of Classification and Clustering Methods in Agriculture

Welcome friends to this NPTEL online certification course of Machine Learning for Soil and Crop Management. And today, we are going to start lecture 16. And we are in the week 4 and this will be the first lecture of week 4, and in this lecture we are going to focus mainly on, in this week, we are going to focus mainly on Application of Classification and Different Clustering Methods for Agriculture.

Now, in the previous three weeks, we have discussed the basics of machine learning. Also, we have discussed the basics of multivariate data analysis and also we have in the previous week we have discussed the different multivariate regression problems like partial least squares regression, also we have discussed principal component analysis, then principal component regression partial least squares regression, random forest regression, support vector regression.

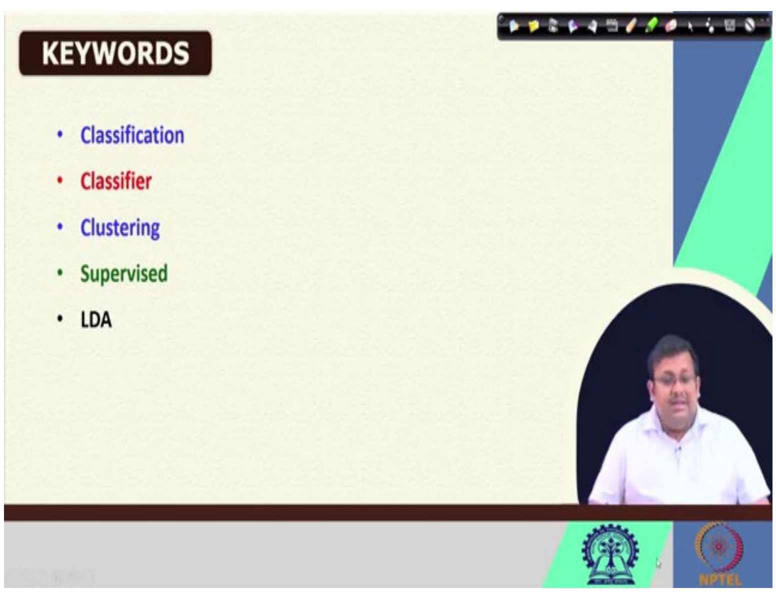
So, we have discussed some of the most widely used regression methods. Today, we are going to start the discussion on different classification and clustering methods. You know the basic difference between the regression and classification when our target or the dependent variable is a continuous numeric variable, then we call it a then it is a regression problem and when our target is categorical variable, then we call it a classification problem.

Classification problem, we are going to start today basically, with the basic classification algorithms, and then, we are going to discuss some of the most widely used classification methods.

(Refer Slide Time: 02:37)

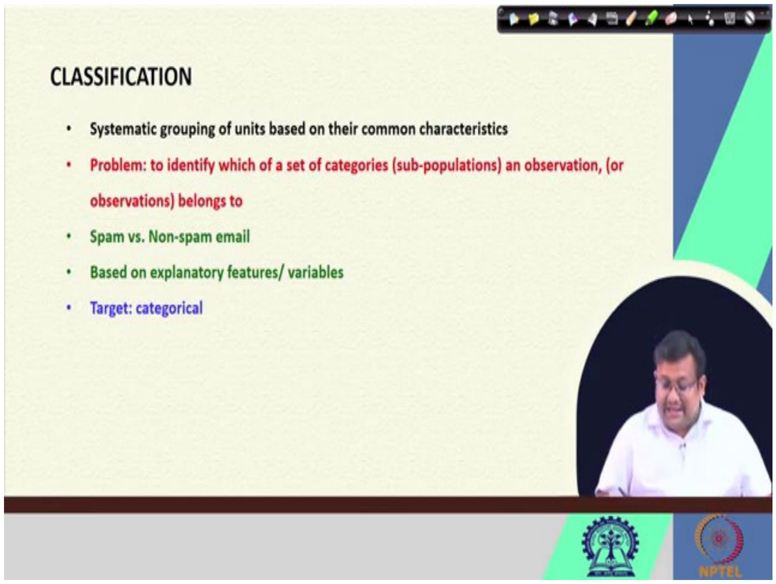
So, let us start in this lecture, we are going to mainly discuss this following concept, we are going to first discuss what is classification and then we are also going to discuss one of the major linear classification algorithm that is linear discriminant analysis or LDA. And then we are going to discuss how to calculate the LDA and what are the advantages and disadvantages of LDA we are also going to discuss.

(Refer Slide Time: 03:10)



So, these are the keywords which we are going to discuss classification also classifier, then clustering then supervised and LDA these are some of the keywords for this lecture number 16.

(Refer Slide Time: 03:27)



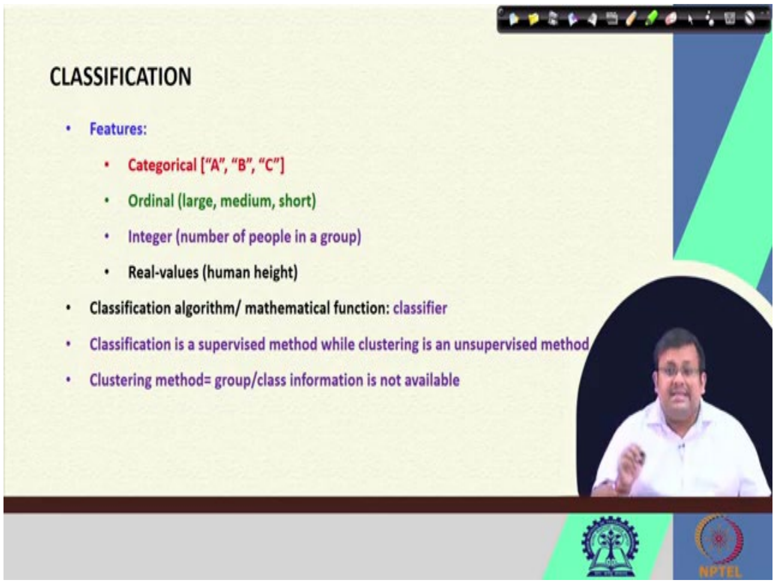
So, what is classification you already know the base that when our target variable is a categorical variable, then we call it is a classification problem. So, in other words, classification denotes the systematic grouping of units based on their common characteristics or features. So, if there are multiple samples and we want to group the samples in a different classes based on some common characteristics or features, then we call this problem as a classification problem.

Now, the major goal or problem in a classification, Statistical Classification is to identify which of the set of the categories or sub populations and observation belongs to. So, basically we have to assign any sample or observation into one of the sub population in the feature space. So, for example, a perfect example of a classification problem we generally encounter each and every day is separating the mails automatic separation of the mails in spam or non-spam folder.

So generally those emails which are generally spam, our mailbox automatically identify them based on certain features or characteristics, maybe the presence of certain words and then they are grouped into a specific folder that is a spam folder. So, they are automatically filtered out. So, this is a classification problem here our target or the dependent variable is of, are generally binary because we are getting either spam or non-spam email.

So, using the algorithm, the mailbox is discriminating all the emails into two sub population that is either spam or no spam. So, these grouping into spam or non-spam email is based on certain explanatory features or variable. In generally classification problem we call them features, they are similar to variables. And remember, here our target variable is always categorical variable.

(Refer Slide Time: 06:22)



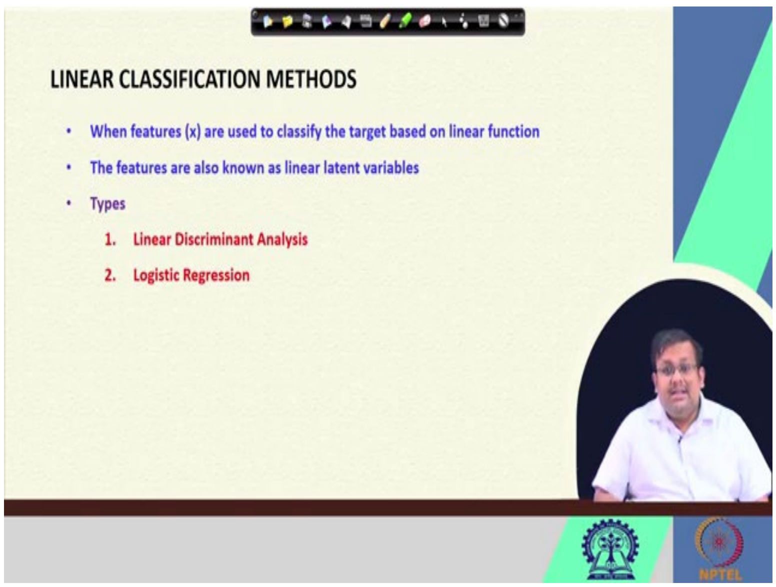
Now, what type of features can be there? There are different types of features either these independent or independent variables or features in any classification problem could be categorical that means we can have multiple classes like A, B, C, D and so on or we have some ordinal values like large medium short. We can also have some integer variables, for example, number of people in a group it is an integer variable we cannot have fraction values here.

And also some real values, for example, human height is a real value because it can take any value. So, these different types of features or variables are there in the feature space and our target is categorical variable, then we go for the classification problem. Now, there are different types of classification algorithm and the mathematical function we generally use for classifying the sample into different groups or sub population is known as classifier.

So, the mathematical function is known as classifier. Now, remember that what is the distinction between classification and clustering? We frequently see these two terms use we can see frequent use of these terms in any statistical methodologies that is one is classification another is clustering, what is the difference, what is the fundamental difference in case of classification, it is a kind of a supervised method and whereas, in case of clustering, it is generally unsupervised method. Now, in case of clustering, we do not have any group or class information.

And in case of a classification problem, we have their groups defined or there are target groups. So, this is the basic difference between classification and clustering.

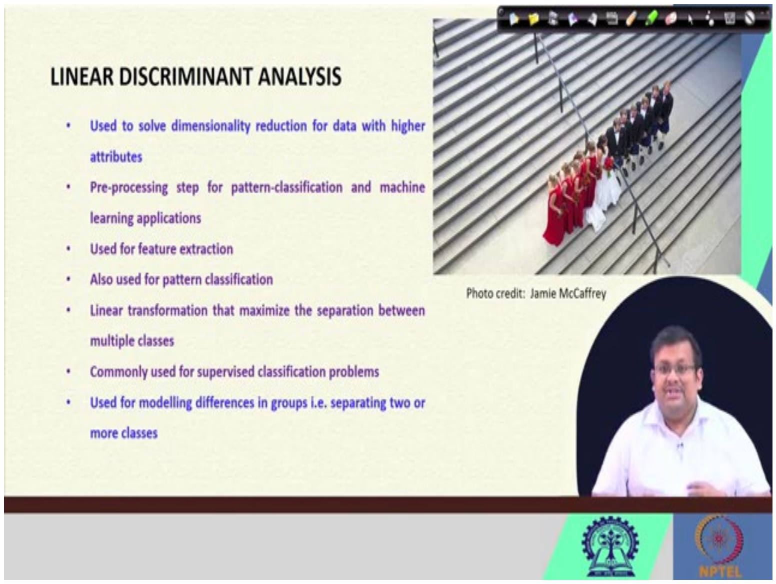
(Refer Slide Time: 08:55)



So, now, we know the difference between classification and clustering and their connection to the supervised and unsupervised method. In this week, we are going to discuss different classification methods. And the first type of classification methods we are going to discuss is the linear classification methods. Now, linear classification methods we call a problem as early as a linear classification when the features or variables are used to classify the targets based on linear function and those features are known as linear latent variables.

So, the two most important linear classification methods are linear discriminant analysis and logistic regression. So, in this lecture we are going to elaborate or we are going to explain the linear discriminant analysis and in the next lecture we are going to discuss the logistic regression.

(Refer Slide Time: 10:08)

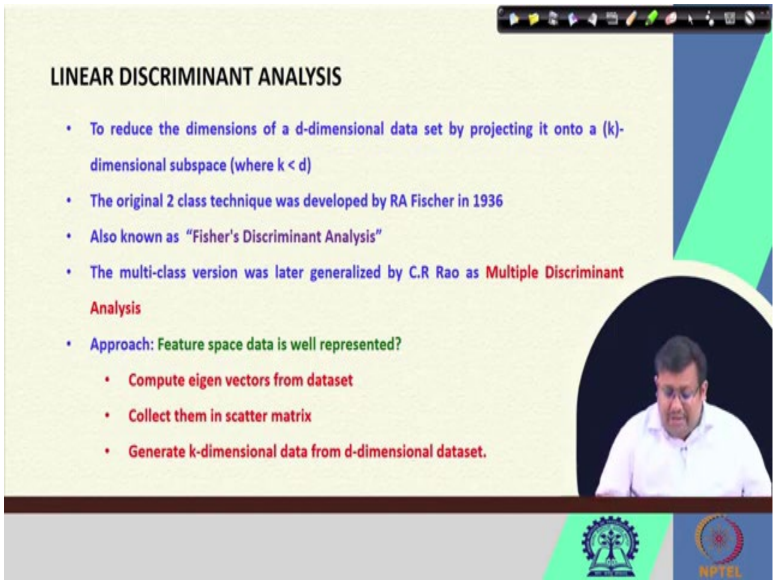


So, linear discriminant analysis is used to solve that dimension reduction again just like principal component analysis in case of LDA. So, we try to reduce the multi-dimensional data into some small number of dimensions. So, generally these LDA are linear discriminant analysis is used for pattern classification and machine learning applications. Generally they are used this is used for feature extraction also.

So, we use this linear discriminant analysis for extracting some important features by combining all the features in the feature space. So, the idea behind linear discriminant analyses is to linearly transform and linearly transform the features and maximize the separation between the multiple classes. Again please pay attention we are linearly transforming the features in the feature space to ensure there is maximum separation between the different classes or sub populations.

So, generally these LDA are linear discriminant analysis is used for supervised classification problems and generally it is also used to model different sense in the groups for example, separating two or more classes. So, here is a movie I found a very good picture photo of linear discriminant analysis you can see how a line on a new axis we can assume this the new axis or direction or axis is separating the samples into two classes based on some common features. So, this is an example of linear discriminant analysis.

(Refer Slide Time: 12:35)

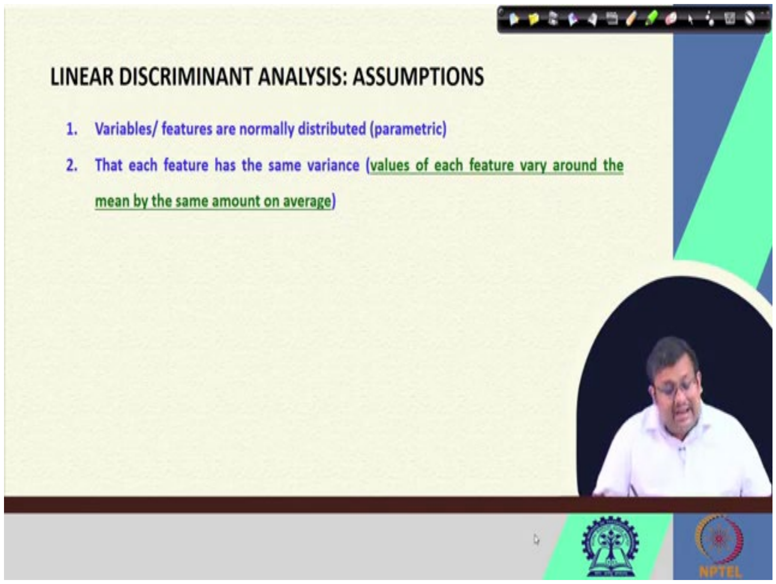


So, in case of linear discriminant analysis, our idea is to reduce the dimension or for example, let us consider there is a d dimension data space. And we want to project, reproject this d dimensional data set into k dimension where k is less than D. So, of course, in this case you can identify you can see that we are reducing the feature space or we are we are reducing the dimension when there are a huge number of variables of course, there are huge number when there are in variable there are in dimensions.

So, when we are reducing the dimension is the dimensionality reduction problem. So, similarly, in case of LDA we are doing the same kind of thing as we have seen in case of PCA, if you recall. Now, the original linear discriminant analysis was proposed by R A Fischer in 1936 and it was originally a two class technique. Later so the another name of LDA is Fischer discriminant analysis also. Now, the multi-class version was later generalized by C. R Rao and also known as the multiple discriminant analysis.

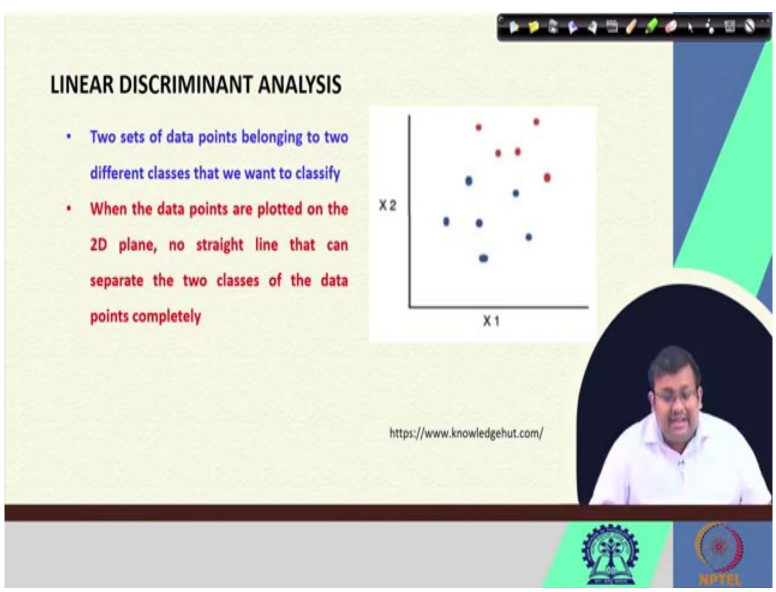
However, both of them are generally termed as discriminant analysis or linear discriminant analysis. Now, what are the general approach for the linear execution of the linear discriminant analysis first of all we calculate from the feature space we calculate the eigen vectors and we collect them in the scatter matrix and then we generate the k dimensional data from the d dimensional data space I will show you how we can do that step by step.

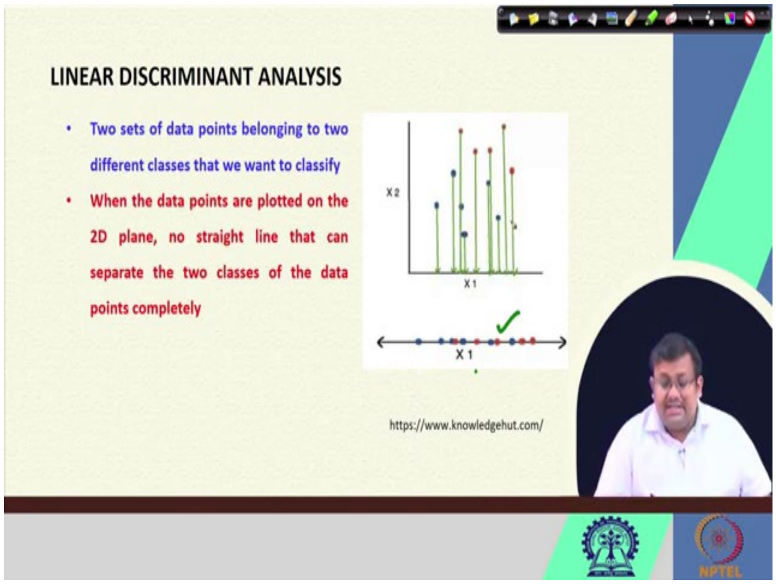
(Refer Slide Time: 14:39)



Now, what are the basic assumption in case of linear discriminant analysis there are major two assumption first of all the variable and features are normally distributed. So, it is a parametric approach. Remember, it is a parametric approach because it assumes that variables and features are normally distributed and each feature has same variance. So, here the variance feature variance is constant. So, these are the two major assumptions in case of linear discriminant analysis.

(Refer Slide Time: 15:16)



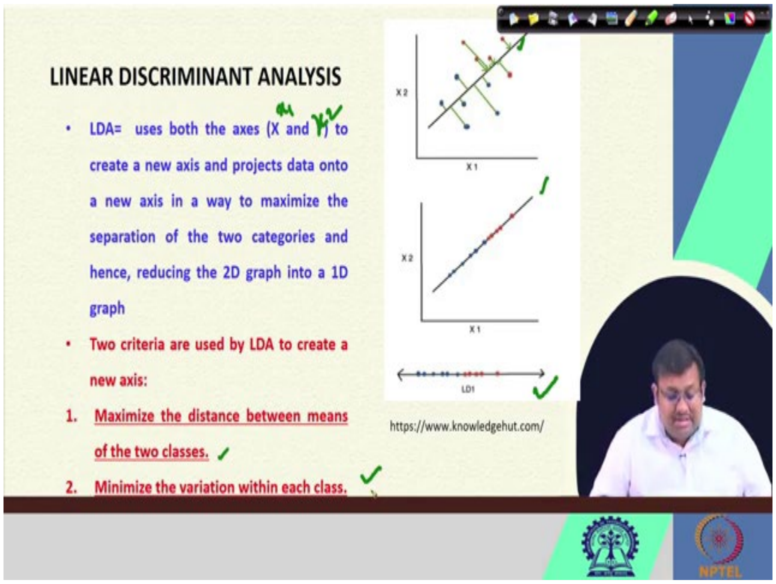


So, let us see one good example. So, it is assumed that there are two feature in a feature space x1 and x2. So, we can see two sets of data points belong to two different classes that we want to classify and here we can see that the samples the different classes are designated by different colors. Here there are two colors, one is blue and other is red. So, essentially we want to discriminate the samples into two sub population or class.

So, when the data points are plotted in the in this a 2D plane, so, we cannot see a two in a straight line that can separate the two classes of the data points completely. Why I am saying so? Because, if we want to reduce this two dimensional data into one dimensional data suppose so, what we will do, we will draw the lines from these individual samples back into the x1 axis, x1 axis.

And so, here you can see the points are projected or reprojected over this x1 axis, and however, we can see that there are something overlapping because here you can see the rate sample comes in between these blue dots. So, there are some overlapping, if we go for the simple reproduction of the samples into one of this dimension. Now, what is the solution?

(Refer Slide Time: 17:08)

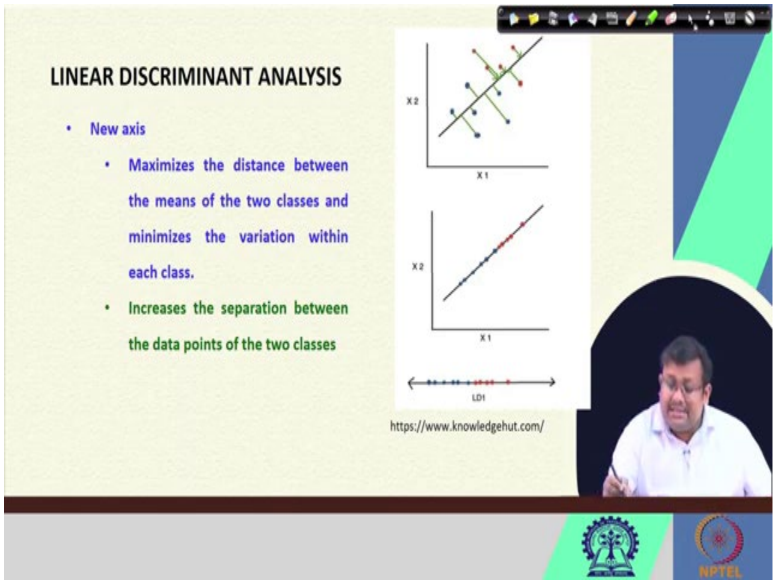


So, the solution is we calculate a new axis and which can maintain the maximum distance between these subpopulations. So, here in this linear discriminant analysis, it uses both axes. So, here the axes are x1 and x2. So, here we are using two axes x1 and x2 to create a new axes. So, this is the new axis, which we are seeing here and projects the data onto these new axes in a way to maximize the separation of the two categories and hence reducing the 2D graph into 1D graph.

So, here we can see it is 1D graph, and you can see here, but, so, this is a 1D graph. So, we are re projecting the samples into these new axes. So, here we can see clear separation between the blue class and the red class. So, this is an example of linear discriminant this as you can see, will this is how we calculate this is the basically the principle of linear discriminant analysis, we are projecting the data into a new dimension, so that we can maintain the maximum separation between the two classes.

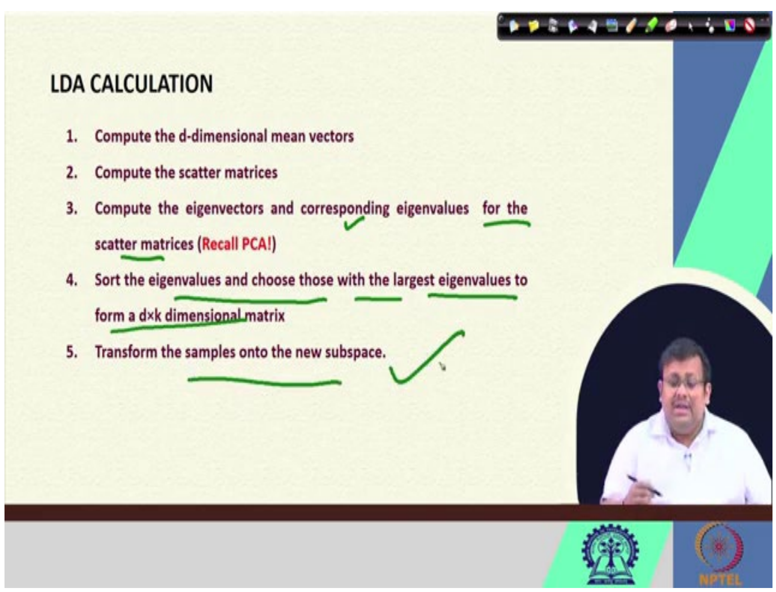
So, the two criteria are used for LDA to create these new axes first of all, maximize the distance between the means of the two classes and minimize the variation within each class. So, these two are the major criteria we need to fulfill while we compute this, the single dimension to reprojects, the simple one dimension to reproject the samples. So, this is a linear discriminant analysis.

(Refer Slide Time: 19:13)



So, this new axes remember, based on these two criteria maximizes the distance between the two class between the means of these two class and minimizes the variation within each class. Again, here, when you draw these new axes, it minimizes the distance between the means of the two classes and maximize the distance between the means of the two classes and minimizes the variation within each class. And in other words, it increases the separation between the data points of the two classes. So, this is how this LDA basically works.

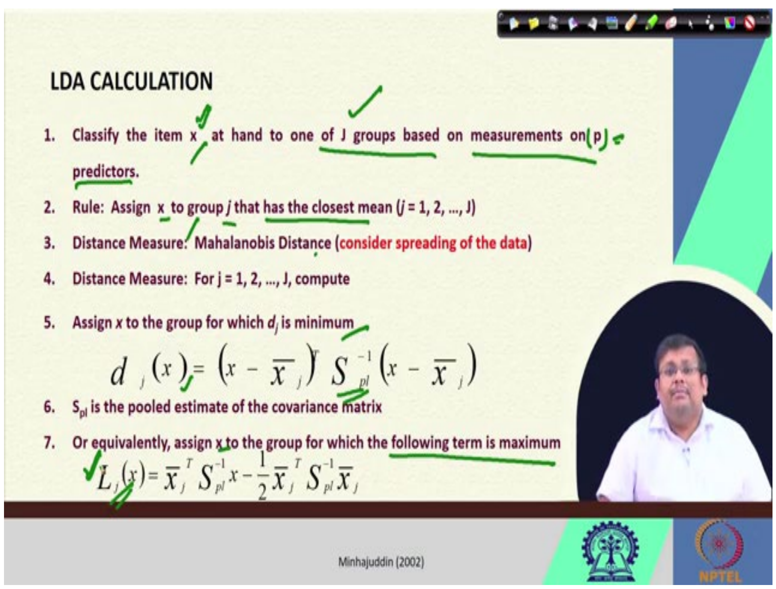
(Refer Slide Time: 20:00)



So, how to calculate the LDA? So, there are some general five steps and in the next slide we are going to see details. So, generally first we compute the D dimensional mean vectors then we calculate the scatter matrix. In the third step we compute the Eigen vectors and corresponding eigen values of the scatter matrix recall PCA we already know what is Eigen vectors and Eigen values. Now, we have to solve the Eigen values and choose the choose those with the largest Eigen values to form a d into k dimensional matrix.

So, basically from the two dimensional data set we have to calculate these vectors and then we have to calculate the scatter matrix once we calculate the scatter matrix we have to calculate the eigen values and eigen vectors then we have to solve the eigen values from large to small and then we want to transform the samples on the new subs-pace. So, this is how we do the LDA calculation.

(Refer Slide Time: 21:13)



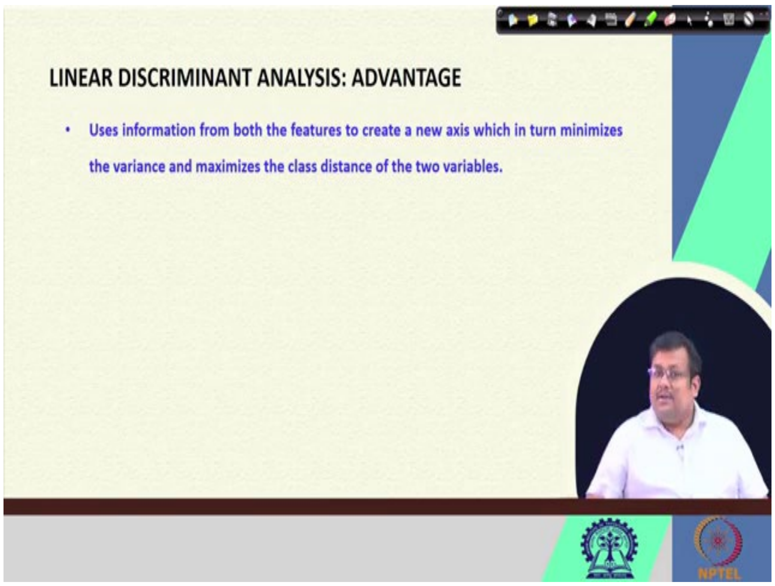
So, in more details, we have to classify if we have to classify the x number of samples x items at hand to one of the J groups let us consider there are J classes or groups based on the measurements on p predictors. So, suppose, we want to classify the item suppose, this is an unknown item x, and there are p predictors or features in this feature space and we want to classify these x into one of these J groups, where this j varies from 1 to J capital J.

So, our rule for LDA is to assign these x to a group j that has the closest mean. So, what distance measure we generally use for this? Generally use this Mahalanobis distance measured which considered the spreading of the data. So, the distance measure from 1 to 2 J groups we need to compute these distances distance measure and then we assign these new variable or new sample x to the group for which this distance d subscript j is minimum.

And how to calculate these d subscript j for a new value new sample then we have to calculate by this formula x minus x j bar transpose into S inverse Spl inverse into X minus X bar j where these Spl stands for the covariance matrix what is covariance we have already discussed or equivalently we can assign x to this group for which the following term is maximum. So, the following term is maximum when we calculate these.

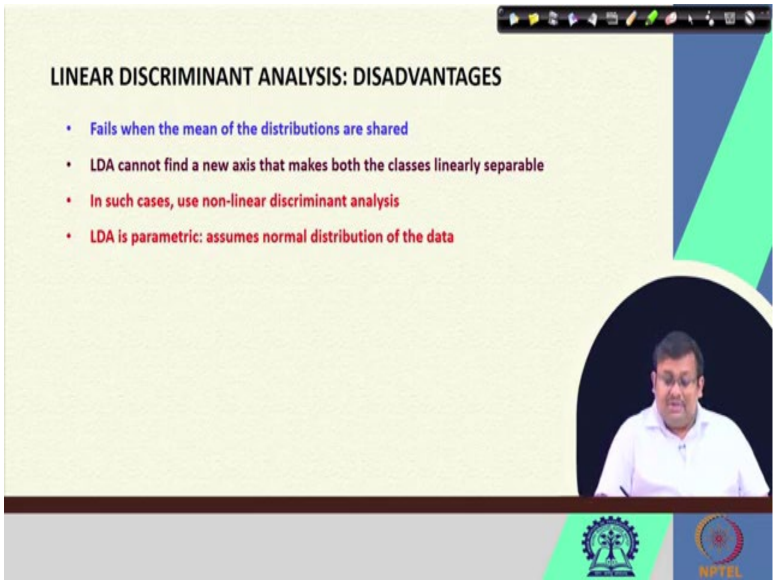
So, we calculate this value by using this formula and we assign the new sample into that group for which this value is maximum. So, this is how we calculate the linear discriminant analysis.

(Refer Slide Time: 23:51)



So, what are the advantages of linear discriminant analysis? Well, the major advantage of using the linear discriminant analysis is it uses information from both the features to create a new axes which in turn minimizes the variance and maximizes the class distance between the two variables. So, this is the major advantage.

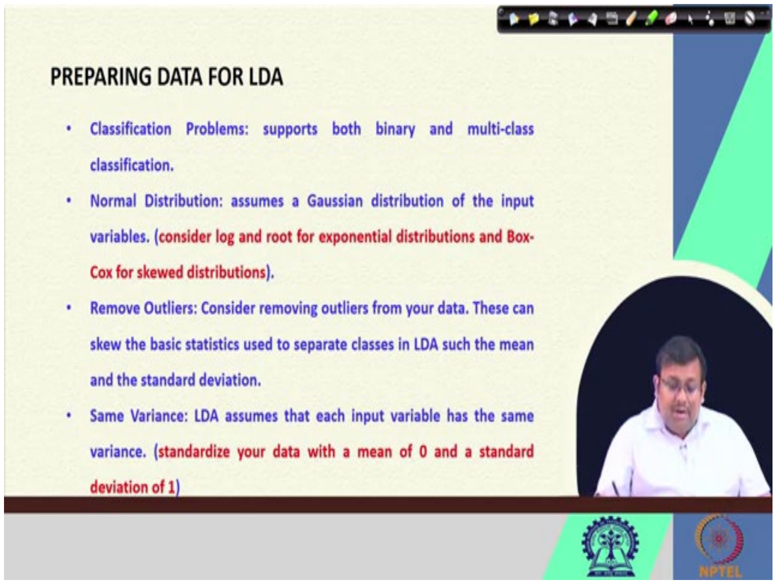
(Refer Slide Time: 24:21)



However, there are certain disadvantages also. First of all the LDA fails when the mean of the distributions are shared. So, in that condition when the mean of distributions is shared LDA cannot find a new axis that can make both the classes separable and in that case, we cannot rely on the linear discriminant analysis then we have to rely on the nonlinear discriminant analysis. And one of the major problem for linear discriminant analysis is, it is a parametric so here the assumption of normal distribution of the features are followed.

So, we have to verify whether our features or variables are normally distributed or not. And we have to do the if they are not normally distributed, we have to do the required transformations Box and Cox transformations we have seen, we have discussed the what Box and Cox transformation previously. So, that type of transformation we have to do.

(Refer Slide Time: 25:33)

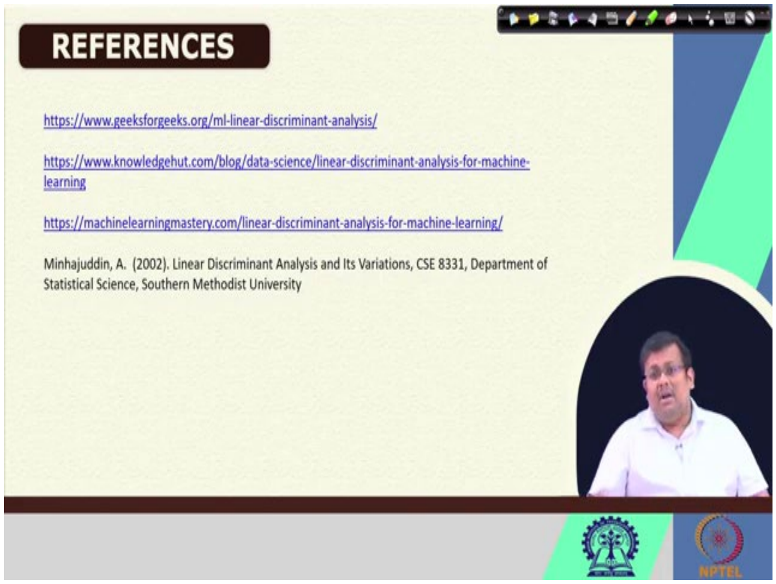


So, how to prepare your data for linear discriminant analysis in case of classification problem, these LDA supports both binary and multi-class classification when we are assuming that the input variables are features are following the normal distribution, if the data or features are not normally distributed then as I have mentioned we need to consider log and root transformation for and also Box and Cox transformations.

We also need to remove the out layers from the data if we want to go for the linear discriminant analysis because this can skew the basic statistics used to separate the classes in LDA such that mean and standard deviation. And when there is a same variance, same variance is important assumption. So, LDA assume that each input variable has the same variance. So, you need to standardize your data with a mean of 0 and standard deviation of 1.

So, that is why we have learned how to standardize scaling and centering, centering scaling of the data we have already discussed. So, we have to do the normality check, we have to do the standardization of the data, we have to remove the out layer. And so, these are the some of the important data I would say manipulation you need to do if we decide to go for the linear discriminant analysis.

(Refer Slide Time: 27:31)



So, guys, these are the references for this lecture. And I hope that now, the linear discriminant analysis is clear to all of you. We have discussed the basics, of course, we can dig into the more detailed mathematical expressions. But we do not have the scope for discussing all those in details, but I hope that I am able to convey the major message from or the major message as far as the principle of LDA is concerned. I hope this is useful for you.

And we will see later some application of LDA in our upcoming lectures. So, thank you guys. Let us meet in our next lecture.