Course Name: **Machine Learning and Deep learning - Fundamentals and Applications**

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-2

Lecture-8

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and application. In my last class I discussed the concept of Bayesian decision theory. I explained the concept of the Bayesian decision theory for discrete features. After this I discuss or I explain the concept of normal distribution. I have shown the expression for the normal distribution that is the Gaussian distribution and one is the univariate normal density and another one is the multivariate normal density. Today I will continue the same discussion.

The concept is actually the Bayesian decision theory for normal distribution. So in this case you know in case of the Bayesian decision theory we have to consider the class conditional density that is the probability of X given omega i that is nothing but the likelihood or I can say it is a class conditional density. If the class conditional density follows normal distribution then what will be the nature of the decision boundary between the classes. So suppose if I consider large number of features corresponding to a particular class then this distribution the distribution is the probability of X given omega i that follows the normal distribution as part of central limit theorem

and this concept is actually nothing but the supervised learning.

So for each and every classes I have training samples I have feature vectors and if I consider large number of feature vectors then this density the probability of X given omega i that is the class conditional density it follows normal distribution. So based on this the Bayesian decision theory for normal distribution I will explain what will be the nature of the decision boundary between the classes. So let us start the class. The class is the Bayesian decision theory for normal distribution.

So in my last class I have shown the normal distribution.

So if I consider the multivariate distribution that is I have two parameters one is the mean vector and another one is the covariance matrix. So corresponding to this my density the normal density that is the multivariate normal density is twice pi D by 2 and determinate of the covariance matrix we are considering an exponential minus 1 by 2 X minus here I can write X bar X minus mu transpose X minus mu. So this is the expression for the multivariate normal density. So for this what we have considered we are considering the D dimensional vector the vector is X.

So if I consider this is the feature vector the D dimensional feature vector and this is my covariance matrix expected value and X minus mu transpose.

So this is my covariance matrix if I consider D is equal to 1 this is not a vector. So X is a random variable and corresponding to this I have the univariate density and in this case I have two parameters one is the mean another one is the variance. In case of the multivariate I have two parameters one is the mean vector and another one is the covariance matrix. And yesterday in the last class I have shown the nature of the covariance matrix the covariance matrix is like this sigma 1 square sigma 2 square like this I have the diagonal I have the diagonal elements like this and what are the off diagonal elements sigma 1 2 sigma 1 D sigma 2 1 sigma 2 D sigma D 1 sigma D 2 like this. So this is the covariance matrix.

So the diagonal elements are the variances of respective Xi and the off diagonal elements are the covariance between Xi and Xj. So if I have only the diagonal elements suppose only I have the diagonal elements and suppose all the off diagonal elements these are 0. So these are 0s suppose this is 0 and all these elements are 0. So that means Xi and Xj are statistically independent they are statistically independent. So that is the case.

So that means the Feature vector if I consider X is a Feature vector the Features are uncorrelated. I am repeating this if I consider X is a Feature vector and X1 X2 Xd these are the components of the Feature vector or I can say these are the elements of the Feature vector or I can say these are the Features individual Features X1 X2 these are the Features. So Features will be uncorrelated if I have the diagonal covariance matrix the off diagonal elements are 0 then in this case I can say Xi and Xj are statistically independent or I can say suppose if I consider the Feature vector that the Features are uncorrelated. And after this I discussed the concept of the Mahalanobis distance.

So this is the distance root over X minus mu transpose.

So this is a famous distance and this distance is called the Mahalanobis distance. So later on I will explain the importance of this distance. And if I consider suppose these are the samples some of the samples for a class suppose the class is omega 1 and these are the some samples for another class that class is suppose omega 2. So I have suppose two classes. So the center of this particular cluster is determined by the mean the mean vector mu 1 and similarly the center or the centroid of the second class is mu 2.

So center of the first class is mu 1 that is the mean and center of the second class is mu 2. So center of the cluster is determined by the mean vector and the shape of the cluster if you see the shape of this cluster the shape of the cluster is determined by the covariance matrix. So last class I discussed about these concepts. So this is the example of 2D Gaussian. So you can plot the 2D Gaussian maybe in the MATLAB also you can plot.

So this is one example of the 2D Gaussian. Now come to the main point. The main point is the Bayesian decision theory and I want to determine the decision surfaces that is a decision boundary between the classes. So what will be the nature of the decision boundaries and in this case we are considering the class conditional density follows normal distribution.

So let us see the mathematical analysis for this decision boundaries the decision surfaces.

So let us move to the next slide. So likelihood function what is the likelihood function is probability of X given omega i. So this is the class conditional density and suppose this density it follows the normal density. So it is supposed twice by d by 2 and this is the covariance matrix for the class i class omega i 1 by 2 and exponential.

So we are considering a d dimensional Feature vector.

So this is the density that is the class conditional density or maybe we can consider likelihood function omega i with respect to X. So the class conditional density or the likelihood function it is a Gaussian distribution and we are considering the C number of classes 1 2 suppose C number of classes. So in this case you in this expression we are considering the d dimensional Feature vector X is a d dimensional Feature vector X 1 X 2 X d. So this is a Feature vector X and we are considering C number of classes. So this is the covariance matrix this is a covariance matrix for the class omega i.

So what is actually this covariance matrix? The covariance matrix for the class omega i is nothing but the expected value X minus mu i X minus mu i transpose. So you know how to determine this. So I should write like this this is the actually for the class omega i and that means actually there is a covariance matrix for the class omega i and what is the mean vector? The mean also you can compute mean is nothing but the expected value of X the mean vector you can determine like this. Now let us consider the discriminate function. What is the discriminate function? Discriminant function that is g i x is equal to l n.

So we know this expression this is the expression for the discriminate function. So considering this because we have considered that this class conditional density or the likelihood function follows the normal distribution. So based on this I can determine the discriminate function based on this condition. So what will be the discriminate function? It is nothing but it is minus 1 by 2 X minus mu i transpose. This is a covariance matrix for the class omega i plus l n plus c i.

c i is a constant suppose so let us consider this as equation number 1. What is actually c i? c i is a constant so it is nothing but this d by 2 l n twice pi minus 1 by 2 l n. So this is c i. So we can find the expression for the discriminate function you can see.

So from this actually from the expression from the class conditional density we can determine the discriminate function.

So let us move to the next slide. So if I expand the previous equation expanding expanding the previous equation so you can see g i x I can write like this g i x is nothing but minus 1 by 2 X transpose. So this is the equation. So just you need to expand the previous equation and that is very simple. So this is suppose the equation number 2.

So this is by expanding the previous equation. So here you can see we have a quadratic term here this is the quadratic term and this equation is nothing but non-linear quadratic form I can consider non-linear quadratic form. So we have the quadratic term is also there. So you can see this term the first term is the quadratic term. So suppose I can give one example suppose d is equal to 2 dimension is 2 then corresponding to this you can determine the covariance matrix the covariance matrix will be simply sigma i square 0 0 sigma i whole square this is the covariance matrix and from equation number 2 from equation number 2 I will be getting the discriminate function g i x is equal to minus 1 by 2 sigma i square x 1 plus x 1 square plus x 2 square plus 1 by sigma i square mu i 1 x 1 plus mu i 2 x 2 minus 1 by 2 sigma i square mu i 1 square plus mu i 2 square plus so I can write this expression for the discriminate function.

$$g_i(x) = -\frac{1}{2}x^T\sigma_i^{-1}x + -\frac{1}{2}x^T\sigma_i^{-1}\mu_i - \frac{1}{2}\mu_i^T\sigma_i^{-1}\mu_i + \frac{1}{2}\mu_i^T\sigma_i^{-1}x + ln\, ln\, P(w_i) + C_i$$

also you should remember that mu into the x t is equal to x mu t. So you have to apply this also to get this equation number 3 from equation number 2 from equation number 2 you can apply this one to get equation number 3 from equation 2 to equation number 3. So this is the general from of the discriminate function this is general from the discriminate function. So in my previous classes also I have shown this from so this is the weight vector is Wi and also I have the bias and I have shown that the decision surfaces or the decision boundaries are the hyper planes. Now I want to show or I want to locate or I want to fix the decision boundary between the classes.

So let us see how we can do this. So let us move to the next slide the case number 1 I am considering. So in the case number 1 we are considering this case diagonal covariance matrix with equal elements. So we are considering this case that is the diagonal covariance matrix with equal elements. So what is the meaning of this? The meaning is this meaning actually the Feature vector is mutually uncorrelated of same variance. So this is the meaning of this diagonal covariance matrix with equal elements.

Sorry it should be equal is there. Diagonal covariance matrix with equal elements. So this point we are considering diagonal covariance matrix with equal elements that means the Feature vector is mutually uncorrelated and of same variance. So corresponding to this my covariance matrix is sigma square I, I is the identity matrix. So it is a d dimensional identity matrix. So corresponding to this from the equation number 3 the equation number 3 I can write like this g i x is equal to 1 by sigma square mu i transpose x plus w i naught.

So I can write like this. Since the one point you should remember sigma inverse is nothing but 1 by sigma square sigma whole square sigma square and i is the identity matrix I can write like this. So equation number 3 I can write like this. So what about the decision hyperplanes? Decision hyperplanes g i j x is equal to g i x minus g j x and that is equal to w transpose x minus x naught that is equal to 0. So this is the equation of the decision boundary. Suppose w transpose x 1 plus w naught this w transpose x 1 plus w naught I can say is equal to suppose w transpose x 2 plus w naught.

So w naught so from this I can write w transpose x 1 minus x 2 is equal to 0. Actually I am applying this to get this one. I am applying this to get this one. So you can see I am getting the equation of the decision hyperplanes and the weight vector is nothing but the weight vector is nothing but the difference between these two means mu i and mu j. The weight vector is nothing but the difference between these two means and also what is x naught? x naught is a point actually x naught is nothing but 1 by 2 mu i plus mu j minus sigma square ln mu i minus mu j.

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 ln(\frac{P(w_i)}{P(w_j)}) \frac{\mu_i - \mu_j}{||\mu_i - \mu_j||^2}$$

So this is x naught. So these are very important these two equations are very important. One is I have determined the expression for the decision hyperplanes that is g i z and another one is the x naught. So this decision surfaces decision surface is a hyperplane passing through the point passing through the point what is the point? The point is x naught passing through the point x naught. So these two equations one is the equation of the decision hyperplane so maybe I can change my color of the ink. So this equation this is the equation for the decision hyperplane and also we are considering the x naught x naught is a point through which the hyperplane is passing.

So these two equations are very important one is the equation of the decision hyperplane and another one is the x naught that is the decision surface is a hyperplane passing through the point the point is x naught. So now I want to determine what will be the decision boundary. So let us move to the next slide. So we obtain these two equations g i z x that is the decision hyperplane that is nothing but W transpose x minus x naught. So this is the important equation this is the equation of the decision hyperplane and what is the weight vector? The weight vector is nothing but the difference between these two means mu i and mu j and what is x naught? x naught is the point through which the hyperplane is passing.

So it is mu i plus mu j minus sigma square ln. So we have these two important equations one is the equation of the decision hyperplane and another one is x naught and one important point is the weight vector is this this is the equation of the weight vector that is nothing but the difference between these two means. Now suppose one condition I am considering suppose this prior probabilities probability of omega i is equal to probability of omega j. So corresponding to this condition what will be the decision boundary? So if I consider this case then x naught is equal to 1 by 2 mu i plus mu j. So x naught will be like this so the vector x naught is 1 by 2 mu i plus mu j that is the expression for this. Then the hyperplane the meaning is actually the meaning of this what is the meaning of this? The meaning is the hyperplane passes through the mean the mean of mu i and mu j.

So that is the meaning of this because I am taking the average of this 1 by 2 mu i plus mu j. So hyperplane passes through the mean of mu i and mu j corresponding to the case the case is if the probability of omega i is equal to probability of omega j and one important sentence I can write that important sentence is the important statement is hyperplane is orthogonal to the vector W that is mu i minus mu j. So W is the weight vector so from this expression you can see because W transpose dot x minus x naught so I have to put d here x naught is a vector W transpose x minus x naught is equal to 0 that is the meaning is so from this expression actually from this expression I can write like this the hyperplane is orthogonal to the vector the vector is the weight vector and that is nothing but the difference between these two means. So this is the important consideration so always you have to follow this one that the hyperplane is orthogonal to the weight vector that is the difference between the means mu i and mu j. Now let us consider the case if I consider so suppose the probability of omega i is less than omega j the meaning is the hyperplane the hyperplane is located closer to closer to the mean mu i so that means if the probability of omega i is less than probability of omega j then hyperplane is located closer to mu i that is it is located closer to the cluster corresponding to the class omega i.

So I have two clusters one is the cluster corresponding to the first class omega i and another cluster corresponding to the class omega j. So corresponding to the first cluster omega i the mean is mu i so if I consider this case the probability of omega i less than probability of omega j then the decision boundary will be located closer to the first cluster that is closer to the mean of the first cluster. Similarly if I consider probability of omega i greater than probability of omega j this second condition then the same thing will applicable the hyperplane is located closer to mu j that means we have to consider the second cluster. So the hyperplane or the decision boundary will be located closer to the second cluster corresponding to the class omega j and if this variance sigma square is small with respect to with respect to the difference between these two means the distance between the two means and we are considering the Euclidean norm the location of the hyperplane location of the hyperplane is rather insensitive to the values of these two prior probabilities probability of omega i and probability of omega j. So that means the sigma square that is the variance is small with respect to the difference between these two means the location of the hyperplane is insensitive to the values of these two probabilities.

What is the meaning of this variance the variance is small the small variance I can say what is the meaning of small variance and small variance means random vectors random vectors are clustered within a small radius around their mean values. So random vectors are clustered within a small radius around their mean values. So that means it is nothing but the compact clusters the clusters are very compact clusters. So if I consider small variance the random vectors are clustered within a small radius around their mean values. So that means the compact clusters for the compact clusters the location of the hyperplane is insensitive to the values of the probability probability of omega i and probability of omega j.

That means the meaning is for the compact cluster you have sufficient independence or sufficient freedom to place the decision boundary between the classes. That means it is easy to place the decision boundary between the classes for the compact clusters. So now let us see how to get the decision boundary based on these conditions. So one condition already I have explained that is the hyperplane is orthogonal to the weight vector.

So this first condition is very important and second condition is the hyperplane will pass through the point the point is x naught.

So second point is x naught. So already we have derived the equation for x naught. So based on these two conditions let us draw the decision boundary between the classes. So let us move to the next slide. So I am drawing this decision boundary. So suppose we are considering this feature space and two features x 1 and x 2 suppose.

Now let us consider this is the mean. The mean is mu i corresponding to the class. I have two classes. The two classes are omega i and omega j. So these two classes we are considering and another mean we are considering suppose this is the mean of the second class.

So it is mu j. Now I want to determine the difference between these two means. The difference between these two means is this. So this is the difference between these two means that is nothing but this vector is mu i minus mu j and that is nothing but the weight vector. That is nothing but the weight vector mu i minus mu j is nothing but the weight vector. And after this I am considering another vector that vector is x naught.

So we have derived the equation for x naught. Now how to draw the decision boundary. So I have to draw the decision boundary. The decision boundary should be orthogonal to the weight vector and it should passes through the point.

The point is x naught. So this is my decision boundary. This is the decision boundary and you can see it is orthogonal to the weight vector. It is orthogonal to the weight vector. Weight vector is nothing but the difference between mu i and mu j. So this is I can say as decision boundary between these two classes and in this case you can see the decision boundary is orthogonal to the weight vector and it passes through the point x naught.

So in this case we have considered that covariance matrix is sigma square i. So already I have explained that the diagonal covariance matrix with equal elements that means the Feature vector is mutually uncorrelated and has same variance. So i is the identity matrix. So this is the procedure to draw the decision boundary. Now let us consider that is we have considered that variance is very small which respect to this difference between these two means.

So this is nothing but the compact clusters. So for the compact clusters what will be the decision boundary. So let us draw the decision boundary for the compact clusters.

The procedure is same. Same procedure we have to apply. So x 1 and x 2. So we are considering the Feature space and first I am considering this vector that is nothing but the mu i and we are considering this is the cluster and another is cluster and that is suppose the cluster is something like this and this is the mu j one is mu i and another is mu j. This is mu i and mu j and that is corresponding to the class omega i and this is the class omega j. So these are compact clusters.

So the procedure is this. So I have to determine the weight vector. The weight vector I can determine that is nothing but the difference between these two means. So this is the weight vector.

So this is weight vector is w and that is nothing but the difference between mu i and mu j.

After this we are considering the point. The point is x naught. This is the x naught and after this I have to draw the decision boundary that is the bisector I have to show. So this decision boundary will be perpendicular or orthogonal to the weight vector and it is passing through the point. The point is x naught. So this is for the compact case.

So you can see it is easy to place the decision boundary between the classes. So these are the compact clusters and actually the compact means samples with high probability. So if the samples with high probability then I will be getting the compact clusters. Samples with high probability means is a compact clusters. So for the non-compact clusters it is very difficult to place the decision boundary.

So I can show that one also pictorially. So the same procedure. So this is suppose one cluster and suppose this is another cluster. So corresponding to the first cluster I have the mean. The mean is mu i and corresponding to the second cluster I have the mean mu j and we can find we can find the weight vector the procedure is same. So this is the weight vector and the point also we have to show the point is x naught.

So this is x naught. Now I have to draw the decision boundary. Now this is the decision boundary that is also orthogonal to the weight vector and it is passing through the point x naught. So this case is the non-compact case. So in the non-compact case the location of decision hyper plane is much more critical. It is very difficult to place the decision boundary between the classes. That means the location of the decision hyper plane is much more critical as compared to the compact case.

So this is the procedure how to get the decision boundary between the classes how to find the location of the decision boundary between the classes. So first point you have to remember that the decision boundary should be orthogonal to the weight vector. The weight vector is W and that is the difference between these two means mu i and mu j.

The second point is the decision boundary should be passing through the point. The point is x naught. This is for case number 1. In case number 2 we will be considering the covariance matrix is not diagonal. The covariance matrix is same for all the classes but it is not a diagonal covariance matrix. In case number 1 we consider the diagonal covariance matrix.

So this case number 2 I will be discussing in my next class. So what is the decision boundary for the non-diagonal covariance matrix. So that we have to discuss and after this I will discuss the minimum distance classifiers based on Euclidean distance and based on Mahalanobis distance. So in this class I discussed the concept of the Bayesian decision theory. I have explained the concept how we can determine the decision boundary between the classes. I have considered that the class conditional density that is the probability of x given omega i it follows the normal distribution.

And based on this I have determined the expression for the discriminate function. After determining the expression for the discriminate function I want to determine the hyperplanes or the decision boundary between the classes. So in the case number 1 I have considered the diagonal covariance matrix. For the diagonal covariance matrix the decision boundary will be passing through the point, point is x naught. And also the decision boundary is orthogonal to the weight vector, the weight vector is w.

The weight vector is nothing but the difference between the mean mu i and mu j. So based on these two conditions the first condition is the decision boundary should be orthogonal to the vector, the vector is the weight vector. And it should pass through the point, the point is x naught. This is for case number 1. The case number 1 is the covariance matrix is same for all the classes and we are considering the diagonal covariance matrix.

In the next class I will be considering the case number 2. In case number 2 we are considering the covariance matrix is same for all the classes but it is a non-diagonal covariance matrix. So for this I have to determine the location of the decision hyperplane and how to determine the decision hyperplanes and the decision boundary. So that concept I will be explaining in my next class. So let me stop here today. Thank you.