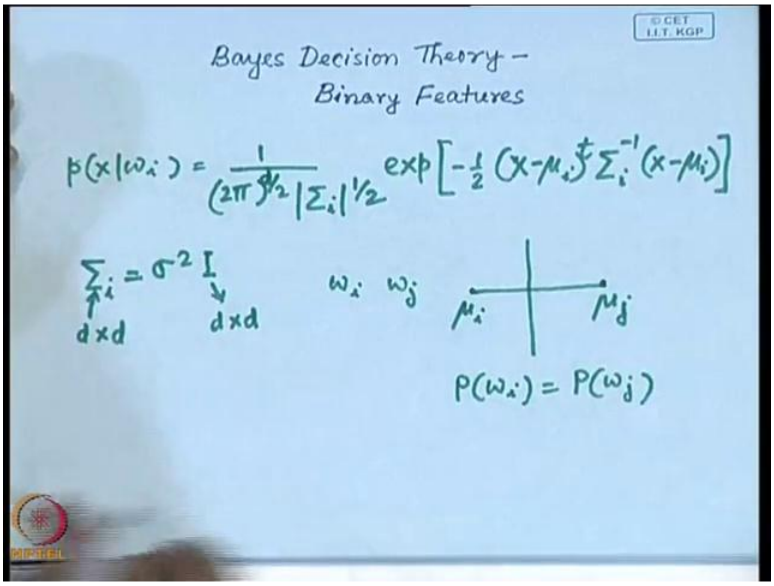
Pattern Recognition and Applications

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

Lecture - 9
Bayes Decision Theory – Binary Features

(Refer Slide Time: 00:41)



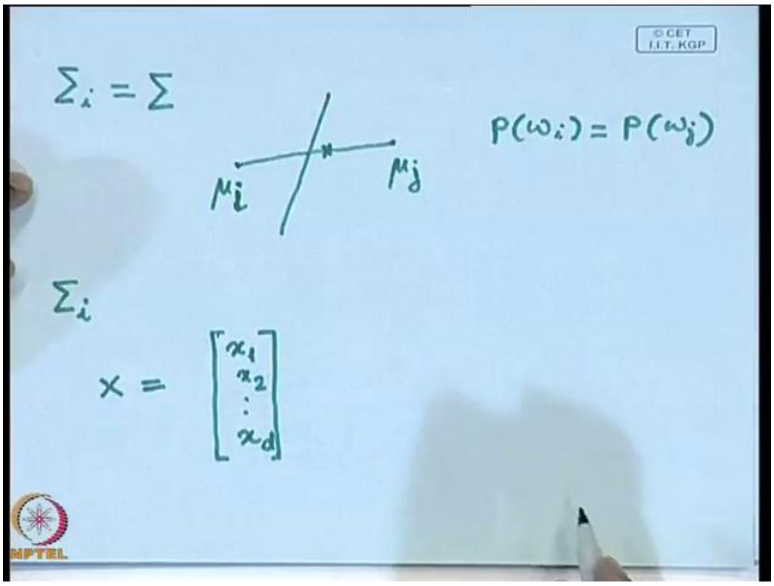
Good morning, so we are going to have, going to discuss today the Bayes decision theory for binary features. So far what we have discussed is assuming that the feature vectors are distributed following some normal distribution of the form P of X given omega i is equal to… So, this is the expression for a multivariate normal distribution. And we have seen that, in this expression where sigma i represents the covariance matrix, that for different conditions of covariance matrices, we can have different types of classifier.

So, the first case we have discussed is, if the covariance matrix for every class i is of the form, sigma square into identity matrix I, where both this covariance matrix is of dimension d by d. Where our feature vectors are of dimension d and this identity matrix is also of dimension d by d. So, because this is identity matrix so this simply says that covariance matrix is a diagonal matrix, where every diagonal element is of value sigma square. That means every component have the same variance or else, off diagonal elements are equal to 0 so because off diagonal elements are 0's so the components, different components of the feature vectors are statistically independent.

So, in such cases we have seen that, the classifier is nothing but a linear classifier or when we talked about the discriminant function, the discriminant function for, function for individual classes, they are also linear functions. And because they are linear functions so the classifier which employs this linear functions to classify and unknown feature vector, that is a linear machine. And in particular, if we want to find out the decision boundary two different classes say ith class, omega i and the jth class, omega j the decision boundary between these two different classes is a hyper plane, which is orthogonal to the line joining mu i and mu j, where mu i and mu j is at the centers of the classes omega i and omega j.

So, effectively I have a situation something like this, that, if I have mu i somewhere over here, which is the mean of the class omega i and mean j is somewhere over here. Then the decision surface is orthogonal is a hyper plane, which is orthogonal to the line joining mu i and mu j. And if the apriori probabilities p of omega i, is equal to p of omega j then this decision boundary or the decision surface becomes an orthogonal bisector of the line joining mu i and mu j. So, this was our simplest case.

(Refer Slide Time: 04:20)



In the second case, we have seen that if sigma i is equal to sigma, that is every class or the samples belonging to every class have the same covariance matrix, but otherwise the covariance matrix is arbitrary, unlike in the first case where, covariance matrix has its specific form like this. In the second case the covariance matrices are arbitrary, but every class have the same covariance matrix, which ideally means that, the points belonging to the same class or the points belonging to different classes, they are clustered in hyper ellipsoidal spaces of same shape and same size. And in such case we have seen that, the discriminating function that we get for different classes they are also linear.

So, in both the cases in the first case, as well in the second case the discriminating functions become linear so the classifier is nothing but a linear machine. However there is some difference between the decision surfaces, that we get between two different classes omega i and omega j. In the first case, the decision surface was orthogonal to the line joining mu i and mu j, in the second case the decision surface is not in general orthogonal to the line joining mu i and mu j.

So, here again if mu i and mu j so this is mu i and this is mu j, which are the centers of the two classes omega i and omega j then the decision surface between these two classes will be something like this. In the previous case, it was orthogonal to the line joining mu i and mu j, in this case in general it is not orthogonal. However the decision surface is a hyper plane or it represents the linear equation and here again, if the apriori probability is p of omega i is same as, p of omega j. Then this decision surface though it is not orthogonal to the line joining mu i and mu j, but it will pass through the point which is midway between the points mu i and mu j.

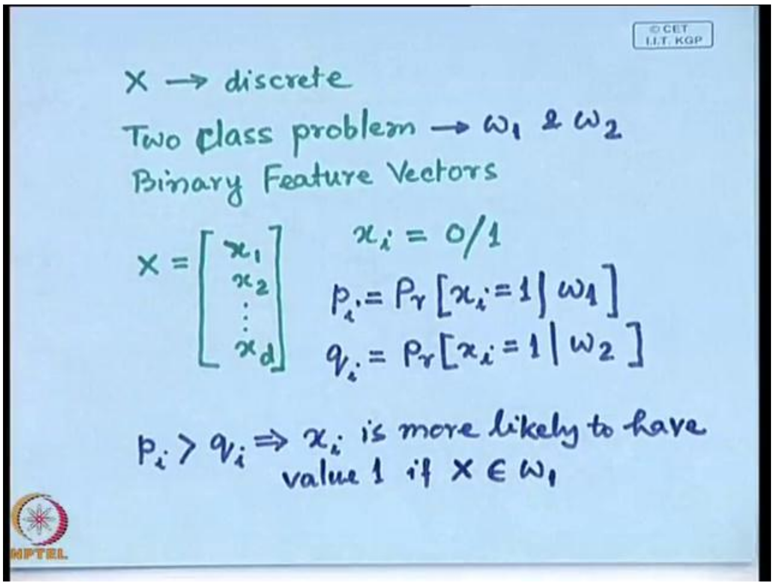
And the third case we have said that, the covariance matrices of different classes are totally arbitrary. So, for ith class I will have one covariance matrix, for jth class I will have another covariance matrix. So, that effectively means that the clusters or the points, vectors belonging to different classes they are clustered into hyper ellipsoidal spaces. In this, in the first, second case the hyper ellipsoidal spaces were of same shape and same size. In this case, the points belonging to different classes will also form hyper ellipsoidal spaces, but these hyper ellipsoidal spaces may not have same shape or may not have same size. So, they will have different shapes as well as different sizes, but the points belonging to the same class they, form a hyper ellipsoidal spaces.

However, in all these three different cases that we have discussed, we have assumed that the feature vector x is continuous or the individual components of the feature vector x. Because our feature vector x is, nothing but a d dimensional vector having the components x1, x2 up to xd. So, it is a d dimensional feature vector so in all this three different cases our basic assumptions was that, the feature vectors are continuous or in otherwise, individual components are also continuous. That effectively means that if I consider a d dimensional feature space then the feature vector can be represented by any point, is represented by any point within that d dimensional feature space. I do not have any specific set of points, from which the feature vectors are drawn.

However, in most of the practical applications and particularly in these days as we are walking with digital computers, all the data that we get are digital data. And the moment we get digital data, the vectors that we generate are no more continuous rather, they are discrete vectors or every component of the feature vector every xi will have a discrete values. Discrete values means, it will assume one half a set of specific values so instead of the continuous variable it becomes a discrete variable.

So, when it is a discrete variable, in that case in all our previous lectures wherever we have talked about integration, the integration is to be replaced by summation. And the summation has to be carried out, over the discrete space. So, we will take a specific case of this discrete feature vectors. So, let us consider a case where, so we will have a feature vector x which will be discrete.

(Refer Slide Time: 10:11)



And we consider a specific case of a two class problem and the feature vectors, we will assume to be binary feature vectors. Binary feature vector means, every component of the feature vector will assume binary value either 0 or 1.So, when I have this feature vector x, which is given by x1, x2 up to xd. Again we assume that, we have d dimensional feature vectors. Every component of the feature vector x1, x2 or x3 can assume either a value equal to 0 or a value equal to 1.

So, that means that, if a feature vector is taken from a particular class it will say whether a particular feature is present or it is absent. So, when I have this sort of binary feature vectors so we will also assume the different components of this feature vectors are conditionally independent. To have something similar to statistical independence of different components inventing most feature vectors. So, here also you will assume that different components of the feature vectors are conditionally independent.

So, every feature value in this feature vector xi, every feature value can have a value either 0 or 1. So, it will say whether the feature is present or the feature is absent and correspondingly, the probabilities will be something like this. So, every feature component will be represented by a probability value where, the probability is something like this, that I will represent pi to represent the probability of the ith component. So, the pi will be equal to probability that, component xi is equal to 1 given that, the true state of nature or the true class is omega 1. We are concentrating a two case problem.

So, we have classes omega 1 and omega 2. So, pi represents probability that xi is equal to 1 given that, the true state of nature is omega 1 similarly, qi is probability that, xi the same component is equal to 1, given the true state of nature is omega 2. So, given this type of probability measures, it simply means that if I have a situation that pi is greater than qi. In such cases, it simply means that ith component xi is more likely to have value 1 if, x belongs to class omega 1.

Because, it is the probability of assuming a value equal to 1, when the true state of the nature of the two classes omega 1 or when the true state of the nature of two classes omega 2. So, if pi is greater than qi that simply means that, if the sample is taken from class omega 1. Then it is more likely that the ith component x1 will have value equal to 1 or xi will have a value equal to 1, more frequently if the vector is taken from class omega 1. And if it is taken from class omega 2 then it is less frequent that the ith component xi will have a value equal to 1.

So, when I have this kind of situation now, let us see that what are the cases in which these kind of feature vectors are more useful. Say for example, if we want to find out the health of a plant say, power plant, if I want to determine the health of a power plant. Then what, we normally do is there are number of sensors which are used to monitor different parameters of the plant. And after monitoring those different parameters, you decide whether the plant is or there is some danger in the plant.

And when you sensor, when you monitor the sensor outputs you just see that, whether the sensor output is above a threshold level or below a threshold level. So, if it is above a threshold level, we set a value equal to 1, if it is below a threshold level we set a value equal to 0. Let me take a more obvious example, when I go to the market to purchase oranges, you must have noticed that even in our tech market you usually get two types of oranges. One type of oranges, which are produced at Nagpur and one type of orange which are coming from Darjeeling. Have you noticed any difference in appearance between these types of oranges.

Student: Color is different

Color is different, if it is from Darjeeling color is more attractive. It is really orangy color, it is more yellowish whereas, oranges from Nagpur they are more greenish and if it becomes quite old, it becomes more reddish. If you look at the surface texture, the oranges which are coming from Darjeeling, they are smooth whereas, the oranges which are taken from Nagpur, they are rough. So, if simply based on these two features I want to determine, I want to have an automated machine which will simply classify, tell me that whether these are Darjeeling orange or a Nagpur orange. So, it will try to take the decision based on the color in the simplest case or it will try to take the decision based on the feature, the texture feature.

So, if I keep some of the feature vectors like, whether the color is yellowish answer will be either yes or no, whether it is greenish either it will be, answer will be either yes or no whether it is smooth, answer will be either yes or no. However, when I get an orange from Darjeeling and I take an orange from Nagpur, there is no guarantee that all the oranges from Darjeeling, will always have smooth texture or will always have orangy or yellowish color. Or if take oranges from Nagpur there is no guarantee that, I will always have greenish color, Nagpur even may produce some oranges which will have yellowish color or which will have smooth textures.

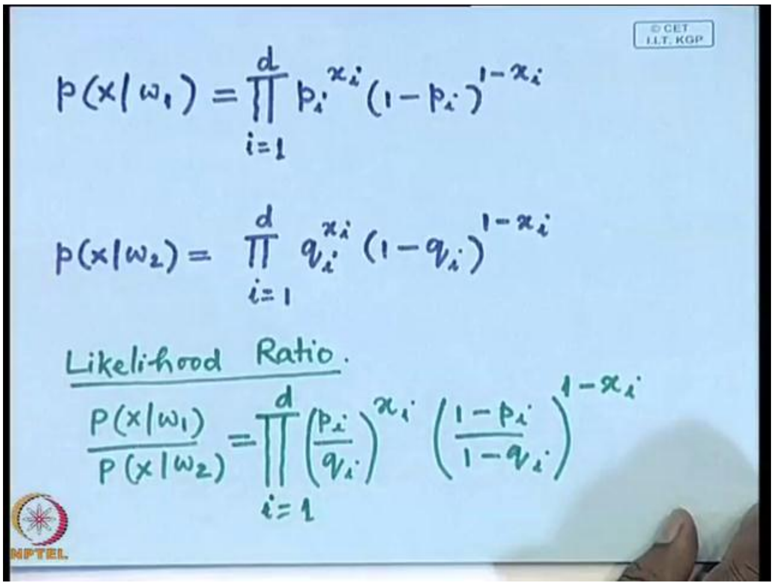
So, there is always a finite probability that an orange produced at Nagpur will have yellowish color. So, it is not necessary if that color to be yellowish, I put that as a binary feature, for all the oranges coming from Nagpur that binary value will always be equal to 0. It is not guaranteed, for some of them I may get values which are equal to 1, but that is less frequent than, the value equal to 1 when the oranges are taken from Darjeeling.

So, simply over here if that feature I put as the ith feature xi then pi will be more frequently equal to 1, xi will be more frequently equal to 1, if this omega i is Darjeeling. And this will be less frequently equal to 1, if omega 2 is from Nagpur so coming over here if pi is greater than qi. So, it simply explains this particular situation, that is for oranges coming from Darjeeling, I will have more number of oranges having yellowish color than, the number of oranges I get with yellowish color from the Nagpur oranges. So, this is a typical situation like this.

And what does conditional independence mean, the texture and the color they are independent. I mean if the texture is rough, that does not necessarily mean that the color will be greenish or if the texture is smooth, that does not necessarily mean that the color will be yellowish. So, coming to the plant if I take two features say, I want to monitor the health of a boiler. If I take two features, one is pressure inside the boiler and temperature inside the boiler.

They are not independent because if temperature increases the pressure will increase. I mean just from the basic laws of gauss so they are not independent they are dependent. So, when I try to select the features I should select in such a way that the features are independent because that solves many of the mathematical problems, I do not have to look for complicated mathematics. So, if I assume that the features are conditionally independent.

(Refer Slide Time: 21:17)



Then, the same probability function I can write as p of x given omega 1, where x is the feature vector, which has d number of binary valued feature components. So, this p of x given omega 1, I can simply write as pi, pi is the probability that xi equal to 1, given the true state of nature is omega 1. But it also has a finite probability that, it may have a value equal to 0 so I have to concentrate on that as well.

So, it is pi to the power xi into 1 minus pi to the power 1 minus xi obviously, if xi equal to 1, 1 minus xi equal to 0, if xi equal to 0, 1 minus xi equal to 1, and because the components are conditionally independent so the overall probability will be product of independent probability values. So, this I have to take for i is equal to 1 to d, as I have d number of components so this is the class conditional probability of a feature vector x, if the state of nature is omega 1.

So, in the same manner I can write p of x given omega 2, that is class conditional probability if the feature vector belongs to class omega 2 is nothing but. Now, the probability of xi equal to 1, then the true state of nature is omega 2 is qi. So, I will have qi to the power xi into 1 minus qi to the power 1 minus xi, take the product from i equal to 1 to d.

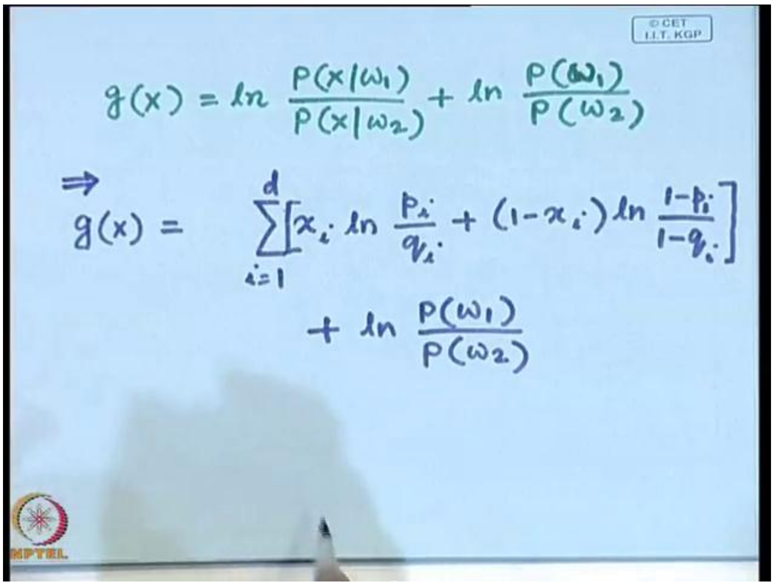
Student: Sir where d is dimension of the.

D is the dimension of the feature vector, so I have d number of components in the feature vector. So, these are the two class conditional probability values now, from here I can find out, what is called likelihood ratio. So, the likelihood ratio is given by p of x given omega 1 upon p of x given omega 2 which is nothing but if I simply multiply these two, it becomes pi upon qi to the power xi into 1 minus pi upon 1 minus qi to the power 1 minus xi. Take the product where, I varying from one to d so this is what is the likelihood ratio right.

Student: Sir xi could be a vector or Xi is a single component?

xi is the single component, it is a scalar, xi is the single component, the capital x is the vector, having d number of components so but the value of i will vary from 1 to d.

(Refer Slide Time: 25:56)

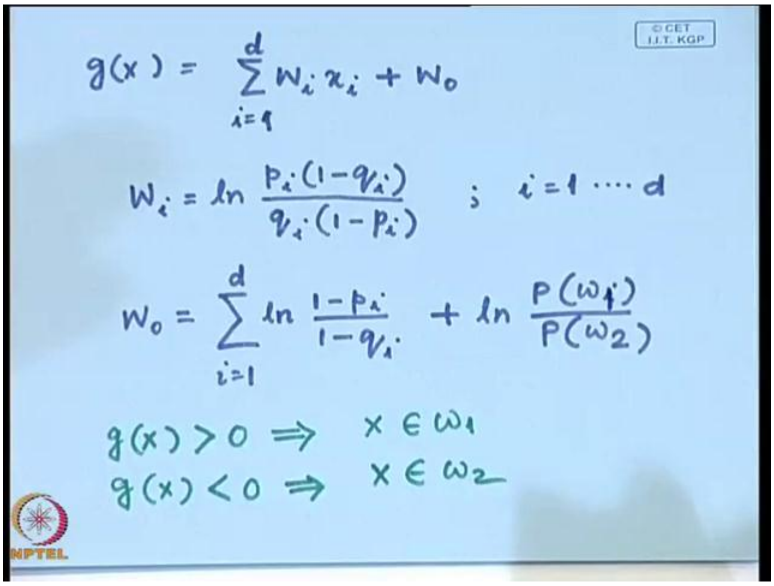


Now, you know that the decision surface or our decision function between the two classes given, two classes is given by gx is equal to log of px, given omega 1 upon p of x, given omega 2 plus log of p omega 1 upon p omega 2. This we derived earlier now, if we put this p of x given omega 1 upon p of x given omega two, which is equal to this expression, if I put this expression into this function.

So, it will give us because we are taking logarithm so the product term over here will be converted to sum of logarithmic functions. So, what I simply get is gx is equal to summation of xi because here xi was a power. So, this will become a product of the, in the logarithmic term so it is xi then log of pi upon qi plus 1 minus xi log of 1 minus pi upon 1 minus qi, where i will vary from 1 to d plus log of p of omega 1, where this p omega 1 is the apriori probability upon p of omega 2. So, this is the decision function and for a two category case we have already seen that, if g of x becomes greater than 0 then our decision was that x belongs to class omega 1. If g of x becomes less than 0 then our decision was that x belongs to class omega 2. If g of x is equal to 0, that actually tells us that what is the decision boundary between the classes omega 1 and omega 2.

Now, if you notice that this equation is a linear equation, isn't it? Because this simply says, the linear combinations of different components xi of the feature vector x, I do not have any xi term, xi square term or xi cube term. So, the equation is a linear equation and this linear equation can simply be written in the form, if I just rearrange this particular linear equation I can write it in the form.

(Refer Slide Time: 29:31)



gx is equal to sum of wi xi plus w0, do not confuse w with omega. So, it becomes sum of wi xi plus w0, where i varies from 1 to d, because I have d number of components in the feature vector. So, you find that this is, nothing but a linear combination of the different components, of the feature vector xi plus a threshold term, which is w0. And this wi is for different values of I, this represents a wide vector and this is, nothing but a dot product or inner product of the wide vector with the feature vector.

So, when I write it like this, over here this wi is simply log of pi into 1 minus qi that is quite obvious. Because, you find that this becomes xi into log of pi minus xi into log of o1ne plus pi or plus xi into log of 1 minus qi. So, this 1 minus qi term that goes to the numerator and 1 minus pi term that comes to the denominator.

So, it simply becomes log of pi into 1 minus qi upon, log of pi into 1 minus qi upon qi into 1 minus pi where, i varies from 1 to d. Because, I have d number of components in the wide vector as well and w naught, that is the threshold is given by log of 1 minus pi upon 1 minus qi, sum of this for i is equal to 1 to d plus log of p of omega 1 upon p of omega 2.

Now if you analyze this, as I said that our decision will be that if, g of x is greater than 0 then we decide that x belongs to class omega 1. If g of x is less than 0 then we decide that x belongs to class omega 2, if this is equal to 0 that is a boundary case. So, if you analyze this expression or what do we get, what does these different components of w, that is wi, that effectively tell us. You find that xi that is the ith component of the feature vector x is the binary value feature vector. It can have a value equal to 0, it can have a value equal to 1.

Now, if it has a value equal to 1 then the contribution of the term wi into xi for that particular component xi, to this function gx is, nothing but equal to the magnitude of wi. Because, xi is equal to 1 so it is nothing but equal to the value, the magnitude of that particular component of wi. So, effectively this wi, magnitude of it simply tells you that what is the importance or what is the relevance of the component xi in decision making that, whether the sample will belong to class omega 1 or the sample will belong to class omega 2. If value of wi is very large then xi has more weightage to decide about the class, if value of wi is small then xi has less weightage to decide about the class.

And in other case, if pi is equal to qi that is value of xi to be equal to 1 is more likely, is same equally likely, even if the x belongs to class omega 1 or the feature vector x belong to class omega 2. So, pi is equal to qi, pi as we said it is the probability that xi will be equal to 1, if the true state of nature is omega 1. And qi is the probability that, xi will be equal to 1 if the true state of nature is omega 2. So, if pi is equal to qi that simply indicates that, whether x belongs to class omega 1 or x belongs to omega 2, xi is equally likely to have value equal to 1. So, that simply means that xi has no relevance in deciding the class.

So if i has, if xi has no relevance in deciding the class then why should the corresponding vector wi be there. I can make wi equal to 0, without hampering my decision. So, if you come to this wi, the expression for this wi you have find that if pi is equal to qi then this expression pi into 1 minus qi upon qi into 1 minus pi. This expression will be equal to 1 log of this is equal to 0 so the corresponding wi is equal to 0. And that is quite obvious because if pi is equal to qi then xi, the particular feature vector xi has no relevance in deciding the class of the feature vector x.

On the other hand if pi is greater than qi, if pi is greater than qi then having value of xi equal to 1 should tell me that, the sample is more likely to belong to class omega 1 than, to belong to class omega 2. Whereas, if pi is less than qi then the sample xi equal to 1 tells me that it is more likely to belong to class omega 2 than, its likelihood to belong to class omega 1.

So, again coming to this particular case, if pi is greater than qi, if pi is greater than qi then obviously 1 minus qi will be greater than 1 minus pi. So, in this expression the numerator becomes larger than the denominator so this value is greater than 1. And when this value is greater than 1, value of wi is positive, if value of wi is positive what happens to my gx.

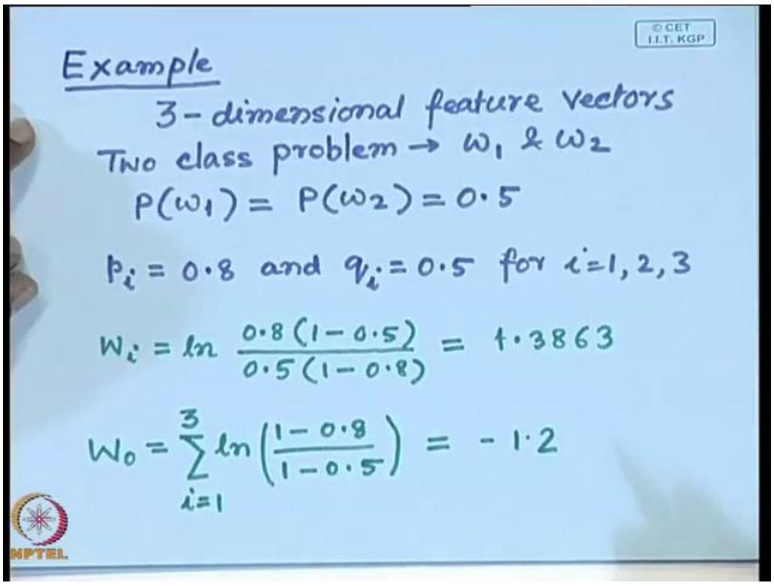
Student: Positive.

Not necessary, it depends on other values of i as well. So, effectively I can say that if xi is equal to 1, that particular component then this component xi gives a vote of value wi to gx, to decide that this feature vector x belongs to class omega 1. So, it is the, it gives the vote equal to the corresponding wide in favor of class omega 1.

On the other hand if pi is less than qi, pi is less than qi so 1 minus qi will be less than 1 minus pi so numerator becomes less than the denominator. So, this term is a fraction which is less than 1, log of this will be negative that means, wi in that case is negative. If wi is negative that means, the corresponding xi into wi is trying to make gx less than 0, trying to make, whether it will be less than 0 or not, that depends upon other wi into xi term. But what xi is trying to do in this case, it is trying to give a vote equal to the modulus of wi, in favor of class omega 2. Because, it is subtracting so it is giving a vote which is equal to modulus of wi in favor of class omega 2.

So if pi is greater than qi, the component xi gives a vote equal to wi in favor of class omega 1, if pi is less than qi then component xi, gives a vote equal to modulus of wi in favor of class omega 2. That means this component is trying to push the decision surface of the decision boundary, either towards omega 1 or towards omega 2.

(Refer Slide Time: 40:15)



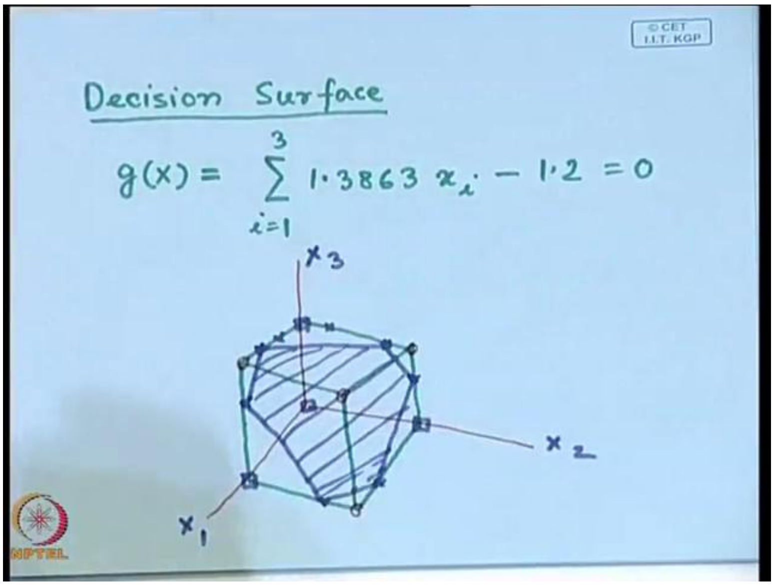
Let us take an example so let us consider a three dimensional space or three dimensional feature vectors. And let us so and let us consider a two class problem that is, I have classes omega 1 and omega 2. So, these are the two classes I have and every feature vector is a three dimensional feature vector where, every individual component can be either 0 or 1.

And let us also assume that, the apriori probabilities p of omega 1 is same as p of omega 2, which is equal to 0.5. And let us assume that value of pi is equal to 0.8 and qi is equal to say, 0.5 for all values of i, that is for i varying from 1, 2 and 3. So, it says that every component of the feature vector has a probability of being equal to 1 is 0.8 if the feature vector is taken from class omega 1. And every component has a probability of 0.5 of being equal to 1, if the feature vector is taken from class omega 2.

So given this pi and qi values a probability values, I can compute the corresponding wide vectors, so simply from this expression that, wi is equal to log of pi into 1 minus qi upon qi into 1 minus pi for different values of i, I get different components of the wide vector wi. So, I get wi is equal to log of pi, that is 0.8 into 1 minus qi, that is 0.5 upon qi that is 0.5 into 1 minus pi that is 0.8. And if I compute this, this becomes a value 0.3863 and the threshold w naught, which is given by this expression, w naught is equal to log of 1 minus pi upon 1 minus qi, take the summation for i is equal to 1 to d plus log of p omega 1 upon p omega 2.

Now over here, in this case p of omega 1 and p of omega 2 they are equal to, they are equal and both equal to 0.5. So, this last term log of p omega 1 upon p omega 2 that will be equal to 0. So, what I have to compute is simply this term, w naught is equal to log of 1 minus pi upon 2 minus qi take the summation over, i is equal to 1 to d. So, this w naught will be simply log of 1 minus pi that is 0.8 upon 1 minus qi that is 0.5. This component takes the summation for i is equal to 1 to 3 and we will find that this value will be something like, it will have a value something like this.

(Refer Slide Time: 45:30)



So given this, the decision surface between the classes omega 1 and omega 2. You verify this values, whether you are truly getting this values or not. So, assuming this our decision surface gx will be given by sum of 1.3863 xi, where this i varies from 1 to 3 because I have three numbers of components. So, this is nothing but 1.3863 x1 plus 1.383 x2 plus 1.383 x3 minus 1.2 this is equal to 0.

So, if I try to plot this decision surface the, in three d one point you might have noticed that, because our features are binary features. Every feature component can assume a value either 0 or equal to 1, so every feature vector will be represented by a vertex of a hypercube in the d dimensional space. It can be either 0 0 0 or 0 0 1 or 0 1 0 or 1 0 0 and so on. So, every feature vector will be represented by a vertex, in the d dimensional space of a hypercube.

So, if I try to plot this decision surface, the decision surface will be something like this. So, let us take a cube in this three dimensional space and if you plot the surface, the surface will come out to be something like this. So, this is our decision surface so you find that, this decision surface says that, these are the points which lie on one side of the hyper plane. And these are the points which lie on the other side of the hyper plane, so it simply says that, if at least two vectors of the feature, if at least two components of the feature vector are equal to 1. Then the point is classified to class omega 1.

Student: Sir it should be at least one because equation says it will be at least one ((Refer Time: 49:34)), which one. The equation from this decision surface that you ((Refer time: 49:40)), this one.

Student: Yes sir.

Student: If at least one is ((Refer time: 49:52)).

I mean it is the other way, these are the points which are put to class omega 1, omega 2 and the other side is put to class omega 1. So, it says if at least one of them is equal to 1 then the decision will be in favor of class omega 1. Otherwise, the decision will be in favor of omega 2.

So, we find that I again get a simple hyper plane in three dimensions, it is just a plane, which is boundary between the two classes omega 1 and omega 2. So, if the probability values are different, if different pis have different other values then the position as well as orientation of this plane may be different. But effectively what it does is, the inter d dimensional space is broke into two halves, one half will be given to class omega 1, the other half will be given to class omega 2. The nature will be a bit more complicated when, the numbers of classes are more than 2, because then we have to think of more than one decision surfaces and how they combine.

Student: Sir, will axis represents x1, x2?

Axis represents x1, x2, x3.So, let us stop here today.

Thank you.