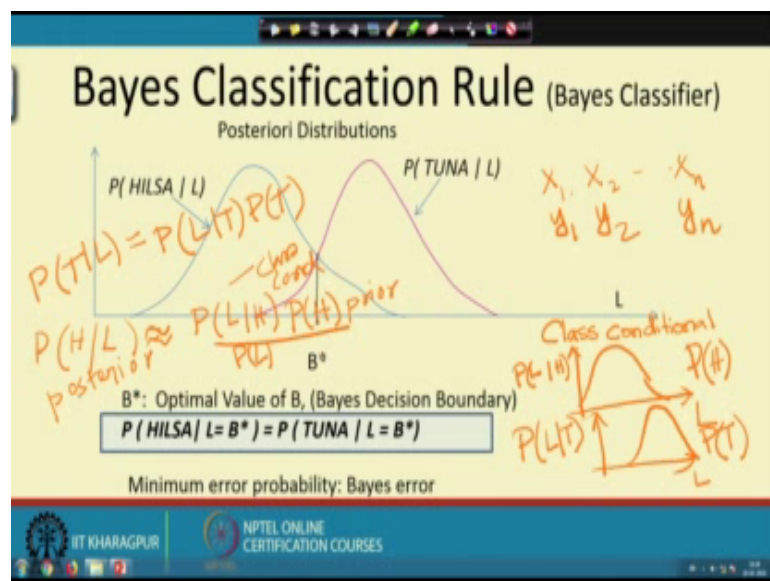


Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 14
Bayes Classifier-III

Let, me quickly summarize the Bayesian classifier that we have designed so far. So, what we did was we obtained examples of two classes.

(Refer Slide Time: 00:30)



It can be multiple classes also, so we had examples which consisted of measurements in this particular case only a single measurement the length of a of some object which is in this case is a fish which we are putting into one of the two classes tuna and hilsa. So, we have observations and we consider the class label, which is h or t again in this case for each of these observations of l, we put them into two groups, one group for all the observations in a particular classes say Hilsa and other all the population in the other class tuna and then we first construct the class conditional distribution.

So, what we did in that case is we find out different values of the length l and count find out the probability from the histogram counting that given a fish belongs to say class h, what is the probability it has a length l, some length l 1, 2, 3, 4 some length l given that it belongs to class eight. So, we construct a probability distribution of this type and we repeat this for both the classes. So, similarly we construct l given t tuna class and you

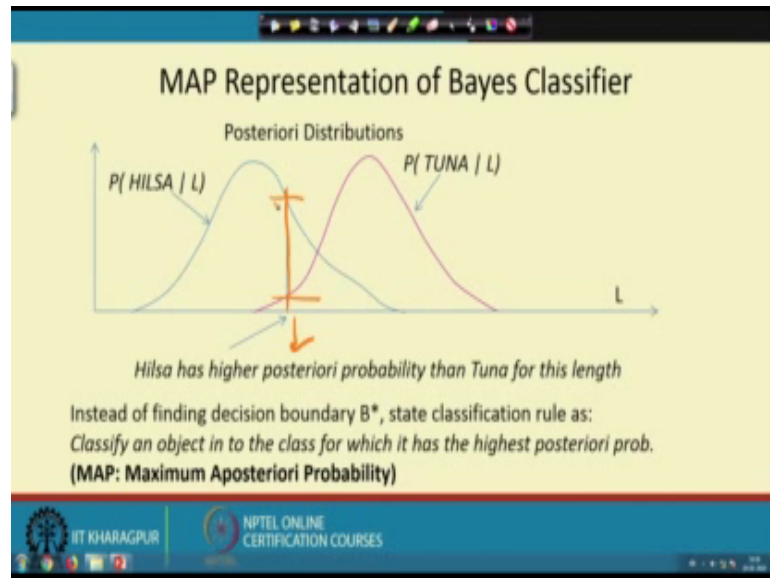
construct another distribution. So, x axis is length y axis is probability of a fish having that length known that it belongs to class h known that it belongs to class t and we found that this is not enough to take care of the bias in the population.

So, we considered prior probabilities of a random fish without knowing what its length is falling to the probability p_h to the class h and the probability a random fish belongs to the class t and what we did was we used the Bayes rule and obtained the posterior distribution, which is probability f is belongs to certain class h given it is l has certain value is actually not exactly equal, but proportional to let, me write it as proportional to probability the class conditional probability that we find out earlier multiplied by the prior probability.

So, this gives the so this is the class conditional and this is the prior we multiply these two to get the posterior distribution applying the Bayes rule. Note that there is a constant denominator $p(l)$. Since, which is same for all the classes we ignore it and say this is proportional, we similarly find the posterior distribution for both the classes t and h using the same formula and plot the posterior distributions plot the posterior distributions of each of the class.

So, this is for h class and this pink one is for the t class and then as we had reasoned earlier we will find that the minimum error would occur, if the decision boundary between both the classes lies at the intersection of the classes, lie at the intersection of the classes. So, what we actually did was kind of I mean summarize these two in a slightly different form which we call the Maximum A Posteriori Probability form of the Bayes classifier, which says that once you construct the posterior probabilities and given a new example with length L .

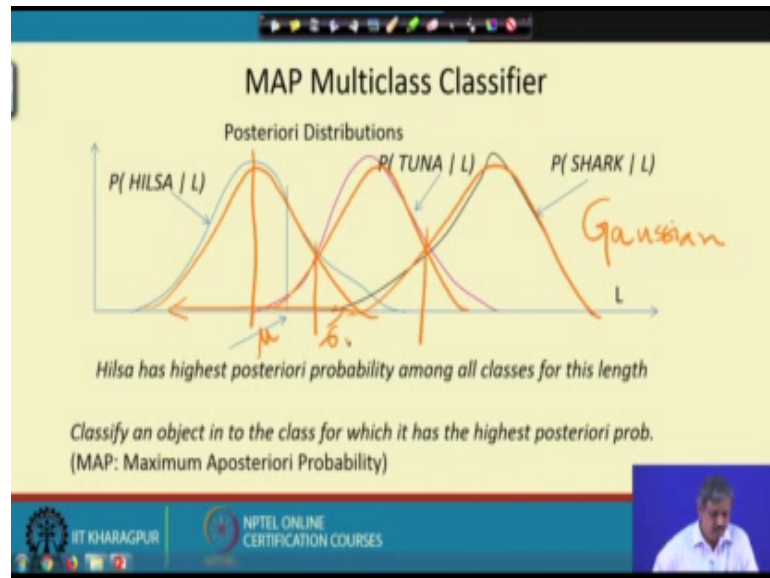
(Refer Slide Time: 04:52)



I just checked the posterior probabilities for each of the classes whichever class provides the higher posterior probability classify that particular example having length L to the higher class higher probability class. So, this principle we call the map principle the maximum posteriori probability principle and this is equivalent to the Bayes classifier. So, the and of course, we showed that this has the minimum error possible.

So, in order to actually use this classifier what we have to do is to kind of know the distribution, the class conditional and the prior probabilities of each of the classes. So, that we can apply it; so we also showed that you can actually extend it to multiple classes applying the same map principle.

(Refer Slide Time: 06:17)



And then we our next job is to actually find out these distributions and then apply the (Refer Time: 06:27) and get the classifier.

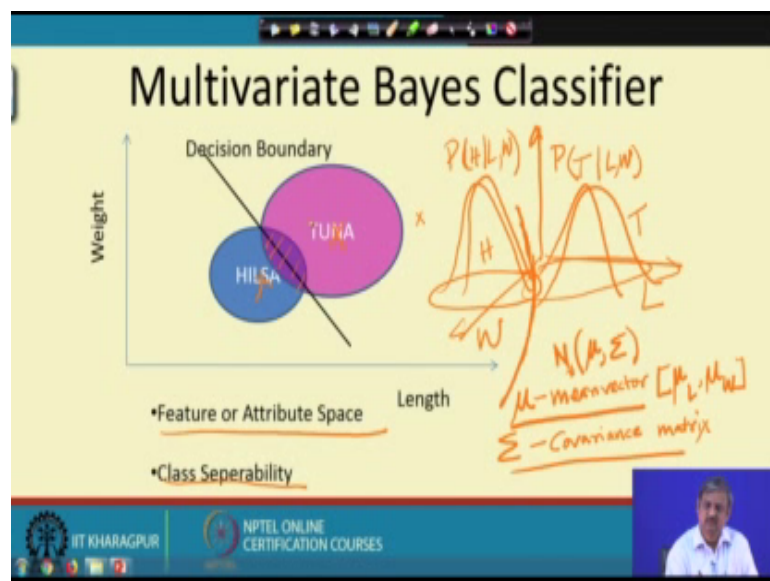
So, now there are two ways of constructing this map rule one of the way is called a parametric method. So, what the parametric method does is that it makes some assumptions on this class distributions of this posterior class distributions we will make some assumptions.

So, one of the most common assumption is that you assume that each of these classes each of this class distributions, follow a normal or a Gaussian distribution. So, often this is a reasonable assumption, because most of the natural random variables which initial length of the fish or height of a person, they would have this kind of bell separate behaviour. So, there would be a mean value and some dispersion around it some dispersion. So, this is the mean value μ and this is the dispersion you call it lets say σ .

So, we make this assumption and we will actually see that of course, the value of the μ and the σ they are unknown and what will actually do is that from the population that you have from the training samples that we have, we will try to estimate the value of μ and σ there are definite ways of doing that. So, you for example, μ is the average value of all the training samples and σ is the average value of the squared difference of each sample from the mean value.

So, standard deviation and mu we can estimate and after estimating (Refer Time: 08:59) we have kind of we know these distributions and if we know this distribution. So, we can find out what is their boundary in terms as a function of the values of mu and sigma, we can find out the boundary. So, all of you know the form of the Gaussian distribution, it is one by sigma into some form of sigma square into e to the power x minus mu by sigma whole square negative power of that so, it decays (Refer Time: 09:37). So, if we work out just as an exercise, what happens when each of these class distributions are Gaussians, but bivariate Gaussians a special case of multivariate Gaussians

(Refer Slide Time: 09:48)



For example, if I measure instead of just length the length and weight of the fish, as I mentioned before each fish would look like a point in two dimensional space and this representation would be called a feature space. So, suppose I draw it in a three d way, and if I draw the posterior probabilities, they would look like bell three dimensional bell separate curves each class would look like that. So, another class maybe will look like this.

So, let us say this is the Hilsa class and this is the tuna class to this actually is and since, this is two dimension the boundary will be not just a single b star I do we had noted earlier in the one dimensional case, it would be that the a curve it would be the locus of all two dimensional points, where this posterior distributions have same value intersect.

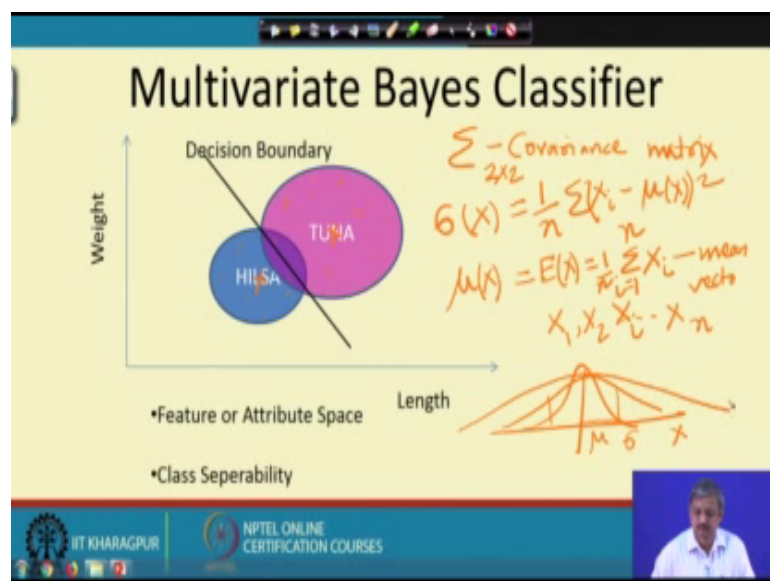
So, actually the if their distributions like this there will be some amount of overlap. So, this region would be called as the overlapping region. So, this picture I am saying is a top projection if we look from the top it is a projection on the LWspace, the I will have a short discussion on the shape of these projections depending on the parameters of the Gaussian distribution.

So, the amount of overlap depends on what is called the class separability. So, how well separated two classes are; So, now what we plan to do is to find out the shape of this locus of intersecting points the class boundaries two dimensional for the different type of Gaussians, we have in the individual class distributions.

So, you know a Gaussian distribution is parameterized by two variables, one is the mean vector, which will actually be like two values mean of the length of a Hilsa fish, say this is the Hilsa distribution and mean of the weight of the Hilsa fish, here is the mean vector and then there will be a covariance matrix, there will be a covariance matrix.

So, this mean vector will determine this kind of the central point it will determine, where my these bells are placed where in the coordinate system our bells about the close are their part where they at least and this covariance matrix, will determine what is the shape of this projections, it will determine what is the shape of this projection. So, let me quickly tell you how these projections look like.

(Refer Slide Time: 14:57)



So, they are determined by the covariance matrix. So, what does covariance matrix means I am telling you, covariance matrix is a matrix whose size if we have d dimension in this case we have two dimension is 2×2 $d \times d$. So, you have $d \times d$ so what are the values of the entry let me first define what is a variance all of you know it.

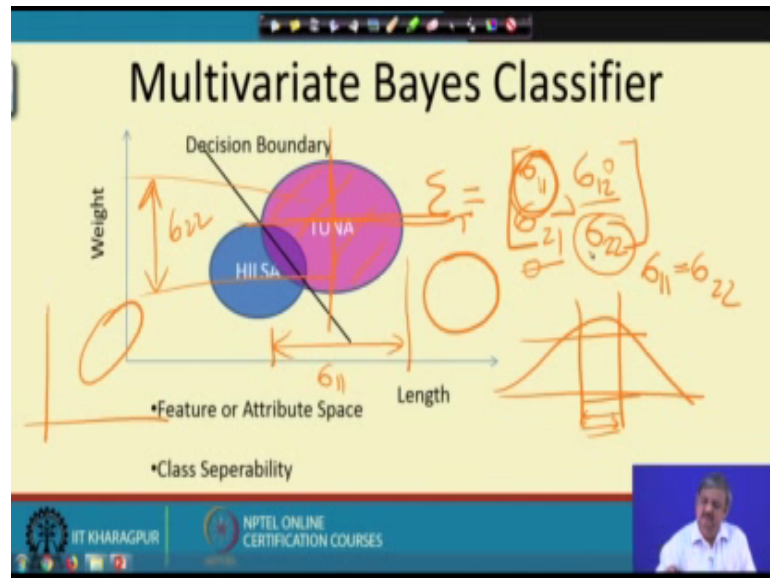
So, variance of a random variable x is what is the mean of a random variable x , mean of a random variable x is the expected value of x , which means that you can estimate it in the following way if you have n samples, x_1, x_2 in plain terms, there are some technical points, which I am not discussing here, if you have $x_1, x_2, \dots, x_i, \dots, x_n$ points all of them vectors each of them vectors, you do is that you take a sum off all these vectors of the all x_i vectors i equal to 1 to n just add them up term by term and then take average value.

So, these would give me kind of a mean vector, meaning it would give me the most centrally located point among this all the tuna, all the tuna fishes that we acquire out all these tuna fishes it would give the most central (Refer Time: 17:00) similarly for Hilsa.

Now, what is sigma? So sigma is the and the estimate of sigma is, if you instead of adding up just x you take the difference of x from the mean value take the difference of I am not putting x this is the mean take the difference of x from the mean value the deviation, x from the central value square it up and add it up square and add it up and take the average value of this dispersion take the average value of dispersion.

So, that is a kind of measure of the dispersion of the distribution. So, this would be if you have a distribution like this on x this is the mean and kind of this is the spread the dispersion. So, you can have sharper distributions with small μ has small sigma and you can help spread distributions with larger sigma ok so now these covariance matrix what it looks like is the following.

(Refer Slide Time: 18:41)



This covariance matrix this sigma as long as it is a univariate random even its fine, but if you have multivariate then sigma is more complex this dispersion is more complex because, in the univariate case dispersion is just kind of sorry, I am better drawing this it is just kind of you take some threshold of the probability measure its distance from the mu and this must interval this much length is my sigma that measures the distribution, but in circles or ellipses it is not.

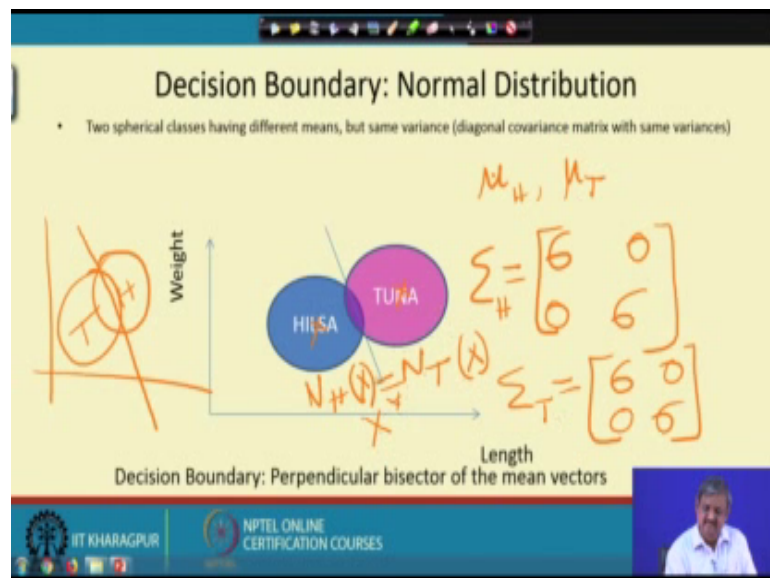
So, easy to measure the distribution how do you measure it. So, it will look like this I have the following quantities, I am not writing down algebraic equations you can you have done the statistics course. So, I assume that you know the definitions of this I am trying to geometrically explain what it means. The sigma 1, 1 it is the dispersion of the x component of the vector from the mean value of this x component is kind of say I am finding sigma for this tuna case, this big ellipse.

So, this sigma 1, 1, would kind of measure the dispersion along the first axis the x axis. So, this is sigma 1, 1 and then sigma 2, 2 would measure the dispersion along the second axis so this would be sigma 2, 2. In this case, if the sigma you can naturally guess that the this axis of the ellipse is longer than the y axis spread along y direction. So, sigma 1, 1 is greater than sigma 2, 2 that is why it is an ellipse one of the axis is longer than the other axis, what happens if they are equal if sigma 1, 1 was equal to sigma 2, 2 then this projections would look like circles all right.

So, now one more thing in this particular case the σ_1, σ_2 , are actually 0 because you see this plate and this plate they are kind of orthogonal to each other they are axis aligned, they are axis aligned the principal axis of the ellipse and the second minor axis are axis aligned. So, the cross dispersion are might achievable here what would happen, if this cross dispersion the off diagonal terms of the covariance matrix are not 0, we would get ellipses which are tilted which are tilted not axis aligned.

So, that is what we would get alright. So, it is clear so if for the two variate case, we will get ellipses projections the axis of the ellipse will depend on the diagonal elements of the covariant matrix and the tilt of the ellipse would depend on the up diagonal element of the covariance matrix. Now, let us look at one special case.

(Refer Slide Time: 22:24)



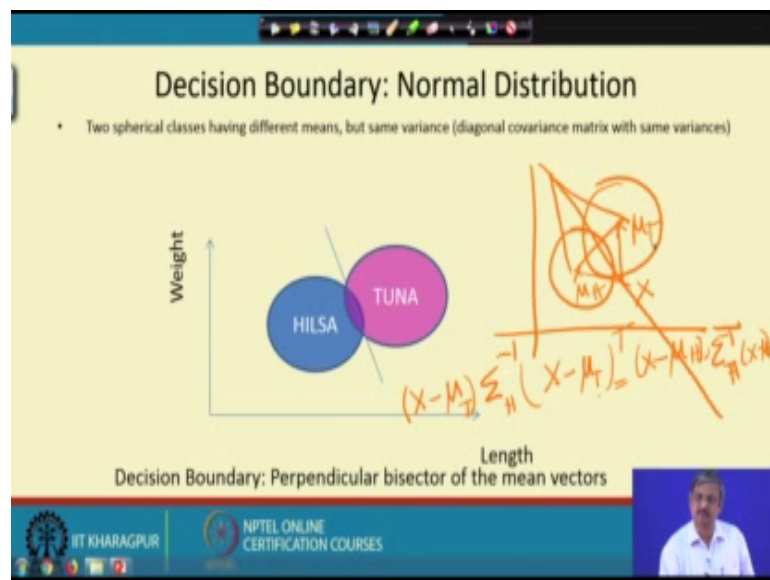
Let us look at a case, where the mean can be anything let us, say μ_H and μ_T . So, your centres of these ellipses can be located anywhere, but the covariance matrix of both the ellipses take this form both the ellipses σ_H is of this form σ_T is also of this form up diagonal elements 0 and diagonal elements are, equal what does this mean basically this means up diagonal element 0 means, the projections of they are axis aligned and same value for both the diagonal elements that this they means that this projections are circular.

So, this picture is actually not correct so I would actually have things like this. So, if this be the case, question is what is the Bayes optimal classifier? What is the locus of points?

Which gives equal probability for both these normal distributions, N_H and N_T and for both these normal distributions what are the probabilities, I give it as exercise to you I am not doing it though it is very easy to do is that if you just write down the equation for the normal distribution and pick up all x , for which $N_H(x)$ equals $N_T(x)$. $N_H(x)$ is the probability of x belonging to the normal distribution formed by h which has mean μ_H and σ_H and another probability of n belonging to μ_T σ_T , thus tuna distribution.

If we equate this and find out all the x stars which satisfy this quantity you will see that the locus of those points which from the decision boundary.

(Refer Slide Time: 24:52)



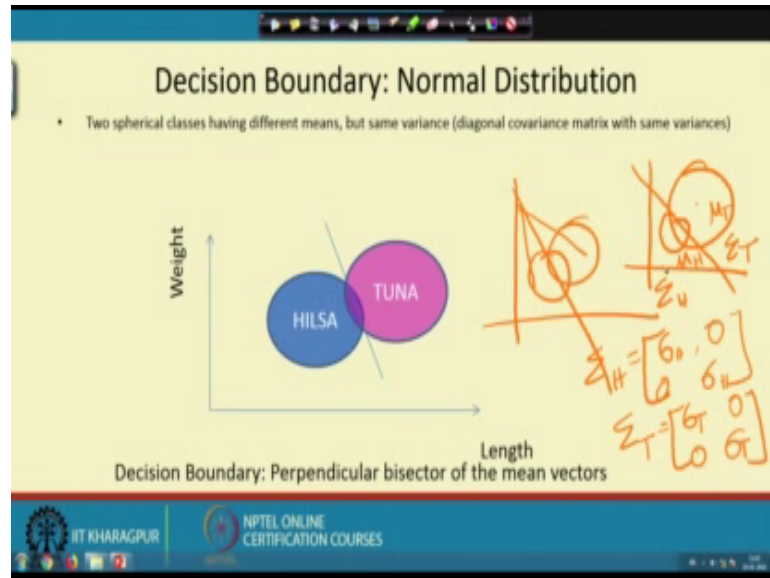
The class boundary will be nothing, but the following. You take the you join both the means μ_H and μ_T and draw its perpendicular bisector ok.

So, what is the property of the perpendicular bisector each point x on the perpendicular bisector a equidistance from μ_T and μ_H . So, take any point in the bisector they are equidistance.

So, basically x minus μ_T sigma is same for both is x minus μ_H by the way sigma will no longer be like this sigma will actually look like this it will look like, the distance will look like x minus μ_T the sigma H inverse and this into transpose of this into μ_H to μ inverse x minus μ_H mu h sorry, you can work it out. So, basically if you work

out the perpendicular bisector all locus of all points equidistance from both the mean, will actually give you the class boundary just work it out. Let us look into the second case, that where the means are ok.

(Refer Slide Time: 26:54)



There you have two means, but individual classes are no longer two circle earlier it has two circles of same size, same sense, same sigma. Now, I have two circles no more no ellipses yet, but two circles one small one large one μ_H μ_T two circle centres, but one which sigma H diagonal sigma H and one with sigma t, where sigma H takes the form sigma H 0, 0 sigma H still circle still axis aligned circle, but this sigma t and sigma H are no longer same so one circle is larger than the other circle.

So, what are the equi probable points now what is the locus of points which are equi probable let us see, so earlier I said for the equal circular it is locus of all points the perpendicular bisectors, which are equidistance from both the centres equi Euclidean distance from the both the centres.

We will see that in the next case, we will actually have to locus as all points, which are equidistance from both the means, but not the Euclidean distance some other distance called the Mahalanobis distance. A simple, extension of the Euclidean distance the Mahalanobis distance and in both unequal sized case, the locus is nothing, but not the perpendicular bisector which is equi Euclidean distance from both centres, but all points which are equi Mahalanobis distance from both the centres ok.

So, in my next lecture soon I will first discuss, what is Mahalanobis distance? And find out the expression. So, this is the end of this lecture thank you, I will after this I will continue with my definition of the Mahalanobis distance ok. I will continue with this.

(Refer Slide Time: 29:23)



Distances

- Two vectors: Euclidean, Minkowski etc
- A vector and a distribution: Mahalanobis, Bhattacharya

Which distribution is closer to x ?

$$d_M = \frac{(x - \mu)^2}{\sigma}, d_M = (X - \mu)\Sigma^{-1}(X - \mu)^T$$

- Between two distributions: Kullback-Liebler Divergence

 IIT KHARAGPUR  NPTEL ONLINE CERTIFICATION COURSES