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-26

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my first class, I discussed the fundamental concept of pattern classification. One important step of a typical pattern classification system is feature selection. So, I have to select the most discriminative features. And if I want to improve the performance of a classifier, I have to consider more number of features. But there is a problem.

Suppose I have limited number of training samples, then what will happen if I increase the dimension of the feature vector that means I am considering more number of features, then the performance or the accuracy of the classifier may not improve. And this is called the curse of dimensionality. And the problem is because of the limited number of training samples.

And another problem is if I increase the dimension of the feature vector, computational complexity increases.

So, if I consider more and more features, then the computational complexity increases. So, to consider this issue, one important point is I have to reduce the dimension of the feature vector. So, I will be discussing two popular approaches. One is PCA the principal component analysis and another one is MDA the multiple discriminate analysis. In the PCA, we have to project the high dimensional data into a low dimensional space.

So, suppose I have a feature vector the D dimensional feature vector and that can be projected into a low dimensional space. And after the projection data will be uncorrelated that is the transform data will be uncorrelated. So, I have to find the best projection direction and that is the main objective of the PCA principal component analysis. In case of the MDA, the objective or the goal is to find a set of vectors which maximizes between class scatter and minimizes within class scatter.

Today I am going to discuss the problem of dimensionality and the next is the concept of PCA to reduce the dimensionality of the feature vector.

So, let us start this class problem of dimensionality. So, suppose I am considering a D dimensional feature vector. So, x_1, x_2 these are the components of the feature vector. So, these are the features and you can see each and every feature convey some information characteristics of the pattern and each feature has different discriminatory power. So, now we have to consider the problem of considering more number of features that means we are considering high dimensional feature vector.

So, what is the problem? So, one problem I told you that is the curse of dimensionality that suppose I have limited number of training samples and if I increase the dimension of the feature vector, the performance of the classifier may not improve and that is the curse of dimensionality. Now in my discussion of the minimum error rate classification, I have shown the concept of the probability of error. So, let us discuss about the probability of error. So, suppose I am considering two classes, the classes ω_1 and ω_2 and corresponding to this I am showing the probability density functions. So, what is the probability density function? This is the probability density function corresponding to a class the class is ω_1 and suppose corresponding to this class the mean is μ_1 and corresponding to the second class I have this distribution that is the class conditional density I am plotting.

So, corresponding to this and this is for the class ω_2 and suppose this is mean μ_2 and what I am plotting here, I am plotting this probability density function $x|\omega$ versus x. So, that is nothing but the class conditional density versus x. So, we are considering two classes and we are assuming that this $P(\omega_1) = P(\omega_2)$ that is the a priori probabilities are same. So, a priori probability for the class ω_1 is the probability of ω_1 and another probability the probability of ω_2 that is for the class ω_2. So, they are equal and corresponding to this how to take a classification decision for taking a classification decision suppose here actually if I consider this region and this green colored region that region is nothing but the area corresponding to the probability of error.

So, this is the area corresponding to the probability of error. So, for taking a classification decision what we can consider suppose we are considering a threshold here threshold is T and in this case what another condition I am considering that the variances of two classes are same. So, another condition here I am considering that $\sigma_1^2 = \sigma_2^2$. So, variances of two classes are same that means these two distribution have same size and same shape and in this case the means are different that the two these two means are different $\mu_1 \neq \mu_2$. So, here you can see that area corresponding to the green colored that is the area corresponding to the error and how to take a classification decision.

So, we have considered a threshold if $P(x|w) > T$ then we have to consider the class ω_2 and if this density $P(x|w) < T$ then I have to consider the class ω_1. So, based on this threshold I can take a classification decision, but while taking this decision I am encountering some errors. So, here you can see this errors this error already I have shown that is the green colored region that is the area corresponding to the error. So, based on this threshold I can take a classification decision if this density is greater than the threshold I have to consider a particular class. And if it is less than that particular threshold then I have to consider the another class.

Now, how to reduce this error? So, one possibility is that if I increase the separation between these two means then I can reduce the probability of error I can reduce the error. So, how to do this I can show it pictorially. So, let us draw the same thing here. So, what I need to do I have to increase the separation between the means of the two classes. So, this is one distribution corresponding to the class ω_1 and this is the mean the mean is μ_1.

So, let us consider the second class second class is this corresponding to the class ω_2 and suppose the mean μ_2. So, here you can see I am increasing the separation between these two means and because of this you can see here this area that this green color area is reduced that means the error is reduced. So, that means I can say that discriminatory power of the features increase if the means are widely apart.

So, this is the important sentence I can repeat this. So, that is the discriminatory power of the feature increases if the means are widely apart.

Now to improve the performance of a classifier we may consider more and more features if the classifier accuracy is not good we may increase the dimension of the feature vector. But if I increase the dimension of the feature vector the problem already I told you the major problem is the increase in the computational complexity. So, that point I am going to discuss because the computational complexity depends on the dimension of the feature vector. So, let us move to the next slide. So, in my earlier class I have derived the expression for the discriminate function the discriminate function is $g_i(\bar{x}) = -\frac{1}{2}(\bar{x} - \bar{\mu}_i) - \frac{d}{2} ln(2\pi) - \frac{1}{2} ln(|\Sigma_i|) + ln(P(\omega_i))$.

So, this is the expression for the discriminate function. So, based on the discriminate function I can take a classification decision. Now I have to determine the amount of computation in calculating the value of the discriminate function $g_i(\bar{x})$. So, in this expression here you can see this ln probability of omega i.

So, this actually it depends on it depends on n, n means the number of samples. So, it depends on number of samples. So, that I am not considering and if I consider this term this is class independent.

So, that means it has no discriminatory power. So, I am not considering that one.

So, now there are two major computations one is the mean another one is the computation of the covariance matrix. So, here n the small n represents the number of samples. Now how to compute the mean. So, for computation of the mean you can see you can determine the mean. So, this is the mean vector is nothing but $\mu\text{-}_i = \frac{1}{n}\sum_{k=1}{}^{n} X\text{-}_k$.

So, that is the Feature vector of dimension d, $X\text{-}_k$ is a Feature vectorof dimension d. So, in this case in computation of the mean how many multiplications or how many additions or how many divisions I have to determine. So, total number of summations you can see. So, n number of summations followed by one division and if I consider this division it takes constant time. So, in the calculation of the mean what is the computation.

So, n number of summations followed by one division and this division takes constant time and we are considering the Feature vector of dimension d. So, that means while computing this mean total number of summations will be nxd. So, that means for this the total number of summations will be nxd. So, this is the amount of computation for calculating the mean. Next I will consider the covariance matrix.

So, this is the expression for the covariance matrix $\Sigma_i = \frac{1}{n}\sum_{k=1}{}^{n}(X\text{-}_k - \mu\text{-}_i)(X\text{-}_k - \mu\text{-}_i)^T$. So, here you can see this is here you can see this is nothing but the a vector of dimension $dx1$ and this is a vector of dimension $1xd$ because we are considering transpose. So, ultimately I am getting a matrix the matrix is dxd matrix. So, for calculating or for the computation of the covariance matrix. So, approximately we need nd square number of computations.

So, that means I can say the order of the computation is nd^2 and suppose if I consider c number of classes. So, for c number of classes c classes the order of the computations the order of the computation is cnd^2. So, we are only considering the computation required for computing the covariance matrix because it is nd^2 the mean is only nd that is not so significant, but we have to consider the computation required for computing the value of the covariance matrix that is in the order of nd^2.

So, for c number of classes it is cnd^2. So, in this case you can see here c and n this is linear c is the number of classes and n is the number of samples.

So, it is linear and this d^2 and this d^2 that is the quadratic term. So, that is the most important computation. So, that is the quadratic term we have to consider d^2. So, d is the dimension of the feature vector. So, if I increase the dimension of the feature vector by a factor of 2 the computational complexity increases by a factor of 4.

And also if I increase the dimension of the feature vector by a factor of 3, the computational complexity increases by a factor of 9. So, you can see the effect of the dimension of the feature vectors. So, if I increase the dimension of the feature vector the computational complexity increases. So, here you can see that the major computation is the computation of the covariance matrix. So, that means, if I incorporate more and more features, it may increases the performance of the classifier, however, increases the complexity.

So, that is the summary of this discussion. So, if I consider more and more features, it increases the computational complexity, but the performance may improve. So, how to reduce the computational complexity that means how to reduce the dimension of the feature vector.

So, one thing we can consider the projection of high dimensional data to a low dimensional space.

So, let us move to the next slide. So, how to reduce the dimension of the feature vector, we can consider the projection of high dimensional data to a low dimensional space. So, we may consider that point, we may consider this technique that is we can project the high dimensional data to a low dimensional space. And after the projection, the feature will be orthogonal that means, the feature should be uncorrelated and they should not affect each other. So, we are projecting high dimensional feature vector into a low dimensional space. And because of this projection, the transform data will be uncorrelated.

So, that means the feature should be orthogonal and feature should be uncorrelated, they should not affect each other. So, I can say that feature should be orthogonal feature should be orthogonal. Now, how to reduce the dimension of the feature vector. So, let us discuss this concept reduction of dimensionality. So, my original dimension is d, d dimensional feature vector and I want to reduce it to d'.

So, in this case, $d' \leq d$. So, already I told you, I will be considering two popular techniques, one is the PCA another one is the MDA. So, the first technique is the principal component analysis. So, that is PCA. So, the fundamental concept of the PCA is the projection of high dimensional feature vector to a low dimensional space.

The second point is MDA that is the multiple discriminate analysis. So, in case of the multiple discriminate analysis, what we can consider? We can consider these two cases, we can increase the separation between the means of the classes that already I have discussed. So, how to increase the separation between the means. So, if I increase the separation between the means, the classification error will be reduced. So, that means, I can say increase the separation between means of the classes.

So, we can consider this one, the second one is we can consider the compact clusters. So that means, what is the compact clusters? That means, we can reduce the variance of the samples. So, we can reduce the variance of the samples. So, that means, I can consider one term and that is called within class scatter.

This is nothing but within class scatter. That means, I can reduce the variance of the samples belonging to a particular class and that is nothing but I can say it is intra class scatter. So, within class scatter or the intra class scatter and this increase the separation between the means of the classes that is nothing but the within class scatter and increase the separation between the means of the classes. That means, it is the inter class separation that is nothing but it is between class scatter. I can say it is the inter class scatter. So, I can increase the separation between the means of the classes and that is nothing but between class scatter or I can say inter class scatter and also I can consider the compact clusters that is nothing but we can reduce the variance of the samples of the classes and that is nothing but within class scatter or I can say it is the intra class scatter.

So, this concept the increase the separation between the means of the classes that already I have discussed in my previous slide. The second point is how to consider the compact clusters. That means how to reduce the variance of the samples. So, if I reduce the variance of the samples then also we can reduce the classification error that I can show you pictorially. Suppose I am showing the probability density function for 2 classes.

So, this is the probability density function. So, again we are considering 2 classes. This is 1 class and suppose this is the mean μ_1 and second class is this. So, it has a mean suppose μ_2.

Now, I want to reduce the variance of the samples that means I am considering the compact clusters.

So, now I am not changing the means. So, mean will be same I am not changing the means of the clusters. So, this will be μ_1 and μ_2 it will remain same, but I am reducing the variance of the samples. So, you can see I am reducing the variance and this is product class ω_1, this is product class ω_2. So, this is ω_1 and this is ω_2. So, we had this area corresponding to the classification error, but in the second case you can see this error is significantly reduced because we are reducing the variance of the samples that will be taken care by within class scatter.

So, this is the concept. So, you can see by increasing the separation between the means we can reduce the classification error and by considering the compact clusters we can reduce the classification error. So, move to the next slide and the principal component analysis. So, we are considering n number of feature vectors n number of d dimensional feature vector feature vectors of dimension d. So, that means we are considering the feature vector $x\text{-}_1, x\text{-}_2, \ldots, x\text{-}_n$. So, we are considering n number of feature vector of dimension d and from this feature vector we can determine the mean of this feature vector.

So, what is the mean of the feature vector? The mean of the feature vector is nothing but $m\text{-} = \frac{1}{n}\sum_{k=1}{}^{n} x\text{-}_k$. So, you can see the n vectors are represented by mean and that is nothing but single representation and that is not a good representation because we are not considering variance. The mean we are considering that is actually 0 variability and this is nothing but 0 dimensional representation 0 dimensional representation. Because n vectors are represented by mean and it is a single representation and we are not considering variance that means 0 variability we are considering. I am considering another representation in that case I will be considering all the n vectors together and we want to find the best representation of all these n vectors.

The mean is not a best representation because it is a 0 dimensional representation and we are not considering the variance. Now how to represent n vectors that means I will be considering all the n vectors together and we are also considering the variability that means the variability will be preserved and that is nothing but the representation of d dimensional feature vector into 1 dimensional line. So, let us consider that a concept. So, how to represent d dimensional feature vector into 1 dimensional line and this line is represented by the unit vector the unit vector is e passing through the mean.

So, the pictorially how can I show you. So suppose these are some samples. So, n number of samples are available. So, n points n number of points we are considering and I want to find the best representation. So that means the representation of the d dimensional feature vector into 1 dimensional line.

So that means I am considering a line. So line is suppose is a line and this line is represented by the unit vector the unit vector is e and it is passing through the mean and it is passing through the mean. So suppose this is the mean. So the mean is m. So that means what we have to consider the mapping of all the points into this line.

So how to do the mapping. So just I have to do the mapping I have to do the mapping of all the points into this line. So mapping of all the points into this line and this line is represented by the unit vector the unit vector is e and this line is passing through the mean. So what is the equation of this line the equation of this line is $x\text{-} = m\text{-} + ae\text{-}$. So this is the equation of the line. So for different values of a we are moving along the line and what is a, a is the position of different points on the line.

So suppose considering this point suppose that this point is $x\text{-}_k$ this point is $x\text{-}_k$ that is mapped onto this line the line is represented by e. So the point $x\text{-}_k$ can be represented like this $x\text{-}_k \approx m\text{-} + a_k e\text{-}$. So what is a_k here a_k is here this is a_k corresponding to this it is a_k. So the point $x\text{-}_k$ is mapped onto this line. So a k is the position of the point on the line the line is represented by the unit vector e and here you can see $x\text{-}_k$ is of dimension d and it is represented as a line.

So there will be some error so that is why I am considering the approximate representation. I am repeating this the $x\text{-}_k$ is of dimension d and it is represented as a line. So we are encountering some errors. So that is why $x\text{-}_k$ is approximately represented as m, m is the mean a_k is the position of $x\text{-}_k$ on the line the line is represented by the unit vector e.

So $x\text{-}_k = m\text{-} + a_k e\text{-}$. So it is the approximate representation. So based on this we are considering one error function. So what is the error? Error is nothing but the difference between the actual value and the approximated value. So what is the approximated value? Approximated value is $m\text{-} + a_k e\text{-}$ and what is the actual value? Actual value is $x\text{-}_k$. So we are considering the error function error function the error function is $J(a_1, a_2, \ldots, a_n, e)$ these are the positions of different points and is nothing but the unit vector showing the direction of the line and $k = 1{:}n$, n number of samples we are considering.

This is the approximate representation $m\text{-} + a_k e\text{-}$ approximate representation of the point and what is the actual value? Actual value is $x\text{-}_k$ and we are considering the sum squared error. So we are considering the sum squared error and you can see for different points I have different a_k's. So just I can write for different points I have different a_k's. So we have defined the error function.

Now I have to minimize this sum squared error in terms of a. So move to the next slide. So we have to minimize the sum squared error in terms of a that means I want to find the value of a_k. So for the time being I am considering e is fixed. So considering that e is fixed suppose.

So for the time being I am considering the direction is fixed. So we can find an optimal set of coefficients a_k by minimizing the sum squared error. So this J is nothing but a_1, a_2, a_n and this is the unit vector. So we have this expression $\sum_{k=1}^{n} \quad ||a_k e\text{-} - (x\text{-}_k - m\text{-})||^2$. So this expression I can write like this $\sum_{k=1}^{n} \quad ||a_k e\text{-} - (x\text{-}_k - m\text{-})||^2$ so we can expand like this. So in this case e is the unit vector so this will be 1 actually and if I see this term it is independent of a.

So that means we need not consider that term. So now because I have to minimize the sum squared error so I have to differentiate J with respect to a_k and it should be equal to 0 and that is the partial derivative with respect to a_k and it is equating to 0. So if I consider this expression then it is equating to 0 so I will be getting from this $2\sum a_k - 2\sum e\text{-}^T(x\text{-}_k - m\text{-}) = 0$. So I am taking the partial derivative and it is equating to 0.

The partial derivative of J with respect to a_k and it is equating to 0. So from this expression you can see I can determine a_k, $a_k = e\text{-}^T(x\text{-}_k - m\text{-})$. So I am getting the value of a_k. So what is the interpretation of this equation? So what is the interpretation of this equation? This is an important equation. The interpretation is this. It is the orthogonal projection of $x\text{-}_k$ onto a line passing through the mean that I can say it is orthogonal projection of $x\text{-}_k$ onto a line passing through the mean and this is the best representation in 1D.

So I can write this is nothing but the orthogonal projection of $x\text{-}_k$ onto a line passing through the mean. The mean is m, the mean of all the samples. So what is the best representation in 1D? So that means we have to project d-dimensional feature vector along a line passing through the mean and this is the orthogonal projection. So I am repeating this. The best 1D representation is I have to project d-dimensional feature vector along a line passing through the mean.

Now after this we have to determine which is the best projection direction. So which is the best e. So next what we have to see which is the best e. That means which is the best projection direction that we want to determine.

So move to the next slide. So we have expressed $J(e\text{-})$ like this. This is the criterion function $J(e\text{-}) = \sum a_k^2 - 2\sum a_k e\text{-}^T(x\text{-}_k - m\text{-}) + \sum||x\text{-}_k - m\text{-}||^2$. So we have this expression. So now we have to find the best projection direction. That means we have to find the best e and this I can represent like this $\sum e\text{-}^T(x\text{-}_k - m\text{-})$ because already we have determined the value of a_k. So from that I am just putting this value, putting the value of a_k, $-\sum[e\text{-}^T(x\text{-}_k - m\text{-})(x\text{-}_k - m\text{-})^T e\text{-}] + \sum||x\text{-}_k - m\text{-}||^2$.

So which can be written like this $-e\text{-}^T[\sum(x\text{-}_k - m\text{-})(x\text{-}_k - m\text{-})^T]e\text{-} + \sum||x\text{-}_k - m\text{-}||^2$

because e is independent of k. So that is why I am taking it out $-e\text{-}^T[\sum(x\text{-}_k - m\text{-})(x\text{-}_k - m\text{-})^T]e\text{-} + \sum||x\text{-}_k - m\text{-}||^2$. So I can get this one. So if you see this expression so this part I can consider as the scatter matrix that actually corresponds to the spread of data.

The scatter matrix is nothing but the spread of data. And if I consider the scale version of this that is nothing but the covariance matrix. What is the covariance matrix? The covariance matrix is nothing but the scale version of the scatter matrix. The covariance matrix is $\Sigma = \frac{1}{n}\sum(x\text{-}_k - m\text{-})(x\text{-}_k - m\text{-})^T$. So that is the covariance matrix. So this $J(e\text{-})$ I can represent like this $J(e\text{-}) = -e\text{-}^T Se\text{-} + \sum||x\text{-}_k - m\text{-}||^2$.

So the second term I need not consider because it is independent of e. So that means this term it is independent of e. So I am not considering that part. So it is independent of e. So this term also we did not consider because it is independent of e.

So we have not considered that term. So we have to minimize this error. So we have to minimize the error. So we have to minimize this. That means if I want to minimize this $J(e\text{-})$ I have to maximize this term. So I have to maximize this because it is a negative term.

So if I want to minimize $J(e\text{-})$ I have to maximize the term $e\text{-}^T Se\text{-}$. So maximize $e\text{-}^T Se\text{-}$. So here S is the scatter matrix. How to find the maximum value? So for this we have to consider the solution and it can be considered by Lagrangian.

So move to the next slide. So we are considering the solution by Lagrangian. So $u = e\text{-}^T Se\text{-} - \lambda(e\text{-}^T e\text{-} - 1)$ and subject to the condition $||e\text{-}|| = 1$. So e is the unit vector. So it will be equal to 1. To find the maximum value we have to differentiate u with respect to e.

So it is $2Se\text{-} - 2\lambda e\text{-} = 0$. So here lambda is nothing but the Lagrangian multiplier. So it will be equal to 0. So from this what we can obtain $Se\text{-} = \lambda e\text{-}$.

So this is nothing but the Eigen value expression. So this is the Eigen value expression. So in this case e is nothing but the vector earlier we considered as the unit vector and it shows the direction of the line passing through the mean and this e is nothing but the Eigen vector of the scatter matrix. So Eigen vector of S, S is the scatter matrix and lambda is the corresponding Eigen value. So since we have to maximize we have to maximize $e\text{-}^T Se\text{-}$ that means we have to maximize $e\text{-}^T \lambda e\text{-}$. So that means we have to take the maximum value of lambda. So if I want to maximize $e\text{-}^T Se\text{-}$ that means it is equivalent to we have to maximize $e\text{-}^T \lambda e\text{-}$ that means we have to take the maximum value of lambda.

So which Eigen vectors we need to consider. The Eigen vectors corresponding to maximum Eigen values. So we can determine the Eigen vectors e corresponding to the maximum Eigen values and after this I have to find a_k and these are called the principal components and if I only consider only one Eigen vector then the dimension is reduced to 1 okay. So that means I can say if only one Eigen vector is considered then the dimension is reduced to 1. So from this expression you can see here already I have explained so I have to select the Eigen vectors corresponding to the largest Eigen values. So move to the next slide.

Suppose I want to reduce the dimension of the Feature vector to d' from d so $d' < d.$ So that means I have to consider d' number of Eigen vectors. The original dimension is d and I am reducing it to d'. So d' is less than d so that means I have to consider d' number of Eigen vectors. So if I consider this dimension is reduced to d' then this $x{-}$ the point $x{-}$ can be represented like this $m{-} + \sum_{i=1}^{d'} a_i e{-}_i$ because I am reducing the dimension to d' so $a_i e{-}_i$. So the point $x{-}$ can be represented like this and this criterion function the criterion function $J_{d'}$ can be represented like this corresponding to this d' dimension.

So $J_{d'} = \sum_{k=1}^{n} ||(m{-} + \sum_{i=1}^{d'} a_{ki} e{-}_i) - x{-}_k||^2$ so this is the criterion function and you know the scatter matrix is a real and a symmetric matrix the scatter matrix is real and the symmetric. In this case what I have shown that dimension is reduced to d' and the point $x{-}$ can be represented like this and corresponding to this I am considering the criterion function that is nothing but $J_{d'}$. So we can obtain like this. So this criterion function is minimized when the vectors $e{-}_1, e{-}_2, e{-}_{d'}$ are the Eigen vectors of the scatter matrix having largest Eigen values and the coefficients a_i are the principal components. So I am repeating this because I am defining the criterion function corresponding to the dimension d' and the point $x{-}$ can be represented like this $x{-} = m{-} + \sum_{i=1}^{d'} a_i e{-}_i$.

So this is the expression for the $x{-}$ the point $x{-}$ and corresponding to this criterion function is $J_{d'}$. So we have this expression and this criterion function is minimized when the vectors $e{-}_1, e{-}_2, e{-}_{d'}$ these are the Eigen vectors of the scatter matrix corresponding to the largest Eigen values. So we have to consider the largest Eigen values and the corresponding Eigen vectors and the coefficients the coefficients are coefficients a_i are nothing but the principal components. So these are the principal components. So you can see I am reducing the dimension of the Feature vector from d to d'. So that means we are considering d' number of Eigen vectors and we are considering the orthogonal projection the corresponding principal component give d' dimensional vectors.

So here you can see we are considering the d dimensional Feature vector and it is reduced to d' dimensional Feature vector. So that means we are considering d' number of Eigen vectors and we are considering the orthogonal projection. So the corresponding principal components give d’ dimensional vectors. So I am repeating this the original dimension is d and it is reduced to d’ and corresponding to this the point x- is represented like this the point x- is represented like this and corresponding to this we are considering the criterion function the criterion function is $J_{d'}$ and this riterion function can be minimized when the vectors the Eigen vectors $e\text{-}_1, e\text{-}_2, e\text{-}_{d'}$ are the Eigen vectors corresponding to the largest Eigen values. So these are the Eigen vectors of the scatter matrix and we have to consider the Eigen vectors of the scatter matrix having largest Eigen values and the coefficients a_i are called the principal components.

So here you can see how we can determine the base projection direction the base projection direction is nothing but the direction of the Eigen vectors corresponding to the largest Eigen values and also how we can determine the coefficients that is the principal component a_i's. In this class I discussed the concept of problem of dimensionality and also I have introduced the concept of PCA the principal component analysis. So if I increase the dimension of the feature vector the computational complexity increases. I have also explained the concept how to reduce the probability of error. So if I increase the separation between the means of the classes the probability of error the error reduces and also if I reduce the variance of the samples of the classes then the error can be reduced.

One is how to increase the between class scatter and how to reduce the within class scatter. So these two concepts I have explained that is by increasing the separation between the means and by considering the compact clusters I can reduce the classification error. After this I discussed the concept of the PCA the principal component analysis and I have shown how to do the orthogonal projections and how to determine the best projection direction. So the best projection direction is given by the Eigen vectors corresponding to the largest Eigen values.

So this is the fundamental concept of the PCA. So in my next class also I will be explaining the concept of the PCA. So let me stop here today. Thank you.