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-3

Lecture-11

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. I have been discussing about the Bayesian decision theory. In the Bayesian decision theory, I have to determine the probability of omega i given x that is the posterior probability. And that can be determined by considering two information one is the likelihood that is the $P(x/\omega_i)$ that is also called the class conditional density. And also another information is the prior probability that is probability of ω_i.

And in the Bayesian decision theory already I told you that in case of the Bayes law, the evidence has no role in classification.

It is same for all the classes. So, it is simply a normalizing factor. So, that should not be considered. So, we need these two information one is the class conditional density the probability of x given omega i and another one is the probability of ω_i.

But in many cases, some information are not available. For example, the class conditional density $P(x/\omega_i)$ the density from is known that is we know what is the density, but we do not know about the values of the parameters. So, for example, if I consider a Gaussian distribution or the normal distribution, it has two parameters, one is the mean another one is the variance. So, if I consider the multivariate case, one is the mean vector and another one is the covariance matrix. So, we do not know about the values of these parameters.

So, we have to estimate these parameters. So, the first problem is the density from is known that means the density of the class conditional density that is known, but we do not know the values of the parameters. So, we have to estimate the parameters and there are two popular techniques one is called the maximum likelihood estimation and another one is called a Bayesian estimation. So, this is called a parameter estimation. In the non-parametric methods, the density from is not known.

The density from of that is the $P(x/\omega_i)$ that density from is not available. We do not know about the density from. So, we have to estimate the density and there are two popular techniques one is the Parzen window technique and another one is the k nearest neighbor techniques. So, that is called a non-parametric techniques. So, now I will be discussing these parametric methods and the non-parametric methods.

So, in the parametric methods, I have to estimate the values of the parameters by these two techniques one is the maximum likelihood estimation another one is the Bayesian estimation. So, first I will discuss about the maximum likelihood estimation and in the next class I will be discussing the concept of the Bayesian estimation. So, let us start this class. So, in case of the Bayesian decision making. So, already I told you because we have to determine the $P(\omega_i/x)$ that we have to determine and that is equal to probability of x given omega i that is the class conditional density and the prior probability the probability of omega i and we have the normalizing factor.

So, i is equal to 1 to c probability of x given omega i and probability of omega i. So, in this case what information we need for determining this probability that is the posterior probability the probability of omega i given x. So, that is the supervised classification. So, in the supervised classification I already told you for each and every classes I have the training data samples. So, that concept I will explain later on, but in the statistical machine learning we have to determine this probability that $P(\omega_i/x)$.

So, for this I need this information that is the class conditional density and another one is the prior probability and maybe I have these number of classes c number of classes. So, this information I need for determination of the this posterior probability. So, all this information I need. So, in case of the parametric method the density from is known that is this density from is known and suppose it is a normal density. So, in the normal density I have two parameters one is the mean vector and another one is the covariance matrix we have to estimate the values of these parameters.

So, this parameter vector the parameter vector is represented by θ. So, suppose these are the components of the parameter vector theta 1 theta 2 suppose theta p. So, this is a parameter vector and in this case suppose the theta 1 is a parameter the parameter is nothing, but the mean theta 2 is another parameter and that parameter may be the covariance. So, in this case what we have to consider that we have to estimate the parameter vector. So, we have to determine the parameters.

So, what information is available the information is the class conditional density the density from is known, but we do not know what are the values of the parameters. So, in case of the supervised learning we have the training samples for each and every classes.

So, for C number of classes I have C number of training data samples and with the help of these data samples or with the help of this training data I have to determine the values of these parameters. That means I have to estimate the parameter vector that is the supervised learning. In case of the unsupervised already I told you I have the Feature vectors and I have to group the Feature vectors based on some similarity.

So, for determination of the similarity I can consider maybe Euclidean distance and based on this similarity measurement I can group the Feature vectors and that is nothing, but the clustering. So, one is the supervised learning and another one is the unsupervised learning. So, these two information I need for determination of the posterior probability probability of omega I given X the first information is the likelihood information that is the class conditional density and another one is the prior information the prior probability. So, for estimation of these parameters already I told you. So, there are two popular techniques one is the maximum likelihood estimation and another one is the Bayesian estimation.

So, in my next slide I will be explaining the basic concept of the maximum likelihood estimation and the Bayesian estimation. So, in case of the maximum likelihood estimation that I can say ML estimation the maximum likelihood estimation. So, we have to estimate the parameters so, in this case the parameters are fixed parameters are fixed, but unknown. So, in case of the maximum likelihood estimation the parameters are fixed, but it is unknown.

Now what we have to consider maximize the probability of obtaining the given training data set.

So, that means the parameter estimation maximizes a likelihood function. So, in this case we have to maximize the probability of obtaining the given training data set. So, that means the parameter estimation technique maximizes a likelihood function. So, mathematically I can say the maximize what I have to maximize the probability of D, D is the training data set and theta is the parameter vector.

So, we have to maximize this and in this case the theta is fixed.

So, this is the fundamental concept of the maximum likelihood estimation. So, I will explain in detail about the maximum likelihood estimation in my next slide. And another technique of parameter estimation I should say that is the Bayesian estimation. So, in the Bayesian estimation we consider the parameters are random variables, parameters are random variables, rb means the random variables with a known a priori information or I can say a priori distribution. So, what is the objective of the training samples? Training samples allows conversion of a priori information into posterior density.

So, in this case the parameters are random variable with a known priori distribution and the training samples convert the priori information into a posterior density. So, in this case the objective is mainly to maximize the probability of theta given D, D is the training data set and theta is the parameter vector. So, we have to maximize theta given D. So, that we have to maximize and theta already I told you that theta is a random variable. In case of the maximum likelihood estimation we consider the theta it is fixed but it is unknown.

But in this case we are considering theta is a random variable and we have to maximize the probability of theta given D. So, you can see the fundamental difference between the maximum likelihood estimation and the Bayesian estimation. In case of the maximum likelihood estimation we have to maximize this. This is for the maximum likelihood and for the Bayesian estimation we have to maximize the theta given D.

This is for the Bayesian estimation one is for the maximum likelihood estimation another one is the Bayesian estimation.

So, this is the fundamental difference between the maximum likelihood estimation and the Bayesian estimation. So, these are the parametric methods. So, already I told you so what I need to discuss the one is the parametric methods. In a parametric methods the density from is known but I do not know the values of the parameters.

So, for this I will be discussing these two techniques one is the maximum likelihood estimation and another one is the Bayesian estimation.

So, in case of the parametric methods we know the density from the density from is known what density from density from is known form is known density means the probability of X given Omega I. So, this density from is known and from this we have to estimate the parameters estimate the values of the parameters. So, this is the concept of the parametric methods and in case of the non-parametric methods density from is not known density from is not known we have to estimate the density. So, for this I will be discussing two important concepts one is called the Parzen window technique Parzen window and another one is a popular technique that is the k nearest neighbor k and n method.

So, these things I am going to discuss.

So, first I will discuss about the parametric methods the maximum likelihood and the Bayesian estimation. So, what I have discussed in maximum likelihood estimation the parameters in the ML estimations are fixed but unknown. So, what is the best parameters the best parameters are obtained by maximizing the probability of obtaining the samples observed that is the the principle that means we have to maximize the probability of D given theta that is the objective and parameters are chosen in a way that the best support describe the training data set. So, parameters are selected in a way that they best describe

the training data set. So, that concept already I have explained that is the concept of the maximum likelihood estimation in case of the Bayesian method the parameters are random variable having some known distribution.

So, the training samples that is the observation of the samples converts these to a posterior density. So, already I have explained. So, the prior information is converted into a posterior density with the help of the training samples and in this case we have to maximize the probability of theta given D and theta is a random variable. So, this we will discuss later on now what is the Bayesian estimation and in this case whenever we maximize this one because we have to maximize this one. So, we will be getting a peak corresponding to we are plotting this one probability of theta given D and this is the estimated value of theta.

So, we will be getting a peak near the true value of the parameters. So, this is called the Bayesian learning. So, we have to take the help of the training samples and ultimately we are estimating the value of the parameter the parameter is the theta theta is the parameter vector. So, now before going to the Bayesian learning I will now discuss the concept of the maximum likelihood estimation and in both the cases the main concept is the Bayes theorem. So, now let us discuss the fundamental concept of the maximum likelihood estimation.

So, maximum likelihood estimation. So, first let us discuss this one this is the Bayesian decision theory we have to determine the posterior probability. So, probability of X given omega J and the prior probability is omega J and we have the normalizing factor. So, J is equal to 1 to C, C number of classes we are considering X omega J probability of omega J. So, we are considering this.

Now, let us consider the theta J that is the parameter vector it is fixed, but unknown.

So, theta J is fixed, but unknown. So, we have to see what is the role of the training algorithm. So, this is my training algorithms. So, my input is the training data set suppose the training data set is D 1, D 2 these are the training data set for C number of classes I have C number of training data set and we know the class conditional density the density from is known that is also my input. So, from these two information I have to determine the values of the parameters that is the parameter estimation.

So, corresponding to this one this class conditional density.

So, already I told you the density from is known. So, what is the density from the probability of X given omega J the density from is known and that is also called the parametric form the parametric from I can write like this probability of X given theta is a parametric from is known and suppose if I consider it is the normal density. So, I have two

parameters one is the mean vector and another one is the covariance matrix. So, this is available. So, that means parameter vector is theta J and it has two components one is the mean vector another one is the covariance matrix.

So, this is the concept of this parametric from. So, let us move to the next slide dependence on the parameter vector dependence dependence on theta J. So, how to write the dependence on $P(x/\omega_j)$ and theta J. So, this is the dependence on theta J. So, what we need to determine we need to determine theta 1 theta 2 from the training data set.

So, we have to determine these parameters. So, what is the supervised training? So, I can show like this suppose this is my training data set D 1 D 2 D I and I have some samples suppose the samples are X 1 X 2 these are the samples and the samples are nothing but the feature vectors these are the samples corresponding to the training data set D I and I have the classes is suppose omega 1 omega 2 omega I and omega J. So, what is the supervised classification? The supervised classification is suppose this training data set D I that is only for the class omega I that is not for the class omega J that is not for the class omega J. So, for each and every classes I have the separate training data set for example, corresponding to the class omega 1 I have the training data set the training data set is D 1. So, you can see this is the fundamental concept of the training data set in the case of the supervised learning and another important consideration in case of the maximum likelihood estimations the samples this is a very important consideration samples in D I D I is the training data set corresponding to the class omega I the samples in D I are drawn independently according to the probability law and the samples are also samples are I I D. So, what is the meaning of I I D? I I D means independent and identically distributed.

So, samples are I I D that is a very important consideration in case of the maximum likelihood estimation the samples are I I D and identically distributed random variables. So, R V means the random variables. So, now let us move to the basic principle of the maximum likelihood estimation. So, let us see how we have to determine the values of the parameters. So, suppose we use a set the training data set is suppose D of training samples drawn independently this is a very important consideration drawn independently in a probability density the probability density is probability of X given theta to estimate the unknown parameter vector theta.

So, in this case we are considering the training data set is the D and suppose this training data set D contains n samples n number of samples. So, what are the samples X 1 the samples are the feature vectors actually X 2. So, these are the samples and already I told you these samples are drawn independently the samples are drawn independently the samples are drawn independently. So, I can write this probability of D given theta the probability of D given theta that is actually called the likelihood of theta this is called

likelihood of the vector theta. theta is the parameter vector with respect to the set of samples.

So, we have to determine this one that likelihood of theta with respect to the set of samples. So, that is probability of D given theta. So, I can write so in the product from because the samples are drawn independently. So, it is in the product from I can write the probability of X k given theta. So, I can write like this to k is equal to 1 to n because we are considering n number of samples that is available in the training data set D.

So, in case of the maximum likelihood estimation in case of the ML estimation. So, what actually we are looking for maximize the probability of D given theta that we have to maximize. So, what actually it is meaning the meaning is so this estimate corresponds to corresponds to the value of theta value of theta that I can say that supports the actually observed training samples. This estimate corresponds to the value of theta that supports the actually observed training samples.

So, that is the meaning of this. So, we have to maximize the probability of D given theta that is that the maximize the probability of obtaining a given data set and D for the vector the vector is theta. So, we move to the next slide. So, we have to maximize probability of D given theta that we have to maximize. So, now let us consider the log likelihood function. So, what is the likelihood function we are considering now that is L theta that is we are taking the natural logarithm.

So, why we are considering the log because from the arithmatic point of view it is also important because the multiplication can be converted into addition and also that I can write the monotonically increasing the estimated value of theta that maximize log likelihood we are considering log likelihood also maximize likelihood. So, that is the likelihood is probability of D given theta. So, this log is a monotonically increasing function. So, that is one important aspect. Another one is from the arithmatic point of view and this multiplication is converted into addition with the help of the log.

So, this I can write like this. So, it is now summation after taking the log k is equal to 1 to n ln probability P X k given theta. Now we have to maximize the likelihood function. So, for this what we have to consider we have to determine or we have to take the derivative. So, this derivative is taken this is the gradient operation I am taking.

So, I have to find a maximum value. So, that is why I am taking the derivative with respect to theta. So, k is equal to 1 to n this is the gradient operation D theta ln probability of X k given theta and because we have to find the maximum. So, it is equating to 0. So, what is this gradient operator because we have to take the partial derivative. So, this partial

derivative it is with respect to theta 1 with respect to theta 2 like this with respect to theta P.

So, for this parameter vector theta it has this component theta 1 theta 2 up to suppose theta P. So, we have to take the partial derivative with respect to theta 1 theta 2 and like this up to theta P. So, this maximum likelihood estimation of theta is obtained from the P number of equations. So, if you see here I have the P number of equations. So, in this case the theta is the parameter vector theta for normal density I have 2 thetas one is theta 1 another one is theta 2.

So, theta 1 corresponds to the mean vector and theta 2 corresponds to the covariance matrix. So, you can see with the help of this equation with the help of this equation is an important equation we can determine the values of the parameters. So, the values of the parameters is nothing but this is the estimated value of theta argmax and this is the likelihood function the log likelihood function L theta. So, corresponding to this you can see I have to determine the this the maximum of this. So, probability of D given theta I am finding the maximum value with respect to theta.

So, I will be getting a peak something like this corresponding to the estimated value of theta. So, this is the estimated value of theta. So, from the these equations from these equations I have this maximum likelihood estimation ML maximum likelihood estimate of theta maximum likelihood estimation of theta is obtained from P number of equations P number of equations. So, move to the next slide.

So, for example, suppose I can estimate the mean. So, this is the estimated value of the mean and that can be obtained like this. So, k is equal to 1 to n x k. So, that I will explain how to determine this one in the next slide and that is nothing but the m, m is the arithmetic mean. So, that is nothing but the arithmetic mean of the samples. Now let us consider another estimator that is called maximum a posteriori maximum a posterior estimation.

So, that is equal to map estimator is nothing but we are considering the likelihood function that already I have defined and we are considering the prior information the prior information of theta. So, this is called a maximum a posterior estimation. So, if I consider for flat prior that means the prior has no information for flat prior the ML estimation is equivalent to the map estimation the same for flat priors ML estimation is nothing but this equivalent to map estimator. So, now let us discuss how to determine the values of the parameters, because we have to determine the mean vector and also we have to determine the covariance matrix if I consider a normal distribution.

So, let us move to the next slide. So, in the next slide, I will be discussing these two cases.

In the first case, I will be determining the mean vector and in the second case, I will be determining the mean vector and the covariance matrix. So, let us move to the next slide. So, case 1 is estimation of the mean vector mean vector is mu of a Gaussian PDF and in this case, we are considering this sigma is known that is the covariance matrix is known this case. So, this case is estimation of the mean mu of a Gaussian PDF and the sigma is known.

So, that is why I am considering mu i is equal to theta i that is a parameter vector and only one parameter we are considering mu and the covariance matrix is known. So, corresponding to this, I can write like this probability of X given mu. So, that is it follows the normal distribution. So, we are considering the multivariate normal distribution.

So, $f(X \mid \mu, \Sigma) = \frac{1}{2\pi^{d/2}\Sigma^{\frac{1}{2}}} exp(-\frac{1}{2}(X-\mu)T\Sigma^{-1}(X-\mu))$. So, the training set normal X k of D i is log we are taking the log likelihood function we are considering this one log likelihood function that is equal to if I take the log of this it will be minus 1 by 2 X k minus mu transpose. So, actually from this we are getting this, this is simply I am taking the log. So, after taking log, I will be getting this one that is a log likelihood function. After this we have to maximize this. So, that is why I have to take the differentiation with respect to the parameter mu that is a mean vector and this mu cap is nothing but the estimated value of the parameter mean.

So, it is k is equal to 1 to n because we have to equate it to 0 because we are finding the maximum. So, that I can write like this k is equal to 1 to n. So, mu estimated that is equal to 0. So, we are getting this expression. After this we have to do some mathematics to get the values of the parameter the parameter is the mean.

So, move to the next slide. So, in the previous slide we obtained this one that we are taking the differentiation of the log likelihood function. This is the estimated value of the parameter. So, we obtained like this. So, k is equal to 1 to n sigma inverse X k minus mu that is equal to 0.

Now after this the next step is we have to do some mathematics. The mathematics is multiplying by sigma and rearranging. We will be getting n mu hat k is equal to 1 to n X k. So, these are the sample.

So, k is equal to 1 to n. So, n number of samples are there. So, from this you can directly estimate the values of the parameter that is a mean vector you can determine and that is nothing but k is equal to 1 to n X k. So, we have obtained this one. So, this is the expression for the estimated value of the parameter the parameter is mean. And already I told you this is nothing but the arithmetic mean of the samples. So, like this you can determine the

values of the parameters from the training data set and move to the case number 2.

So, I am moving to the case number 2 unknown mean vector and also the covariance. So, corresponding to this the same procedure can be applied and we can determine the mean vector again the mean vector is nothing but it is a sample mean. So, k is equal to 1 to n X k. So, this is nothing but a sample mean sample mean we can determine and also we can determine the covariance matrix that is equal to 1 by n k is equal to 1 to n X k minus mu X k minus mu transpose. So, we can determine the covariance matrix this is nothing but the arithmetic average of n number of matrices n matrices.

So, this is this part is nothing but the matrix 1 matrix. So, we have n number of matrices and what is the matrices X k minus mu hat and X k minus mu hat transpose. So, that means we are taking the average of this arithmetic average of this matrix. So, like this you can determine the values of these two parameters one is the mean vector another one is the covariance matrix. So, in this class I discussed the fundamental concept of parameter estimation there are two techniques popular techniques one is the maximum likelihood estimation another one is the Bayesian estimation. So, I have explained how to determine the values of the parameter by considering the maximum likelihood estimation.

So, we have to maximize the probability of D given theta and based on this we can determine the values of the parameters. So, in my class I will be discussing the second technique that is called the Bayesian estimation. So, that concept also I will be explaining in my next class. So, in that case I have to maximize the probability of theta given D and theta is a random variable.

So, that concept I will be explaining in my next class. So, let me stop here today. Thank you.