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-8

Lecture-30

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. Up till now I have been discussing about the concept of supervised learning techniques. Now I will be discussing the concept of unsupervised techniques. Today I am going to discuss one important concept that is the Gaussian mixture model and the expectation maximization algorithm. In most of the engineering and other applications, like in signal processing, communication, machine learning, and many other applications, we consider a Gaussian distribution forrepresenting a data distribution. That means, a particular data distribution can be approximated by a Gaussian distribution.

It has many advantages. So one advantage is that Gaussian distribution is very closer to the natural distribution. That is the first advantage. Another important point is the mathematical manipulations for Gaussian function is easy.

So if I want to determine the derivative of a Gaussian function, it is easy to determine this derivative. I can determine the nth order derivative of a Gaussian function. Like this there are many advantages of considering a Gaussian function. In case of a complex data set, or maybe the distribution, it is not possible to represent this data set by a single Gaussian. So for this, I have to consider multiple Gaussians, that means the mixture of Gaussians.

So for a complex data distribution, it is not possible to represent the distribution by a single Gaussian, we have to consider multiple Gaussians and that is the mixture of Gaussians. Now in case of a Gaussian distribution, or in case of the Gaussian density, I have two parameters. One is the mean, another one is the variance. If I consider the multidimensional case, one is the mean vector and another parameter is the covariance matrix. So these parameters I can determine by some estimation techniques like maximum likelihood estimation technique I can apply to determine or to estimate the values of the parameters.

The mean vector and the covariance matrix. In case of a Gaussian mixture model, I cannot apply the maximum likelihood estimation technique, because it cannot give a close form of solution. So for this, I have to apply some iterative algorithm and one algorithm is very popular algorithm is expectation maximization algorithm. So in this class, I will be discussing first the concept of the Gaussian distribution, that is the Gaussian density. And after this, I will be discussing the concept of mixture model, that is the Gaussian mixture model.

And after this, to determine the values of the parameters, I will show one iterative algorithm that is the expectation maximization algorithm. So let us begin this class. So GMM that is the Gaussian mixture model using EM, EM means the expectation maximization technique.

So first let us consider a Gaussian distribution. So it is one dimensional distribution.

So this density the Gaussian density I can write like this. It has two parameters, one is the mean another one is the variance. So it is $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. So this is the univariate Gaussian density.

If I consider the multivariate normal distribution or the Gaussian distribution that I can write like this $N(x|\mu, \sigma^2)$ is a vector and

I am considering the mean vector and the covariance matrix.

This concept already I have explained in my previous classes. So $(2\pi)^{\frac{d}{2}}$. So what is d? d is the dimension of the vector x-, the exponential. So this is the expression for the multivariate density, multivariate normal density. So we need to estimate the parameters of a distribution and for this we can consider the maximum likelihood estimation.

So this is a very popular estimation technique. So we may consider the maximum likelihood estimation technique. So in this case I want to show you suppose I have some data points. So this distribution I can represent by a Gaussian function that means by a Gaussian distribution. So if I consider this one that is nothing but the Gaussian distribution.

If I consider the multidimensional case I may get this type of ellipse and this is the centroid. That means it is the mean of this Gaussian. So that means this data distribution is represented by a Gaussian distribution or Gaussian function. So now how to determine the values of the parameters? So already I told you we have to apply the ML technique the maximum likelihood estimation technique. So move to the next slide.

So what we need to consider the log of Gaussian distribution we have to consider in case of the maximum likelihood estimation and take the derivative and equate it to 0. So what I need to consider log of Gaussian distribution. First I have to take this one after this take the derivative and equate it to 0. So after this I will be getting the values of the parameters. So mathematically I can write like this I am taking the $lnP(x_i|\mu_i, \Sigma) = -\frac{d}{2}ln(2\pi) - \frac{1}{2}ln|\Sigma| - \frac{1}{2}(x_i - \mu_i)^T\Sigma^{-1}(x_i - \mu_i)$.

This expression already I have explained in my discussion of Bayesian decision theory. So we are taking the log of the Gaussian distribution and after this I have to take the derivative with respect to the parameter μ_i and equating it to 0 and for the another parameter that is the covariance matrix again I am taking the derivative with respect to the covariance matrix and equating it to 0. So corresponding to this one I will be getting the mean the estimated mean by the maximum likelihood estimation technique. So that is $\mu_{i_{ML}} = \frac{1}{N}\sum_{n=1}^{N} X_{i_n}$. So we have N number of samples or the data points X_{i_n} and corresponding to this one I have the estimate for the covariance matrix.

So $\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_{i_n} - \mu_{i_{ML}})(x_{i_n} - \mu_{i_{ML}})^T$. So I have the estimate for the covariance matrix. So here N is nothing but number of samples or data points. So this is about this is about a Gaussian distribution and how to determine the or how to estimate the values of the parameters. Now let us discuss about the Gaussian mixture model. So why actually we need a mixture model that concept I am going to explain in my next slide.

So why actually we need a mixture model. Suppose I have some data points these are the data sample points. So we are considering this data set. So this data distribution it is not possible to represent by a single Gaussian that means I cannot consider a single Gaussian distribution to represent this data distribution. So we have to consider multiple Gaussians.

In this case we may consider three Gaussians to represent the data distribution. So maybe we can consider these all these points we can consider this will be one cluster and this cluster can be represented by one Gaussian. Similarly I may consider all these data points these data points or this data distribution can be represented by another Gaussian. So this is one Gaussian and similarly the last one if I consider all these data distribution data points that can be also represented by another Gaussian. So three Gaussians I can consider.

So that means I will be getting this one. So this will be one Gaussian. So this Gaussian approximate this data distribution I have another Gaussian. So all these are data points corresponding to the second Gaussian and I have another Gaussian. So these are the data this data distribution is represented by the third Gaussian.

So this is number one Gaussian one number two and number three. So three Gaussians we are considering. So we have three clusters. So we can also represent like this. So this is one cluster This is another cluster. This is another cluster. So three clusters we can consider.

So these are actually the isocontours. So that means the distance from the mean is same for the isocontours that is the Mohalanobis distance is same for all the isocontours. So we are getting three Gaussians Gaussian one Gaussian two Gaussian three and that is the mixture model. That means the entire data distribution is represented by a mixture model the mixture of three Gaussians. So in one dimension also we can show this one.

So in one dimension suppose how to show this one. So first suppose I have these data points. So corresponding to this data point I can consider one Gaussian. Next I have these data points corresponding to this distribution I can consider another Gaussian maybe I can consider another Gaussian for this green color data points.

So in 1D I can show like this. So you can see that is the concept of the mixture model. So in this case I am giving one example here. Here I am showing two Gaussians the first is represented by $f_O(x) = N(x; 2,2)$.

So that is the first one. The first one is it is shown by the black color.

So the first one is this is the first Gaussian number one Gaussian and in this case the mean is 2 and the variance is you can see it is 2. And the second function Gaussian function is $f_1(x) = N(x; 10,5)$. So for this I can show the Gaussian is the blue color. So the second Gaussian is number 2 Gaussian and that is represented by the Gaussian function $f_1(x)$ and you can see the mean is 10 for this Gaussian.

And if you see the red one that is actually the combination of these two.

The red one is the combination of these two. The red one is nothing but this 1+2. So we have to consider the linear superposition of Gaussians. So this red one is nothing but the linear superposition of these two Gaussians. So this is one example of a Gaussian mixture model.

So how to represent it mathematically. So move to the next slide. So suppose this Gaussian mixture model GMM how to represent mathematically this density I can represent like this because the mixture model is nothing but the linear superposition of Gaussians. So $P(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \ldots + \pi_K f_K(x)$. So here K means we are considering K number of Gaussians. So what is mixture model? The mixture model is the weighted sum of the mixers of the PDFs where the weights are determined by a distribution the distribution is

π.

That means I can say the mixture model is nothing but the linear superposition of Gaussians. So here the mixing coefficient the mixing coefficient that is the weight $\sum_{k=1}^{K}\pi_k = 1$. So this can be written like this summation $\sum_{k=1}^{K}$ π_k π is nothing but the mixing coefficient π is nothing but the mixing coefficient $f_k(x)$. So this expression I can write in this form and if I consider this $f_1(x)$ that is nothing but the Gaussian density that is the normal density this density I can write like this π_1 normal density.

So it is $N(x|\mu_1, \Sigma_1)$ I am not giving the vector sign here π_1 is nothing but a vector but I am not giving the vector sign.

So like this I can consider number of Gaussians. So K number of Gaussians I can consider K number of Gaussians we are considering. So actually this expression also I can write like this. So in this case the mixing coefficient the mixing coefficient is π_k so mixing coefficient π_k so this $k = 1{:}K$. So k number of Gaussians we are considering and finally this expression I can write in this form this is the mixture of Gaussian.

So this is nothing but the mixture model. In this case we are considering K number of Gaussians. So now we have to discuss the concept of the expectation maximization algorithm because in case of the GMM that is a mixture model I cannot apply maximum likelihood estimation technique to estimate the values of the parameters because I will not be getting the close from of the solution. So let us move to the next slide. So now we are discussing this GMM the Gaussian mixture model. So from the previous slide how actually we are representing the density we are representing like this $\sum_{k=1}^{K}\pi_k f_k(x)$ K number of Gaussian this is the width that is the mixing coefficient and we are considering K number of Gaussians I am not giving the vector sign this x is a vector mu is also a vector.

So from the previous slide I can show like this. So in this expression this k is nothing but it represents number of Gaussians. This π_k it represents mixing coefficients or I can say widths mixing coefficients or widths and we have to consider some conditions that means we have to consider this normalization and also another condition is positivity.

So these conditions I can write like this the π_k that is the mixing coefficient lies between 0 and 1 and this $\sum_{k=1}^{K}\pi_k = 1$. So suppose if I want to determine the log likelihood how to determine the log likelihood of this.

So this log likelihood also I can determine in maximum likelihood estimation we determine the log likelihood in this case also I am determining the log likelihood. So you can see here I have 3 parameters one is mean one is the covariance matrix another one is the mixing

coefficient. So 3 parameters we are considering mean so you can see these 3 parameters I am considering mean covariance matrix and the mixing coefficient. $lnP(X|\mu, \Sigma, \pi) = \sum_{n=1}^{N} lnP(X_n) = \sum_{n=1}^{N} ln\{\sum_{k=1}^{K} \pi_k N(X_n|\mu_k, \Sigma_k)\}$.

So this is the log likelihood. So this ML technique I cannot apply in this case because there is no close form solution. So these parameters can be calculated by considering EM algorithm. So move to the next slide.

So we can think of the mixing coefficient as prior probabilities for the components.

So this sentence I can better to write. So we can think of mixing coefficients as prior probabilities. For the components so for a given value of x we can evaluate the corresponding posterior probabilities and which is called the responsibilities. So in the EM algorithm we are defining one term that is the responsibilities. That means for a given value of x we can evaluate the corresponding posterior probabilities and that is the responsibility. So from the Bayes rule we are considering this $\gamma_k(x)$ that is the responsibility and this is actually the latent variable we are considering.

So by applying the Bayes rule you can see I can write like this $P(k|x) = \frac{P(k)P(x|k)}{P(x)}$. So that is the unconditional prior. So this is by using the Bayes rule. So we can determine the posterior probability, probability of x given x we can determine with the help of this formula.

So now this class conditional density I can write as a mixture of Gaussian. That means what we are considering this class conditional density we are considering as a mixture of Gaussian and if I consider this denominator we are considering the summation of all the Gaussians. So $\sum j = 1^{K} \pi_j N(x|\mu_j, \Sigma_j)$ that is nothing but summation of all the Gaussians we are considering. So here $\pi_k = \frac{N_k}{N}$. So here N_k is nothing but number of points assigned to cluster k and N is nothing but total number of sample points all the data points.

So in case of the expectation maximization algorithm this term is very important the term is the responsibility. So we can determine the responsibility like this. This EM algorithm that is actually the iterative optimization technique which is operated locally. So let us move to this slide. So this expectation maximization algorithm that is the iterative optimization technique which is operated locally.

So expectation maximization algorithm this is the iterative optimization technique. So in the figure you can see this is the iterative technique we are applying. So you can see that we have initial point and after successive iterations we are getting the optimal point. So

the final optimal point I am getting after successive iterations. So in these iterations there are two steps one is the estimation step another one is the maximization step.

So in the EM algorithm I have two steps one is the estimation step. So in the estimation step what I need to do for given parameter values we can compute the expected values of the latent variable that is nothing but the responsibilities that is the expectation step. Next one is the maximization step. So in the maximization step update the parameters of the model based on the latent variable calculated using ML method.

So this latent variable is nothing but the responsibility. So update the parameters of the model. So what are the parameters of the model? The parameters of the model already I told you I have the parameter is the mean one is the covariance matrix and another one is the mixing coefficients. So these are the parameters of the model. So these are the parameters. So the parameters of the models are mean covariance matrix and the mixing coefficient.

So first I have to determine the latent variable. First I have to determine the latent variable and after this update the parameters of the model based on the latent variable calculated using ML method. So that is the concept of the expectation maximization algorithm. So now let us write the algorithm that is the expectation maximization algorithm for the GMM. So EM algorithm for GMM Gaussian mixture model.

So I want to summarize this algorithm. The first step is number 1 initialize the means μ_j covariance matrix Σ_j and the mixing coefficient π_j. So evaluate the initial values by likelihood. So one technique is randomly we can select these values. So initialize means initially we can select randomly the random values of these parameters.

And suppose if I consider a data set or the data distribution. So I may consider the mean of this entire data distribution as the initial guess or initial value that also I can consider. But one technique is randomly also we can consider the values of these parameters initially. After considering this initialization we have to go for the E-step that is the expectation step. What is E-step? We have to evaluate responsibilities using the current parameter values. So this is the latent variable already I told you this responsibility I can determine like this.

This derivation I have not discussed in this class. This can be also derived. So how to get this expression? So this responsibility I can determine like this. So we are considering K number of Gaussians. So this is the expression for the responsibility.

So we can evaluate responsibilities using the current parameter values. So this is the second step. After this move to the next step that is the start step that is the M-step that is the

maximization step. We can calculate the parameters using the current responsibilities. Re-estimate the parameters using the current responsibilities.

The parameters are mean, mean vector, $\mu_j = \frac{\sum_{n=1}^{N} \gamma_j(X_n) X_n}{\sum_{n=1}^{N} \gamma_j(X_n)}$. So this expression also it can be derived but I have not shown the derivation. This is the expression for the mean. And finally I can determine the mixing coefficient that is $\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(X_n)$.

So these three parameters I can determine from the responsibilities. And finally I have to converse. So the final step I have to go for the convergence. So move to the final step. So for this what we are considering for the convergence for testing the convergence of this algorithm because this is the iterative algorithm.

So evaluate log likelihood. I have three parameters mean, covariance matrix and the mixing coefficient. So n=1:N, N number of data points. And we are considering K number of Gaussians. So we can determine the log likelihood.

If there is no convergence, return to step 2. So how to actually determine the convergence? Suppose I can consider the parameters mean, covariance matrix and the mixing coefficient. In two successive iterations if there is not much changes of these values of these parameters that means I can stop the algorithm. I am repeating this I can determine the mean, I can determine the covariance, I can determine the mixing coefficients from the responsibilities. And in two successive iterations if there is no significant changes of the values of these parameters then I can stop the algorithm that is the condition for convergence.

Another way also I can do I can evaluate this log likelihood. Here you can see I am evaluating the log likelihood. So in two successive iterations if there is not much changes of this log likelihood value then also I can stop the algorithm that is the condition for the convergence. So based on the parameter values I can take a decision and based on the log likelihood also I can take a decision.

So when to stop the algorithm. So because this is the iterative algorithm. So let us now consider one illustration of the expectation maximization algorithm. In this figure you can see I am showing a data distribution. So these are data points distribution of data and with the help of single Gaussian it is not possible to approximate this data distribution. So that is why we have to consider number of Gaussians. In this case I am considering two Gaussians one is the red color another one is the blue colored Gaussian clusters.

And suppose this is the initial position of these two Gaussians one is the blue another one

is the red and all the data points I have shown as a as green color data points. So move to the next slide here you can see some of the points data points are assigned to the blue cluster that is you can see some of these data points are assigned to this that is the blue colored Gaussian and some of the data points are assigned to the red colored Gaussian. So this is the initial assignment. So here you can see this is the $l = 1$ means this is the first iteration after doing this you can see I am getting these two new Gaussians because I am assigning the sum of the data points to the blue colored Gaussian and some of the data points are assigned to the red colored Gaussian and after this assignment I am having these two Gaussians after the first iteration. and In iteration number two you can see the position of these two Gaussians in the data set so in the entire data distribution you can see the position of this two Gaussian after the second iteration in the feed iteration you can see the position of this two Gaussians one is the blue colored Gaussian and other one is the red color Gaussian and after twenty eighth iteration you can see the final position of this two Gaussian so this blue color Gaussian can approximate this all these data points that is all these data distribution and the red color Gaussian can approximate all these data points so that means by considering these two Gaussians I can approximate this entire data distribution.

So after this iteration I will be getting the convergence because not much changes in the values of the parameters the parameters are already I told you the mean vector covariance matrix and the mixing coefficients so based on this I can determine the convergence condition that is I can stop the algorithm. So all the steps here I am showing together already I have explained so initially I have shown two Gaussians and after this after the first iteration you can see the position of this Gaussians after the second iterations you can see the position of this Gaussians like this after the twenty eighth iterations you can see the position of this Gaussians and after this I can stop the iteration I can stop the algorithm. So this is the concept of the mixture model and this is the concept of the expectation maximization algorithm. So one application of this algorithm I can show here so this is applied in computer vision this computer vision application is we can consider the segmentation problem that is the segmentation of foreground and the background that is also called a background modeling and another one is the tracking. So the tracking of the object or the tracking in this case I am considering the tracking of a leaf and in this case you can see suppose corresponding to this video if I consider a this particular pixel or this pixels because in a particular video I have number of frames and this pixel values the distribution of this pixel values I can represent by a mixture of Gaussian.

Similarly in the second case also if I consider this point or any one of the points the pixel values because I have number of frames in the video so the distribution of this pixel values I can represent by a mixture of Gaussian and based on this I can develop a background modeling algorithm or maybe the tracking algorithm I can consider. So this is one

application of Gaussian mixture model in case of computer vision application. So one is the background modeling another one is the tracking. So here you can see I am playing this video and you can see just I am doing the segmentation separation of the foreground and the background and in this case for a particular pixel we are considering the Gaussian mixture model the mixture of Gaussians because single Gaussian cannot approximate the pixel distribution the pixel value distributions. Similarly in the second case also I can play this video this is the tracking in this case also we have to consider the mixture model the Gaussian mixture model.

So in this class I discussed the concept of the mixture model that is the Gaussian mixture model and why this Gaussian mixture model is important for approximating a data set or a data distribution that concept I have explained. After this I explained what is the problem of the maximum likelihood estimation to determine the values of the parameters in case of the mixture model the Gaussian mixture model because I will not get the close from of the solution. So that is why I have to consider the iterative algorithm that is the expectation maximization algorithm. In the expectation maximization algorithm first I have to determine the responsibility and after this with the help of this responsibility I can determine the values of the parameters and this is the iterative algorithm and after this based on some conditions I can stop the algorithm that is the convergence. So maybe I can consider the parameter values the parameter means the mean covariance matrix and the mixing coefficient and based on this I can stop the algorithm.

In two successive iterations if there is not much changes in the values of the parameters then based on this I can stop the algorithm. Also I can consider the log likelihood also for the convergence. In two successive iterations if there is not much changes of the log likelihood value then based on this I can stop the algorithm. So this is one important discussion the discussion on Gaussian mixture model and the expectation maximization algorithm. So let me stop here today. Thank you.