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-9

Lecture-33

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of the fuzzy K means clustering. Today I am going to discuss the concept of another two popular clustering techniques. The first one is the hierarchical agglomerative clustering technique. And the second one is also very popular that is the mean shift clustering technique. In case of the agglomerative clustering, each data point is considered as a cluster.

And these two data points or two clusters can be merged if they are very similar. The similarity condition can be mathematically defined and based on the similarity measure, I can find a similarity between two data points or two clusters. If they are very similar, I can merge these data points or these two clusters. And this procedure I have to repeat until some convergence conditions are not satisfied.

And finally, I will be getting the clusters. So, this is the fundamental concept of the agglomerative clustering. In case of the mean shift clustering, the data points or the sample points in the d dimensional feature space are considered as an empirical probability density function PDF. And the dense region in the feature space corresponds to the mode of the PDF. And after this, I have to apply a gradient SN procedure on the PDF until convergence.

The output of this procedure is the stationary points. So, I will be getting some stationary points that is nothing but the modes of the PDF. And the data points associated with a particular stationary point are the members of a particular cluster. So, that is the fundamental concept of the mean shift clustering.

So, let us discuss about these two clustering techniques.

The first one is the agglomerative clustering. And second, I will explain the concept of the mean shift clustering. So, in case of the hierarchical clustering, that is the agglomerative clustering. So, start with every point in its own cluster.

So, that means, each data point I can consider as a cluster.

So, repeatedly merge the closest two clusters. So, I can merge two clusters based on similarity, if two points are very similar or the two clusters are very similar I can merge. So, different definitions of closes give different algorithms. So, we will discuss about this concept. So, how to determine the similarity between two data points or two clusters and that point I am going to discuss in my next slide.

So, this is the concept of the hierarchical clustering. So, that means, we have to consider every data point as a cluster. And after this, we can merge these data points or maybe two clusters, if they are very similar.

This similarity I can mathematically define. So, how to define the similarity.

Let us move to the next slide. So, here I am showing this example, that is the agglomerative clustering. So, you can see here the data points I am showing in the figure and in this case is a very simple clustering algorithm. So, first I have to define a distance between the clusters, because I have to find the similarity between the clusters. So, how to do this, compute the distances between all clusters and merge two clusters, which are very closest.

So, merge two closest clusters. So, we can find the similarity between the clusters. If they are very similar, we can merge two clusters, save both clustering and sequence of cluster operations. And finally, I will be getting different clusters and output of this algorithm, I will be getting a dendrogram. So, what is the dendrogram I will show in my next slide.

So, this is the procedure of the agglomerative clustering. So, based on the similarity, I can merge two clusters, the closest clusters and we will be getting a dendrogram. So, what is the dendrogram you can see in my next slide. So, here you can see I am merging two data points, because they are very similar, you can see two red data points and they are very close, we can apply some distance measure to determine the similarity between these data points or maybe two clusters. And since they are very closest, they are very similar.

So, we can merge. Again, we can see another two points here in the figure that two blue colored data points, they are also very similar. So, we can merge. And again, I am considering the another point, the third point, the red point.

So, that point is also very close to the previous two red points.

And that can be also merged. So, that can be also merged in the cluster. So, like this, I have to do this iteration.

Finally, I am getting this dendrogram, if you see the dendrogram. So, this is a tree like structure.

So, if I do the merging based on the similarity, I can find the dendrogram. In this case, I am showing the dissimilarity, the similarity or the dissimilarity you can define mathematically. And if the two points are similar, I can merge. So, finally, I am getting the dendrogram. So, it is similar to a tree like structure.

And you can see this is the fundamental concept of the dendrogram. So, now how to define cluster similarity. So, we may consider some mathematical measures like average distance between points. So, maybe we can consider average distance between the points, maximum distance we can consider, minimum distance we can consider, distance between the means we can consider. So, based on this measure, we can find the similarity between the data points or the similarity between the clusters.

So, because of the clustering, I am getting a dendrogram that is a tree like structure. So, whenever we do the clustering based on this definition, so, I will be getting a dendrogram and we can consider a threshold based on maximum number of clusters that is the threshold based on the maximum number of clusters. And based on this, I can stop the algorithm. So, some threshold I can consider or maybe the distance between the merges that measure also we can consider to stop the algorithm. So, one is the maximum number of clusters we can consider.

So, that threshold we can consider and based on this threshold, I can stop the algorithm or maybe I can consider the distance between merges. So, that point also I can consider to stop the algorithm because it is the iterative algorithm and we have to repeatedly do this and I am getting a dendrogram, it is very similar to a tree and based on this I am getting the clusters. So, for measuring the similarity between the data points, we can consider some mathematical measures.

So, maybe the cluster distances we can define like this. So, if I consider two clusters, suppose C_i and C_j, two clusters, and we can find the minimum distance between these two clusters.

So, x is a point corresponding to the cluster C_i. So, x is a point, the data point corresponding to the cluster C_i and y is the data point in the cluster C_j. And we can find the distance between x and y. So, this distance we can determine that is the Euclidean distance and the minimum distance we can determine and based on this we can find a similarity between

two clusters or two data points. So, pictorially I have shown here.

So, mainly we are considering the nearest neighbor. So, it produces minimal spanning tree. So, if I consider this distance measure, I will be getting the minimal spanning tree. So, this minimum distance I am considering or maybe I can consider the maximum distance between the clusters. So, if you see the second distance, the distance between two clusters C_i and C_j and we are considering the maximum distance between x and y.

So, x is the data point of the cluster C_i and y is the data point of the cluster C_j and we are finding the distance between x and y and we are considering the maximum distance. So, in the figure also I have shown this condition that is the maximum distance between x and y we can consider. And if I consider this maximum distance, it can avoid the elongated clusters. So, the second figure also I have shown the concept of the maximum distance or maybe we can consider the average distance between the clusters. The D_{avg} we can consider the average distance between two clusters also that we can consider for determining the similarity between the data points or the clusters.

And finally, I can determine the distance between two means. So, the mean μ_i corresponding to the first cluster and μ_j that is the mean corresponding to the second cluster. So, we can determine the distance between these two means and based on this we can find the similarity between the clusters. So, based on the similarity measure, I can merge two data points or two clusters. So, that is the concept of the distance measure.

So, finally, it is a very simple algorithm that is the agglomerative clustering. So, what are the pros of this algorithm simple to implement and there are many applications of the agglomerative clustering clusters have adaptive shapes. So, that is also another advantage and provides a hierarchy of clusters because I am getting the dendrogram and I am getting a hierarchy of clusters and what are the cons of these algorithms. So, I may get imbalanced clusters. So, may have imbalanced clusters and in this case also I have to consider the threshold that is we have to select the number of clusters or the threshold I have to select.

So, that is the disadvantage of this algorithm. So, we have to select the number of clusters or maybe the threshold to terminate this algorithm because we have to consider the convergence condition. So, based on the threshold so that also I have to define. So, these are the disadvantages and advantages of the algorithm that is the agglomerative clustering algorithm.

So, now let us move to the second clustering technique that is the mean shift clustering technique.

So, here I am showing one illustration. So, in this case you can see I am considering some data points in the feature space and first what I have to consider I have to consider a region of interest. So, if you see in the figure I am showing a region of interest and the center of this region of interest I have shown. So, this is the center of the region of interest. So, corresponding to this region of interest you can see the sample points within this particular region of interest. So, all the data points within this region of interest you can see.

So, corresponding to these data points I can find the modes. So, if you see so this is the dense region that is the mode of the data points corresponding to this region of interest that is the center of mass we can determine from these data points in the region of interest because we have to consider the region of interest and the data points within this particular region of interest. So, that is the center of mass corresponding to these data points and based on these two information I can determine the mean shift vector the mean shift vector I can determine. So, that means I have to move the region of interest to the new center of mass and initially I have considered the center of the region of interest. So, that point I have to shift to the new position the new position is the center of mass.

So, based on the mean shift vector I can shift the position of the region of interest. So, you can see I am changing the position of the region of interest and corresponding to this position again you can see the center of the region of interest and corresponding to this region of interest you can see the data points within this particular region of interest and corresponding to this again I can determine the center of mass. So, this is the center of mass corresponding to these data points and based on this I can determine the mean shift vector and after this I have to change the position of the region of interest. So, you can see I am changing the position of the region of interest based on this mean shift vector and this is another position of the region of interest and corresponding to this position you can see all the data points within this particular region of interest and again you can determine the center of mass and you can move the region of interest and finally, you can see this is the position and based on the mean shift vector I can change the position of the region of interest. So, this algorithm I have to do iteratively until some convergence condition is not satisfied and corresponding to this position you can see I am getting a cluster a particular cluster.

Suppose this is the final position of the algorithm or the final step of the algorithm and corresponding to this position I am getting a particular cluster. So, this is a simple procedure because we have to determine the modes of the data points because we are considering the distribution of data and that I can consider as a pdf the probability density function and we have to find the modes of the pdf the probability density function. So, finally, I am getting the clusters. So, this is the cluster center corresponding to all these

data points. So, what is the mean shift if you see here in this figure I am showing the distribution of data.

So, the non-parametric density gradient estimation technique is nothing, but the mean shift algorithm. So, in my previous classes I discussed one important concept that is the Parzen window technique if you remember the Parzen window technique. So, in the Parzen window technique we can determine the density in this case you can see I am showing the data distribution and we can estimate the density that is the non-parametric density gradient estimation that is nothing, but the mean shift. So, this pdf if you see this pdf here this pdf actually represents the data distribution. So, this is my original data distribution and the pdf if you see that pdf corresponds to the data distribution.

So, I have to estimate the density. So, earlier I discussed the concept of the Parzen window. So, with the help of the Parzen window I can determine the density. So, in this case pictorially I have shown here I have shown the data points. So, if you see the real data samples in the figure and corresponding to this real data samples I am showing the pdfs. So, if you see the dense region corresponds to the mode of the pdf.

So, if I consider this is one pdf and this is another pdf and maybe another pdf I can consider here. So, the peak actually it corresponds to the modes of the pdf. So, the real data samples I have shown and that is actually represented by the pdf. So, that is the empirical probability density function. So, that means treating the points in the d dimensional feature space as an empirical probability density function.

And the dense regions in the feature space corresponds to the local maximum or the modes of the distribution that is the concept of the density estimation. So, I am repeating this the dense region in the feature space corresponds to the local maxima or the modes of the distribution. And for each data point a gradient SN procedure can be applied on the local estimated density until convergence and the stationary points of this procedure gives modes of the distribution. The data points associated with the same stationary points are considered as members of the same cluster.

So, this procedure already I have explained. So, how to determine the modes and the data points associated with the same stationary points that means the modes are considered members of the same cluster. So, here in this figure I have shown this concept. So, these data points or the distribution of this data I am showing as a pdf the probability density function. So, in this case the problem is the kernel density estimation. So, if you see the concept of the Parzen window that is also the same concept the kernel density estimation.

So, we can estimate the density. So, here the px is the density and we can consider this formula the $P(x) = \frac{1}{n}\sum_{i=1}^{n} K(x - x_i)$. So, we are considering number of data points x_1 x_2 x_3 these are the data points and based on this formula I can determine the density. So, that concept already I have explained in my class of Parzen window. So, this data distribution I can consider as a probability density function. So, we may also consider the kernel function something like this.

So, the product form is considered. So, in this case I am getting the same function on each dimension or maybe I can consider this from the kernel function. So, it is nothing but the radially symmetric kernels. So, this is one example of the radially symmetric kernels and function of vector length only. So, we are considering the length of the vector K(x) is the kernel profile and c is nothing but the normalizing constant.

So, here the c is nothing but the normalizing constant. So, this c is the normalization constants. So, mostly we can consider the radially symmetric kernels or maybe we can consider this kernel also, but in most of the applications we consider radially symmetric kernels, because it is a function of vector length only. So, these kernel functions I can consider. So, if I consider this is the expression for the density and we are considering n number of data points. Now, we can consider this type of kernel functions the first one is the triangular kernel functions.

So, that is represented mathematically like this. The second one is the uniform kernel. So, this is $K_U(x) = cif||x|| \leq 1$ otherwise it is 0. So, corresponding to this you can see the density. So, if I consider the triangular kernel. So, corresponding to this, this is the density representation or maybe I can consider the normal kernel that is nothing but the Gaussian kernel and c is the normalization constants.

So, corresponding to the this normal kernel I am getting this pdf. So, these type of kernels I can consider for the density estimation, because I have to estimate the density the P(x) and we need the kernel functions. So, maybe we can consider the normal kernel uniform kernel or triangular kernel. So, these type of kernels I can consider.

So, now, what I am considering I am taking the gradient of the pdf. So, in this expression I am taking the gradient of the density. So, that is nothing but the kernel density gradient estimation. So, that point we are considering the kernel density gradient estimation we are considering. And if I consider this kernel form $K(x - x_i) = ck(||\frac{x-x_i}{h}||^2)$ and h is nothing but the size of the window and actually it is also called the bandwidth parameter. So, it is also called the bandwidth parameter that is actually the radius of the kernel.

So, this kernel we can consider in this example and if I take the gradient of this. So, from this expression actually if I take the gradient of this kernel you can see I am taking the gradient of this. So, if I do some mathematics then I will be getting this expression. So, $g_i = -k'(||\frac{x-x_i}{h}||^2)$.

So, that I am getting this expression. So, in this case we are considering $g(x) = -k'(x)$ that is actually the derivative of the kernel profile derivative of the kernel profile. So, what we are actually doing in this case we have defined the density after this I am taking the gradient of this density. So, I am finding the gradient estimation. So, ∇k because k is the kernel functions and if I do the mathematics. So, mathematics I will be getting this expression and g_i is nothing but this.

So, $g(x) = -k'(x)$ that is actually the derivative of the kernel profile. So, this expression I am getting that is the kernel density gradient estimation. So, from the previous slide I am getting this gradient I am getting this expression. So, you can see this expression actually I am calculating the gradient of the density function. So, this expression I am getting this expression I can consider for computing the mean shift in my illustration I had explain what is the mean shift vector.

So, mathematically I have to explain what is the mean shift vector. So, in this expression you can see I have 2 terms the first term is this and second term is this. So, second term is actually the mean shift vector. So, in my next slide I will be explaining what is the first term and what is the second term the second term is nothing but the mean shift vector. So, move to the second slide. So, now I will show what is actually the first term and what is the second term the first term is nothing but the kernel density estimation and if I consider the second term the second term is nothing but the estimation of the mean shift vector.

So, what is the mean shift vector I can show in my next slide here I am showing what is the mean shift vector in this expression I am determining the mean shift vector m(x). So, I have to compute the mean shift vector and that is nothing but the m(x) that is the mean shift vector. So, in this expression if I see this expression this x is nothing but the original position of the region of interest that is the x the original position of the region of interest and if I consider this term up to this term that is nothing but the center of gravity of the data points within this particular region of interest or I can say this is the mode of the distribution the modes of the density function. So, the second term corresponds to the second point that is nothing but the center of gravity or the modes of the density function and the difference between these two is nothing but the mean shift vector the difference between these two is the mean shift vector. So, we can determine the mean shift vector by considering these two points the first is x is the original point and we can determine the center of gravity the difference between these two is the mean shift vector.

So, what is the mean shift clustering the mean shift algorithm finds modes of the given set of data points. So, what is the procedure first I have to select the kernel and the bandwidth the bandwidth is the h is the bandwidth and for each point center a window on that point and compute the mean of the data in the search window and after this center the search window at the new mean location and I have to repeat the steps b and c until the convergence condition is not obtained. This is the algorithm of the mean shift clustering and after this in the third step assign points that lead to nearby modes to the same clusters. So, what is the meaning of this the data points associated with the same stationary point that is the mode are considered members of the same cluster. So, that is the meaning of this that means the data points associated with the same mode or the stationary point are considered members of the same clusters.

So, that is the concept of this mean shift clustering. So, I am repeating this we have to compute the mean shift vector translate the kernel window by the mean shift vector recompute the weighted mean and after this I have to stop the iteration if gradient is closest to 0. So, that is the convergence condition. So, these steps I can write again. So, number one step is the first I have to compute the mean shift vector.

So, mean shift vector is suppose $m(x_i^T)$. So, first I am determining the mean shift vector number one number two translate density estimation window density estimation window. So, how to translate this one.

So, x_i^{t+1} iteration. So, this should be t. So, in the t+1 iteration. So, $x_i^t + m(x_i^t)$ this is a mean shift vector I am considering. So, based on the mean shift vector I can do the translation of the window and after this number three iterate step one and two until convergence. So, that means the convergence I can consider like this the gradient of the density function should be equal to 0.

So, that condition I can consider. So, these are the steps of the mean shift algorithm. So, this mean shift algorithm has many applications. So, here I am explaining one application in image processing. So, for image segmentation I can apply the mean shift clustering algorithm. So, we have to find the modes of the non-parametric density.

So, in the image you can see I am showing the pixel distributions. So, if I consider the grayscale image intensity values or I can also consider the color values, but here I am showing the grayscale intensity values and corresponding to this you can see the distribution of data points this distribution you can see. So, corresponding to this distribution you can see the modes of the density. So, here you can see I am getting one mode another mode I am getting. So, these all these modes I am getting corresponding to

this data distribution.

So, these are the modes of the density function. So, this I can represent like this. So, corresponding to these modes I have clusters. So, this is one cluster. So, this corresponds to another cluster this corresponds to one cluster this corresponds to one cluster. So, I will be getting all these clusters based on this probability density function. So, you can see the modes of this distribution function or the modes of the density function and corresponding to this I will be getting the clusters like this.

And based on these clusters I can do the segmentation image segmentation is nothing, but the partitioning of an image into connected homogeneous region. So, this homogeneity I can define in terms of the grayscale value or maybe in terms of color value or maybe in terms of texture, but in this case I am considering the grayscale intensity value. So, I am repeating this the partitioning of an image into connected homogeneous region and the homogeneity is defined in terms of the grayscale value. So, these are some examples of image segmentations. So, we are applying the mean shift segmentation algorithm, the mean shift clustering algorithms and you can see the segmented outputs.

So, the first one is one result and segmented landscape number two that is also another result. And this is also another example, the mean shift segmentation results corresponding to these images. So, we are getting some results. So, this is one result and also we are getting these results based on this mean shift clustering techniques. And this is also another result.

In this case also we are applying the mean shift clustering technique. So, this is one application of mean shift clustering technique. This mean shift clustering technique can be used in image processing for image segmentation. In this class I explained the concept of two popular clustering techniques. The first I explained the concept of the agglomerative clustering technique and second I explained the concept of the mean shift clustering. In case of the agglomerative clustering, two data points or two clusters can be merged together if they are very similar.

The similarity I can mathematically define and based on the similarity I can merge two clusters or two data points. So, this process I have to repeat iteratively until some convergence conditions are not satisfied. And finally, I will be getting the clusters. In case of the mean shift clustering, I have to determine the mean shift vector and I have to move the region of interest to a new position based on the mean shift vector.

And finally, I will be getting the clusters. So, already I have explained this concept and the mean shift algorithm has many applications. So, one application already I have

explained that is the image segmentation. So, I can do the image segmentation with the help of the mean shift clustering. Another application is the tracking, object tracking.

In computer vision one popular application is tracking. So, I can do the tracking with the help of this algorithm that is the mean shift algorithm. So, this is about the agglomerative clustering and the mean shift clustering. So, let me stop here today. Thank you.