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-2

Lecture-7

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of Bayesian decision theory. And after this, I discussed the concept of the probability of error and risk. And based on this, I discussed the concept of zero one loss function. With the help of this function, I can take a classification decision. After this, I discussed the concept of the discriminate function.

The discriminate function is $g_i(x)$. So, for c number of classes, I have c number of discriminate function. And I have to pick the largest discriminate function.

And based on this, I can select a particular class.

Today, I will continue my same discussion that is a Bayesian decision theory. But in this case, I will consider discrete features. So, what is continuous feature and what is discrete features I am going to explain. And based on this, I will be explaining the concept of Bayesian decision theory. So, let us see the Bayesian decision theory for discrete features.

So, Bayes decision theory for discrete features. So, let us start this concept. So, in case of the continuous case, if I consider continuous case, Feature vector, the Feature vector is x, Feature vector x could be any points in d dimensional euclidean space.

So, that is actually R^d. So, this is for a continuous case.

Feature vector x could be any points in d dimensional euclidean space R^d . So, if I consider discrete case, that means we are considering discrete features. So, in case of the discrete case, this Feature vector x can assume only one of the m discrete values. So, x can assume any only one of the m discrete values.

So, maybe discrete value we can consider V1, maybe V2, these are the discrete values we can consider.

So, in case of the continuous we considered integration, this we considered the integration of the class conditional density, that is the likelihood. And in case of the discrete case, this is replaced by summation. This is the summation over x for all the Feature values and ω_j. So, corresponding to this our Bayes formula will be already I have explained the Bayes formula. You know this formula, this probability of $\omega_j|x$ that is nothing but the posterior probability and this probability $x|\omega_j$ that is the likelihood or you can say the class conditional density that is the prior probability and denominator is the evidence.

So, evidence already you know how to write the evidence, evidence is nothing but for c number of classes, it is probability of $x|\omega_j$into probability of ω_j. So, you know this one. Based on the Risks, what is the decision rule? The decision rule already you know what is the decision rule? I have to select a particular action, I have to select a particular action, action is suppose α_i, this action I am selecting for which my Risks is minimum. What is my Risks? My conditional Risks is the conditional Risks is R the action is α_i and that is taken for the Feature vector, the Feature vector is x is minimum. So, this is my decision rule.

So, select a particular action, the action is α_i for which the conditional Risks that is the R $\alpha_i|x$. So, that should be minimum. So, this is the decision rule. Now, let us consider the independent binary Features. So, what is independent binary Features? Independent binary Feature.

So, what is independent binary Features? So, Feature values or maybe you can write the Feature vector, the Feature vector are binary value. And we are considering, we are considering they are conditionally independent, conditionally independent. They are conditionally independent. So, the Feature vector already you know how to write this is the D dimensional Feature vector and these are the components of the Feature vector.

So, these are D dimensional Feature vector.

So, this is a Feature vector. So, this Feature vectors are binary valued and conditionally independent. So, this actually this is the concept of the Naive Bayes classifier. This is the concept of the Naive Bayes classifier.

So, we are considering the Feature vector are binary valued and conditionally independent.

So, now, let us consider a two class problem. Two class problem we are considering now.

Suppose I am considering one Feature xi, this is a binary we are considering the binary valued. So, either it may be 0 or 1 this value that means 0 means no or yes, no or yes.

So, we are considering the Feature value xi, the Feature is xi either it may be 0 or 1 that we are considering.

So, the probability Pi we can determine the probability xi is equal to 1 corresponding to the class Omega 1. So, that means this Feature gives the yes answer, the answer is 1 for the class Omega 1. And what is the probability Qi, Qi is the probability Pr and xi is equal to 1. So, it gives the answer yes for the class ω_2. So, I am considering these two probabilities, probability Pi and Qi corresponding to the Feature, the Feature is xi.

So, xi is equal to 1 that means it is giving the answer the answer is yes, corresponding to the class that class is ω_1 or ω_2, because we are considering two classes. So, if Pi is greater than Qi, suppose this probability Pi is greater than Qi that means what is the meaning of this, the ith Feature we are considering the ith Feature give ith Feature gives a yes, the answer is the yes, yes answer more frequently, more frequently when the state of the nature is Omega 1 that means the class is ω_1 when it is ω_2. So, that means if the probability Pi is greater than Qi and here we are considering ith Feature, this ith Feature gives yes answer more frequently when the state of the nature is ω_1. So, that means it is favoring the class ω_1. This is the meaning of this probability.

So, we are determining the probability Pi and Qi and based on this probability we can take a classification decision. So, move to the next slide. So, now, because we are considering the conditional independence, conditional independence we are considering. So, what is the class conditional probabilities? So, what is the class conditional probabilities? That is the probability of $x|\omega_1$that is the class conditional probability and since we are considering this binomial distribution. So, binomial distribution is i is equal to 1 to d because the dimension of the Feature vector is the probability Pi Xi 1 minus Xi that we are considering this probability we can determine and similarly we can determine the probability of X given Omega 2 that also we can determine.

So, this is nothing but the binomial distribution. So, Qi so, I can determine the probability of $x|\omega_2$ after this we can determine the likelihood ratio. So, likelihood ratio already I have defined in my previous classes. So, likelihood what is the likelihood the probability of $x|\omega_1$and probability of $x|\omega_2$. So, this ratio we can determine that is nothing but i is equal to 1 to d.

$$\frac{P(x|w_1)}{P(x|w_2)} = \pi_{i=1}^{d} \frac{P_i}{q_i}^{x_i} \frac{1-p_i}{1-q_i}^{1-x_i}$$

$$g(x) = \sum_{i=1}^{d} \left[x_i ln\frac{p_i}{q_i} + (1 - x_i) ln\frac{1 - p_i}{1 - q_i}\right] + ln\frac{P(w_1)}{P(w_2)}$$

So, this expression I know. So, this expression I am getting just I am putting the value of the likelihood that I am getting. So, from this gx, this gx already I have determined. So, I can write the gx like this gx is equal to summation i is equal to 1 to d and this is Wi Wi that Wi is the weight not the class in product class we are considering Omega i but in this case we are considering the weight, weight is Wi plus W naught. So, what is Wi? Wi is nothing but ln that is the weight. So, this is the weight Wi is the weight.

So, it is i is equal to 1 to so we are considering the d dimensional so it is like this and W naught is the bias is a bias. So, bias is nothing but summation i is equal to 1 to d ln. So, this should be 1 minus Pi 1 minus Pi 1 minus Qi plus ln probability of Omega 1 and probability of Omega 2. So, this actually we have obtained from this, this we have obtained from this, these two I am defining like this. So, this Wi that is the weight we can represent like this weight is nothing but ln Pi into 1 minus Qi divided by Qi into 1 minus Pi that is the weight and W naught is the bias.

So, this W naught is the bias and this Wi is the weight. So, bias also I can write like this. Now based on this weight Wi and also based on the bias W naught I can take a classification decision.

So, here you can see I am repeating this the class conditional probabilities I am representing like this. This is the binomial distribution and from this we can determine the likelihood ratio.

So, this is the likelihood ratio we have determined. So, from this expression I am determining the discriminant function. So, this is the discriminant function.

So, discriminant function I am representing like this gx is equal to Wi xi plus W naught and summation I am taking from i is equal to 1 to d because we are considering d dimensional Feature vector.

So, this is the expression for gx. Now, let us discuss how you can take a classification decision based on this Wi, Wi is the weight. So, I am moving to the next slide. So, what you have obtained gx already I have obtained that is nothing but the summation i is equal to 1 to d the weight vector is Wi is the weight and this is xi plus W naught, W naught is the bias and what is Wi, Wi is equal to ln Pi 1 minus Qi and Qi 1 minus Pi. So, in this case i is equal to 1 to up to d and W naught already I have defined summation over that is from i is equal to d ln 1 minus Pi 1 minus Qi plus ln probability of omega 1 and probability of omega 2 that expression already I have defined. Now, let us see how to take a classification decision.

Decide the class omega 1 if gx that is a discriminate function gx is greater than 0 and you can select the class omega 2 if the discriminate function gx is less than 0. So, based on this condition you can take a classification decision. Suppose if Pi is equal to Qi these two probabilities are equal. So, what feature we are considering i feature we are considering. So, xi gives no information about the about the class or you can write the state of the nature state of the nature or maybe the class.

And in this case corresponding to this this Wi, Wi is the weight Wi is equal to 0 that means it is called Feature independence. So, you can understand that if I consider these two probabilities Pi and Qi they are equal then the i th feature the i th feature is the xi is the i th feature gives no information about the state of the nature and corresponding to this Wi. So, from this expression Wi will be 0. If I consider Pi is equal to Qi then Wi will be 0 and this condition is called Feature independence. The second condition is if Pi is greater than Qi this probability Pi is greater than Qi then what will happen from this expression then 1 minus Pi will be less than 1 minus Qi and in this case what will be the width the width Wi that will be positive.

So, that means the meaning is decision, decision will be will be in the favour of in the favour of omega 1. So, this is the case if Pi is greater than Qi then from the expression you can see 1 minus Pi will be less than 1 minus Qi then this weight if you see this weight the expression for the width Wi that will be positive and decision will be in the favour of the class the class is omega 1 and similarly if Pi is less than Qi. So, what will happen this weight Wi that will be negative that will be negative and the decision and decision will be in favour of in favour of the second class the second class is omega 2. So, this is the case and if this probability omega 1 is greater than probability of omega 2 that is the prior probability. So, that means it increases bias, bias is W naught.

So, this is the expression for the bias you can see this is the expression for the bias. So, it increases the bias W naught. So, that means decision in favour of omega 1 and if the second condition, second condition is this probability of omega 2 is greater than probability of omega 1. So, that means it decreases the bias and whenever it decreases the bias. So, decision is in favour of that class that class is omega 2.

So, here you can see from this discussion from this weight the weight is Wi and the bias is W naught you can take a classification decision. So, in this discussion what we are considering, we are considering discrete features. So, that means the feature vector x can assume only one of the m discrete values, the discrete value V1, V2, V3 we have already explained and also we have considered the feature vectors are binary valued. So, maybe it may be either the value is 1 or maybe 0 and also we are considering the concept of conditionally independent. So, based on this we have determined the probability Pi and the probability Qi and after determining the probability Pi and Qi what we have determined, we have determined the likelihood ratio and from the likelihood ratio we have determined the discriminate function.

So, the discriminate function is represented in this from the from is the weight is Wi and the bias is W naught. Now, after determining this discriminate function based on this Wi and based on the bias W naught, we can take a classification decision. So, this is about the Bayesian decision making or Bayesian decision theory for discrete features. So, up till now, I discuss the concept of the Bayesian decision theory and that is the fundamental concept of Bayesian decision theory and in this class I discuss the concept of the Bayesian decision theory for discrete features.

So, in my previous classes, I discussed the concept of performance evaluation of a classifier.

So, for this I considered like the confusion matrix. So, how to determine the confusion matrix and from the confusion matrix you can determine the percentage of accuracy, percentage of misclassification and also the rejection percentage. So, all these parameters you can determine from the confusion matrix.

After this I discuss the concept of ROC that is the receiver operating characteristics.

So, you can determine true positive, false positive. So, all these parameters you can determine and these parameters are required for performance evaluation of a classifier. So, in continuation of this, I want to explain it again. Now, let us see how to evaluate the performance of a classifier. So, evaluation of a classifier.

Evaluation of a classifier that is the performance evaluation of a classifier.

So, in my previous discussion that is in the discussion of the probability of error, I have shown how to plot the class conditional density with respect to the feature vector. I am plotting it again. So, and this is my class conditional density, the probability of X given omega i that we are plotting with respect to this X. And suppose I am considering two classes. For the first class, suppose I am drawing the distribution and that is nothing but the Gaussian distribution, the Gaussian distribution for the first class.

So, this is omega 1 for the first class. Similarly, I can consider another distribution that is the Gaussian distribution for the second class. So, the second class is omega 2. So, both are Gaussian distribution and corresponding to the first class, suppose corresponding to this Gaussian, the mean is suppose mu 1 and corresponding to the second Gaussian, the mean is suppose mu 2. Now, I want to determine the performance of the classifier. So, for this suppose I am considering one threshold value of X.

So, this is suppose X star we are considering that threshold we are considering. Now, there may be these cases. The first case is suppose the probability of X greater than the threshold, the threshold is X star. And in this case, we are considering X is assigned to the class omega 2. That means it corresponds to true positive or I can say it is hit.

So, the concept is if X is greater than the threshold, the threshold is X star, then X is assigned to the class, the class is omega 2. So, that means this portion we can consider, there is a true positive. This is the first case. In the second case, what I can consider, the X is greater than the threshold, the threshold value is X star. And in this case, what I am considering, X is assigned to the class omega 1.

So, X is assigned to the class omega 1. That means I have to consider the region, the region I can consider like this. Suppose I can consider this region. So, this region is like this, this is the region. So, that means this region, this region is nothing but false positive.

Or maybe I can say the false alarm. So, corresponding to this case, the second case, the probability we are determining X is greater than the threshold and X is assigned to the class omega 1. And that is nothing but a false alarm or false positive, false positive. Or I can say it is alarm, false alarm. Next, I am considering another condition, the probability of X less than the threshold, the threshold is X star.

And in this case, X is assigned to the class omega 2. So, corresponding to this you can see that is nothing but if I consider this portion, this portion that I can consider as false negative. So, actually the class should be omega 1, but I am considering it as omega 2. So, in this portion that is nothing but the false negative. Or that means I can say it is miss, that is a miss classification. So, actually it should be omega 1, but I am considering X is assigned to the class omega 2.

So, that means I can say it is a false negative. Okay, so finally I am considering another case that X is less than the threshold X star. And in this case, X is assigned to the class omega 1, that class. So, that means it is nothing but the correct rejection, the correct rejection So, I can say another word correct rejection true negative. So, that means this portion I can say, this portion is true negative.

Or I can say rejection. So, you can see I am considering all these four conditions. In the first case, you can see if X is greater than the threshold, X is assigned to the class, the class is omega 2, that is actually the true positive. But in the second case, if you see this case, if the X is greater than the threshold, but X is assigned to the class omega 1, actually it should be omega 2. So, that is why I can say it is a false positive. And similarly, in the third case, if X is less than the threshold, X is assigned to the class omega 2.

So, that means nothing but it is a false negative. Actually, I should consider X should belong to omega 1. So, X should be assigned to the class omega 1. But wrongly I am considering X is assigned to the class omega 2. And finally, what we are considering, if the X is less than the threshold X star, X is assigned to the class omega 1.

And that is nothing but true negative. So, all these parameters we can determine based on these conditions. So, now for performance evaluation, one parameter, that parameter I can consider as Discriminability ratio, that ratio we can consider. That is suppose I am defining like D, D is the discriminability ratio. And that is nothing but the separation between mu 2 and mu 1.

And also I am considering this sigma, sigma of these two Gaussians. So, suppose this is the sigma for this Gaussian. And also I am considering same sigma, the spread of the Gaussian is determined by the sigma, the parameter sigma. So, sigma we are considering. The sigma is same for both the Gaussians. And in this case for this parameter, that is the parameter is the Discriminability ratio, we are considering the separation between two means divided by sigma.

So, that means, if the separation between these two is high, then what I can consider the accuracy will be increase. Otherwise, the misclassification will take place. The separation between these two means.

So, that means I can write high D is desirable. So, a high D is desirable. Because if I maximize the separation between these two Gaussians, then what will happen? My false alarm will be less, the misclassification will be less. But if I consider suppose mu 1 is equal to mu 2, then corresponding to this, discriminability ratio will be zero. So, suppose if I consider mu 1 is equal to mu 2, that means these two Gaussians, the two Gaussians will be like this. This will be overlapping. So, this is one Gaussian and suppose another Gaussian is suppose these two will be overlapping.

These two means will be same mu 1 is equal to mu 2. Then in this case, this is the worst performance of the classifier, then you will be getting the misclassification. So, this is not desirable. So, we have to increase the separation between mu 1 and mu 2. That means, if I increase the separation between the mu 1 and mu 2, this parameter that discriminability ratio that will increase. So, corresponding to D is equal to zero, the performance is very bad for this Bayesian classifier.

So, based on this, I can define one characteristics already you know what is the characteristics that is the receiver operating characteristics based on this discriminability ratio. So, move to the next slide. So, in the discriminability ratio, so, we consider the parameter D, that is nothing but the separation between the two means mu 2 minus mu 1 divided by the sigma, the parameter sigma. And based on this, we can consider receiver operating characteristics, receiver operating characteristics is nothing but ROC. So, what is the receiver operating characteristics? So, I am plotting that is I am plotting between what that is true positive, true positive, true positive means hit and false alarm, that is the false positive, false positive or I can say the false alarm I am plotting.

So, corresponding to this discriminability ratio. So, if I consider suppose D is equal to zero. So, I will be getting the curve or something like this, this is for D is equal to zero, the discriminability ratio zero. If I increase the discriminability, this ratio if I increase, so, this is the curve, suppose the corresponding to D is equal to one. And suppose I can see if I increase the separation between the means of these two Gaussians.

So, this is the curve corresponding to D is equal to two. And like this, if I increase the separation between these two means, then I will be getting the ROC curve corresponding to D is equal to three. So, that means I am increasing the separation between these two means. So, in this case, what we are considering, suppose we are varying the threshold, we are varying the threshold x star. So, what will happen, the true positive, that is the true positive probabilities and false positive probabilities will vary with respect to the threshold x. So, that means based on the threshold, what I can say that is the true positive, true positive and false positive will vary with the threshold x star.

So, you can see, if I vary the threshold x, you can control the true positive and the false positive, because the true positive and the false positive depends on x. So, in my previous slide, I have shown, so this is the threshold, if you see this is the threshold. So, based on this threshold, I can adjust this true positive, true positive means the hit probabilities and also the false positive, false positive means alarm probabilities, alarm probabilities I can sense that depends on the threshold x star. So, here you can see based on the discriminability ratio, I am plotting the ROC curve.

So, this curve is nothing but the ROC that is the receiver operating characteristics. So, this concept already I have explained in my previous classes, but in this case, what I am considering, I am considering the bayes decision theory to explain this concept. So, how to determine the performance of a classifier. Now, after this, I am discussing that the concept of the Bayesian decision surfaces, that is what is the decision surface between two classes or maybe the more classes, that concept I am going to explain. So, before explaining this, I want to explain the concept of the normal and the Gaussian distribution.

So, what is normal and Gaussian distribution? So, let us see what is the normal distribution. So, normal distribution, this density, I can write like this 1 by twice pi sigma square. So, you know about this normal density, this expression for the normal density, 1 by 2 x minus mu, that is the mean and the variance also we are considering. So, x is a random variable and it follows a normal distribution. So, your normal distribution, you know, a normal distribution is something like this.

This is a normal distribution and this is the mean, mean of this distribution. This is the mean of the distribution. I am not going to explain what is the normal distribution. I think you know this one. Now, this x we are considering as a random variable.

So, x is a random variable, x is a random variable. Now, the expected value of x, what is the expected value of x, the expected value of x of this random variable is nothing but minus infinity to infinity x px dx and that is nothing but the mean, the mean of the normal distribution. And what is the variance of x? Variance of x, x is a random variable. So, variance of x is nothing but expected value of x minus mu, I can write like this. So, that is nothing but minus infinity to plus infinity, x minus mu whole square, this density and dx and that is nothing but sigma square.

So, that is the variance, variance of the normal distribution. Okay. So, this x we are considering as a random variable. This concept I think you know, because already you have studied the course on probability and random process. So, now what is multivariate Gaussian distribution? So, move to the next slide. So, what is the multivariate Gaussian distribution? Multivariate Gaussian distribution.

So, previously I considered only the univariate Gaussian distribution. Now, what is the multivariate Gaussian distribution? So, now suppose x is a vector and suppose these are the components of the vector x 2, or these are the elements of the vector and this is a d dimensional vector. Now, this density, there is a multivariate Gaussian density, I can write like this twice pi d by 2, d is the dimension of the vector x. This sigma is nothing but it is called a covariance matrix and I am taking the determinant of the covariance matrix. So, it is 1 by 2 x minus mu transpose.

This is sigma inverse x minus mu, mu is also a vector, is a mean vector. So, in this case, the mean vector is nothing but the expected value of x, x is a vector. So, what is the expected value of x? That is nothing but expected value of x 1, expected value of x 2, like this, the expected value of x d, because we are considering the d dimensional vector, the vector is x. So, corresponding to this, I have the mean mu 1, mu 2, like this, mu d. So, this is the mean vector, the mean vector we can determine like this.

And this is nothing but, this is the covariance matrix. So, it is the d cross d covariance matrix, covariance matrix. These are d cross d covariance matrix. So, let us move to the next slide. What is this covariance matrix? So, this covariance matrix, this is the sigma, this is actually the square matrix. And for the square matrix, i z element is sigma i j, i th, j th element is the sigma i a, and this is nothing but the covariance of x i and x j.

So, we are considering this sigma i j, sigma i j is nothing but the covariance between x i and x j. So, what is mathematically the sigma i j, that is the covariance between x i and x j, that is nothing but the expected value, we are considering, x i minus mu i, and x j minus mu j. So, this covariance matrix I can write like this. And in this case, i n j, it is from 1, 2 up to d, because we have considered a d dimensional vector x. So, corresponding to this, this sigma, the covariance matrix, I can write like this, this expected value x 1 minus mu 1, and x 1 minus mu 2.

So, this is the first element of the covariance matrix. What is the second element, the second element is expected value x 1 minus mu 1, and x 2 minus mu 2. Like this, if I move to this, so this is the last one is expected value of x 1 mu 1 into x d minus mu d. So, if I go to the second row, so first element of this matrix is expected value of x 2 minus mu 2, and x 1 minus mu 1. What is the second element, the second element is expected value of x 2 minus mu 2 into x 2 minus mu 2.

That is the second element in the second row. And finally, what is the last element, the last element is x 2 minus mu 2 x d minus mu d. So, this is the last element. And like this, I can move and what is the final, finally, I am getting this element in the last row, that is x d minus mu d into x 1 minus mu 1 expected value x d minus mu d into x 2 minus mu 2. And finally, the last element of this matrix is x d minus mu d x d minus mu d.

And this is the last element of this matrix. So, this is the matrix, I am getting that is the covariance matrix. So, that I can write like this. So, if I see the sigma 1 1, sigma 1 2, like this, the sigma 1 d. So, sigma 2 1, sigma 2 2, like this sigma 2 d. And the last will be sigma d 1, sigma d 2, sigma d d. So, that I can write like this, sigma 1 square, sigma 1 2, sigma 1 d, and sigma 2 1, sigma 2 whole square, sigma 2 d, sigma d 1, sigma d 2, and sigma d whole square.

So, this is the expression for the covariance matrix. So, let us move to the next slide. So, for the multivariate, this multivariate Gaussian density is represented like this, it is n that is the normal distribution. So, I have the mean vector and the another one is the covariance matrix. So, these two parameters, one is the mean vector, another one is the covariance matrix. And this sigma inverse that is nothing but I am taking the inverse of the covariance matrix.

So, this is the inverse of the covariance matrix, the inverse. And what is this, this is nothing but the determinant of the covariance matrix. So, this expression for density already I have shown, the density expression is 1 by twice pi d by 2, because we are considering the d dimensional Feature vector 1 by 2 exponential minus 1 by 2 x minus mu transpose, the mean and the inverse of the covariance matrix x minus mu. And this is the expression for the multivariate Gaussian density. And if I consider d is equal to 1, suppose, in the previous case, we are considering the d dimensional Gaussian density, d dimensional vector, the vector is x, suppose d is equal to 1, then this multidimensional Gaussian, it is converted into the univariate Gaussian density. So, this if I consider d is equal to 1, this multivariate Gaussian density is converted into the univariate Gaussian density.

So, univariate Gaussian density already I told you know it is twice pi sigma exponential minus 1 by 2 x minus mu sigma square. So, this is the expression for the univariate Gaussian density. So, this is the univariate Gaussian density and this is the multivariate Gaussian density. So, this is the univariate density, normal density.

So, it has two parameters, one is the mean and another one is the variance. In case of the multivariate Gaussian density, I have two parameters, the parameters are mean vector and the covariance matrix. In case of the univariate density, I have two parameters, one is the mean another one is the variance. Let us see how to draw this Gaussian.

Suppose, I am plotting this one, this is the density with respect to x. So, I am considering two Gaussians. So, first Gaussian is suppose something like this and second Gaussian suppose something like the flat and the mean of these two. So, mean of these two is suppose the same mean, the mean is mu. So, for the first Gaussian, for the first Gaussian, this Gaussian the variance is sigma 1 square and for the second Gaussian, the variance is sigma 2 square. So, in this case, this variance actually controls the spread of the Gaussian.

So, that means in this case, sigma 1 square is greater than sigma 2 square. So, this spread of the Gaussian is controlled by the parameter, the parameter is the variance. So, in this case, the sigma 1 square is greater than sigma 2 square. In my last slide, I have shown that what is the expression for the covariance matrix, if I consider the multivariate Gaussian, if I consider the multivariate Gaussian, the expression for the covariance matrix is sigma 1 square sigma 1 2 sigma 1 d. In my previous slide, I have shown like this sigma 2 1 sigma 2 whole square sigma 2 d and finally, sigma d 1 sigma d 2 sigma d whole square.

So, this is the expression for the covariance matrix. So, if I see here, this diagonal elements, diagonal elements, sigma ij, that is the variance, these are nothing but the variances of respective, respective xi. So, that is actually I can write sigma i whole square. So, diagonal elements of this matrix, the matrix is the covariance matrix.

So, these are the variance of respective xi of respective xi. So, that is actually I can write sigma i whole square. So, diagonal elements elements, these are the off diagonal elements. xi. So, these are the off diagonal elements. So, off diagonal elements, elements, these off diagonal elements are nothing but the covariances, off diagonal elements are sigma ij and that is nothing but the covariance, covariances of xi and xj, respective xi. And suppose if I consider the sigma ij, sigma ij is equal to 0, what is the meaning of this? If I consider sigma ij is equal to 0, then xi and xj are statistically independent.

So, I can write xi and xj are statistically independent. So, this covariance matrix is quite important. So, I want to repeat this one, if I consider the multivariate Gaussian density, then I have two parameters, one is the mean vector, another one is the covariance matrix. And if I consider d is equal to 1, that is the dimension of the vector x is equal to 1, this multivariate Gaussian density is converted into the univariate Gaussian density. And corresponding to this univariate Gaussian density, I have two parameters, one is the mean, another one is the variance, you have seen here. And after this, I have shown the expression for the covariance matrix. And what are the diagonal elements? The diagonal elements are sigma ij, that is nothing but sigma i square, that is the variance of respective xi.

And if I consider the off diagonal elements of covariance matrix, that is the covariance of xi and xj. And if I consider sigma ij is equal to 0, that is the condition, that means xi and xj are statistically independent. So, this is the case. In the summary of this, what we have considered that in the multivariate density already you know, and this is the expression for the multivariate density. So, that is the 1 by twice pi d by 2, this is the covariance matrix and exponential 1 by 2 x minus mu transpose sigma inverse x minus mu.

So, this is the expression for the multivariate Gaussian density already I have explained. Now, let us define one distance. So, distance is R square. So, distance is defined like this x minus mu transpose that inverse of the covariance matrix x minus mu. So, this is a very popular distance in machine learning.

This distance is called, it is squared Mahalanobis distance. So, this is a very popular distance, the squared Mahalanobis distance. We can take also squared root then I will be getting the simple Mahalanobis distance. Professor Mahalanobis is from ISI Kolkata. So, he is a very famous statistician. So, he formulated this distance, the distance is the Mahalanobis distance. Suppose if I consider the distance from x to mu, so my vector is x and distance from x to mu, you can determine with the help of this Mahalanobis distance.

And what is the actually the Euclidean distance you know what is the Euclidean distance or Euclidean norm. Euclidean distance you know already the distance between x and mu that is the Euclidean norm x minus mu. So, this is the Euclidean distance between x and mu. So, in case of the multivariate distribution, the normal distribution, suppose if I have some clusters, some of the samples are available, these are the samples and suppose I have another clusters. So, this cluster corresponding to the class omega 1, this cluster corresponding to the class omega 2, two clusters.

This is the center of the cluster, two clusters. The center of the cluster is determined by the mean vector. So, this is the mean vector is suppose mu 1 and another one is mu 2. So, I am considering two clusters and I am considering these are the samples corresponding to the first cluster and that corresponds to the class omega 1 and corresponding to the second cluster I am showing the samples these are the samples of the second class, the class is omega 2. The center of the cluster is determined by the mean vector and shape of this cluster the shape of the cluster may be something like this or maybe like this. So, shape of these clusters are determined by the covariance matrix. So, I am repeating this if I consider suppose these clusters, the clusters corresponding to different classes, then the center of the cluster is determined by the mean vector and shape of the cluster is determined by the covariance matrix.

So, that is the case. So, my shape of the clusters may be like this, these are the shape of the clusters or maybe that these or maybe that these so these type of shapes we can consider or maybe the circular shape. So, like this we can consider the shape of the clusters and that is determined by the covariance matrix and the center of the cluster is nothing but the mean vector. So, this is the concept of the normal distribution. One is the univariate normal distribution and another one is the multivariate normal distribution. So, in this class, I discussed the concept of Bayesian decision theory for discrete features and after this I discussed the concept of the performance evaluation of a classifier.

So, based on these parameters, one is the true positive false positive true negative. So, for all these parameters, how we can determine that discriminability ratio that is nothing but the separation between two means of two Gaussian, the two Gaussians corresponding to the class conditional density. If I consider two classes, the class is omega 1 and omega 2. So, corresponding to these two classes, suppose if I consider the distribution is the Gaussian distribution. So, this discriminability ratio can be defined like this the separation between these two mean divided by sigma the parameter sigma of the Gaussian distribution.

And after this, I considered the ROC the receiver operating characteristics. And finally, I discussed the concept of normal distribution. So, one is the univariate distribution and another one is the multivariate distribution. So, in my next class, I will be discussing the concept of the Bayesian decision theory, the same thing I will discuss, but the main concept I will be discussing that concept is how to determine the decision boundary between the classes. So, suppose I have two classes. So, what will be the nature of the decision boundary? Suppose if I consider multiple classes, so what will be the decision boundary between the classes, whether it is a plane or whether it is a straight line, whether it is a ellipse.

So, like this, we have to decide that is the decision boundary between the classes. So, in my next class, I will discuss all these concepts. So, let me stop here today. Thank you.