Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-7

Lecture-28

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class I discussed the concept of KL transformation and the PCA the principal component analysis. In the KL transformation from the input vector $X\text{-}$ I can determine the mean vector and the covariance matrix. From the covariance matrix I can determine eigenvalues and the corresponding eigenvectors and with the help of this eigenvectors I can determine the transformation matrix A of the KL transformation. So, what is KL transformation $Y\text{-} = A(X\text{-} - \mu_{X\text{-}})$ that is the KL transformation.

My original data they are highly correlated, but after the transformation the transform data will be uncorrelated. So,that is the objective of the KL transformation. So, I am projecting data along the direction of the eigenvectors and because of this projection the transform data will be uncorrelated.

And after this I discussed the concept of the truncated transformation matrix.

We are not considering all the eigenvectors in case of the truncated transformation matrix and with the help of these eigenvectors K number of eigenvectors I can determine the truncated transformation matrix A_K. A_K is the truncated transformation matrix. And after this I can determine the transformation the transformation is $Y\text{-} = A_K(X\text{-} - \mu_{X\text{-}})$.

So, that means I am considering the largest eigenvalues and the corresponding eigenvectors and these are the principal components and that is the concept of the principal component analysis. So, one problem of the KL transformation is that in KL transformation the transformation kernel is not fixed.

It depends on the statistics of the input data that is the main problem of the KL transformation. Unlike other transformation like DFT, DCT the transformation kernel is fixed, but in case of the KL transformation the transformation kernel that is the transformation matrix depends on the statistics of the input data that is the problem of the KL transformation. And also in case of the PCA you have seen that I am projecting data along the direction of the eigenvectors and there is no class information. So, if I consider suppose classes different types of classes. So, in case of the PCA class information we are not considering only we are determining the best projection direction and that means we are reducing the dimension of the input vector.

So, the class information is not available and that is why the PCA is an unsupervised technique. So, the problem of the PCA is that class information we are not considering we are only projecting data along the direction of the eigenvectors. To consider that issue we are considering another technique that is called the LDA linear discriminant analysis. In this case we are determining the best projection direction considering the class information. So, the objective of the LDA is to find a set of vectors which maximize between class scatter and minimizes within class scatter that is the goal of the LDA the linear discriminate analysis.

So, let us discuss about LDA and already I told you what is the difference between the PCA and the LDA in case of the PCA class information is not available, but in case of the LDA we have class information that is why LDA is a supervised technique. So, let us now discuss about the concept of the LDA the linear discriminate analysis. So, in the LDA already I told you the goal is to find a set of vectors that maximizes the between class scatter while minimizing the within class scatter. So, now let us discuss about the LDA the linear discriminate analysis LDA. So, what is the goal?

So, LDA uses class information in case of the PCA we are not considering the class information and the goal is to find a set of vectors that maximizes that maximize the between class scatter while minimizing the within class scatter.

So, that is the goal of the LDA and we are considering class information in LDA. So, in this case LDA considers two criteria's. So, what are the criteria's one is maximize the distance between means of the classes. So, that means one objective is we have to maximize the between class scatter. So, that is one objective the number one maximize the distance between means of the classes and another one is minimize the variations within each class.

That means I have to minimize the within class scatter. So, that is why we are considering one quantity the quantity is this the difference between these two means. So, $\mu\text{-}_1^2$ - $\mu\text{-}_2^2$. So, for two classes we are considering this and $S_1^2 + S_2^2$ that concept I will be explaining later on but this is the quantity I have to maximize. So, this quantity I have to maximize.

So, I will explain you later on how to get this quantity but to fulfill these two conditions I have to maximize this quantity and that is the objective of the LDA. So, let us explain the concept of the LDA and already I told you the difference between the PCA and the LDA. So, in this figure you can see in the first figure we are projecting data along the direction of eigenvectors and that is nothing but the PCA the principal component analysis. So, this is the PCA that means we are projecting data along the direction of the eigenvectors. So, what is missing in this case the missing is the class information is missing.

So, the class information is missing here. So, let us see what is the good projection what is the good projection. So, in this figure you can see I am considering two classes suppose this is ω_1 and another one is ω_2. So, you can see the samples belonging to the class ω_1 and the samples belonging to the class ω_2. So, if I consider the projection direction suppose 1, 1 is the projection direction another 1 is 2.

So, if I consider the projection direction 1 then you can see the two classes overlap that means if I consider the projection direction 1 you can see the two classes overlap and if I consider projection direction 2 the second direction the two classes are well separated. So, we can consider the projection direction direction 2 but corresponding to the projection direction 1 you can see the two classes overlap here the overlapping take place here these overlapping but corresponding to the projection direction 2 the two classes are separated well separated. So, that means the projection direction 2 is better as compared to the projection direction 1. So, we have to find the best projection direction we have to find so move to the next slide. So, what information we have to consider? So, one is the between class scatter or the between class distance we have to consider.

So, what is the between class distance? The between class distance is nothing but distance between the centroids of different classes. So, in this figure you can see we are considering two classes. So, these are the samples corresponding to the class ω_1 and these are the samples corresponding to the class ω_2 and you can see the centroid of the class ω_1 and centroid of the class ω_2 and you can see that this is the distance between these two centroids and actually this is the measure of between class distance. So, that is nothing but the between class distance and if you move to the second figure in the second figure we have considered within class distance. So, that means what is the within class distance accumulated distance accumulated distance of an instance to the centroid of its class.

So, here you can see what we are considering corresponding to this class ω_1 I have the samples and corresponding to the second class I have the samples. So, I am finding the distance if you see this is the centroid. So, I have two centroids centroid 1 corresponding to the first class and centroid 2 is the centroid of the second class the samples of the second class. So, what we are finding we are finding the distance between the samples and the centroid that is nothing but the accumulated distance of an instance to the centroid of its class that is nothing but the accumulated distance of an instance to the centroid of a particular class and that is the meaning of the within class distance.

And what is the objective of the LDA the objective of the LDA is to maximize between class distance and minimize within class distance.

So, this is the concept of the between class distance and within class distance. So, that linear discriminate analysis finds most discriminate projection by maximizing between class distance and minimizing within class distance. So, if I consider in the figure I am showing the samples belonging to two classes. So, I am showing the samples belonging to two classes ω_1 and ω_2 and you can see the centroid 1 corresponding to the samples of the class 1 ω_1 and centroid 2 that is the centroid of the samples of the class ω_2. So, what is the objective of the LDA we have to find the most discriminate projection by maximizing between class distance and minimizing within class distance.

So, in this the second figure you can see I am showing two projection directions. So, already I have explained that one. So, if I consider the projection direction 1 and another one is the projection direction 2. So, you can see in case of the projection direction 1 you can see the samples are overlapping the samples belonging to two classes they are overlapping. But in case of the projection number 2 they are well separated.

So, that means I have to consider the projection direction 2 we have to consider the projection direction 2 projection direction 2 we have to consider. So, one is not good because in case of the one you can see the overlapping of the samples belonging to two different classes. So, in case of the one you can see the overlapping take place between the samples of two classes. So, now let us discuss the mathematical concept behind LDA. So, what is the mathematics? So, let us consider what is LDA the linear discriminnant analysis.

So, suppose we have C number of classes C classes we are considering and suppose each class and each class has N_i number of samples N_i samples and this is m dimensional samples m dimensional samples. So, where $i = 1,2,...,C$. So, we have C number of classes and you can see each class has N_i number of samples and this samples are m dimensional. So, how can I write the m dimensional samples m dimensional samples samples I can write suppose $\{X\text{-}^1, X\text{-}^2, ..., X\text{-}^{N_i}\}$. So, these are the samples the m dimensional samples and I have all together N_i number of samples.

So, if I stack these samples from different classes into one big fat matrix X- such that each column represents one sample. So, I am repeating this. So, stacking these samples from different classes into one big fat matrix X- such that each column represents one sample. So, I will be getting one matrix the matrix is X-. So, what is the objective of the this LDA.

So, we want to obtain a transformation that means to obtain a transformation of X- we are doing the transformation of X- to Y- through projecting the samples in X-, X- is the matrix because how to get the matrix already I have explained. That means, I have to stack the samples from different classes into one matrix the matrix is X- such that each column represents one sample. So, like this I am getting this matrix X-. So, to obtain a transformation of X- to Y- through projecting the samples in X- onto onto a hyper plane with dimension C-1 C is the number of classes. So, let us see what does this mean.

So, move to the next slide. So, suppose we assume m dimensional samples. So, m dimensional samples the samples are $X\text{-}^1$ $X\text{-}^2$ suppose I have N number of samples. So, out of this N_1 number of samples that belongs to N_1 number of samples that belongs to the class ω_1 and N_2 number of samples that belongs to the class ω_2. So, m dimensional samples we are considering $\{X\text{-}^1, X\text{-}^2, \ldots, X\text{-}^N\}$ and out of this suppose N_1 number of samples belong to $\omega_1$and N_2 number of samples belong to ω_2. So, we want to obtain a scalar Y by projecting the samples X- onto a line.

So, that means what we want to obtain to obtain to obtain a scalar scalar is Y- by projecting the samples X- onto a line. So, that means I am doing the projections and suppose if I consider $C = 2$ that means $C - 1$ space corresponding to $C = 2$. So, that means if I consider 2 number of classes. So, the space is $C - 1$. So, that means because of this projection dimension is also reduced.

So, what is this projection this is nothing, but $Y = W\text{-}^T X\text{-}$. So, where X- is the input vector and you can see these are the components of the vector x_1 x_2 up to x_m. So, this is my input vector and what is the weight vector the weight vector is W- is the weight vector and these are the coefficients w_1 w_2 up to w_m. So, W- is the weight vector. So, I have this transformation that means I am doing the projection like this, that is nothing but the dot product.

So, W- is the projection vector I can say this is the W- is the projection vector this is the projection vector or I can say the weight vector or the projection vector projection vector is W- and that is used to project X- to Y-. So, in this figure you can see I am showing two projection directions in the first figure if you see in the first figure the figure number 1 I am showing the two dimensional feature space and you can see these are the samples suppose the samples belonging to the class ω_1 and the rate samples that is the samples belonging to class ω_2. If I consider this projection direction direction is suppose 1 you can see the samples are overlapping and if I consider the second figure I am I am considering the projection direction 2 corresponding to this projection direction 2 the samples of two classes are well separated. That means which one is the best projection direction the best projection direction is then direction 2 one is not a good projection direction because overlapping take place. So, we have to consider the projection direction 2.

Now how to get the optimum projection direction which one is the best projection direction there may be many projection direction but out of which one is the best or which one is the optimum projection direction. That means we have to find the objective is to find the optimum direction given by actually the projection vector $W\text{-}$. So, that is this optimum I can write this optimum value this optimum $W\text{-}$ I can write $W\text{-}^{*}$. So, I have to find the optimum direction of $W\text{-}$ $W\text{-}^{*}$ we have to determine.

So, that is the objective of the LDA. So, we have to find the best projection vector and for this we have to see the separation between the two classes that we have to observe the separation between the classes. So, already I told you we have to maximize between class scatter and we have to minimize within class scatter. So, how to define this scatter so move to the next slide. So, to get the best projection direction how to get the best projection direction to get the optimum $W\text{-}$ that is the projection direction. So, the mean vector so what I have to do the mean vector of each class $X\text{-}$ and y feature space is we can obtain like this.

So, suppose mean is defined like this $\mu_{i\text{-}} = \frac{1}{N_i}\sum_{X\text{-}\in\omega_i} X\text{-}$. So, the mean vector of the all the input vectors belonging to the class ω_i we can determine and after the projection of this I can also determine the mean and after the projection I can also determine the mean. So, after the projection that this mean vector is nothing but $\mu_{i\text{-}}^{\tilde{}} = \frac{1}{N_i}\sum_{y\in\omega_i} y$. Which can be represented like this $\frac{1}{N_i}\sum_{X\text{-}\in\omega_i} W\text{-}^{T} X\text{-}$. So, that is equal to $W\text{-}^{T}\frac{1}{N_i}\sum_{X\text{-}\in\omega_i} X\text{-}$. So, I am getting this one. So, this $\mu\text{-}_{i}$ that is before the projection before projection and this $\mu_{i\text{-}}^{\tilde{}}$ that is the after the projection.

So, that means what is the interpretation of this projecting $X\text{-}$ to y will lead to projecting the mean of $X\text{-}$ to the mean of y. So, I am repeating this what is the interpretation of this projecting $X\text{-}$ to y will lead to projecting the mean of $X\text{-}$ to the mean of y that is the interpretation of this. So, we can determine the distance between the projected means and that I can consider as my objective function. So, what is the objective function I can consider suppose objective function function objective function is nothing but I can consider suppose $J(w\text{-})$ I can consider and that is nothing but the distance between the projected means. So, distance between the projected means $\mu\text{-}_{1}^{\tilde{}} - \mu\text{-}_{2}^{\tilde{}}$ and you remember we are considering only 2 classes in this example.

So, that is $|\vec{w}^T\vec{\mu}_1 - \vec{w}^T\vec{\mu}_2|$and that is equal to $|\vec{w}^T(\vec{\mu}_1 - \vec{\mu}_2)|$. So, this objective function $J(\vec{w})$ we can consider and you can see we are considering the measured the measure is the distance between the projected means. So, that should be maximize. So, after the projection of the samples belonging to 2 classes we are getting the means corresponding to the first class and corresponding to the second class we are getting 2 means and these 2 means should be well separated and that is why we are considering this objective function $J(\vec{w})$ that we are considering. So, now the problem is we are considering this that distance between the projected means we are considering as an objective function, but it is not a very good measure because it does not take into consideration of the standard deviation within the class.

So, that we have to consider the standard deviation within the class we have to consider I am repeating this projected means we are considering as an objective function, but that is not a good measure because we are not considering the standard deviation within the class. So, that can be illustrated in the next slide. So, here I am showing and these 2 projection directions. So, in the first projection direction if you see in this direction direction suppose 1 the 2 means are well separated you can see the separation between these 2 means are very good, but still the samples are overlapping, but if I consider the second projection direction that is number 2 here. So, in case of the projection direction 2 the distance between these 2 mean is not significant it is not big, but still the separability is very good.

So, that means, you can see I can say this axis has larger distance between means, but separability is not good because the samples will be overlapping. So, that means, the direction 1 is not good and the second direction this axis gives better class separability. So, that means, what is the interpretation of this that means, we considered the criteria function as $J(\vec{w})$ and that is the difference between the 2 projected means and that is not a very good measure because we are not considering the standard deviation within the class. So, this is the example in this case. So, to consider this issue what we have to consider the method proposed by Fisher and that we have to consider.

So, to consider this issue you can see in my next slide to consider this issue. So, what I can consider maximize a function that represents the difference between and the means normalized by a measure of the within class variability. So, that is actually it is called a scatter and that is called a scatter. So, for each class we can define the scatter and that is equivalent of the variance. So, what is the scatter? So, scatter $\tilde{S}_i^2 = \sum_{y\in\omega_i}(y - \tilde{\mu}_i)^2$.

So, for each and every class I have to consider this scatter and that is actually equivalent of the variance and that is nothing, but the sum of square difference between the projected samples and their class means. So, we can determine the scatter like this. So, what is $S_i^{\tilde{2}}$? So, that is nothing, but the variability that is the measure what is $S_i^{\tilde{2}}$ that is the variability within class ω_i after projecting it onto the y space after projecting it on the y space that is a scatter. So, we can consider $S_1^{\tilde{2}} + S_2^{\tilde{2}}$. So, that is the measure of the variability within the two classes after the projection.

So, that is I can say this is nothing, but the variability within two classes after projection and this is nothing, but within class scatter of the projected samples I can write of the projected samples. So, you can see that $S_1^{\tilde{2}} + S_2^{\tilde{2}}$ and that is nothing, but the within class scatter of the projected samples that we can determine. So, move to the next slide. So, the linear discriminant is defined as a linear function $W\text{-}^{T}X$ that maximize the criterion function. What is the criterion function? The distance between the projected means normalized by the within class scatter of the projected samples.

So, that means we are considering this the criterion function. What is the criterion function? The criterion function is $J(w\text{-})$ that is nothing, but the difference between the projected means square and $\frac{|\mu^{\tilde{}}{}_1-\mu^{\tilde{}}{}_2|^2}{S_1^{\tilde{2}}+S_2^{\tilde{2}}}$. So, that means what is the criterion function? We are considering the distance between the projected means normalized by the within class scatter of the projected samples. So, what actually we are now looking for? We are looking for a projection where samples from the same class are projected very close to each other and at the same time, the projected means are as far as possible. So, that means what we are looking for? We are looking for a projection direction where samples of the same class are projected very close to each other and at the same time, the projected means should be far away from each other.

So, that is the objective. So, in this figure what we are considering, I am showing the projection direction 1 suppose this is a projection direction1. In this projection direction what actually we are looking for that means the first condition is the projected means should be far away from each other. So, that is one condition and also the samples from the same class are projected very close to each other. So, you can see this is the projection of the samples belonging to one class and this is the projection of the samples belonging to another class and you can see the separation between these two means the projected means μ_1 and μ_2. So, I am repeating this, I am looking for a projection direction where the samples from the same class are projected very close to each other and at the same time, the projected means should be far away from each other.

So, these are the conditions and based on these conditions, we are determining this criterion function that is $J(w\text{-})$. So, we have to find the optimum projection direction. So, we have to find the optimum projection is $w\text{-}^*$ that we have to determine. So, how to find this one? So, we have to find the optimum projection direction and how to find this optimum projection direction you can see. So, we will define a measure of the scatter in multivariate feature space $X\text{-}$.

So, we have to find the optimum projection direction that is $w\text{-}^*$. So, for this what we can consider, we will define a measure of the scatter in multivariate feature space $X\text{-}$ and which are denoted as scatter matrix. So, what is the scatter matrix? The scatter matrix I can consider $S_i = \sum_{x\text{-}\in\omega_i}(X\text{-} - \mu\text{-}_i)(X\text{-} - \mu\text{-}_i)^T$. So, that means, what we are considering. So, we will define a measure of the scatter in multivariate feature space $X\text{-}$ and which can be denoted by a scatter matrix. So, what is S_w now? $S_w = S_1 + S_2$ we have to determine.

So, this S_i is nothing but is a covariance matrix of class ω_i and we have obtained S_w. So, what is S_w? S_w is nothing but $S_1 + S_2$ and that is actually S_w is called the Within class scatter matrix. Within class scatter. So, we have defined the Within class scatter matrix. So, the scatter of the projection y can be expressed as a function of the scatter matrix in feature space $X\text{-}$.

So, how to do this? Move to the next slide. So, we are representing $S_i^{\sim 2} = \sum_{y\in\omega_i}(y - \mu_i^{\sim})^2 = \sum_{X\text{-}\in\omega_i}(w\text{-}^T x\text{-} - w\text{-}\mu_i)^2$ I can write like this. So, in this expression this y is nothing but this already we have defined that is the projection of $X\text{-}$ onto y and this one is this the $\mu_i^{\sim}$ that is nothing but the projection of the mean. So, that means this $S_i^{\sim 2} = \sum_{X\text{-}\in\omega_i} w\text{-}^T(x\text{-} - \mu_i)(x\text{-} - \mu_i)^T w\text{-} = w\text{-}^T(\sum_{X\text{-}\in\omega_i}(x\text{-} - \mu_i)(x\text{-} - \mu_i)^T)w\text{-} = w\text{-}^T S_i w\text{-}$.

So, what is the this value $S_1^{\sim 2} + S_2^{\sim 2} = w\text{-}^T S_1 w\text{-} + w\text{-}^T S_2 w\text{-} = w\text{-}^T(S_1 + S_2)w\text{-} = w\text{-}^T S_w w\text{-}$. This is $w\text{-}^T S_w w\text{-}$ and this is nothing but $S_w^{\sim}$ within class scatter. So, here in this case if you see here this $S_1^{\sim 2} = w\text{-}^T S_1 w\text{-}$ and what is $S_w^{\sim 2}$ that is nothing but this. So, in this case we are getting S_w. So, what is actually this S_w? S_w tilde we are getting and that is nothing but what is $S_w^{\sim}$ that is nothing but within class scatter matrix of the projected samples y.

So, we are getting the within class scatter matrix of the projected samples y. Similarly, the difference between the projected means in y space can be expressed in terms of the means in the original feature space $X\text{-}$ space that is the $X\text{-}$ space. So, let us move to the next slide. So, we can determine $(\mu\text{-}_1^{\sim} - \mu\text{-}_2^{\sim})^2 = (w\text{-}^T\mu\text{-}_1 - w\text{-}^T\mu\text{-}_2)^2 = w\text{-}^T(\mu\text{-}_1 - \mu\text{-}_2)(\mu\text{-}_1 - \mu\text{-}_2)^T w\text{-}$. So, in this case you can see here this $\mu\text{-}_1^{\sim}$that is nothing but this one and what is $\mu\text{-}_2^{\sim}$ what is $\mu\text{-}_2^{\sim} = w\text{-}^T\mu_2$.

So, corresponding to this we will be getting $w\text{-}^T S_B w\text{-}$ and that is called $S_B^{\sim}$ and this $S_B^{\sim}$ that is called between class scatter. So, in this expression you can see we are determining $\mu\text{-}_1^{\sim}$ that is nothing but $w\text{-}^T \mu_2$. So, μ_1 is the original mean of the samples and $\mu\text{-}_1^{\sim}$ that is the projected means and similarly μ_2 is nothing but the mean of the original samples and $\mu\text{-}_2^{\sim}$ that is nothing but the mean of the projected samples. So, from this you can determine the between class scatter that is the S_B you can determine. So, in this case you can see the $S_B^{\sim}$ is the between class scatter of the projected samples samples y and what was S_B? S_B is nothing but $S_B^{\sim}$ is nothing but the between class scatter of the projected samples y and what is S_B? S_B is the between class scatter between class scatter of the original samples original samples means $X\text{-}$.

So, we can determine the within class scatter matrix and also the between class scatter matrix and based on this we can determine the criterion function that is the criteria function is $J(w\text{-})$ already we have defined. So, this $J(w\text{-}) = \frac{|\mu\text{-}_1^{\sim} - \mu\text{-}_2^{\sim}|^2}{S_1^{\sim 2} + S_2^{\sim 2}} = \frac{w\text{-}^T S_B w\text{-}}{w\text{-}^T S_w w\text{-}}$. So, what is this $J(w\text{-})$? It is a what is $J(w\text{-})$? It is a measure of the difference between class means normalized by a measure of the within class scatter matrix that is $J(w\text{-})$. So, I am repeating this $J(w\text{-})$ that is the criterion function is a measure of the difference between class means normalized by a measure of the within class scatter matrix that is $J(w\text{-})$.

Now we have to find a maximum $J(w\text{-})$. So, for this we have to differentiate and equate to 0. So, our objective is to find the optimum value of $w\text{-}$ that is a projection vector. So, that means, we have to maximize $J(w\text{-})$ and we have to equate it to 0 because we have to find the maximum value. So, how to get this one? So, move to the next slide.

So, to find maximum of $J(w\text{-})$ this criterion function. So, we have to differentiate and equate to 0. So, that means, $\frac{d}{dw\text{-}} J(w\text{-}) = \frac{d}{dw\text{-}} \frac{w\text{-}^T S_B w\text{-}}{w\text{-}^T S_w w\text{-}} = 0$. So, how to differentiate this one? So, this will be equal to $(w\text{-}^T S_w w\text{-}) 2 S_B w\text{-} - (w\text{-}^T S_B w\text{-}) 2 S_w w\text{-} = 0$. So, that is equal to $(w\text{-}^T S_w w\text{-}) 2 S_B w\text{-} - (w\text{-}^T S_B w\text{-}) 2 S_w w\text{-} = 0$ and dividing by $2 w\text{-}^T S_w w\text{-}$.

So, dividing by this $2 w\text{-}^T S_w w\text{-}$. So, what we will get? We will be getting $\frac{w\text{-}^T S_w w\text{-}}{w\text{-}^T S_w w\text{-}} S_B w\text{-}$. So, this is not a difficult mathematics only you have to do the differentiation and do some mathematics. So, $\frac{w\text{-}^T S_w w\text{-}}{w\text{-}^T S_w w\text{-}} S_B w\text{-} - \frac{w\text{-}^T S_B w\text{-}}{w\text{-}^T S_w w\text{-}} S_B w\text{-} S_w w\text{-} = 0 \Rightarrow S_B w\text{-} - J(w\text{-}) S_w w\text{-} = 0 \Rightarrow S_w^{-1} S_B w\text{-} - J(w\text{-}) w\text{-} = 0$.

So, to solve this equation we are considering the generalized Eigen value problem. So, move to the next slide for solving we are considering the generalized Eigen value problem. So, $S_w^{-1} S_B w\text{-} = \lambda w\text{-}$ where $\lambda = J(w\text{-})$ and λ is nothing but a scalar is a scalar λ is a scalar. So, corresponding to this Eigen value problem we can determine the optimum value of $w\text{-}$ that is the projection vector $w\text{-}^{*} = arg\ max_{w\text{-}} J(w\text{-}) = arg\ max_{w\text{-}} \frac{w\text{-}^{T} S_B w\text{-}}{w\text{-}^{T} S_w w\text{-}} = S_{w\text{-}}^{-1}(\mu_{1\text{-}} - \mu_{2\text{-}})$. So, we are getting the optimum projection direction that is $w\text{-}^{*} = S_{w\text{-}}^{-1}(\mu_{1\text{-}} - \mu_{2\text{-}})$.

So, we are getting the best projection direction by using this equation. So, this is the equation. So, with the help of this equation you can determine the best projection direction. So, considering this case because we are determining the best projection direction with the help of this criterion function the criterion function is $J(w\text{-})$ you can see we have determined the projection direction the best projection direction $w\text{-}^{*}$. So, in this class I discussed the concept of LDA linear discriminant analysis I considered only two classes and based on these two classes I have determined the best projection direction and the same concept can be extended for C number of classes that is called multiple discriminant analysis. One concept is pretty important that is the between class scatter and within class scatter and based on this we have determined the best projection direction.

So, the goal of the LDA is to find a set of vectors that maximizes between class scatter and simultaneously it minimizes within class scatter that is the goal of the LDA the linear discriminant analysis. So, let me stop here today. Thank you.