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-31

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of Gaussian mixer model and the expectation maximization algorithm. In case of the Gaussian mixer model, I fit k number of Gaussians into data. That means, a particular data distribution is represented by k number of Gaussians. And after this I discussed the concept of expectation maximization algorithm.

So, in this algorithm, first I have to determine responsibility and based on the responsibility, I can determine the values of the parameters. So, the first step is the expectation step in case of the expectation maximization algorithm. So, from the current parameter values, I can determine the responsibility. And after this in the maximization step from the responsibility, I can determine or I can recompute the parameter values.

So, parameters are mean vector covariance matrix and the mixing coefficients. So, this is the iterative process. So, I have to do the iterations until the convergence condition is not satisfied. So, yesterday I discussed about this concept. And today I am going to discuss the concept of another clustering technique that is the k-means clustering.

So, in unsupervised clustering technique, we group feature vectors based on similarity. So, to find a similarity we can consider some distance measures, maybe we can consider Euclidean distance or other distance measure. And this is the fundamental concept of clustering. So, in the k-means clustering, we mainly consider the distance of the data points to the centroids. So, based on this distance measure, I can assign the data points to different centroids.

So, I have K number of centroids. And we can do the clustering based on determining the nearest distance. So, the actual concept I am going to explain now. So, now let us begin this class. So, what is clustering? Clustering means group together similar points and

represent them with a single token that is the definition of clustering.

And in my discussion, I will be considering three popular clustering techniques. One is the k-means clustering and that is the iteratively reassigned points to the nearest cluster center. So, it is an iterative algorithm. And I will be explaining this concept after some time. Another popular technique is agglomerative clustering.

So, start with each point as its own cluster and iteratively merge the closest clusters. So, that concept I will be explaining in my next class. And finally, I will be discussing another popular clustering technique that is the mean shift clustering. So, it is mainly to estimate modes of a PDF. So, how to determine the modes of a PDF and based on this, I can consider it I can consider this clustering technique that is the mean shift clustering.

Now, let us discuss the algorithm that is the k-means clustering algorithm. So, first one is decide on a value for k that means how many clusters we need to consider that we have to decide. After this initialize the k cluster centers. So, k number of cluster centers I have to initialize and it can be done randomly. After this decide the class membership of the N objects that means N number of data points by assigning them to the nearest cluster center.

So, I have N number of data points and I am assigning them to the nearest cluster center based on some distance measure. So, I have to determine the distance between the data points and the cluster centers and based on the minimum distance I can assign these data points to different clusters. So, I have k number of clusters and corresponding to this I have k cluster centers. After this re estimate the k cluster centers by assuming that the membership found above are correct. So, after this I have to re-estimate the k cluster centers by assuming the membership found above are correct.

So, this is the fourth step and if none of the N objects that means if none of the N number of data points since membership in the last iteration then I can stop this algorithm. So, this is the concept the fundamental concept of the k means clustering. Now let us consider one illustration of this k means clustering. So, in this case in the step number one you can see I have these data points all these data point I am considering in the feature space and we can consider the Euclidean distance measure to find a similarity that means I have to find the nearest distance between the data points and the centroids. After this what I need to consider I need to consider the centroids the initial centroids I have to consider.

In this example I am considering three centroids k_1 k_2 and k_3. So, randomly I am selecting this k_1 k_2 and k_3 these are centroids. So, this is about the step number one. So, in the step number two I have to see the data points which belongs to a particular centroid. So, based on the minimum distance I can determine this.

So, I have to compute the distance between the data points and the centroids and based on the nearest distance I can assign the data points to the centroids different centroids. So, here you can see I am considering green data points these data points are assigned to the cluster center k_1 this is based on the minimum distance. And if you see the red points the red data points these are assigned to the centroid k_2 because these are closer to the centroid k_2. If you see the blue data points these are assigned to the centroid k_3. So, this is the step number two.

So, you can see I am getting three clusters corresponding to the cluster centers k_1 k_2 and k_3. And after this in the step number two here you can see I am recomputing the centroids because after the assignment of the data points to different centroids different cluster centers I have to recompute the centroid. So, based on all these data points if you see the green data points I am recomputing the centroid and you can see I am getting a new the position of the centroid for k_1 k_2 and k_3.

So, for all the centroids I have to recompute because based on the new data points I can compute the new value of k_1 k_2 and k_3. So, you can see here in the step number three I am getting the new positions of the centroids k_1 k_2 and the k_3.

So, this is the step three. So, in the step number three this green point there is a green point here the green data points this was assigned to k_1 and these data points if you see here this was assigned to k_3. Now in the step number four you can see these data point is now assigned to k_3. So, this is assigned to k_3 because it is near to k_3 as compared to k_1 and k_2. So, we have to do this for all the data points.

So, that means I am assigning these data points to the centroids k_1 k_2 and k_3 and after this I have to recompute the cluster centers. That means I have to recompute the centroids because after the assignments of the new data points the centroid value will be changing. So, I have to recompute the centroids k_1 k_2 and k_3 and you can see I am changing the position of k_1 k_2 and k_3 and after doing all the iterations you can see this is the final position in step number five of the centroid k_1 k_2 and k_3. So, if I do more and more iterations then there will not be any changes of the position of the centroids k_1 k_2 and k_3 or there may be some minimum changes then in this case I can stop the algorithm that is the k means clustering algorithm. So, in two successive iterations if there is not much changes in the position of the centroids k_1 k_2 and k_3 then I can stop the iteration.

So, this is the fundamental concept of k means clustering and in this case also I have shown the same concept the k means clustering this is taken from the Wikipedia. So, you can see

I have shown the data points and I have to initialize cluster centers. So, one is the red color you can see this is the one cluster center and another one is the green colored one another one is the blue colored one after this I have to assign points to the cluster based on the nearest distance and after doing this I have to recompute the means that already I have explained.

So, if I recompute the means then the position of the cluster centers will be changing. So, that means after the computing of the means the position of the cluster centers will be changing and the process one and

two that I have to repeat until the convergence condition is obtained.

The conversion condition is already I have explained in two successive iterations if there is not much changes in the position of the centroids or the means of the clusters then I can stop the algorithm. So, I have to repeat one and two until the convergence condition is not satisfied. So, this is the k means clustering. So, I can write the pseudo code for the k means clustering.

So, k means clustering. So, what is the pseudo code for this clustering. So, begin the algorithm initialize N is the number of data points k is the number of clusters and μ_1 μ_2 up to μ_k these are the cluster centers. So, randomly I can select the cluster centers these are the means μ_1 μ_2 μ_k these are the cluster centers. After this in the next step do classify N number of samples according to nearest distance to the mean to the cluster center.

So, after doing this, I have to recompute the cluster centers I have to recompute the means until no changes in the position of the means no changes in the value of the centroids and what is the output of the k means clustering the output of the k means clustering is nothing but the cluster centers.

So, it is μ_1 μ_2. So, these are the cluster centers. So, I have k number of cluster centers. So, that means I will be getting k number of clusters and finally end. So, in this case, it is the unsupervised technique because we are not considering class information. So, we are not considering the class information and that is why it is an unsupervised technique.

So, this is a pseudo code for k means clustering. Now, the mathematically how I can consider this problem that k means clustering algorithm. So, the goal is goal of the k means clustering cluster to minimize variance in data given clusters. So, by preserving information what we need to consider to minimize variance in data given clusters. So, that means what I can consider here, I am considering these 2 parameters C^* and δ^* arg minimum corresponding to these 2 parameters C and δ and I have N number of data points

samples.

So, $C^*, \delta^* = argmin_{c,\delta} \frac{1}{N} \sum_j{}^N \sum_i{}^k \delta_{ij} (C_i - X_j)^2$. So, in this expression just I want to explain this expression this C_i it is actually the cluster center. So, in this case you can see I have considered k number of clusters k number of clusters we are considering.

In this case X_j that is actually the data. So, that is a data vector X_j is nothing but the data vector or the feature vector and this parameter the δ_{ij} what I am considering herethe δ_{ij} to seek whether this data X_j is assigned to the cluster center C_i. So, this parameter delta I am considering to seek whether the data X_j is assigned to C_i, C_i is the cluster center because we are assigning a particular data point to a particular cluster center based on the nearest distance that concept I have explained. So, that is the parameter delta we are considering. So, whether a particular data point is assigned to a particular cluster center.

So, this is the objective of clustering. Now let us write the k-means clustering algorithm based on this mathematical expression. So, the k-means clustering algorithm I can write like this k-means clustering. So, the first step is initialize cluster centers. So, randomly I can consider that t=0. So, mainly we are considering number of iterations.

So, to consider this number of iterations I am considering this variable. So, it is initialized. So, t=0. Number 2 assign each point to the closest center. So, based on the minimum distance I have to assign each of the data points to the closest cluster centers.

So, for this we are determining these parameters the δ parameter because I have to see whether a particular data points belongs to a particular cluster center. So, based on the minimum distance I can determine this. So, $\delta^t = argmin_\delta \frac{1}{N} \sum_j{}^N \sum_i{}^k \delta_{ij} (C_i^{t+1} - X_j)^2$. After this what we need to do we have to update the cluster centers as the mean of the points after assigning each point to the closest cluster center I have to update the cluster centers as the mean of the data points.

So, update cluster centers as the mean of the points. So, C is the cluster center and I am updating the cluster centers. So, update the cluster center as the mean of the points and finally, what I need to consider I have to repeat the steps 2 3 2 and 3 until no points until no points are reassigned. So, we have to increase the iteration number. So, mathematically I can write the k-means clustering like this.

So, this is a k-means clustering algorithm. So, in this k-means clustering what is the important point the first important point is the initialization. So, the initialization is important. So, this point is important initialization. So, we can randomly select k number

of points as initial cluster center initial cluster center and the distance measure what we can consider because we have to find the nearest distance we can consider Euclidean distance.

This is a popular distance measure. So, we may consider Euclidean distance or maybe we can consider other distance measure and you can see finally, I am getting all the cluster centers and that means, we are doing the optimization that means, it will converge to a local minimum. So, I can say it is the optimization what optimization will converge to converge to a local minimum. So, this is about the k-means clustering. So, what are the pros and cons of the k-means clustering the pros and the cons pros and the cons of the k-means clustering.

So, for the pros of what I can tell you. So, we have to find cluster centers that minimize conditional variance that means, it is a good representation of data good representation of data because what we are considering we are finding cluster centers that minimize conditional variance. So, that is why it is a good representation of data and also it is simple to see this algorithm is a very simple process simple and I can say it is a first and it is easy to implement easy to implement. So, I can say these are the pros of the k-means clustering. What are the cons of this algorithm. So, the cons I can highlight like this first the problem is I have to select the k number of centroids.

So, how to select this that is also one important research problem. So, how to select the k number of centroids on what basis you can select the k number of centroids and this algorithm is sensitive to outliers. So, this concept I can show pictorially also. So, if I consider suppose two clusters. So, these are some data points that belongs to this particular cluster and I have some other data points this belongs to another cluster and suppose I have one outlier that is another data point.

So, I have this outlier data points. So, this is the outlier point. So, this is the ideal cluster but practically sometimes it may not happen. So, what will happen practically I can show you also. So, I may get this type of cluster and these two clusters may be very close and this outlier point. So, these two clusters may come closer to each other and also these outlier points may come to the clusters.

So, what will happen sometimes it may happen these two clusters may come closer to each other and this outlier point may be included in one of the cluster because we are considering the minimum distance. So, practically it may happen like this. So, the first one is the ideal cluster you can see the outlier points and two clusters are quite widely separated. This is a very good condition.

So, two clusters are very nicely separated. But in the second case the two clusters are very

close to each other and also one outlier point is included in one of the cluster. So, that is not a good case. So, that is the sensitive to outlier. So, another disadvantage is this algorithm is prone to local minimum. So, it may stuck to a local minimum because the entire algorithm is based on the minimum distance.

And earlier I told you that is a fast algorithm. But however, if you see the computational complexity is iteration how many computation I need for each iteration, the k number of clusters and we are considering N dimensional data points. So, Nd dimensional data points. So, we are considering the computational complexity for each iteration k is the number of clusters N is the number of data points and this d that is the d dimensional points that is the dimension of the feature vector. So, for each iteration you can see the computation of complexity, the order of the computational complexities, you can see kNd.

So, it depends on the dimension of the feature vector. So, these are the disadvantages of the k means clustering. And if I want to compare the k means clustering with the Gaussian mixer model and the expectation maximization model. So, what will happen in case of the Gaussian mixer model already I told you. So, in case of the GMM what we are doing actually we have data points suppose these are the data points.

So, all these are different data points. So we are fitting a Gaussian for this data distribution. So, we are fitting multiple Gaussians. So, like this we are considering these Gaussian distribution. So, like this we are considering the four Gaussians number one, number two, number three, these four Gaussians we are considering and these Gaussians I am fitting into the data set. That means, the entire data set is approximated by four number of Gaussians.

But in case of the k means clustering, we are considering the concept of nearest centroid. So, that means we are identifying clusters by nearest centroid in case of the k means clustering. So, k means clustering what we are considering we are considering the nearest centroid. So, it is based on the minimum distance we are finding the nearest centroid. But in case of the GMM the Gaussian mixer model, we are fitting a set of k number of Gaussians to the data points and we are applying the maximum likelihood over a mixer model.

So, you can see the fundamental difference between the GMM and the k means clustering. I am repeating this in the case of the k means clustering we are identifying clusters by nearest centroids. But in case of the GMM, we are fitting a set of k Gaussians to the data and we are applying maximum likelihood over a mixer model. So, that is the difference between the GMM and the k means clustering.

In this class, I explained the concept of k means clustering. So, in the k means clustering algorithm, first I have to randomly select k number of centroids corresponding to k number of clusters. After this, I have to assign data points to the centroids based on the minimum distance. The distance measure I can consider the Euclidean distance measure I can consider. After this I have to recompute the centroids. And this process I have to do iteratively until the convergence condition is not obtained.

And this is the concept of the k means clustering. And if you see the fundamental difference between the k means clustering and the GMM, the k means clustering makes hard decision because the data points are assigned to the cluster centers based on the nearest distance. GMM or the expectation maximization algorithm makes soft decisions because there is a possibility that a particular data point may belong to some other clusters. So, that means, the GMM or the expectation maximization algorithm makes soft decisions. And the k means clustering is a special case of expectation maximization algorithm. So, you can see the similarity between the GMM EM that is the expectation maximization algorithm and the k means clustering.

So, this is about the k means clustering. So, let me stop here today. Thank you.