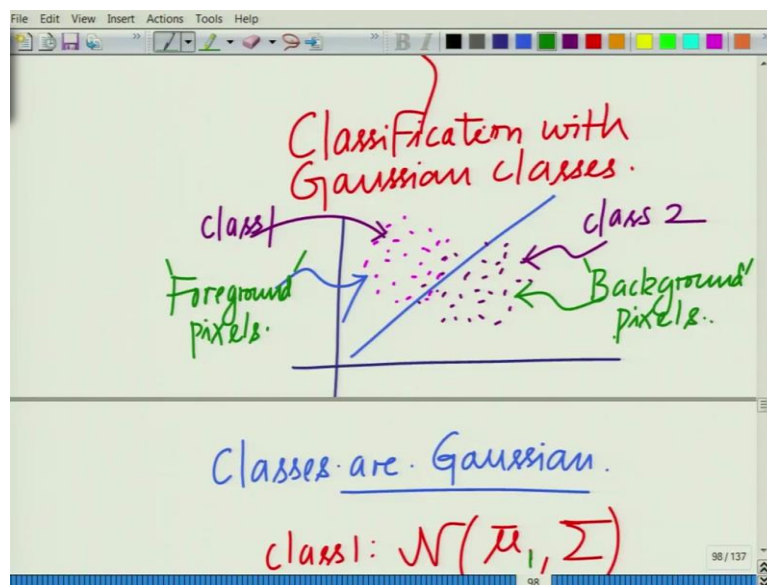
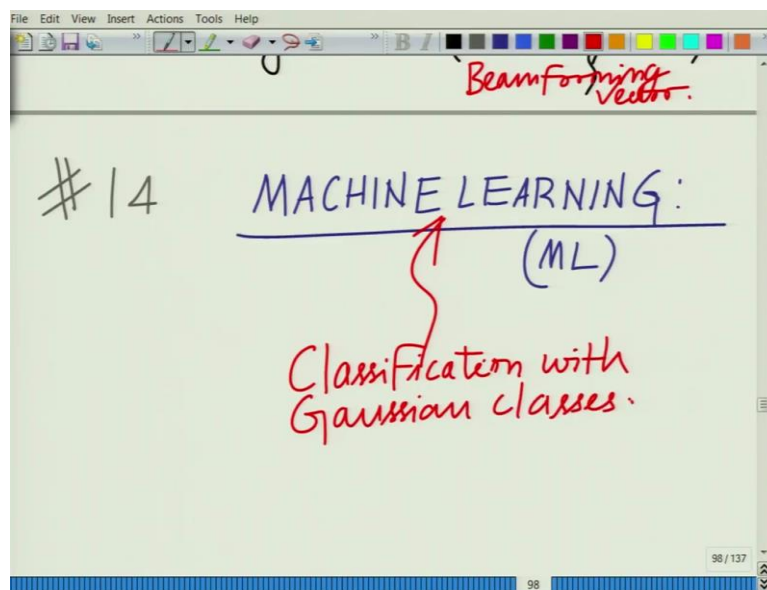


Applied Linear Algebra for Signal Processing, Data Analytics and Machine Learning
Professor. Aditya K. Jagannatham
Department of Electrical Engineering
Indian Institute of Technology, Kanpur
Lecture No. 14
Machine Learning Application: Gaussian Classification

Hello, welcome to another module in this massive open online course. So, in this module let us look at an interesting application of Gaussian's Gaussian random vectors in the context of machine learning and in particular linear algebra linear transformations of Gaussian random vectors in the context of machine learning.

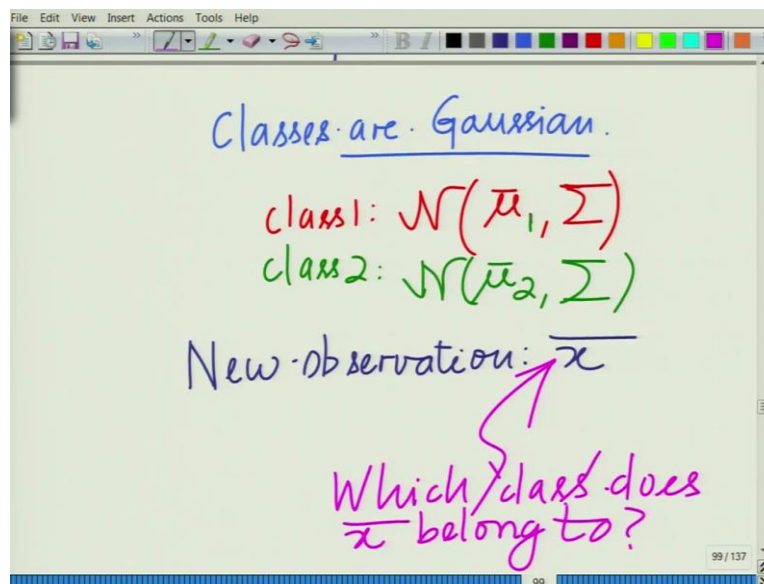
(Refer Slide Time: 00:33)



So, we want to look at an interesting application, one of the first in a series of applications we are going to look at in this course: Machine Learning (ML). The particular problem that we want to look at is what is termed as classification; classification with Gaussian classes. What the problem is essentially very simple. You have the data belonging to 2 classes. So, you have class A and you have the data; another set of data belong to and of course there can also be some kind of a sparse overlap because they are these are after all Gaussians.

And what you are trying to do is you are trying to classify this or basically separate these into 2 different classes. So, you have class let us call this as Class 2 and let us call this as Class 1 and you are basically separating these 2 classes. Or you are basically classifying the data into these 2 classes. So, partitioning the data into 2 classes. And now, let us say these classes are Gaussian. Now, in this problem and this frequently arises: these classes are Gaussian.

(Refer Slide Time: 2:42)



Class 1 is Gaussian with mean $\bar{\mu}_1$ covariance Σ . So, we are simplifying this by assuming this by assuming the same covariance Σ . So, Class 2 is Gaussian with mean $\bar{\mu}_2$ or $\bar{\mu}_2$. So, the means are different. First one has mean $\bar{\mu}_1$; second one has mean $\bar{\mu}_2$ but the covariance is same. So, this is Class 1, Class 2. So, these 2 classes both these classes are Gaussian.

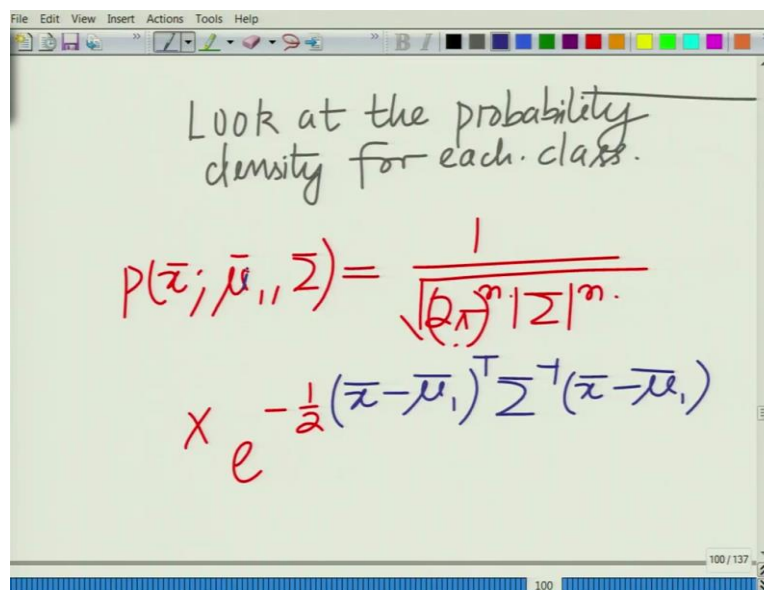
Now, the point is if there is a new point. Now, the question is if there is a new observation \bar{x} . Which class does \bar{x} ; the question is which class does \bar{x} belong to? Which class does \bar{x} belong to that is an interesting one. For instance, let us take a simple example. You can have 2 classes. We can have for instance an image. Image can contain 2 classes: one class

belongs to a certain object; other class of pixels belong to a certain object. You are trying to find a new pixel which object does it belong to?

And let us take this simplified even further. Let us say there is an image with a single object. You have the object which is we can term as the foreground object or a person that is a foreground. And then you have the background. So, you can think of these 2 classes as one being the foreground; other being the background. This is an example. This can be anything. This can basically be the presence of an object, foreground-background pixels. So, foreground pixels, background pixels; foreground pixel background.

Now, we have new pixel vector \bar{x} . Which does it belong to? \bar{x} does belong? Does it belong to the foreground? Does it belong to the background? In general, we are saying that these 2 pixels are basically distributed as Gaussians: foreground Gaussian with mean $\bar{\mu}_1$, covariance matrix Σ . Background mean $\bar{\mu}_2$, covariance matrix Σ . How do you classify a new pixel, \bar{x} ? Let us look at this problem.

(Refer Slide Time: 05:32)



Look at the probability density for each class.

$$P(\bar{x}; \bar{\mu}_1, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|^m}} \times e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_1)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_1)}$$

LOOK at the probability density for each class.

$$P(\bar{x}; \bar{\mu}_1, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}}$$

Determinant

$$e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_1)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_1)}$$

Likelihood of class 1

Let us now write the probability density function of the foreground. So, look at the probability density for each class. Let us look at the Class 1. Let us look at Class 1. What is the probability density for the Class 1? P of x bar which is basically parameterized by mu 1 comma sigma. This is the multivariate Gaussian random probability density function with mean mu 1 bar. In fact, I have to write mean u 1 bar, mean mu1 bar 2 pi sigma n.

This is the determinant as you know. This is the determinant of sigma; determinant of the covariance; this is a probability density function; times e raised to the power of minus half x bar minus mu1 bar transpose sigma inverse x bar minus mu1 bar. So, this is basically your PDF corresponding to. So, this is the determinant that is what I was saying earlier. This is the determinant, there is no n here. This is the determinant: minus half x bar minus mu1 bar sigma inverse minus x bar minus mu1 bar and this is a Gaussian PDF and this is now termed in the context of classification or detection, this is termed as the likelihood corresponding to Class 1.

Because we are looking at it as a function of the different classes. x bar is fixed. So, probability density function is basically a function of x bar. Now, we are looking at a point x bar. x bar is fixed. And you are looking at it's with respect to the different classes or the different hypothesis. So, therefore, this is known as the likelihood. So, this is what is known as the different, the likelihood of Class 1. This is an interesting terminology. This is the likelihood of Class 1 which is nothing but a probability density function evaluated at point x bar corresponding to that particular class.

(Refer Slide Time: 08:11)

A screenshot of a digital whiteboard showing the handwritten equation for the multivariate normal distribution likelihood function. The equation is $p(\bar{x}; \bar{\mu}_2, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_2)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_2)}$. A green arrow points from the variable \bar{x} in the exponent to the \bar{x} in the function's argument.

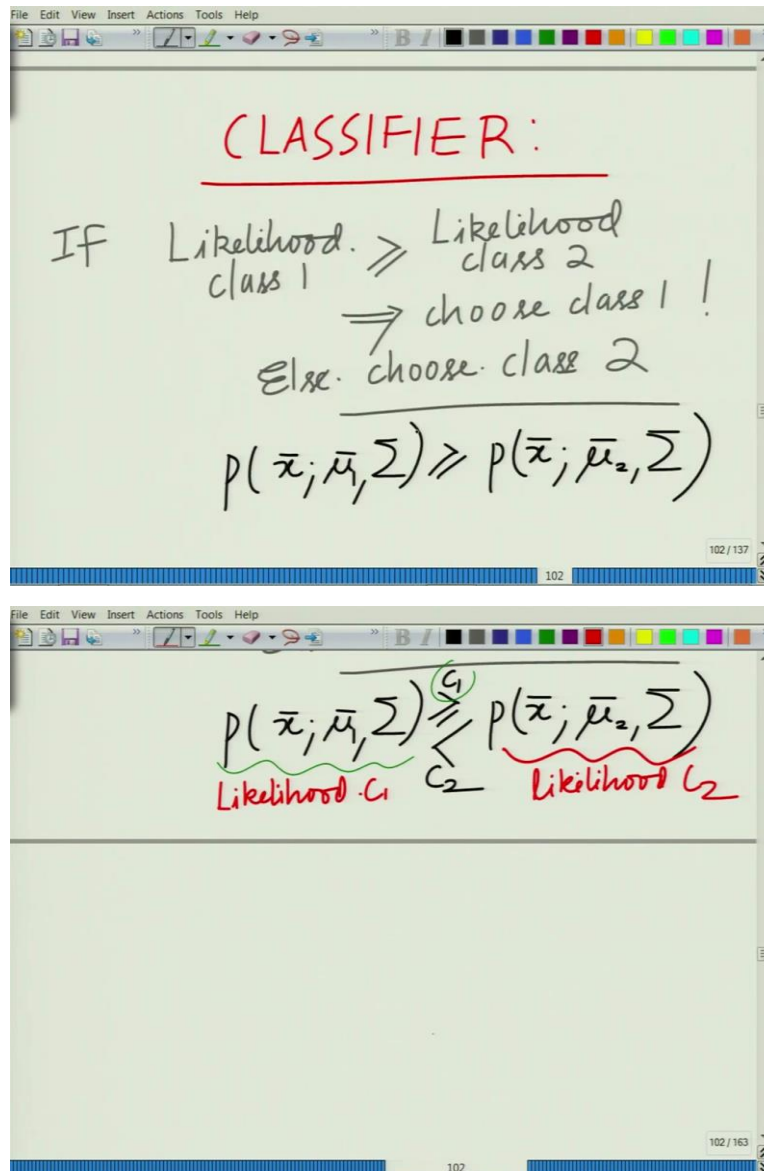
Now, let us look at the likelihood of the Class 2 which is nothing but P of x bar mu 2 bar sigma. This is equal to 1 over square root of 2 pi raised to the power of n magnitude sigma; that is the determinant of the covariance matrix times e raised to minus half x bar minus mu 2 bar transpose sigma inverse x bar minus mu 2 bar.

(Refer Slide Time: 8:53)

A screenshot of a digital whiteboard showing the same handwritten equation as the previous slide: $p(\bar{x}; \bar{\mu}_2, \Sigma) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_2)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_2)}$. A green arrow points from the variable \bar{x} in the exponent to the \bar{x} in the function's argument. The phrase "Likelihood of class 2" is written in purple below the equation.

Now, so, this is basically what we are calling as; this is basically your likelihood. Or this is basically; let me just write it with a different color. This is basically the likelihood, this is the likelihood of Class 2, this is the likelihood of Class 2. Now, how do we make the decision or how do we do the classification?

(Refer Slide Time: 09:32)



Now, what is the classification rule or what is the classifier? This is what we call in the machine learning: what is the classifier? Or what is the classification rule? The classification rule is very intuitive. If the likelihood corresponding to Class 1 is higher, choose Class 1. If the likelihood corresponding to Class 2 is higher, choose Class 2. If both likelihoods are same which arises with 0 probability, you can choose any class. So, if the likelihood; so, the classifier is very simple.

The likelihood corresponding to Class 1 if is greater than or equal to likelihood Class 2, if this is the case, then choose Class 1. That is your simple decision rule. Else, so we can write it if likelihood of Class 1 greater than equal to the likelihood of Class 2, choose Class 1 else choose Class 2. This is your simple classifier, that is your simple classifier. You compare the two classes. If the likelihood that is $P \bar{x} \bar{\mu} \bar{\sigma} \bar{x} \bar{\mu}_1, \bar{\mu}_1 \bar{\sigma}_1$

bar greater than or equal to $P \times \bar{\mu}_2$ bar comma sigma. It is not sigma bar. This is sigma. greater than equal to sigma, then you choose.

So, I can write this both as a single statement. If this is greater than or equal to this, choose Class 1. Or I can write this as over the greater than and if this is less than this, choose Class 2. This is the way you write it in a compact fashion. If it is greater than or equal to the likelihood corresponding to Class 2, so over the greater than or equal to symbol I am writing class 1 or the less than symbol I am writing Class 2; below the less than symbol I am writing class 2.

So, this automatically tells you what is the classification rule. If the likelihood corresponding to Class 1 is greater than or equal to likelihood corresponding Class 2, choose Class 1. Else naturally choose Class 2. So, essentially this is the simple problem or this is the simple rule for this classification problem. So, we were saying choose C1 if the likelihood corresponding to class C1 is the likelihood corresponding to C1 which is given by this. This is the likelihood for C1. So, if the likelihood corresponding to C1 is greater than or equal to this likelihood corresponding to C2.

And of course, if otherwise is less than equal to this. The likelihood corresponding to C1 is less than likelihood corresponding to C2, then choose C2. Now, let us simplify this. Let us substitute the expressions over here.

(Refer Slide Time: 13:18)

Choose C_1 if:

$$\frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_1)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_1)}$$

$$\geq \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(\bar{x} - \bar{\mu}_2)^T \Sigma^{-1} (\bar{x} - \bar{\mu}_2)}$$

So, that essentially gives me choose C1 if well if you substitute those likelihood expressions, if you go back and take a look at it, let me write these expressions anyway. So, this

essentially implies that 1 over square root of 2π raised to the power of n magnitude or the determinant of the covariance e raised to minus half \bar{x} minus μ_1 bar transpose sigma inverse \bar{x} minus μ_1 bar. If this is greater than or equal to, choose C_1 . If this is greater than or equal to 1 over exactly identical constant that is square root of 2π determinant of sigma e raise to minus half \bar{x} minus μ_2 bar sigma inverse \bar{x} minus μ_2 bar.

And now you can see from this that these constants these go away. And now you only are left with the exponential. And essentially you can choose C_1 if this exponent on the first is larger than the quantity in the exponent on the right. But there is a negative sign in front of it. So, essentially what it means is choose C_1 if and the half is essentially its scaling factor. So, choose C_1 .

(Refer Slide Time: 15:15)

choose C_1 if:

$$(\bar{x} - \mu_1)^T \Sigma^{-1} (\bar{x} - \mu_1) \leq (\bar{x} - \mu_2)^T \Sigma^{-1} (\bar{x} - \mu_2)$$

$$\Rightarrow \cancel{\bar{x}^T \Sigma^{-1} \bar{x}} - 2\mu_1^T \Sigma^{-1} \bar{x} + \mu_1^T \Sigma^{-1} \mu_1 \leq \cancel{\bar{x}^T \Sigma^{-1} \bar{x}} - 2\mu_2^T \Sigma^{-1} \bar{x} + \mu_2^T \Sigma^{-1} \mu_2$$

So, you can equivalently; you can easily check this C_1 , this reduces to choose C_1 if \bar{x} minus μ_1 bar transpose sigma inverse \bar{x} minus μ_1 bar, this is less than or equal to \bar{x} minus μ_2 bar sigma inverse \bar{x} minus μ_2 bar transpose sigma inverse \bar{x} minus μ_2 bar. And therefore, now if you simplify this further, this implies if we expand this, this will give you \bar{x} transpose sigma inverse \bar{x} minus $2\mu_1$ bar transpose sigma inverse \bar{x} plus μ_1 bar transpose sigma inverse μ_1 bar. This is less than or equal to.

Again the same thing, \bar{x} transpose sigma inverse \bar{x} minus $2\mu_2$ bar transpose sigma inverse \bar{x} plus μ_2 bar transpose sigma inverse μ_2 bar. And now, you can see the \bar{x} transpose sigma inverse \bar{x} , this cancels from both sides.

(Refer Slide Time: 16:57)

$$\leq \cancel{\bar{x}^T \Sigma^{-1} \bar{x} - 2\bar{\mu}_2^T \bar{x} + \bar{\mu}_2^T \bar{\mu}_2}$$

$$\Rightarrow (\bar{\mu}_2 - \bar{\mu}_1)^T \Sigma^{-1} \bar{x} \leq \frac{1}{2} \bar{\mu}_2^T \Sigma^{-1} \bar{\mu}_2 - \frac{1}{2} \bar{\mu}_1^T \Sigma^{-1} \bar{\mu}_1$$

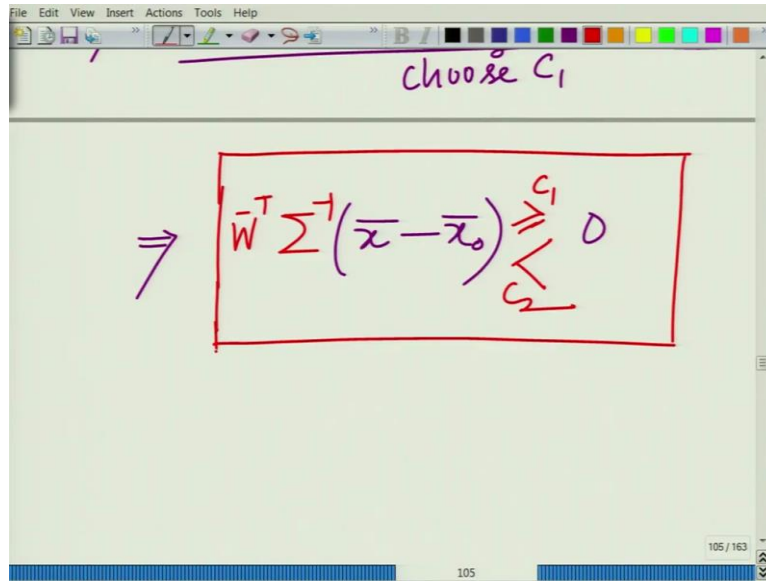
$$\Rightarrow \underline{(\bar{\mu}_1 - \bar{\mu}_2)^T \Sigma^{-1} \left(\bar{x} - \frac{\bar{\mu}_2 + \bar{\mu}_1}{2} \right) \geq 0}$$

This implies essentially that now, if you simplify it, you bring this on to the right, this implies $\bar{\mu}_2^T \Sigma^{-1} \bar{x} - \bar{\mu}_1^T \Sigma^{-1} \bar{x}$, this is less than or equal to $\frac{1}{2} \bar{\mu}_2^T \Sigma^{-1} \bar{\mu}_2 - \frac{1}{2} \bar{\mu}_1^T \Sigma^{-1} \bar{\mu}_1$ which now again once again if you simplify this, this implies that it is not very difficult. After some manipulation you can show this. This essentially $\bar{\mu}_1^T \Sigma^{-1} \bar{x} - \bar{\mu}_2^T \Sigma^{-1} \bar{x}$. Not very difficult to see this, $\bar{\mu}_1^T \Sigma^{-1} \bar{x} - \bar{\mu}_2^T \Sigma^{-1} \bar{x}$, well $\bar{\mu}_2^T \Sigma^{-1} \bar{x} + \bar{\mu}_1^T \Sigma^{-1} \bar{x}$ divided by 2 greater than or equal to 0. So, choose C_1 if this condition holds. If this condition holds, otherwise you choose C_2 .

(Refer Slide Time: 18:24)

$$\Rightarrow \frac{(\bar{\mu}_1 - \bar{\mu}_2)^T \Sigma^{-1} \left(\bar{x} - \frac{\bar{\mu}_2 + \bar{\mu}_1}{2} \right) \geq 0}{\text{choose } C_1}$$

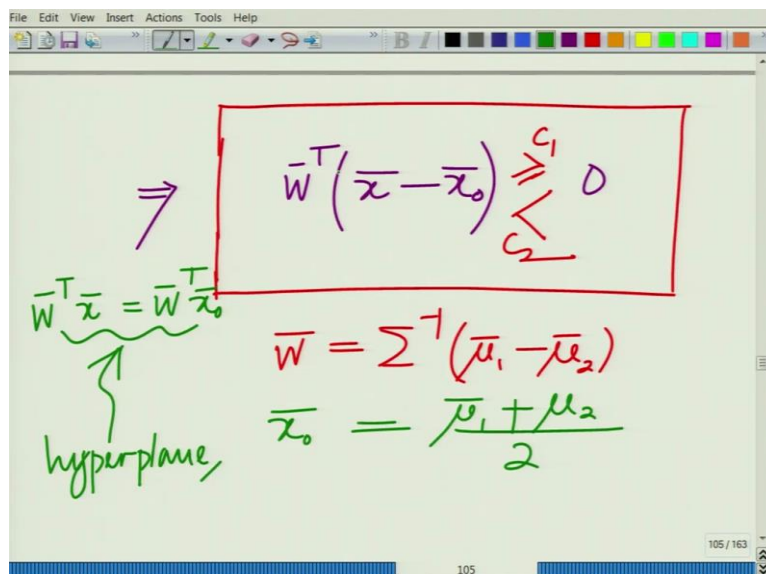
$$\Rightarrow \boxed{\bar{w}^T (\bar{x} - \bar{x}_0) \begin{matrix} \geq C_1 \\ < C_2 \end{matrix} \geq 0}$$

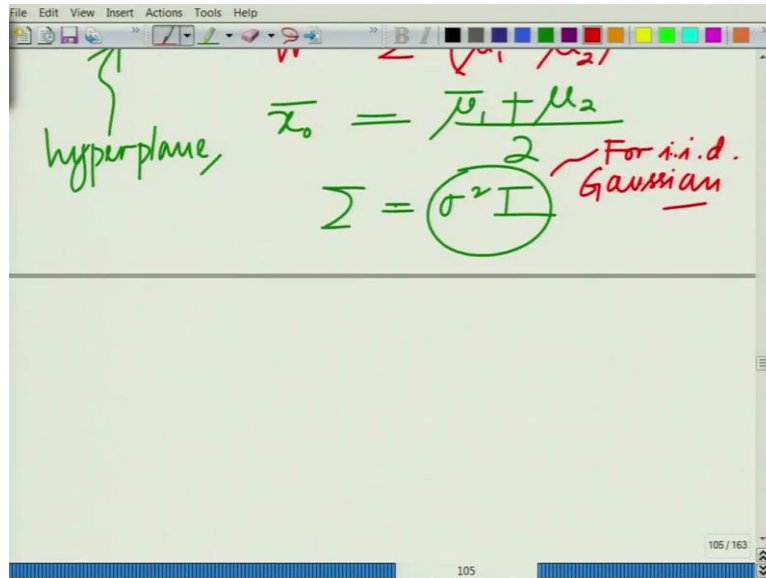


And this finally implies; I can write this conveniently as $\bar{w}^T (\bar{x} - \bar{x}_0) \geq 0$. That is choose C_1 if this holds, greater than or equal to 0. Else if this is less than 0, choose C_2 . Remember this is how we write this condition. This is a very interesting and I am going to explain more about this. This has to be the sigma square; sigma inverse has to be there. So, $\bar{w}^T \Sigma^{-1} (\bar{x} - \bar{x}_0)$; sigma inverse.

This and or you can, you can write this as $\bar{w}^T (\bar{x} - \bar{x}_0)$. I think this is fine. It is better to define it this way. So, where \bar{w} is well where this quantity \bar{w} equals sigma inverse μ_1 bar minus μ_2 bar. So, \bar{w}^T will I once again give you yeah μ_1 bar minus μ_2 bar sigma inverse. In fact, I can write this as \bar{w}^T .

(Refer Slide Time: 20:00)





And \bar{x} which is $\bar{\mu}_1$ plus $\bar{\mu}_2$ by 2. So, this is the condition. And you can see you will later see that this is essentially $w^T \bar{x}$ greater than $w^T \bar{x}_0$. So, this $w^T \bar{x}$ if you look at this, this $w^T \bar{x}$ equals $w^T \bar{x}_0$. This is actually a hyperplane, this is an n -dimensional plane. We are going to look at; the set of all vectors that satisfy this is essentially a hyperplane. We are going to look at that.

So, that is essentially what this is saying. And so, this is the condition. So, choose C_1 if $w^T \bar{x}$ and is also known as a discriminant. Now, let us come. Now, why is this very interesting? Let us look at a simplification. Let us set $\Sigma = \sigma^2 I$; let us set a capital matrix Σ equals σ^2 times identity that is considering independent identically distributed Gaussian random variables. We know that when the covariance is essentially proportional to identity that corresponds to independent identically distributed Gaussian that is a Gaussian random vector with independent identically distributed Gaussian random components. Now, therefore, this essentially for i.i.d or i.i.d Gaussian components. This has a very interesting.

(Refer Slide Time: 21:48)

$$\bar{w} = \frac{1}{\sigma^2}(\bar{\mu}_1, -\bar{\mu}_2)$$

$$\frac{1}{\sigma^2}(\bar{\mu}_1, -\bar{\mu}_2)^T(\bar{x} - \bar{x}_0) \begin{matrix} \geq 0 \\ < 0 \end{matrix} \begin{matrix} C_1 \\ C_2 \end{matrix}$$

$$(\bar{\mu}_1, -\bar{\mu}_2)^T(\bar{x} - \bar{x}_0) \begin{matrix} \geq 0 \\ < 0 \end{matrix} \begin{matrix} C_1 \\ C_2 \end{matrix}$$

$$\bar{x}_0 = \frac{\bar{\mu}_1 + \bar{\mu}_2}{2}$$

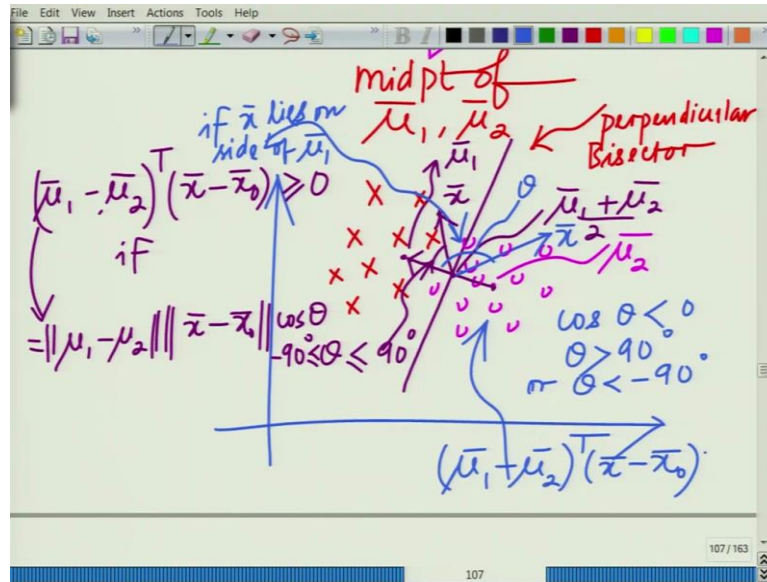
midpt of $\bar{\mu}_1, \bar{\mu}_2$

So, i.i.d Gaussian components, you have your w bar becomes which is essentially w bar becomes 1 over σ square μ_1 bar minus μ_2 bar. And your condition becomes if you look at this 1 over σ square times 1 over σ square times μ_1 bar minus μ_2 bar transpose x bar minus x naught bar greater than equal to less than 0 . So, you choose C_1 . If it is greater than equal to 0 , C_2 is less than 0 .

And of course, since there is 0 on the right, 1 over σ square can be removed and net you get the condition that is choose C_1 if μ_1 bar minus μ_2 bar transpose x bar minus x naught bar greater than or equal to 0 . So, C_1 is greater than or equal to 0 ; C_2 if it is less than 0 . And realize that this x naught bar, this is nothing but if you look at this x naught bar this is equal to μ_1 bar plus μ_2 bar divided by μ_1 bar plus μ_2 bar divided by 2 . This is equal to essentially the midpoint: μ_1 bar comma μ_2 bar. This is the midpoint of these two means

that is $\bar{\mu}_1$ comma $\bar{\mu}_2$. And therefore, the condition is something that is very interesting.

(Refer Slide Time: 23:43)



If you look at these two Gaussian classes, let me just draw the picture again and show you. So, you have these two Gaussian classes and this one has mean that is $\bar{\mu}_2$ and you have the mean that is $\bar{\mu}_1$. Now you join these two; just write this appropriately. So, so this is your $\bar{\mu}_1$. And now you look at the midpoint. Let us draw this hyperplane that bisects these two. So, this is the perpendicular bisector. This is a perpendicular bisector. And what you will see is that what you will see is that this is the midpoint which is essentially $\bar{\mu}_1$ plus $\bar{\mu}_2$ divided by 2.

Now, look at any point \bar{x} or look at any point \bar{x}_0 . Now, $\bar{\mu}_1$ minus $\bar{\mu}_2$ is this vector, $\bar{\mu}_1$ minus $\bar{\mu}_2$ is basically the difference vector. So, this is your $\bar{\mu}_1$. This is your $\bar{\mu}_2$. So, $\bar{\mu}_1$ is $\bar{\mu}_2$ is a vector starting at $\bar{\mu}_2$ pointing ending at $\bar{\mu}_1$ pointing towards $\bar{\mu}_1$. Now, \bar{x} minus \bar{x}_0 is this vector. So, \bar{x} minus; so now if you look at it \bar{x} .

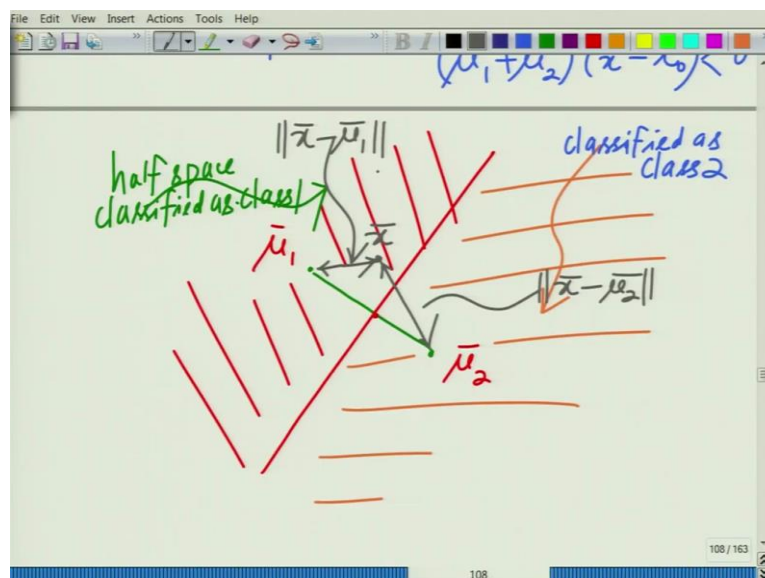
So, now, if you look at it, your $\bar{\mu}_1$ minus $\bar{\mu}_2$ transpose \bar{x} minus \bar{x}_0 which is the dot product between these 2 vectors is greater than or equal to 0 if this angle θ is lies between 0 and 90 degrees. Because remember this is nothing but this is essentially equal to the magnitude of $\bar{\mu}_1$ minus $\bar{\mu}_2$ times or norm of $\bar{\mu}_1$ minus $\bar{\mu}_2$ times norm of \bar{x} minus \bar{x}_0 times cosine of θ where cosine of θ is the angle between \bar{x} minus \bar{x}_0 that is the midpoint $\bar{\mu}_1$ plus $\bar{\mu}_2$

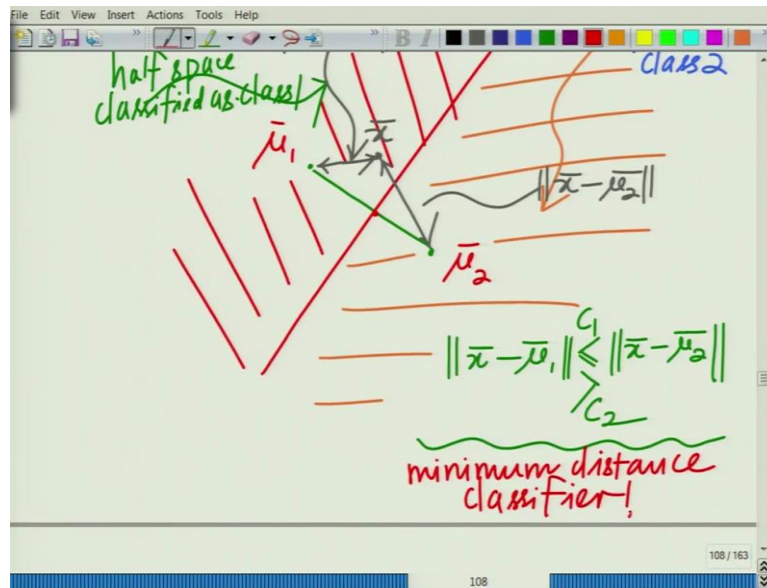
by 2 and $\mu_1 - \mu_2$. If this angle lies between 0 and 90, this is positive which means it will be classified as C1.

And you can clearly see this angle. I have to say this angle less than minus 90, minus 90 degrees less than or equal to theta less than equal to 90. That is, it lies anywhere \bar{x} lies anywhere in this side of the hyperplane. So, this is greater than or equal to 0 if \bar{x} lies on; \bar{x} lies on side of you can say a side of μ_1 . On the other hand, for any point here this is your \bar{x} is here. And now, if you look at again once again this angle, the angle between $\mu_1 - \mu_2$ this angle, this angle theta cosine theta is equal to less than 0 or 90 for theta which is greater than 90 degrees or theta less than minus 90 degrees.

So, you can say theta greater than 90 degrees or theta less than minus 90 degrees. And when cosine theta and when this cosine theta is negative, then for all points on this side of the hyperplane you will have $(\mu_1 - \mu_2)^T (\bar{x} - \mu_0)$. This will be essentially less than 0. So, essentially if you look at this. Let me just draw this figure again.

(Refer Slide Time: 28:53)





You have $\bar{\mu}_1$. This is the midpoint. This is the perpendicular bisector so, this entire region so, this is your $\bar{\mu}_1$. This is your $\bar{\mu}_2$. So, this entire half space we can call it. In fact, this is known as a half space. This entire half space is classified as; this entire half space is classified as basically Class 1. And if you look at this side of the half space; this half space, this is classified Class 2. That is if it lies on the side of the perpendicular bisector that is, you take 2 means: $\bar{\mu}_1$, $\bar{\mu}_2$, draw the perpendicular bisector.

If the point \bar{x} lies on the side of $\bar{\mu}_1$; if the observation x_1 lies on the side of $\bar{\mu}_1$ rise on the side of $\bar{\mu}_1$, it is classified as belonging to class 1 if the observation \bar{x} lies on the side of $\bar{\mu}_2$, it is classified as belonging to class 2. All right. So, that is the interesting. This is very intuitive. So, it says essentially that you are classifying as belonging to the class corresponding to the mean which is closest to the observation vector \bar{x} . That is, if you have \bar{x} ; if you have any point \bar{x} , you look at the distance between \bar{x} and each of the means. So, this is $\bar{x} - \bar{\mu}_2$ and this is $\bar{x} - \bar{\mu}_1$.

And essentially what you are doing is if norm of $\bar{x} - \bar{\mu}_1$ is less than or equal to norm of $\bar{x} - \bar{\mu}_2$, then you are choosing; you are choosing Class C_1 else you are choosing C_2 . So, this is also essentially known as the minimum distance classifier. So, this is a very interesting thing. So, with Gaussian covariance proportional to identity; covariance both covariance are equal. Both covariances of both the classes are equal and proportional to identity. This essentially reduces to minimum distance classifier.

That is, basically you draw the perpendicular bisector between the two means, if it is on the side of the new observation \bar{x} is on the side of μ_1 ; that is it is closer to μ_1 than μ_2 , classify it as Class 1. If it is on the side of μ_2 that is, if it is closer to μ_2 than it is to μ_1 classified as belonging to Class 2. So, this is a very interesting and a very simple classifier and a very interesting application of the principles of linear algebra.

It combines a lot of principles as you can see, it combines the principles of the multivariate Gaussian, multivariate Gaussian probability density function, the covariance of a Gaussian. Covariance of a Gaussian when the random when the components are independent, identically distributed. And finally, this notion of these likelihoods and comparing the likelihoods belongs to both the class of the likelihood corresponding Class 1 is greater than likelihood corresponding Class 2, choose Class 1, else choose Class 2.

And once you simplify it, you will see that it reduces to something that is very intuitive and interesting as well as intuitive that is nothing but the minimum distance classifier. Basically, the half space that is closer to μ_1 is classified as Class 1. Half space that is closer to μ_2 is classified as belonging to Class 2.

So, this is a very interesting application of linear algebra that we studied so far in the context of Machine Learning. And thus, it can be used to build very sophisticated Machine Learning algorithms. Alright, so let us stop here. We will continue this discussion in the subsequent module. Thank you very much.