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-12

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of parameter estimation. I highlighted the concept of the Bayesian estimation and I explained the concept of the maximum likelihood estimation. In the maximum likelihood estimation, the parameters are fixed, but they are unknown. So we have to maximize the probability of obtaining a given training data set. So that means I have to maximize the P(D/θ).

And based on this condition, I can determine the values of the parameters. In case of the Bayesian estimation, the parameters are random variable and I have to maximize the probability of theta given D. So that concept already I have explained. Today I am going to explain the concept of the Bayesian estimation.

So let us start this class and the estimation. The estimation is the Bayesian estimation. So today I will discuss the concept of Bayesian estimation. So in my last class, I explained briefly what is the Bayesian estimation.

So parameters are random variable and I have to maximize the P(θ/D).

Now let us consider this is the Bayes law. So we have explained this in the discussion of the Bayesian decision theory. So this is the posterior probability. I am putting the D that actually represents the data set and it is the dependence on D. So that is the meaning of this.

That is equal to this posterior probability is equal to the likelihood P(x/ω_i).and also I am considering D because dependence of D we are considering and the prior probability that D is also I am considering that is the dependence of the training data set and suppose I am considering the C number of classes. So probability X omega i D and this is omega D. So already and this was explained. So now the priori information is known. So that is why I can consider like this the probability of omega i that depends on D but actually it is known.

So it is I can write probability of omega i the priori information is known. So then this equation this the posterior probability I can write like this omega i D. So I can write like this. Now what I have to determine in case of the Bayesian estimation I have to determine this that is the P(x/D).that I have to determine. So how to determine this P(x/D).

So I will explain in my next slide. So we have to determine the P$(x/D.$ D is the training data set and what information is available this probability of X that is unknown and this is unknown. This is unknown and the parametric from is known parametric from probability of X given theta this is known and this the probability of theta, theta is the parameter vector that is also known that is known.

So this training data set converse the prior information into the posterior information.

So that means this is the known D is available. So P theta is converted into P theta given D. So it is converted to this. So now this I can write like this P(x/D) I have to determine that is nothing but the probability of X given theta probability of theta given D theta.

So I can determine like this.

So actually how to get this one. So I can show how to get this one this probability of X given D that I can write like this probability of X theta given D theta this is nothing but what this is nothing but a joint density. This is nothing but a joint density and this one I can write like this. So I can write like this.

So based on this two I am getting this one based on these two I am getting this one.

So P(x/D) I am getting. So that means this expression again I am writing X given D is nothing but same thing I am writing X theta P theta D D theta. So this expression same thing I am writing here. Now what we have to consider in the Bayesian estimation we have to maximize probability of theta given D we have to maximize this one. So if I maximize this one the theta given D so this theta given D that approaches the Dirac delta function.

So you can see if I maximize then I can determine probability of X given D. So this probability of X given D will be approximately equal to probability of X and theta. So this is the estimated value of theta theta hat theta hat is the estimated value of theta. So I can write like this. So that means what I am doing in this integration in this integration what I am taking the taking average of the probability of X given theta.

So I am taking this one. So this is this theta that is the estimated value of the parameter estimated parameter vector estimated parameter vector. So that means what you can see I am maximizing what I am maximizing probability of theta given D. So that means it approaches the Dirac delta function corresponding to this the theta estimated that is the theta hat. So it approaches the actual value of the parameter.

So this is the fundamental concept of the Bayesian estimation. So what equation is important here this is the important equation in case of the Bayesian estimation. So we can determine the P(x/D) and you can see when P(θ/D) it is maximum that means it approaches the Dirac delta function then you can see from this I will be getting probability of X given D that is approximately equal to probability of X given theta that is the estimated value of the theta. So this concept I will explain in my next slide that how to determine the parameters for the Gaussian case for the normal density. So in my next slide I will explain this concept that is the Bayesian parameter estimation for Gaussian case for the normal density.

So I will be considering the univariate case. So move to the next slide. So what actually we are doing now that probability of X given D is equal to the probability of X given theta probability of theta given D and D theta. So this already I have explained. Now let us consider Bayesian parameter estimation and in this case we are considering the Gaussian case Gaussian and

also we are considering the univariate case just we are considering the Gaussian and or univariate.

So let us see how we can employ the Bayesian estimation technique for determining the values of the parameters for normal distribution and if I consider the univariate case I have two parameters one is the mean another one is the variance. So now this probability of theta given D I have to determine that means it is nothing but the mu given D I have to determine. So mu is the mean and what information is available the parametric form is available. So what is the parametric form it is X mu. Now we are considering the univariate case.

So X is not a vector. So it is normal density. So that parametric form is available probability of X given mu that information is available that is the normal density. Also this probability of mu and that is the prior information that is also known and suppose it is the normal density. So mu naught sigma naught square. So mu naught sigma naught squared what is the mu naught this is the mean.

So I am just considering as initial guess because I have to estimate the parameters. So initial guess is mu naught because I have to estimate the values of the parameters mu and sigma square. So initial guess is mu naught and this variance it represents the uncertainty about the guess uncertainty about the guess. So my initial guess of the mean is mu naught and what is the uncertainty about the guess that is sigma naught square. So now I have to determine the probability of theta given D this I have to determine.

So which I can write like this probability of D theta probability of theta probability of D theta probability of theta theta. So this is I am determining. So what is the this probability of mu given D now the probability of mu given D is equal to so probability of mu given D is nothing but probability of D given mu and probability of mu and integration probability of D given mu and probability of mu D mu. So we are considering this the probability of mu given D. So that is the probability of the mean given that D is the training data set and already I told you the samples are drawn independently and

that is actually the supervised training.

So for each and every classes I have the training data set and that is the concept of the supervised learning. So since we are considering that training samples are drawn independently so I can write the probability of D given mu in this form in the product form the samples are drawn independently. So that concept I have explained in the maximum likelihood estimation the probability of X k given mu and probability of mu. So I can write like this. So now move to the next slide because what information is available the information is probability of X given mu that is the normal density.

So the parameters are mean and the variance and we are considering the initial guess of the mean is mu naught and uncertainty about the guess is sigma naught square. So move to the next slide. So what we have obtained the probability of D given mu is equal to this is the k is equal to product k is equal to 1 to n the probability of X k so kth sample mu. So this probability of mu given D I can write like this alpha. So we are not considering the normalizing factor k is equal to 1 to n so it is X k mu so X k mu and

mu we are obtaining this one.

So the normalizing factor is not considered so that is why I am considering a constant the constant is alpha the alpha constant is considered. So now what is the this probability of mu given D is equal to alpha this product from k is equal to 1 to n because you know this probability of X given mu that is what density it is following because in the previous slide already I have mentioned the probability of X probability of X given mu. So it follows the normal density and also I mentioned this the prior information is available and this density is also the normal density. So in my previous slide I have explained this one. So this probability of mu given D I can write like this alpha that is the proportionality constant because I am not considering the normalizing factor and is 1 by twice pi sigma exponential.

Because we are considering the normal density minus half the kth sample minus mu sigma whole square into 1 by twice pi sigma naught so exponential minus 1 by 2 mu minus mu naught sigma naught whole square. So I can write like this. So here you can see this is the normal density we are considering and corresponding to probability of mu that is also the normal density. So which I can write like this the alpha dash so I am considering the proportionality constant alpha dash and exponential minus 1 by 2 summation k is equal to 1 to n mu minus x k sigma plus mu minus mu naught sigma whole square. And which can be also written like this alpha double dash that means what we are considering the factors that do not depend on the mu have been considered as constant.

So I am repeating this the factors that do not depend on the value mu, mu is the mean have been considered as constant. So that is why I am considering this the constant alpha alpha dash alpha double dash exponential minus 1 by 2 n by sigma square plus 1 by sigma naught square mu square minus twice 1 by sigma square summation k is equal to 1 to n x k plus mu naught divided by sigma naught square. So this mathematics you have to verify again. So in this case what we are considering the factors the factors that do not depend on mu have been considered as constant.

So based on this I am getting alpha alpha dash and alpha double dash.

So if you see the expression so I will be moving to the next slide. So this equation the probability of mu given d I will be writing again in the next slide.

$$P\left(\frac{\mu}{D}\right) = \alpha'' exp[-\frac{1}{2}\left[\frac{n}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right]\mu^2 - 2[\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{{\sigma_0}^2}]\mu]$$

So this expression already I have derived in the previous slide.

So if you see this expression that expression for probability of mu given d this is actually if you see observe carefully. So this is nothing but the exponential function of quadratic function of mu. So if you observe it carefully you can see and this probability of mu given d it is the nature is exponential function of quadratic function of mu. So that means this probability of mu given d it is nothing but the normal density.

So I can write the probability of mu given d is a normal density because it is an exponential function of quadratic function of mu.

So I can write this mu n sigma n square I can write. So this normal density I can write probability of mu given d is equal to this is the Gaussian distribution

$$P\left(\frac{\mu}{D}\right) = \frac{1}{\sqrt{2\pi}\sigma_n} exp\left[-\frac{1}{2}(\frac{\mu - \mu_n}{\sigma_n})^2\right]$$

So we are considering this is the normal distribution. So I will be getting this one and I will be getting another one that mu n divided by sigma n square is equal to n by sigma square and this mu n hat that is the estimated value of the mean and mu naught divided by sigma naught square.

So from this I can determine because we know what is the sample mean. The sample mean already you know the sample mean is nothing but mu n hat is equal to 1 by n and we are considering n number of samples. So k is equal to 1 to n x k. So n number of samples we are considering.

So this is nothing but a sample mean. So now I have to determine the mu n and also sigma n square. So what is mu n and what is the sigma n square that I have to determine. So move to the next slide. So we have to solve for mu n that is actually what is mu n.

This is actually the best guess. So our initial guess was mu naught. So which one is the best guess. So best guess is mu n that is the value of the parameter. So solve for this. So I will be getting this from the previous equations n sigma naught square plus sigma square mu n hat that is nothing but the sample mean sigma square n sigma naught square plus sigma square mu naught and the sigma n square that is what that is actually the sigma n square means uncertainty about this guess.

So we consider sigma naught square that is the uncertainty about the initial guess. Now the best guess is mu n and uncertainty about this guess is sigma n square that is equal to sigma naught square sigma square n sigma naught square plus sigma square. So if you see here this expression actually it has two information one is the prior information and another information is the information obtained from the training samples. So it is a combine this is a combine I can write a combine prior information prior information and the information and the information and the information obtained from the obtained from the training samples.

So this is the guess. So here you can see we are considering sigma n square that is the uncertainty about the guess. So here you can see the sigma n square decreases monotonically with n. So just better to write here the sigma n square decreases from this expression I can write the sigma n square decreases monotonically with n because we are considering n number of samples and it approaches sigma square divided by n as n tends to infinity. When we are considering the large number of samples then it approaches sigma square divided by n. So what is the physical interpretation of this? That means each additional observation decreases uncertainty about the true value of mean.

So I can also write this one this uncertainty about the true value of mu. So that means as n increases probability of mu given d it approaches the Dirac delta function. And that is the concept of the Bayesian learning this is called a Bayesian learning. So you can see the importance of the training data set. So if I consider more and more training data set then you can see that uncertainty about the true value of the parameter it decreases.

So because of the training data set so if I have the more number of training data set the estimation will be more accurate. So that means you can see if I plot suppose this probability of mu given d which so initially suppose I am estimating this one. So this is the estimated value of the parameter. So this probability of mu given these samples x1, x2 so samples are considering like this all the samples are considered up to xn number of samples. So you can see with each and every observation it will be like this.

So it will approach the Dirac delta function and that is the estimated value of the parameter is mean. So it approaches the Dirac delta function. So that means that is the importance of the training data set.

So each additional observation decreases uncertainty about the true value of the parameter.

The parameter here we are considering the mean. So you can see that is the fundamental concept of the Bayesian learning. So move to the next slide. So in the one dimensional case 1D case what we consider the probability of X given D that is the probability of X given mu probability of mu given D. So that is in the first slide I have explained. So this distribution probability of X given mu that is the normal density the mean is mu and the variance is sigma square and also this probability of X given D this is the normal density the mean is mu n and sigma square plus sigma n square.

So corresponding to X given D I have the normal distribution and the estimated value is mu n and uncertainty about this is sigma square plus sigma n square. So what is actually this probability of X given D? The probability of X given D is equal to how to get this? So it is root over twice pi sigma exponential minus 1 by 2 X minus mu sigma whole square 1 by root 2 pi sigma n exponential minus 1 by 2 mu minus mu n sigma n square d mu. So this I can write like this twice pi sigma sigma n exponential minus 1 by 2 X minus mu n sigma square plus sigma n square X minus mu n whole square sigma square plus sigma n square and we are considering a function the function is sigma sigma n. So what is this function? So where this function is actually the sigma n sigma n is nothing but the integration exponential minus 1 by 2 X minus mu n square sigma square plus sigma n square into mu minus sigma n square X plus sigma square mu n sigma square plus sigma n square whole square d mu that is function f sigma sigma n. So that is why that you can see this probability of X given D that is proportional to exponential minus 1 by 2 X minus mu n square divided by sigma square plus sigma n square.

So that is the observation that probability of X given D that we have obtained here so that is actually this. So that means it is a normal distribution the probability of X given D that is a normal distribution with mean mu n and the variance is sigma square plus sigma n square this is the variance. So in this case you can see how to determine this probability of X given D. So in this case you can see here this probability of X given mu that is the normal density.

So here you can see this is the normal density. So what I have to determine the probability of X given D that is already determined. So that is what we have to consider replace mu by mu n. So you can see that I am replacing mu by mu n. So from this actually I am getting this one replace mu by mu n and sigma square by sigma square plus sigma n square. So I can get this the probability of X given D that density we can determine.

So this is for the normal density and we are considering the univariate density. So you can see we have determined the probability of X given D we have determined and the mean is mu n and the variance is sigma square plus sigma n square. And if I consider multidimensional case so move to the next slide. So for a multidimensional case you can also verify that one or I can say the multidimensional case this I will be getting this one. So sigma n to the power minus 1 mu n is equal to n sigma to the power minus 1 this is the sample mean this is the sample mean plus sigma naught to the power minus 1 mu naught.

So probability of X is a vector now X given D is a normal distribution with mean vector mu n and a covariance matrix is sigma plus sigma n this is a covariance matrix. So that means in summary I can write like this this probability of X given mu minus that is the normal distribution with mean vector mu and a covariance matrix sigma. And this is also available the probability of mu that is also the normal density. So initial guess is mu naught and also the covariance matrix is sigma naught. So now I want to determine the probability of mu given D that is also the normal density.

So one is the mean vector mu n is the mean vector and a covariance is sigma n. So we can determine this. So this is the final observation. So we can determine probability of mu given D we can determine. So you can see how we can determine the values of the parameters by considering the Bayesian estimation method and you can see the importance of the training samples.

So if I consider more and more number of training samples the estimation approaches the actual value of the parameters. So that means the uncertainty about the initial guess it decreases and that is the importance of the training data set. Now in the next slide I will be explaining the general theory of the Bayesian estimation. So what is the general theory of the Bayesian estimation.

So I can show like this. General theory of the Bayesian estimation. So I can say the Bayesian estimation. So this is the algorithm for the Bayesian estimation. So what is available this probability of theta that is known and the parametric from is also known x given theta. So it is also known and we have considered the supervised learning and the supervised training that means the independent training samples for each and every classes that is the concept of the supervised training. So probability of x given D is nothing but the probability of x given theta probability of theta given D D theta.

So we know this expression. So what is the probability of theta given D that is from the Bayes law I can write like this probability of D given theta probability of theta and this is the normalizing factor probability of D given theta probability of theta and D theta. So here this integration if you see this integration because we are considering the multidimensional. So x is a vector. So this integration is the multidimensional integration.

So that is that point you have to note down. So this integration is the multidimensional integration and we have to compute this one and this probability of D given theta that we have obtained in case of the maximum likelihood estimation that can be represented like this k is equal to 1 to n because we are considering k number of samples given theta. So here you can see in this expression we are determining we are determining this one probability of D given theta in this case. So to determine the probability of theta given D we have to determine the probability of D given theta and in the maximum likelihood estimation we are determining the probability of D given theta. So you can see that is a relationship relationship with the maximum likelihood estimation. So for determining the probability of theta given D this is one the probability of theta given D we have to determine the probability of D given theta and the probability of D given theta is expressed like this.

So if you understand the concept of the maximum likelihood estimation and if you understand the concept of the independent training samples and the training samples are drawn independently then you can understand this concept. So in supervised learning you know for each and every classes we have the training data set. So this training data set I can write like this dn is equal to x1 x2 xn. So dn training data set I can write like this. So the probability of dn given theta because we are considering the samples x1 x2 xn that these are the samples.

So I can write like this probability of x1 xn theta probability of dn minus 1 given theta. So that I can write like this because I am considering all the samples x1 x2 x3 up to xn. So here if I write xn given theta then we have to consider the remaining samples the remaining sample is dn minus 1. So probability of xn given theta so we are considering the nth sample and after this I can write probability of dn minus 1 given theta because we have to consider the remaining samples.

So if I considered suppose this is the equation number 1 this is the equation number 1. So this equation number 1 can be represented by considering this this case. So in my next slide I will be showing this how to represent the probability of theta given d. So from equation 1 from equation 1 I can write probability of theta given dn so probability of xn probability of xn given theta probability of theta dn minus 1 so all the remaining samples. So probability of xn given theta probability of theta dn minus 1 d theta.

So it is you can see this equation is actually the recursive equation. So why it is the recursive equation recursive estimation that is the Bayesian estimation is the recursive estimation that is actually the Bayesian estimation. So why it is recursive if I consider the probability of theta given d naught from this I can determine probability of theta. So if I use repeatedly this equation the above equations. So repeated use of the equation of the above equation gives densities what density I will be getting first initially I will be getting the P theta after this next sample we are considering we are estimating probability of theta but sample is x1 next if I use this equation repeatedly so I will be considering theta given x1 and x2 two samples we are considering x1 and x2 and like this we can determine all these parameters. So this equation this equation is a recursive equation this equation is a recursive equation that is the recursive estimation and if I repeatedly use the above equation I will be getting P theta P theta given x1 x1 is that sample first sample P theta given x1 comma x2 second sample we are considering.

So we are considering more and more samples like this and if I consider more number of training data set then you can see our estimation approaches the actual value of the parameters that is the beauty of the Bayesian estimation. So with large number of training samples our estimation will be more perfect more accurate so that is the concept of the Bayesian estimation. So in this class I discussed the concept of the Bayesian estimation. So the main concept is I have to maximize the probability of theta given D that I have to maximize in case of maximum likelihood estimation I have to maximize the probability of obtaining a given training data set that is the probability of D given theta. So in the Bayesian estimation I have estimated the values of the parameters in my example I have considered the univariate normal density and I have estimated the value of the parameter that is the mean and also I have shown the variance.

Now if I want to compare the maximum likelihood estimation and the Bayesian estimation the first point is the computational complexity. In case of the maximum likelihood estimation I have to maximize the probability of D given theta and for this I have to consider the partial derivatives. But in case of the Bayesian estimation I have to determine the multi-dimensional integration that is very difficult to determine and it is computationally very complex. The multi-dimensional integration. So for this we can employ some techniques like the Monte Carlo simulation and maybe some other techniques are also there for determining this multi-dimensional integration.

So from the computational point of view the Bayesian estimation is computationally more complex as compared to the maximum likelihood estimation. The second point is the accuracy in case of the Bayesian estimation we are considering the prior information. So that is why the Bayesian estimation is slightly more accurate as compared to the maximum likelihood estimation. But for the flat prior the Bayesian estimation is similar to the maximum likelihood estimation.

So they will give the same results. So for the flat prior the Bayesian estimation is very similar to the maximum likelihood estimation. So this is about the parameter estimation. In the parameter estimation I have discussed these two popular techniques. One is the maximum likelihood estimation and another one is the Bayesian estimation.

In my next class I will be explaining the concept of the non-parametric estimation. So we have to estimate the density unknown density. So there are two popular techniques one is called the Parzen window technique and another one is the k nearest neighbor technique. So in my next class I will be explaining these techniques that is the non-parametric techniques. So let me stop here today. Thank you.