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-6

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of Bayesian decision theory, how you can use the Bayes theorem for decision making. I discuss the concept of posterior probability. So, you can determine posterior probability, probability of omega j given x from the information, the information is likelihood information and the prior information. And in case of the Bayes law, that evidence has no role in classification. It is simply a scaling factor, because it is same for all the classes.

So, this term has no importance the evidence. After this, I discussed the concept of probability of error. So, you can take a classification decision based on this parameter, the parameter is nothing but the probability of error. After this I discussed the concept of loss.

So, suppose x is a Feature vector, x belongs to a particular class, suppose class is omega i. So, for this I am taking a particular action. So, action is considered and corresponding to this action, I have determined the loss. And today I will discuss from the loss how to actually determine the Risks. With the help of these Risks, I can take a classification decision.

So, let us start this class. So, first I will discuss the concept of loss. And after this I will discuss how to determine Risks. So, in my last class, if you see what I have considered suppose what is actually the Risks. So I have discussed this is the Bayes theorem probability of omega j given x, that is the posterior probability and this probability of x given this omega j that is the likelihood and this is the prior probability and this is the evidence.

So, already I told you that evidence has no role in classification, it is simply a normalizing factor. It is same for all the classes. So, in this example, I am considering only C number of classes. Now, let us consider this x, x is a feature vector, it may belongs to belongs to the class omega 1 or maybe it may belongs to the class omega 2. So, for class omega 1, I am taking some actions or decision is taken that is nothing but alpha i decision I am considering and corresponding to this the loss I can define like this lambda action alpha i is taken corresponding to the class omega j.

So, what is the meaning of this, the loss means action alpha i I am taking for the class omega j. So, that I can consider as lambda i j. This is the loss, the loss is defined like this. So, in my last class, I have also shown this example, suppose I have this class omega 1, omega 2, omega k, k number of classes. These are the classes and I am taking some actions.

So, action is suppose $\alpha 0, \alpha 1, \ldots, \alpha k'$ and outcome is $\Omega 1, \Omega 2, \ldots$

So, corresponding to this the loss will be $\lambda k'1$, this is the loss. So, for a pattern classification, I can consider all the section, suppose this section if I consider as the reject option reject. So, what is the meaning of reject?

So, in the previous classes, I discussed the concept of the confusion matrix.

So, if you see this is the confusion matrix. So, these are the actual class labels suppose 1, 2, 3, 4, 5 like this and these are the predicted class 1, 2, 3, 4, 5. So, you can see in the confusion matrix, this is the confusion matrix. So, how many times 1 is recognized as 1, suppose if I consider some value suppose 20 and how many times 1 is recognized as 2, so suppose 2 times. So, I can put this value and similarly how many times 2 is recognized as 2, suppose it is 20 1 times and how many times 2 is recognized as 1 suppose 1 time.

So, like this I can determine the confusion matrix, this is a confusion matrix. So, I will be getting a diagonal matrix. So, from this what are the parameters I can determine? One is suppose the accuracy rate I can determine this accuracy percentage I can determine another one is the misclassification rate misclassification rate that also I can determine. So, how many times 1 is recognizes 2, 3, 4, 5 days.

So, from this I can determine the misclassification how many times it is correctly recognized from this I can determine the accuracy percentage.

How many times it is not correctly recognized from this I can determine the misclassification rate the percentage I can determine. And also suppose there is another parameter that rejection rate what is the rejection rate suppose in alphabet recognition suppose a b c d I want to recognize a b c d alphabet suppose I am giving the input something like this. So, this cannot be recognized this is not a this is not b this is not c or this is not d. So, for this the classifier output should be reject the classifier output should be reject corresponding to this input input is this this is my input.

So, corresponding to this input the output should be reject.

So, that is why we are considering this option that reject option we are considering. So, how many times it is rejected. So, based on this I can determine the rejection rate the rejection percentage I can also determine. So, from the confusion matrix you can determine accuracy percentage misclassification percentage and rejection percentage. So, all these you can determine from the confusion matrix.

So, come to this point. So, you can see the loss we can determine by this the lambda i j. So, that means the action alpha i I am considering corresponding to the class the class is omega j. So, now I have defined the loss. Now let us define what is the conditional risk.

So, in my next slide I will explain what is the conditional risk. So, move to the next slide the conditional risk I can determine r i. So, that is called a risk that is risk r action is alpha i and this is taken for the Feature vector the Feature vector is x action alpha i is considered for the Feature vector the Feature vector is x and we are considering c number of classes. So, j is equal to 1 to c lambda i j lambda i j is nothing but the loss and we are considering the probability that the true state of nature is omega j this probability is nothing but it represents the probability that the true state of nature is omega j and that is called a risk. Now from this you can determine total risk you can determine total risk for all x all the Feature vectors.

So, you can determine so total risk for all the Feature vectors you can determine r is nothing but if I have to take the integration. So, you can determine the total risk like this. Now we have to minimize risk. So, how to minimize risk? So, we have to minimize we have to minimize minimize risk the risk is r. So, if this probability omega j given x is greater than probability of omega i omega is a class another class for i is not equal to j two classes we are considering omega j and omega i decide the action alpha j.

So, this action we are considering. So, this is the procedure. So, we have to minimize the risk. Now here you can see we consider this is the conditional risk the conditional risk is ri rx on alpha i is considered for the Feature vector the Feature vector is x.

This is nothing but the conditional risk.

So, it is nothing but the sum total of all the losses. So, that is nothing but j is equal to one to C lambda alpha i omega j probability of omega j given x. So, this already I have defined if you see sum total of all the losses that we have considered. So, this is nothing but j is equal to one to C lambda ij probability of omega j given x. So, already I have defined this one this is actually nothing but the risk.

Now we have to minimize the risk. So, how to minimize the risk and so like this how to minimize? Minimize risks r alpha i given x is r alpha j x for i is not equal to j for i is not equal to j. So, that means I have to minimize the risk. So, for minimization of the risk we are considering this for this we have to consider the action the action is alpha i we have to consider because the risk alpha i given x is less than risk alpha j given x.

So, that means we have to consider the action alpha i.

So, this is the definition of the loss and this is the definition of the risk. So, this we have considered. Now let us move to one example suppose I have two classes for two classes how to determine the risk how to take a classification decision. So, let us move to the next slide.

So, let us consider a two class problem two classes.

So, class omega 1 and omega 2 two classes and we have considered the actions alpha 1 and alpha 2. So, corresponding to the class omega 1 I have lambda 2 1 this loss is lambda 2 1 if I consider the class omega 1 the action is suppose alpha 2 and corresponding to this you can see my loss is lambda 2 1 and corresponding to the class omega 2 if I takes action is suppose alpha 1. So, corresponding to this the loss I can consider as lambda 1 2. So, that means I can take or I can consider these losses lambda 1 1 I can consider lambda 1 1 is nothing but alpha 1 given omega 1. So, action alpha 1 is considered for the class omega 1 and similarly I can determine lambda 1 2 lambda 2 1 and also lambda 2 2.

So, all these losses I can determine. So, after determining these losses I can determine the risk. So, how to determine the risk R action alpha 1 is considered given the x x is the feature vector. So, this lambda 1 1 probability of omega 1 given x plus lambda 1 2 probability of omega 2 x. So, I can consider the risk like this and also another risk I can determine that is for the action alpha 2.

So, Risks alpha 2 given x. So, it is lambda 2 1 probability of omega 1 x plus lambda 2 2 probability of omega 2 x. So, this Risks also we can determine. Now how to take a classification decision. So, decide omega 1 if the Risks alpha 1 x is less than Risks alpha 2 x. So, already this concept I have discussed.

So, based on the Risks I can determine a particular class. So, we are considering the class omega 1 that is the meaning is lambda 2 1 minus lambda 1 1. So, from these two equations I can write like this

$$(\lambda 21 - \lambda 11) \cdot P(x \mid \Omega 1) \cdot P(\Omega 1) > (\lambda 12 - \lambda 22) \cdot P(x \mid \Omega 2) \cdot P(\Omega 2)$$

So, we can write like this. So, that means what we can consider we can consider omega 1. So, this is considered if I select omega 1. So, that means the total condition I am considering for the class omega 1. Otherwise, I have to decide omega 2.

So, this expression also I can write like this.

So, lambda 2 1 is greater than lambda 1 2 minus lambda 2 2 lambda 2 1 lambda 1 1. So, this I am considering as a threshold. And also we assume that lambda 2 1 is greater than lambda 1 1 this condition we are considering. So, that means this part this we can consider as a threshold and this ratio I can consider as likelihood ratio.

So, this is the likelihood ratio I can write like L R.

So, what is the decision rule? Decision rule is if the likelihood ratio is greater than threshold then decide that class omega 1 otherwise the class is omega 2. So, this is the decision rule. So, in this case you can see what is the advantage of the likelihood ratio. So, advantage is this ratio is independent of x independent independent of the Feature vector x.

So, only we have to consider the likelihood. So, based on the likelihood ratio, we can take a classification decision and this likelihood ratio that is independent of the Feature vector x. So, here you can see how we can decide or how we can take a classification decision based on the likelihood ratio. So, based on this concept I am going to explain another concept that is called the minimum error rate classification. So, move to the next slide the minimum error rate classification. So, for minimum error rate classification we are considering one function that function is called 0 1 0 1 loss function which is defined like this lambda alpha i omega j.

So, this is the expression for the loss x 1 alpha i is considered for the class omega j. So, that is equal to 0 if i is equal to j and it is equal to 1 the loss is maximum if i is not equal to j and in this case we are considering the c number of classes c number of classes. So, in this case what is the meaning of alpha i alpha i is the action when the true state of the nature is omega i I am repeating this alpha i is the action when the true state of the nature is omega i. So, for alpha i if the class is omega j decision is correct when i is equal to j otherwise the decision is wrong if i is not equal to j. So, I can write like this for alpha i if the class if the class is omega j the decision decision is correct when i is equal to j and otherwise error if i is not equal to j you can see here if i is not equal to j this loss function value is equal to 1.

So, if alpha i we are considering and class is omega j the decision is correct when i is equal to j otherwise decision is not correct if i is not equal to j if i is not equal to j you can see the loss function value is equal to 1. So, how to define the Risks now Risks are action is alpha i given the feature vector is x. So, j is equal to 1 to c lambda alpha i omega j probability of omega j given x. So, which I can write like this j is not equal to i if j is not equal to i then the loss function value is 1.

So, this value will be 1. So, I can write like this probability of omega j given x. So, that means I can write like this 1 minus probability of omega i given x I can write like this. So, to minimize error or to minimize Risks to minimize Risks select maximum this probability. So, from this expression you can see if probability of omega i given x is maximum then the Risks will be minimum.

So, that means to minimize Risks we have to select the maximum probability of omega i given x if this probability is maximum then the Risks will be minimum.

So decide omega i if probability of omega i given x is greater than probability of omega j given x for all i is not equal to j. So, this is the condition this is the decision rule. So, we can take a decision based on this condition and this minimum error rate classification it is very important. So, for minimum error rate classification we have defined one function and that function is called 01 loss function and based on 01 loss function I have determined the Risks this is the Risks and you have seen how we can take a classification decision by considering this Risks. So, this is about the minimum error rate classifications.

Now in my first class also I discussed the concept of the discriminate function. So, in statistical machine learning or in statistical pattern classification we considered this function the discriminate function for taking a classification decision. So, in my next slide I will explain what is the discriminate function. So, now the discriminate function. So, discriminate function is represented like this g i x and we are considering c number of classes.

So, here you can see for c number of classes I have c number of discriminate function. So, this x can be assigned to a particular class or x will be assigned to that particular class or to that class. So, x will be assigned to that class for which g i x is maximum. So, meaning is x will be assigned to that class for which g i x is maximum. So, here you can see for c number of classes we have to determine c number of discriminate function and we have to find the largest or the maximum discriminate function out of c number of discriminate function and that corresponds to that particular class.

So, the classifier is said to assign a Feature vector x to a class omega i that means x is assigned to the class omega i if g i x that is the discriminate function is greater than g j x for all i is not equal to j. So, here you can see based on the discriminate function we can take a classification decision for all c number of classes we have c number of discriminate function and we have to select the largest one. So, now because we have to select the largest one. So, this discriminate function I can write like this in terms of risks this is nothing but the conditional risks already I told you this is a conditional risks.

So, that means what is the meaning of this equation the maximum discriminate function corresponds to minimum risks.

So, I can write this is a maximum discriminate function is a maximum value of the discriminate function corresponds to minimum risks. So, you can see the maximum discriminate function there is a maximum value of the discriminate function corresponds to minimum risks that minimum risks corresponds to maximum posterior probability the probability of omega i given x. So, that means I can write this maximum discriminate function discriminate function corresponds to corresponds to maximum posterior probability maximum posterior density. So, maximum discriminate function corresponds to maximum posterior density.

So, you can see this I can write in this from probability of x given omega i probability of omega i probability of x by using the Bayes law I can write like this.

So, gix I can write like this because the evidence we are not considering I can write like this. So, gix is nothing but the multiplication of the likelihood and the prior. So, move to the next slide. So, gix we have determined like this gix is the discriminate function probability of x omega i probability of omega i. So, now, we can take the natural logarithm about the size and one thing is important the scaling of gix does not sense the decision making because decision is taken with the help of the discriminate function.

So, scaling will not affect this one. Now if I take the logarithm in this equation the both the sides. So, what I will be getting gix that is the discriminate function and based on this discriminate function I am taking the classification decision. So, this is the expression for the discriminate function based on this discriminate function how you can take a classification decision suppose I have a Feature vector the d dimensional Feature vector I can write like this the Feature vector is x and is a d dimensional Feature vector. Now corresponding to this Feature vector how to take a classification decision. So, these are the components of the Feature vector x2 x3 and xd the d dimensional Feature vector and already I told you for c number of classes I have c number of discriminate function.

So, this is g1 x g2 x and this is gc x c number of discriminate function and you can see I am just giving the input from the Feature vector. Similarly, for g2 x also I am giving the information that means I am giving the input the input is nothing but the Feature vector and similarly for gc also I am giving the input. After this we have to find the largest discriminate function. So, what I can consider we can determine the cost. So, which one is the largest we have to determine and based on this we have to take classification decision classification action we have to take based on this.

So, you can see how we can decide based on the discriminate function. So, the meaning of the discriminate function is we have to divide the Feature space into c regions. So, that means we have to divide the Feature space we are dividing the Feature space into c decision regions. These regions are like this r1 r2 so r c because I have c number of classes. So, how to take a classification decision if gi x is greater than gj x for all i is not equal to j then x this vector is in the region the region is Ri. So, the meaning is x is assigned to the class the class is omega i the x is assigned to the class omega i and what is the equation of the decision boundary the equation of the decision boundary is gi x is equal to gj x this is the equation of the decision boundary.

So, here you can see how we have taken the decision with the help of the discriminate function. Now, let us consider two classification problem. So, move to the next slide. Suppose let us consider two classes g1 x and g2 x.

So, I have two discriminate function one is g1 x another one is g2 x. So, corresponding to this I have two regions what are the regions in the Feature space one is r1 and another one is r2. So, if g1 x is greater than g2 x then what I have to consider then x should be assigned to the class the class is omega 1 otherwise we have to consider otherwise x should be assigned to the class omega 2 and what is the equation of the decision boundary the equation of the decision boundary is g1 x is equal to g2 x this is the equation of the decision boundary. So, that means this equation I can write like this g1 x minus g2 x is equal to 0. So, that means I can simply write like this gx is equal to 0. So, this is nothing but equation of the curve equation of a curve gx is equal to 0 equation of a curve or maybe a straight line or maybe a circle or maybe a curve.

So, what is the nature of the decision boundary we discussed in the next classes. Now, for the time being you can see this is the equation of a curve. So, it may be linear decision maybe I can consider a straight line or I may consider a plane like this we can consider. So, how to show the decision boundary. So, this is a feature space and I have two regions region r1 and region r2 region r1 corresponds to the class omega 1 and region r2 corresponds to the class omega 2.

So, if gx is greater than 0, then I have to consider the class omega 1 and if the gx is less than 0, I have to consider the region r2 and that corresponds to the class omega 2 and this is the decision boundary. So, it is a straight line. So, gx is equal to 0. This is the equation of the decision boundary.

So, for a 2D vector, it is nothing but the equation of a plane. So, you can see this is the decision boundary the equation of the decision boundary gx is equal to 0. So, this gx already I have shown that gx is nothing but g1x minus g2x that I can write like this ln x omega 1 probability of x omega 2 plus ln probability of omega 1 probability of omega 2. So, this is a very important equation for two classes we have shown and that is the discriminant function gx is equal to g1x minus g2x which can be written like this. So that is gx you can write like this again I am writing this equation ln probability of x omega 1 probability of x omega 2 plus ln probability of omega 1 and probability of omega 2.

So, this is the discriminant function for two classes. So, up till now I have discussed this concept one is the concept of the loss and from the loss I have discussed the concept of the Risks and with the help of the Risks we can take a classification decision. So, what is the summary of to this class the summary of that to this class is so first we considered the Risks the Risks minimization framework. So, we are considering C number of classes. So, briefly I am explaining here because already I have explained and for these classes I am considering some of the possible actions alpha 1 alpha 2 alpha a these actions we have considered and based on this we have defined a loss the loss is lambda alpha i given omega j. So, that loss we have considered after this we have considered what is the expected loss that is the equation for the expected loss that is nothing but the conditional Risks.

So, this is the conditional Risks the conditional Risks you can determine like this and this is the expected loss the conditional Risks you can determine by considering this equation and if I consider all the Feature vectors for all x. So, the total Risks you can determine by considering this equation. So, already I have explained this one after this how to take a classification decision. So, if I consider two class problem two category classification we can consider the action alpha 1 or maybe we can consider alpha 2 alpha 1 corresponding to the class omega 1 if I decide omega 1 alpha 2 if I decide the class omega 2 and based on this I can determine the loss and after this I can determine the conditional Risks because I have four losses lambda 11 lambda 12 lambda 21 and lambda 22 already I have defined. So, I can determine the conditional Risks after determine the conditional Risks what is the decision rule if R alpha 1 given x is less than R alpha 2 given x then we have to consider the x on alpha 1 and what is the meaning of this we have to consider or we have to decide the class omega 1 otherwise we have to consider the class omega 2.

So, this is equivalent to decide the class omega 1 if this particular condition is satisfied. So, this already I have explained. So, based on this condition I can decide a particular class this is the concept of the Risks minimization and after this we considered this ratio that is nothing but the likelihood ratio. So, based on this likelihood ratio also you can take a classification decision. So, if this probability of x given omega 1 divided by probability of x given omega 2 that is the likelihood ratio is greater than this value this is the value then I have to consider the x on alpha 1 that means I have to decide the class omega 1 otherwise I have to consider the x on alpha 2 and that corresponds to the selection of the class omega 2.

So, this is the loss we have considered and this we are defining as a threshold the threshold I have already discussed. So, we can decide a particular class if this ratio the likelihood ratio is greater than this threshold the threshold is theta lambda. So, if the likelihood ratio exceeds a threshold value independent of the input pattern x we can take optimal actions this is the summary of this. So, based on the likelihood ratio that is actually the independent of the pattern x then based on this likelihood ratio we can take a classification decision.

After this the most important topic is the 0 1 loss function. So, we have defined the 0 1 loss function and in this case alpha I action is taken if the true state of the nature is omega j the decision is correct if I is equal to j otherwise it is error if I is not equal to j. So, corresponding to this condition this is the 0 1 loss function lambda alpha I given omega j that we have considered and it is equal to 0 if I is not equal to j and it is equal to 1 if I is not equal to j and based on this we have determined the conditional risks and here you can see we have to maximize this to minimize the risks if the probability of omega I given x is maximum then the risks will be minimum. So, this is the concept of the 0 1 loss function. So, minimize the risks requires the maximizing the probability the probability of omega I given x we have to maximize for minimum error rate decide omega 1 if the probability of omega I given x is greater than probability of omega j given x for I is not equal to j.

So, this is the decision rule. So, this is the concept of 0 1 loss function after this I discussed the concept of the discriminant function. So, g I x is the discriminant function and for c number of classes I have c number of discriminant function and we can take a classification decision the decision is x is assigned to the class omega I if g I x is greater than g j x for I is not equal to j. So, based on the discriminant function I can decide. So, this concept also I have explained and after this discriminant function because we have to find the largest discriminant function out of c number of discriminant function. So, maximum discriminant function corresponds to minimum risks and also I can write the maximum discriminant function corresponds to maximum posterior probability.

So, g I x is equal to probability of omega I given x. So, this g I x can be written like this g I x is equal to probability of x given omega I into probability of omega I that is from the Bayes law and after this I can take the natural logarithm both side and I will be getting this the expression for the discriminant function. So, already I told you the discriminant function do not change the decision when scaled by some positive constant k the decision is not affected when a constant is added to all the discriminant function. So, this already

I have explained. So, Feature space now it is divided into c decision regions. So, I have c number of decision regions r 1 r 2 r 3 r c and how to take a decision rule if g I x is greater than g j x if I is not equal to j then the Feature vector x will be in the region r I, r I means x should be assigned to the class the class is omega I for 2 class case the same principle we can extend and the classifier is called dichotomizer that has 2 discriminant function 1 is the g 1 another 1 is g 2.

So, g x we can write like this and decide omega 1 if g x is greater than 0 otherwise we have to consider omega 2. So, what is the dichotomizer dichotomizer is g I x is equal to ln probability of x given omega I plus ln probability of omega I. So, we are considering the natural logarithm after this we can determine g x and with the help of the Bayes theorem we can write the g x in this form.

So, this is the expression for the g x. So, up till now we have discussed about this. So, for taking a decision classification decision we can consider the discriminant function and based on the discriminant function I can take a classification decision. Up till now I discussed the concept of the Risks. So, from the loss how I can determine the Risks after this I discussed the concept of 0 1 loss function. So, with the help of 0 1 loss function I can take a classification decision after this I discussed the concept of discriminant function for c number of classes I have c number of discriminant function.

So, with the help of this discriminant function I can take a classification decision. In my next class I will continue these concepts and mainly I will consider some of the other concepts like what is the expression for the discriminant function for the normal density Gaussian density. So, all these concepts I will be explaining in my next classes. So, let me stop here today. Thank you.