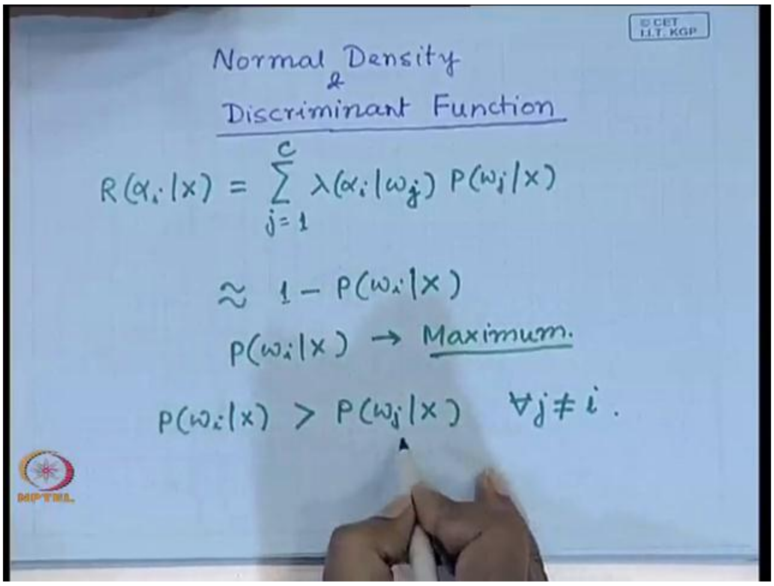
Pattern Recognition and Applications
Prof. P. K. Biswas
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 7
Normal Density and Discriminant Function

Good morning. So, our today's topic of discussion will be normal density and discriminant function. So, in the last class, we have talked about different types of classifiers, which were actually derived from Bayes classifier called Bayes minimum risk classifier. In Bayes minimum risk classifier, we said that I have to have a number of functional units, which will compute the risk involved for taking any particular action.

(Refer Slide Time: 00:58)

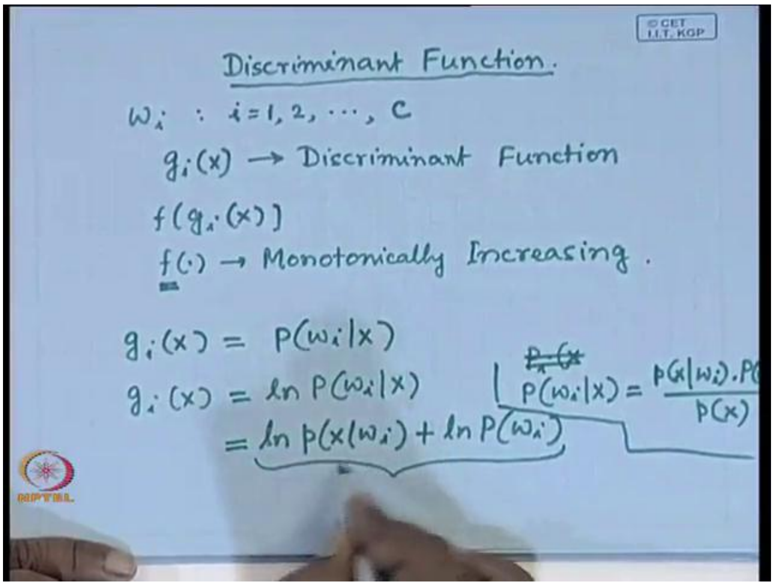


So, the risk function that we said was of this form R of alpha i given x that means given of which a feature vector x, the risk involved in taking an action alpha i which was of this form, lambda alpha i given omega j into P omega j given x. Take the summation for j is equal to 1 to c as c is the total number of classes. So, for each action alpha i, when I have say, P number of actions in my system, then there will be P number of such risk functions computed and that corresponding action alpha will be taken for which this risk function is minimum. So, that is what Bayes minimum risk classifier is.

Then, we have said that we can derive another classifier, which is minimum error rate classifier from this Bayes minimum risk classifier. This risk function, we have said that it can be reduced in the form of 1 minus P omega i given x. So, when this is minimum, the corresponding we have to put x in the corresponding class omega i. As this has to be minimum, correspondingly what we get is this probability P omega i given x that has to be maximum. So, this is nothing but the basic classification rule that we discussed about when I started talking about this decision theory that if P of omega i given x is greater than P of omega j given x for all j not equal to i.

So, this is the posterior probability that given of feature vector x, what is the probability that x will belong to class omega i. This is what is the probability that x will belong to class omega j. So, for that particular omega i where P of omega i given x is maximum, we have to put or we have to classify x as belonging to that particular class. Then we have talked about the discriminant functions.

(Refer Slide Time: 03:50)



So, there what we said is for every class i or every omega i for i varying from 1 to c, as we have considering a generalised case that as if we have c number of classes, so for every class, we define a function say g ix. So, this g ix for a feature vector x, will be computed for every ith class and for whichever value of i, the g ix is maximum, we put x or classify x to that particular class omega i. So, this is what we are calling as discriminant function. We have also said that discriminant function g ix is not really unique.

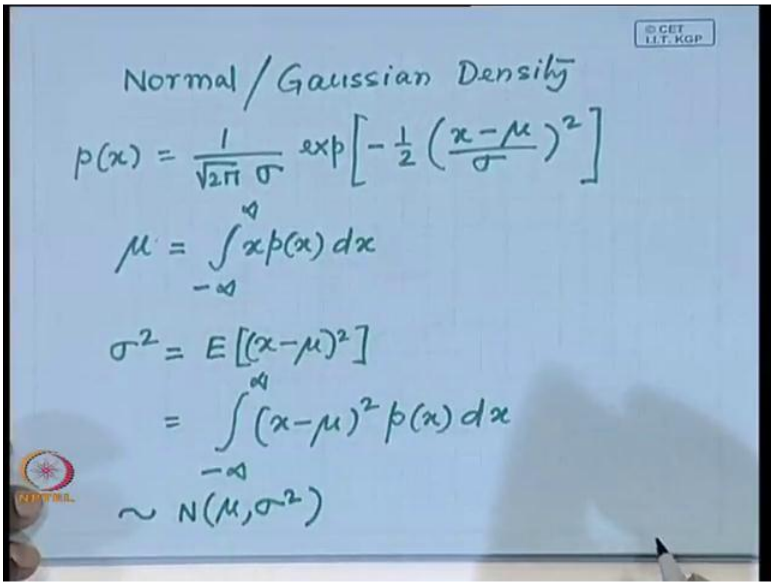
So, if I can identify any function f, which is monotonically increasing, then the functional f on g ix, where this function f this is a monotonically increasing function. So, whenever this f is monotonically increasing, then f of g ix that also serves the same purpose as discriminating function because for whichever value of i, g ix is maximum, it is for the same value of i, f of g ix will also be maximum. Then why do we put this functional f? It is only for convenience of replication in different applications may be g ix itself is not very convenient, but if take a function of g ix, then that will be more convenient for our application.

So, coming to this particular case of Bayes minimum risk classifier or minimum error rate classifier, while we have said that our g ix can be the posterior probability P of omega i given x because as per our minimum error rate classification, it was 1 minus P i P omega i given x. This is because whenever this 1 minus P omega i has to be minimum, P omega i given x has to be maximum. So, I have to put this x into the class omega i for which P omega xi is maximum. So, this itself, I can take as my discriminant function g ix.

Then, we said that logarithm being a monotonically increasing function, so I can also define g ix as log of P of omega i given x and because P of omega i given x, this is nothing but p of x given omega i into P of omega i upon p of x. But, we said that P of x, we can remove because for every value of i, P ix will always appear in the denominator. So, that does not give you any discriminating point. So, I simply give this, p of x given omega i into P of omega i, so there this function will be expanded to l n log of p x given omega i plus l n P of omega i. So, this becomes our discriminant function to be used for classification of an unknown sample x into one of the classes or one of omega i.

So, naturally here you find because the probability density function is involved or the accurate probability is involved, so the structure of this basian classifier will depend upon what kind of probability density that we are making use of. So, we can have various types of probability densities, we can have normal density, we can have laplacian density, we can have exponential density and so on, we can have Poisson density. So, depending upon what kind of density we make use of for a particular application, what kind of probability density function is more appropriate according to that, our structure of the classifier will be different.

(Refer Slide Time: 09:27)

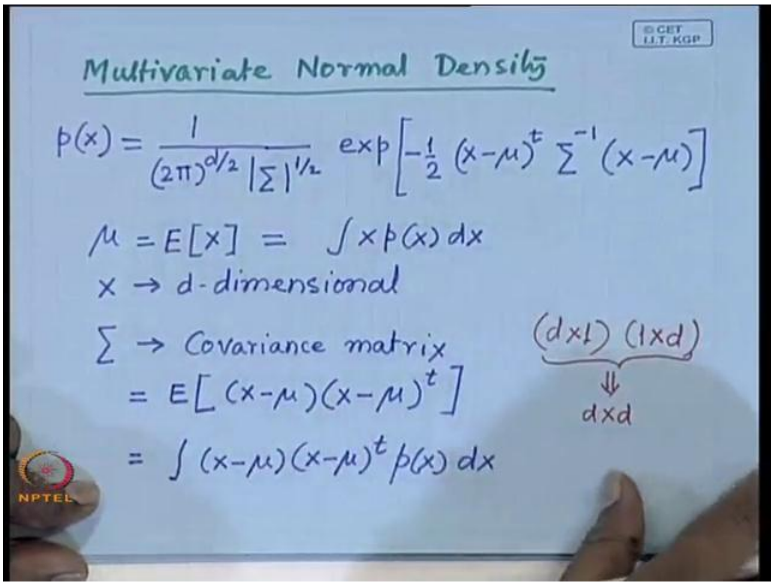


So, the most common probability destiny function, which is in use, is the normal or Gaussian density. So, this is the most common probability density function. So, this normal and Gaussian density for a single variable is simply given as p of x is equal to is, all of you know this expression square root of 2 pi into sigma exponential minus half into x minus mu upon sigma square. So, this is the extension for normal or Gaussian density or the mu. Mu is nothing but the expected value of x or mean value of x, which is simply given as integration of x p x d x.

The integral has to be taken over the limit minus infinity to plus infinity and sigma square, which is the variance or sigma, which is the standard deviation is simply given by the expected value of x minus mu square. It is nothing but integration x minus mu square p x d x. Again, you take the integral from minus infinity to plus infinity. So, this is what the expression of the normal or Gaussian density is in a single variable case. You find that this particular probability density function is specified only by two parameters. One is the mean value mu and the other one is the standard deviation sigma or the variance sigma square.

So, if I know only the mean value or the variance, then I know what this probability density function is. So, in short, this p d f, normal p d f is also written as N mu sigma square. So, this means that it is normal density with the mean value of the variable as mu and variance of the signal as sigma square, but we are talking about feature vectors that mean multiple numbers of components or multiple numbers of features. So, normal density of single variable is not much useful for us, but what is useful for us is multivariate normal. So, let us see what that multivariate normal density is.

(Refer Slide Time: 12:56)



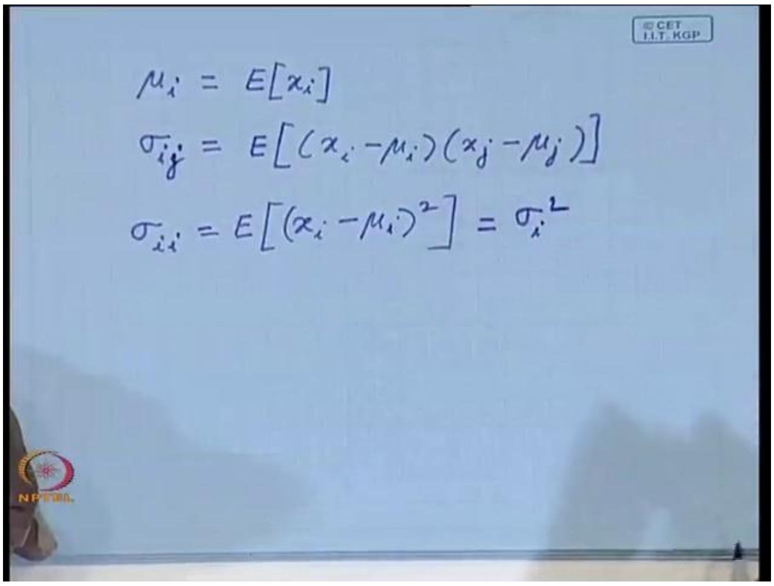
So, here instead of a single feature x, we have a feature vector and the feature vector is again, we use it in lower case x, but instead of writing it as x, we write it as capital X. So, this P X in this case for a multivariate case will be simply given by 2 pi to the power of d by 2. So, this is the expression for the multivariate probability density function, where X is the feature vector, a variable representing the feature vector. Mu is the mean vector. So, mu is nothing but as before, the expected value of the feature vectors X, where X is a d dimensional vector.

This is because we have assumed that our feature vector is d dimensional that is d number of individual features are concatenated together to give you the feature vector. So, X is a d dimensional feature vector. So, accordingly the mean vector mu vector will also be of dimension d. So, this 2 pi to the power d by 2, this d is nothing but the dimension of the feature vector and this sigma is what is called the covariance matrix. So, as mu is the expectation value of X, so as before as in case of single variable case, we can write this as X p X d X, integral of X p X d X and this covariance matrix because this is the expectation value of the covariance of the different components.

So, this sigma covariance matrix, it is nothing but the expectation value of X minus mu into X minus mu transpose. So, note carefully that it is not X minus mu transpose into X minus mu in which case you get a scalar because X is a d dimensional vector, mu is also a d dimensional vector. So, if make it X minus mu X minus mu transpose that becomes a scalar quantity or a dot product of two vectors, rather it is X minus mu into X minus mu transpose. So, this becomes a d dimensional vector. So, X minus mu is of dimension d by 1 because it is a column vector and X minus mu is a vector or a row vector of dimension 1.

So, this is actually the outer product of two vectors. When I take the outer product of two vectors, the result is a d by d dimensional matrix. So, when I take the expectation value of X minus mu into X minus mu transpose, what I get is a d by d dimensional matrix. That is what is nothing but your covariance matrix or expectation value of this is becomes the covariance matrix. So, as before, this can also be written as in the integral form X minus mu into X minus mu transpose p X d X. Take the integral of this over the limit minus infinity to infinity. Then what I get is this covariance matrix sigma. Now, from here, if I try to compute what are the expected values of individual components?

(Refer Slide Time: 18:30)

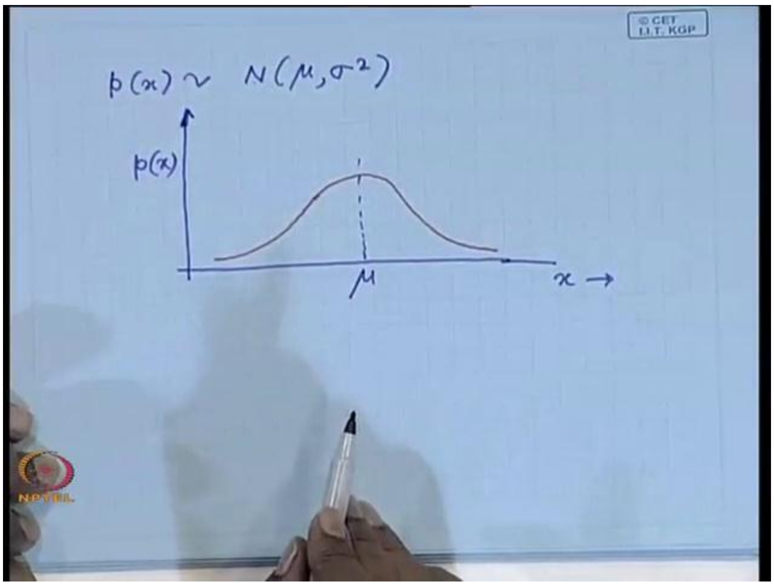


So, the expected value of the ith component mu i which is nothing but the expected value of xi which is the ith component of a feature vector capital X. Similarly, sigma ii which is the iith component in my covariance matrix sigma, capital sigma, sigma ii, this will simply become expected value of xi minus mu i into, sorry i jth component. Let us make it more general sigma ij ijth component in my covariance matrix that will simply become expected value of xi minus mu i into xj minus mu j.

So, you find that what the variance is when I take these two different components xi and xj that is what it gives you the covariance between the ith component and jth component. When I compute this sigma ij which is i jth component in my covariance matrix that is given by this expression. So, obviously if I make i is equal to j that is sigma ii which are nothing but diagonal components, the components on the diagonal of the covariance matrix that simply becomes expected value, expectation value of xi minus mu i square. This is nothing but our sigma i square which is the variance of the ith component of feature vector.

So, when I talk about this covariance matrix, the diagonal elements in the covariance matrix actually give you the variance of the individual components of the feature vector. The off elements, i jth elements actually give you the covariance when I consider the ith component and the jth component of the feature vector together. So, the covariance matrix is a more general form of the variance that we usually use in case of a single variable matrix. Now, what does this multivariate normal density or the multivariate Gaussian density actually tell you? Let us consider a case of the single variable normal density.

(Refer Slide Time: 21:32)

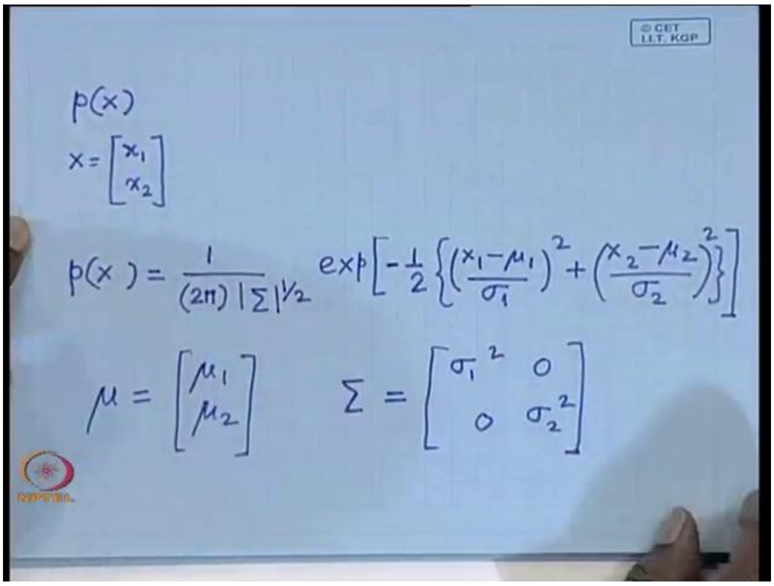


So, I have p of x which is given by the normal density, this with mean value mu and the variances sigma square. So, this simply means if I plot this normal density, plot of the normal density will be like this where the peak is, so I put x along the horizontal direction and p x along the vertical direction.

So, when I have this single variable normal density, this simply tells us that if you take the samples from the same population, then how those samples are going to be placed? How those samples are actually distributed? So, this distribution shows that most ofm maximum of the samples will be around this mean value mu. The other sample values will be distributed according to this. As you go away from the mean value, the population density will go on decreasing at a certain limit after which I can actually neglect the population density, which is given by say plus minus 2 sigma that is the twice of the standard deviation on this side as well as this side.

If I go beyond that, I can actually neglect the population density. So, this is what is meant by this normal density function. Now, it is interpretation in the multivariate density case. Before I go to this d dimensional case, let me just see what will happen in a 2 dimensional case that is bivariate normal density that possibly you were asking in the last class yesterday.

(Refer Slide Time: 23:56)



So, suppose I talk about only two variables x and y or I talk about the normal density of a vector x. This vector x has only two components, x1 and x2. So, this is a bivariate normal density. More generalization of this is multivariate normal density. Now, coming back the same expression, we will find that in such bivariate normal density, this expression as we have written in case of multivariate normal density that is this, this expression can be simplified for a bivariate normal density. It is not simplified.

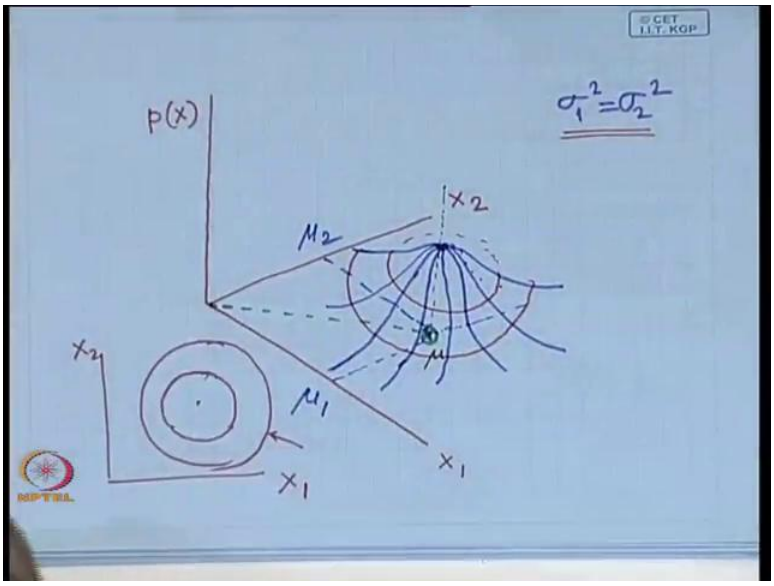
If I simply expand this exponential term assuming that this x is on the two components x1 and x 2, so by considering that, this p x when x is two dimensional will simply be written as 1 over 2 pi. Now, value of d is equal to 2 becomes 1, so 1 over 2 pi into sigma to the power half. When I say sigma within modulo form, it is nothing but determinant of the covariance matrix. So, this modulo means it is determinant of the covariance matrix because I cannot take square root of a matrix.

However, I can take square root of a determinant because determinant has a value, matrix does not have a value. Matrix has to be interpreted. So, this into exponent of again minus half, then this term x minus mu sigma inverse into x minus mu x minus mu transpose into sigma inverse into x minus mu, this can be written in the form. Now, I have two components. One is x1 and other is x2. So, x1 minus mu 1 upon sigma 1 whole square plus x2 minus mu 2 upon sigma 2 whole square, this is the simplified expression because I had assumed that the components x1 and x2 were statistically independent.

So, that is why, I could have a simplified expression only in terms of sigma 1 and sigma 2 only the diagonal elements in a covariance matrix. It is only because my assumption that x1 and x2 are statistically independent. So, sigma 12 or sigma 21 will be equal to 0 because x1 and x2 are statistically independent. If they are not statistically independent, then in this exponential term, I will also have terms corresponding to sigma 12 and sigma 21. So, for explanation, let me simplify. Let me have a simplified assumption that if x1 and x2, they are statistically independent.

However, x1 component has a mean value of mu 1 and x 2 component has a mean value of mu 2, but hence my mean vector mu is nothing but mu 1 mu 2. The covariance matrix sigma is nothing but sigma 1 square sigma 2 square 0 0. So, this is my covariance matrix and this is my mean vector. So, what is the physical interpretation of this particular bivariate normal density function? Let us draw this two dimensional space or three dimensional space because x1 and x2 are the axis of the sample points and the density function is p x, which is of third dimensions.

(Refer Slide Time: 28:45)



So, let me draw this third dimensional space. So, I put x1 along this direction, x2 along this direction and p x along this direction. Suppose mu 1 is of the mean vector, mu is somewhere over here. So, if mu is at this position, so that means the components, this is the location of mu 1 and this is the location of mu 2. Now, find that if I break this expression, this term exponential of minus half x1 minus mu 1 by sigma 1 square plus x 2 minus mu 2 by sigma 2 square, this is actually product of two exponential terms exponential minus half x1 minus mu 1 by sigma 1 square into exponential minus half x2 minus mu 2 by sigma 2 square.

So, when I consider forgetting this part, exponential minus of x1 minus mu 1 by sigma 1 square, this is actually normal density along its direction with a mean value at mu 1 and the variance sigma 1. When I consider only this component exponential minus half x 2 minus mu 2 by sigma 2 square, this becomes the normal density along x2 direction with mean value mu 2 and standard deviation sigma 2. So, this over all exponential term is actually a product of two normal density functions of two question density functions.

So, if I plot this in these two dimensions, I will have something like this. I will have a peak of the density around this mean value mu, mean position mu and the densities will be, I will have Gaussian along this. I will also have Gaussian along this. If it is a single Gaussian, if I assume that my sigma 1, let me take a simplified case that sigma 1 square is equal to sigma 2 square that means the variance along x1 and the variance along x2

Gaussian. So, the kind of shape that I will have over here is if I have a single normal density, so that singe normal density profile, you rotate around the vertical axis.

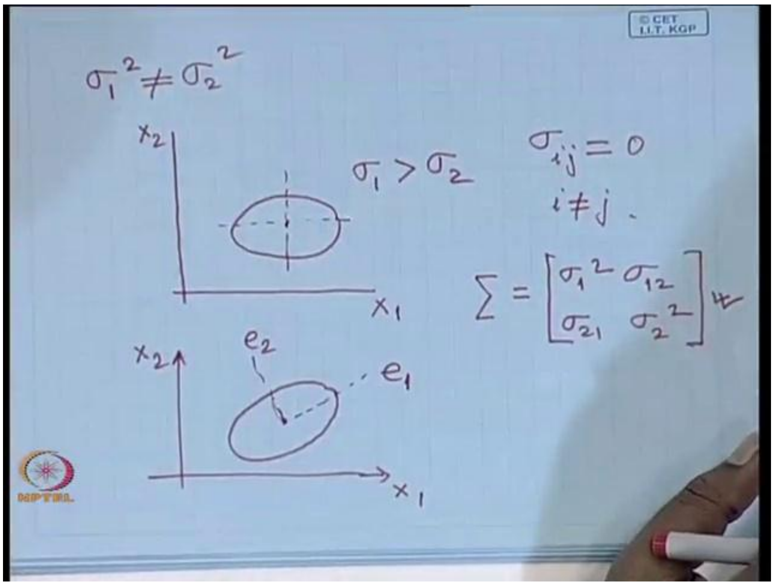
So, whatever surface it will chase; that surface will be the surface of this probability density function p x. So, if I plot over this thing, the surface will be something like this. It will go on this side. It will come like this or this form. Now, given this surface of this bivariate probability density function, if I now try to chase the loci of points of constant density that means all those values of x for which p x is constant. So, those loci will be nothing but the circles, something like this because along this value of p x will be constant.

So, if I take the footprint of this loci on this x1 x2 plane, so on x1 x2 plane, I will have a number of concentric circles like this or these circles. Each of these circles represents the loci of points having constant density. As I move inside the circle, move towards the centre of the circle, these loci of the density value of the density will increase. As I move away from the centre, the value of density function will go on decreasing. So, that indicates that along the circle, I have more probability of occurrence of the points, which are drawn from a single population arbitrate or drawing of one sample does not depend on the drawing of another sample.

What does it mean? When I simply take up the first object, I do not make use of any information of what was the previous object that I have taken. So, just blindly go on taking samples from a population and then the points will be distributed in this form.

So, this is a simple case that I have taken when sigma 1 square is equal to sigma 2 square. That means the variance along x1 direction and the variance along x2 dimensions are the same. What happens if the variances are different? In that case, this circle will become variance.

(Refer Slide Time: 35:21)



So, this circle that I have shown and I have say sigma 1 square is not equal to sigma 2 square, so one will be greater than the other. So, when sigma 1 square is equal to sigma 2 square, I had the footprints of the loci of the constant density as circles. Now, the loci of constant density will be ellipses. How these ellipses will be formed? It will be something like this. So, this is a case when sigma 1 is greater than sigma 2, but hence the spade of points along x1 axis is more than the spade of points along y1 axis.

If sigma 2 becomes greater than sigma 1 that is spade of points along y2 axis is more than spade of points along x2 axis is more than spade of points along x1 axis, then I will have the major axis aligned towards x1 direction and the minor axis aligned towards x2, major axis along x2 direction and the minor axis along x1 direction.

So, this is the nature of the loci of the points of constant density when my assumption was that different components of the feature vectors are statistically independent that means sigma ij was equal to 0 for i not equal to 0. What happens if sigma ij is not equal to 0? That means the samples are not or the components of the feature vector are not statistically independent. So, over there, I will have my covariance matrix sigma. In this case, co variance matrix is used where diagonal matrix or only diagonal elements are non zero elements. All the off diagonal elements are 0. If the components are not statistically independent, then even the off diagonal elements will also be non zero.

So, I will have the situation that sigma 1 square sigma 2 square, here I will have sigma 21, here I will have sigma 12 where had sigma 12 and sigma 21 are non zero values. So, if it so happens here, you find that the major axis and the minor axis of these ellipses they are aligned along x1 dimension and x2 dimension. In this case, when the components are not statistically independent, then the direction of the major axis and minor axis of the ellipse will be given by Eigen vectors of this covariance matrix. The links of the major axis and minor axis will correspond to the Eigen values of the covariance matrix.

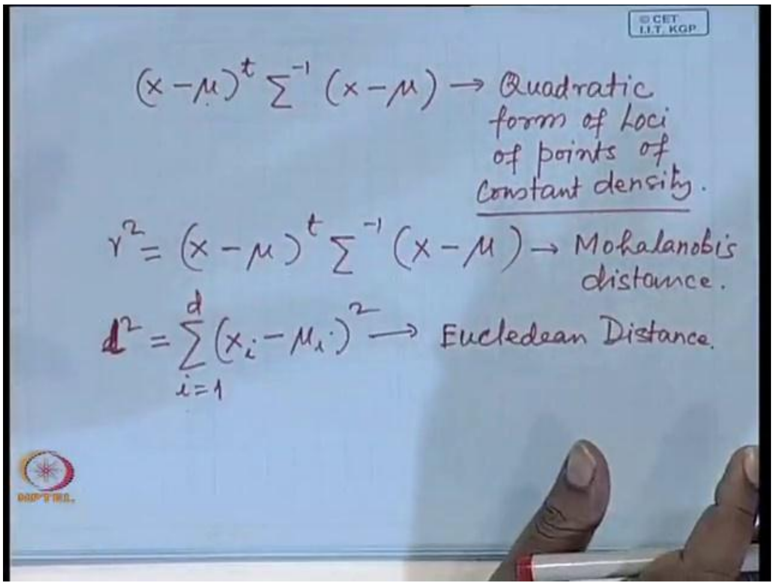
So, instead of having a footprint of the point, of the loci points of constant density like this, the footprints will be something like this. Again, it will be centred around mu, but the directions of major axis and minor axis will be given by the Eigen vectors, this covariance matrix say e1 and e2. So, this is my x1 direction this is my x2 direction and as I said that these directions actually tell you how the points are distributed. I will have maximum density along the, around mean. As I go along this direction, the density will go on reducing. As I move along this direction also, the density will also go on reducing, but here, the rate of decrease of the density function will be more, here the rate of decrease of the function will be less.

Now, what happens in case of mutivariate normal density? I have taken the example of bivariate normal density, but because still up to two variables, I can visualize the movement. See; consider the case of two variables and three variables. So, I will have three axes, one corresponding to x1, one corresponding to x2 and other corresponding to x3. I have to have a fourth axis which corresponds to p x, the probability density function. I cannot draw it on a two dimensional plane, up to two dimensional, I can easily draw and it goes to three dimensional plane, still I can draw, but some difficulty. The moment it goes beyond three, I cannot draw it on a two dimensional plane, but I can think of how the nature can be. So, when it becomes more than two variables that is when multivariate of normal density for a multi variable multivariate Gaussian density, again if I draw the sample of points from that multi dimensional space, the points will form clouds. Here also points are forming clouds or clusters of points, but the maximum density will be around mu 1 or around mu.

As I move away from mu, the density of the points goes on reducing. When I go for this multivariate normal density, then also I can say that those multi dimensional points will form point clouds in a multi dimensional space in such a way that again around mu, the density will be maximum. As I move away from mu, the density will go on reducing and they will be clustered here. This clustering is in the form of an ellipse. In a multi dimensional space, it will be an ellipsoid. So, the points will clustered in an ellipsoid. So, again in three dimensional spaces, if I simply rotate this ellipse along a particular axis, I will have volume.

So, the points will be clustered in that volume. At that centre of the volume, the ellipsoid volume, the density of the point will be maximum. As I move away from ellipsoid from the volume, the density will go on reducing again following that normal density function or Gaussian density function. So, over here as in this case, the loci of points of constant density forms ellipses in multi dimensional space, the loci of points of constant density form ellipsoids.

(Refer Slide Time: 43:05)



This ellipsoid will have a quadratic form, which is given by x minus mu transpose sigma inverse into x minus mu. So, this is the quadratic form of loci of points of constant density. The principle axis of this ellipsoid as in this case, the principle axes of the ellipses are given by the Eigen vectors of the covariance matrix. So, in the same way, the principle axis of the ellipsoids in this case will be given by Eigen vectors of the covariance matrix. The links of the axis will be given by the corresponding Eigen values.

This term, the distance function actually, this locus is nothing but the value of the distance values from the centre.

So, mu gives you the centre of the centroid and x is the points lying on the peripheral. This term actually gives you the distance of x from mu. So, distance function of square, I can define which is given by x minus mu transpose sigma inverse into x minus mu. In previous case, I can define d square is equal to xi minus mu i square, take the summation where i is equal to instead of calling d, let me call it say l here, i is equal to 1 to d because we are assuming vectors d dimensional.

So, this l square is equal to x square minus xi minus mu i square. Take the summation from i equal to 1 to d. This is the Euclidean distance. This distance function if I define as x minus x minus mu transpose sigma inverse x minus mu, this is what is called mahalanobis distance, named after great statistician pc mahalanobis. So, this is what is called mahalanobis distance.

Student: Sir, it is a square or its root should be...

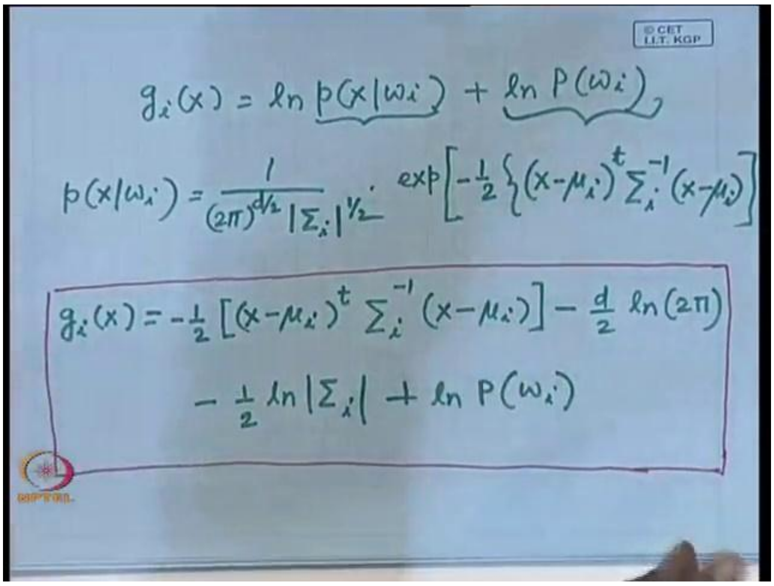
No square, it is a quadratic term x minus mu transpose minus x minus mu.

Student: Sir, I am talking about is r square or r.

The distance is r. The expression is given by this. That is quiet obvious.

Now, let us see that how this can be utilized. So, so far what I have discussed is about the probability density functions, univariate case, bivariate case, multi variate case. I have not gone beyond that, but we started with how these probability functions actually influence the structure of the decision surface. We started with that because g ix was l n p of x given omega i plus l n p omega i probability density and the conditional density function. So, we started with this. So, our purpose is to find out g ix, which is the discriminant function.

(Refer Slide Time: 47:48)



So, as given g ix is nothing but l n p of x given omega i plus l n p of omega i. Now, the expression is p of x omega i. What is said is p of x omega i means that we are taking the samples from class omega i and finding out the sphere density function. So, the expression will remain the same. The mu and sigma will be replaced by mu i sigma i. That means the mean vector for the samples taken from class omega i covariance matrix computed from the samples taken from class omega i that is all, expression will remain the same.

So, the expression for the probability density function if I write p of x given omega i, it is nothing but the same explanation 1 over 2 pi to the power d by 2, then determinant sigma i to the power half. So, you find that sigma r is replaced by sigma i into exponential minus half x minus. Now, mu has to be replaced by mu i. This mu ii do not put it as ith component of mu, rather this is the mean vector of class omega i. So, x minus mu i transpose sigma i inverse into x minus mu i, this is my probability density function or class conditional probability density function.

So, by using this particular probability density function or multivariate probability density function, now I can define g ix to be logarithm of this term plus logarithm of p of omega i that is the priory probability. So, it will simply become, this is an exponential. So, if we take the logarithm, it will be simply this. So, it will simply become minus half into x minus mu i transpose sigma i inverse x minus mu i. This becomes minus d by 2 logarithm of 2 pi, this term. This term becomes minus half log of determinant of sigma i. So, all the terms of this expression are taken care plus I have to put this term, which is log of priory probability of omega i.

So, this is the expression for discriminant function corresponding to claas omega i. So, from here, you will find that this is actually a quadratic expression. Is it not? I have x minus mu transpose into x minus mu with sigma i inverse sandwiched between these 2 terms.

So, actually this is a quadratic equation. So, base discriminator is actually a quadratic discriminator in general or when I want to compute the decision surfaces of two classes between omega i and omega j, the decision surface will actually give quadratic surface between two classes omega i and omega j, the decision surface will actually be a quadratic surface. It is not a linear surface. So, this classifier can take care of linearly non separable classes. However, for specific cases, this can be converted to linear classifier. So, we will talk about those things in the next class. Let us stop here.