Mine Automation and Data Analytics

Prof. Radhakanta Koner

Department of Mining Engineering

IIT (ISM) Dhanbad

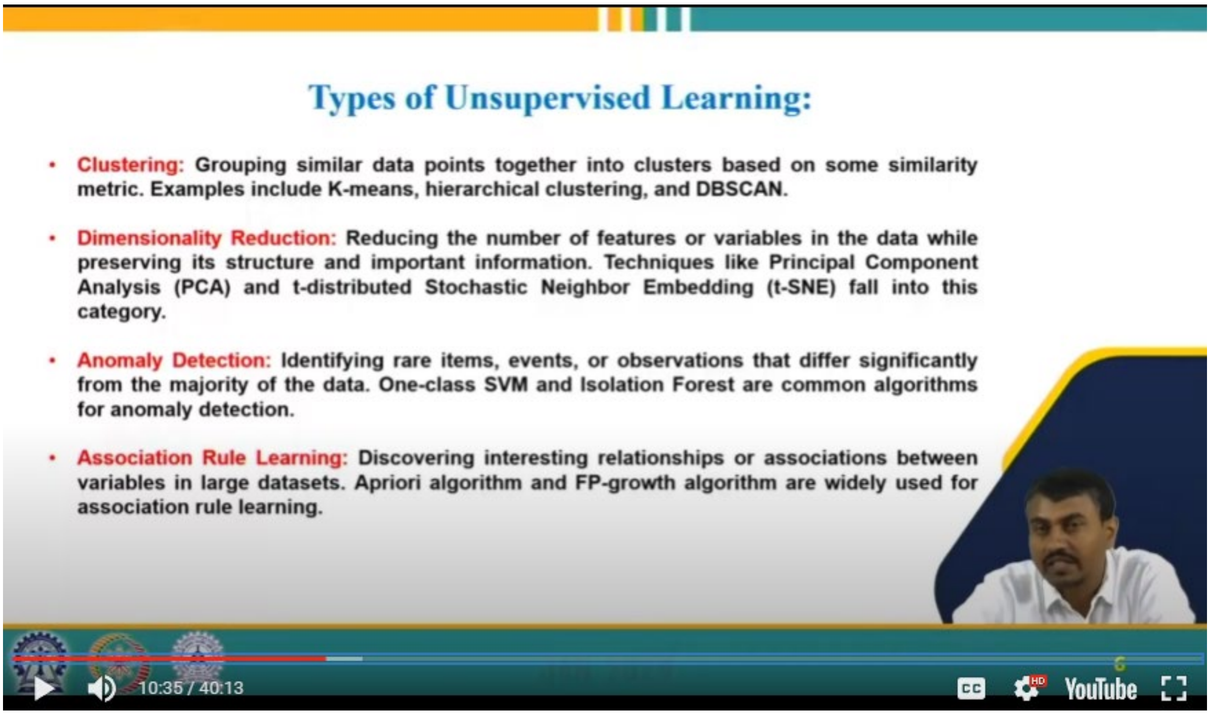
Week 11

Lecture 52: K Means Clustering

... Welcome back to my course, Mine Automation and Data Analytics. Today, we'll discuss the unsupervised machine learning approach. In the last few lessons, you have seen that we were discussing the supervised machine learning approach. And those approaches are primarily focused on regression. And with the ANN, we are also doing some sort of classification as well. So today's topic is completely unsupervised learning. And out of that, we'll discuss it today, and our discussion will be more centered on the clustering method. And that is the K-means clustering method. So in this lesson, we'll first discuss what unsupervised learning is. And then we'll discuss the different applications and challenges associated with unsupervised learning. And we'll introduce the concept of clustering and the different algorithms we use for this clustering method. And we'll introduce K-means clustering with a detailed mathematical explanation. Then the assumptions required for these methods and their advantages and disadvantages, as well as the applications we'll discuss subsequently, So what is an unsupervised machine learning approach? The unsupervised machine learning approach and the difference between the supervised learning approach and the unsupervised learning approach can be represented by these two figures, figures 1 and 2. So these data points in the supervised learning are basically making an alignment, or a linear alignment, corresponding to these particular lines. And we know very well the efficiency of these alignments and how far they are basically located from these central data.

So based on that, the performance of this kind of learning is assessed. If these distances are very small, then we can say the regression is better achieved for this data set. Whereas in comparison to that, here we have a data set, and they are specially located on the x1 and y2 axes. And from the spread of these data points, it can be well observed that these data points are the red color, the blue color, the green color, and basically the violet color. So these points are basically forming clusters. Cluster in what sense? Cluster in the sense that they are basically located in close proximity to each other. The similarity here is their close proximity to each other. And that is the basis for subdividing these data sets into three clusters. So we have subdivided these data sets into three clusters: two and three. Interestingly, we do not have any level data. We do not have prior information that reads that this type of data belongs to that. Based on the spatial locations, their spread, and their closeness to each other, we have come to the conclusion that these data sets represent primarily three major clusters.

So this is the distance, or, as you can say, the difference, between supervised learning and unsupervised learning. So in unsupervised learning, we train the model without any explicit supervision or labeled data set. So what is the advantage? The advantage is that the labeled data set is not present; the data set is present for training. From there, we are basically finding out some patterns. We are basically extracting the hidden pattern in the data set and the structural relationship between these data. And based on that, we are predicting that these data sets mostly represent three clusters or three types of data sets. So the major advantage of unsupervised learning is that it extracts and discovers some hidden information or hidden insight about the association of these data points. So that is the big advantage compared to supervised learning.



Types of Unsupervised Learning:

- **Clustering:** Grouping similar data points together into clusters based on some similarity metric. Examples include K-means, hierarchical clustering, and DBSCAN.
- **Dimensionality Reduction:** Reducing the number of features or variables in the data while preserving its structure and important information. Techniques like Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) fall into this category.
- **Anomaly Detection:** Identifying rare items, events, or observations that differ significantly from the majority of the data. One-class SVM and Isolation Forest are common algorithms for anomaly detection.
- **Association Rule Learning:** Discovering interesting relationships or associations between variables in large datasets. Apriori algorithm and FP-growth algorithm are widely used for association rule learning.

We have to remember that supervised learning and unsupervised learning have their own pros and cons. It is on the user; it is on the user who is using. He or she has to decide, or the community has to decide, which machine learning model is appropriate for the purpose. Because ultimately, whatever we are discussing has to be used in real life or in analyzing our live data. For example, if we analyze the data related to the accident in a mine and collect detailed data, suppose this detailed data is collected. Not only by accident, a kind of unwanted situation arises; their time, their locations, their shift, whatever—detailed information. So if this kind of data is collected, suppose 7 weeks, 12 months, or maybe 2-3 years, if this data is now looked into, that may represent that if these mines are operating with the same kind of conditions, then this data will represent some clusters. That at what point in time or at what locations there was much more proneness to this kind of accident might be. So based on that, we can cluster those sensitive areas, and these areas are very sensitive.

So this is basically an evidence-based approach to classifying this area as more prone to failure or accident. Maybe in that area there is a bending, and that bending is because the elevation was different, so maybe the truck operator is not able to see. So some unwanted collisions or situations might arise repeatedly. So these unsupervised machine learning approaches may help us come to a new conclusion that might be helpful for our decision-making process. And let us start with these assumptions. So here, there is no supervision, and we do not require any labeled data sets. The machine learns from the data itself alone without any external feedback. And that is why it is more suitable for exploratory analysis. So what are the types of unsupervised learning that are available? We have already shown you this particular figure at the start of the machine learning approach. So you can see that clustering, dimensional reduction, marketing, target marketing, and customer segmentation are some of the popular approaches of the unsupervised machine learning approach.

So clustering is a common form of application for unsupervised learning. So what does clustering do? Clustering is the grouping of similar data points together into clusters based on their similarity metric. This similarity metric might be of Euclidean distance, which means closeness with each other, or might be some kind of qualitative closeness with the data. I am talking about the quantitative distance in the coordinate system, and that includes k-means clustering, hierarchical clustering, and DBSCAN. The second example is the dimensionality reduction methods, which basically reduce the number of features or variables in the data while preserving its structure, main structure, and important information in the data set. So that is the basic tenet of these dimensionality reduction methods, and principal component analysis is one of the most popular methods of PCA.

Another method is the t-distribution stochastic neighbor embedding, which is t-SNE. This is an example of a dimensionality reduction method. The third is anomaly detection. It basically identifies the rare item in the data set or observation that differs significantly from the majority of the data. That might be required sometimes, and these anomaly detections—this kind of approach—might be very useful for the treatment of some kind of disease based on certain observations. So in the medical science, these anomaly detection methods have a lot of applications, which means they have applied applications, and nowadays, the medical community is using these methods as well. Association rule learning basically discovers the interesting relationship or association between the variables in large data sets.

So the apriori algorithm, or FP-growth algorithm, is an example of this unsupervised machine learning approach. What are the applications of this unsupervised learning? I have already given one example of medical science anomaly detection. So this is very useful out of the major features because something is deviating from that that might lead to some kind of situation, and that might be the possible reason for this kind of situation. Suppose the baby grows okay during the pregnancy. So there is a wide pattern with the large data set for a few years, 20 years, 40 years, and 30 years. So doctors basically measure whether the growth of the child is according to the natural tendency of the child. So if it is not, then the doctor take some kind of action to ensure that this child will be in good health in this particular lab. So there are a wide range of applications for this kind of technology and approach. So let's focus more on the engineering aspect.

One is market segmentation, which is one approach to identifying the group of customers, their behavior, their search, their likes, and their choices. Based on that, you advertise them on those related products. Image and text clustering involves grouping similar kinds of images or documents together based on their contents. This is very, very useful. Anomaly detection, I already said. In recommendation systems, you recommend to the user those movies based on the preferences of that particular user or customer. This is a very good method. Dimensional reduction is a very useful method, particularly in Industry 4.0. We want to transmit the data in real time. So we have to reduce the data size. So we have to reduce the unnecessary features while preserving the principle features so we can transform the network and not overload it. So for those kinds of situations or applications, these dimensional reduction methods are one way to help maintain the system.

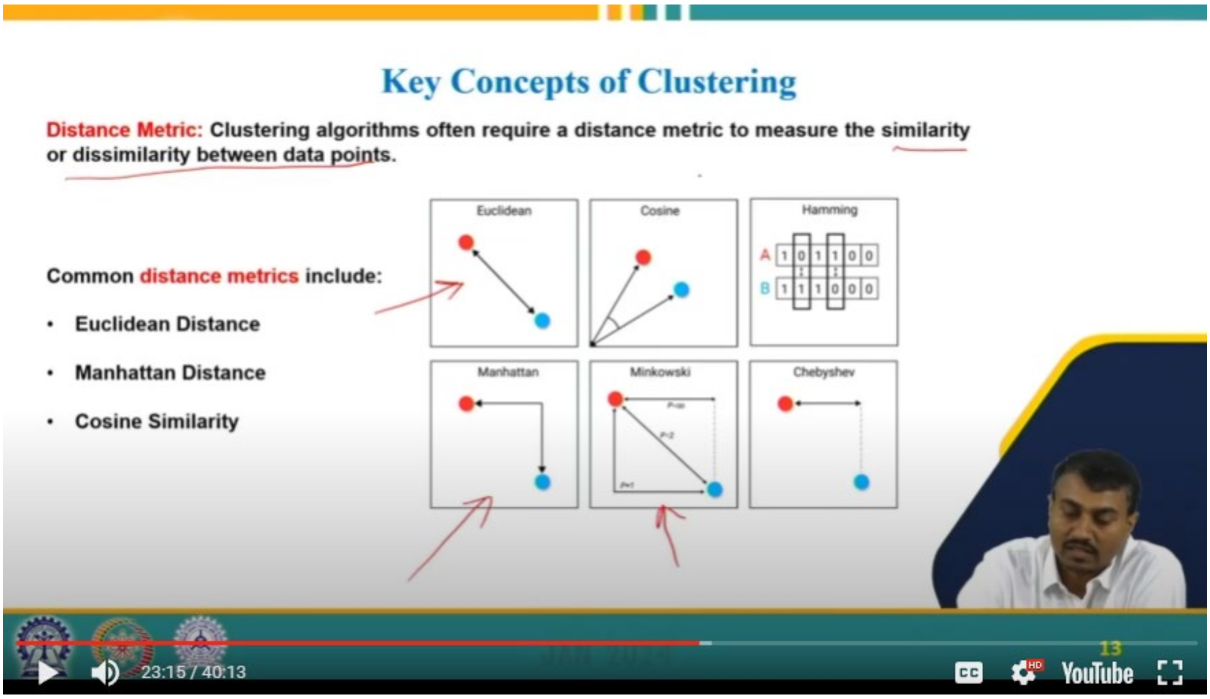
Next are challenges. The major challenge that appears in the unsupervised machine learning approach that you have seen in the supervised method, as you know, is with the level of data set that determines the end output. What should it be? And you match and find out the difference between the outputs you found from this algorithm or approach, and you know what the outputs would be of this data because it is the observed data. So based on that, we finetune, and then again, in the artificial neural network, you have seen the back propagation methods that basically reduce the loss functions using the gradient descent method. So finally, the machine and the approach are basically optimized. Weights are optimized. Weights are updated. They are updated rules. But here we do not have any level output. So how do we assess the performance of this method? So evaluation is a problem in this kind of situation.

Interpretability: interpreting these results is also sometimes difficult, particularly for very high-dimensional data. And scalability is also a fine problem for large or large data sets when there is an internal complexity in these kinds of situations. Unsupervised learning faces some problems. So now that we understood the basic tenets, we understood where to apply the unsupervised learning and where to apply the supervised learning. For example, there are some kinds of patterns on the flowers. Flowers petal length and the dimensions petal length and width and sepal length and width. And based on that, if we are able to classify if the end-level data set is not available, for example, this is the type of flower. So in those cases, we can classify this as a type 1 type of flower, this as a type 2 type of flower, and this as a type 3 type of flower. The same example can be converted into supervised learning: yes, if these dimensions are within this range and maybe their ratios, sepal width, sepal length, and petal width and petal length, this is the typical type of flower. So this target variable is known to us. So we can apply supervised learning as well.

So these are the small differences between the approaches of the unsupervised learning approach and the supervised learning approach. So let's start the clustering. Clustering—you might have heard this particular word too many times during the COVID-19 pandemic-free earth. And there were good applications for these clustering methods. So you have seen that clusters, cluster 1, cluster 2, have a higher have a higher degree of clusters where more patients were detected to be COVID positive, and when the rate of positives is false, they are shifting to another cluster. So based on that, the local administration was dealing with. So it is a very useful application, and it is a data-driven, evidence-based approach, and globally, this particular practice is followed. So clustering is similar to that, similar to that in the sense that here you see these data points are located in the 3-dimensional space. Visibly, their locations are known. So based on their similarity in position, similarity in their positions in the coordinate space or 3-dimensional space, it is very easy for us to subdivide these data sets into four clusters. So basically, in the choice-based system for customers, you might see that a certain kind of customer has certain kinds of preferences and their own choices.

So you can subdivide those customers into those areas so that you can do the business very well. So this is basically clustering. So clustering basically represents togetherness, closeness, or togetherness in quality, togetherness in spatial locations, and a similarity in structures in the pointed space or in the quality is nothing but the representation of the clusters. There are now a now a few types of clustering. So when we are basically clustering the data set, we are basically subdividing the data set into a few clusters based on the similarity and dissimilarity of the data points. This is done basically based on the distance, which is the Euclidean distance. You know, x1-x2 whole square plus y1-y2 whole square plus z1-z2 whole square root over is the distance between the two points in the 3-dimensional space, which is the Euclidean distance. There is also cosine similarity; you might remember the Minkowski distance, and there are a number of things we have already covered; we will also repeat that. So now let us start with the hard clustering.

So this hard clustering represents each data point belonging exclusively to one cluster. And that example is k-means clustering. Soft clustering may mean that the data points can belong to multiple clusters with varying degrees of membership. So here, this clustering is done based on the distance metric, and this distance metric measures the similarity or dissimilarity between the data points. Euclidean distance, that is, Minkowski distance; this is Manhattan distance; this is cosine; this is Chebyshev; this is Hamming. So there are a number of varieties based on the mathematical formulation and the complexity of the data set; we have to choose the suitable methods. K-means clustering is one of the common clustering algorithms that we use globally. So it is a very popular and widely used algorithm in the unsupervised machine learning approach. So what is the fundamental assumption? The fundamental assumption of k-means clustering is that you assume the k number of clusters is there. Initially, at the start of the method, and based on that, you basically assign each data point to the nearest cluster centroid. Then update the centroid based on the mean of the data points assigned to each cluster.



Key Concepts of Clustering

Distance Metric: Clustering algorithms often require a distance metric to measure the similarity or dissimilarity between data points.

Common **distance metrics** include:

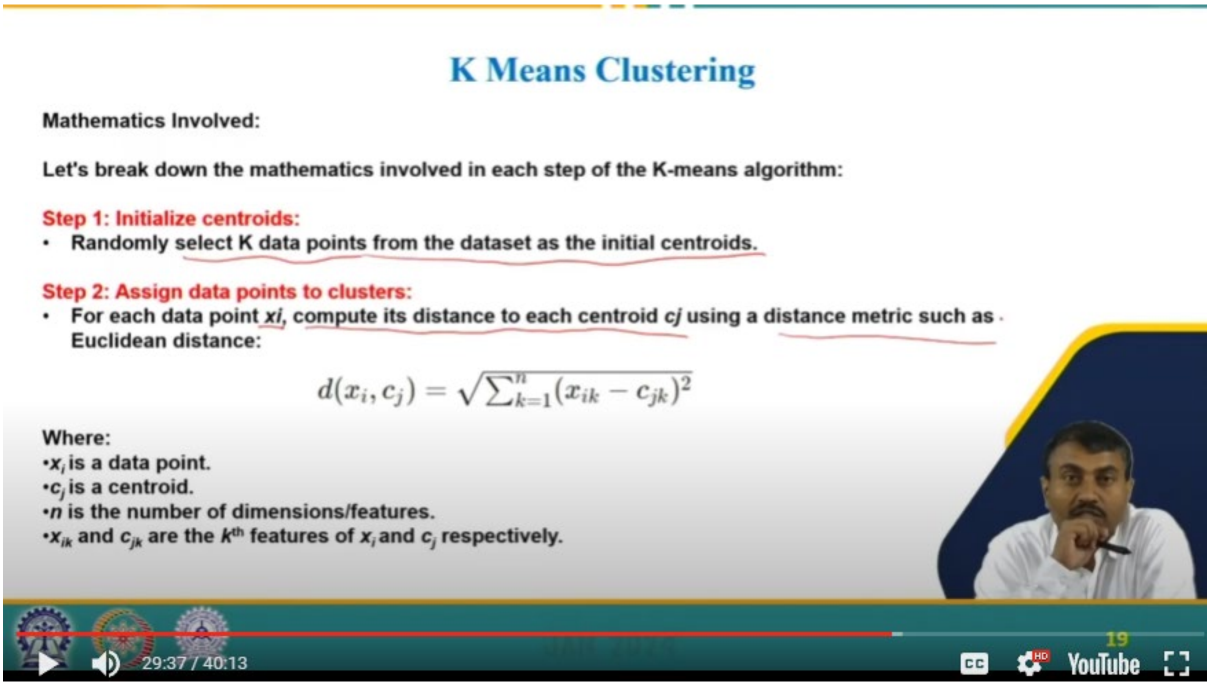
- Euclidean Distance
- Manhattan Distance
- Cosine Similarity

Then there is hierarchical clustering. This clustering builds its tree-like structure by either iteratively merging the smaller cluster into the larger clusters, that is, angglomerative, or splitting them into a smaller one, which is a divisive kind of thing. This basically depends on the purpose. This is another kind of clustering that is density-based clustering, DBSCAN, particularly in the point cloud data set for dimensional measurement. This kind of clustering method is very useful. And this is basically based on the algorithm that uses density-based clustering to group together the data points, which are closely packed together and mark the out layer as the noise. So here, it does not require specifying the number of clusters. This is the beauty, and it can discover clusters of arbitrary shape and size. So for random kinds of samples, irregular samples, and size detection, this kind of algorithm is very, very useful. The Gaussian mixture model is another example of a clustering algorithm.

What are the evaluation criteria? We have already discussed that there are some issues or challenges in measuring the efficiency or metric of the model. So, silhouette score is one of them that is used for performance analysis of these methods. This measures how similar a data point is to its own cluster compared to other clusters. The Davies-Bouldin index measures the average similarity between each cluster and its most similar cluster, taking into account both the compactness and separations of clusters. The Calinski-Harabasz index basically computes the ratio of cluster dispersion to within-cluster dispersion, providing a measure of cluster tightness and separation. So, the main points are that understanding the principle and algorithm are very, very essential. Where to use and where not to use these kinds of methods are important. But it has a lot of potential to uncover the hidden information that is present in the dataset. So, let us start with K-means clustering.

So, K-means clustering is an iterative algorithm that partitions a dataset into K clusters. So, this is the first assumption before you start this method, and it aims to minimize the sum of the squared distance between the data points and their corresponding cluster centroids. How does it work? You can see these data points. So, these data points are specially located, and the distances are far away, but a few points are close together. So, now you assign that four-digit number K. So, this is centroid calculation is done, centroid calculation. So, now you categorize it into four clusters. So, we initialize the K cluster centroid randomly. Then we assign each data point to the nearest centroid, forming a K-number of clusters. Then we update, and then we update the centroid by computing the mean of all data points assigned to each cluster. And then we will repeat steps 2 and 3 until convergence, when the centroid no longer changes significantly. So, what is the mathematics behind it? So, first, we have to initialize the centroid. We have to randomly select the K data points from the dataset as the initial centroids. And for each data point xi, we have to compute its distance to each centroid cj using a distance matrix such as Euclidean distance. So, d (xi, cj) is nothing, but the square root of summation K is equal to 1 to n within the first bracket: xik minus cjk, the whole square.

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2}$$



K Means Clustering

Mathematics Involved:

Let's break down the mathematics involved in each step of the K-means algorithm:

Step 1: Initialize centroids:
- Randomly select K data points from the dataset as the initial centroids.

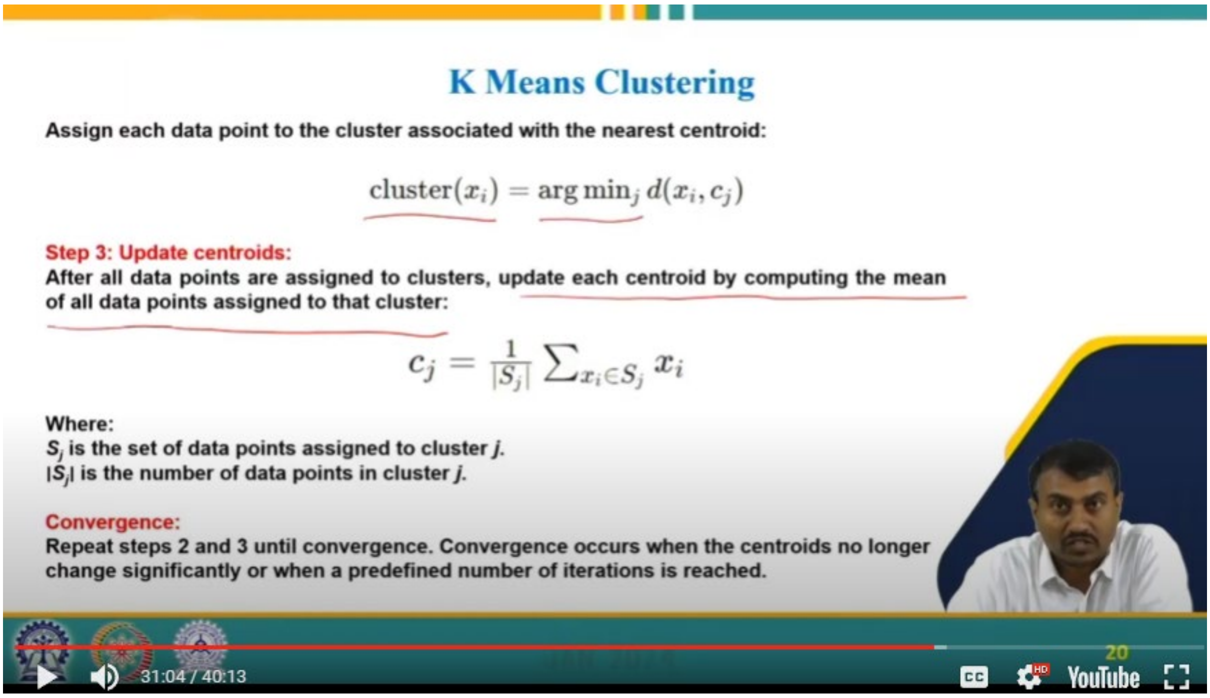
Step 2: Assign data points to clusters:
- For each data point *xi*, compute its distance to each centroid *cj* using a distance metric such as Euclidean distance:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2}$$

Where:
•x_i is a data point.
•c_j is a centroid.
•n is the number of dimensions/features.
•x_{ik} and c_{jk} are the k^{th} features of x_i and c_j respectively.

So, this estimate is the distance summation over and root over. xi is the data point, cj is the centroid, and n is the number of dimensions or features in the data space on the data point. And xik and cjk are the Kth features of xi and cj, respectively. Now, for each data point, we have to assign the cluster associated with the nearest centroid. So, cluster xi, you will return arg min j, dxi, and cj that will be assigned to that near cluster. And we have to update the centroid now. After all these data points are assigned up to step 2 to clusters, update each centroid by computing the mean of all the data points assigned to these clusters. So, cj is equal to 1 by Sj mod summation over xi, which is one of the elements present in the Sj summation of xi. So, Sj is the set of data points assigned to the cluster, Sj is the number of data points in the cluster, and xi is the data point. So, we have to repeat steps 2 and 3 until convergence is achieved, and convergence occurs when the centroid no longer changes significantly or when a predefined number of iterations is reached.

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$$



K Means Clustering

Assign each data point to the cluster associated with the nearest centroid:

$$\text{cluster}(x_i) = \arg\min_j d(x_i, c_j)$$

Step 3: Update centroids:

After all data points are assigned to clusters, update each centroid by computing the mean of all data points assigned to that cluster:

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$$

Where:

S_j is the set of data points assigned to cluster j.

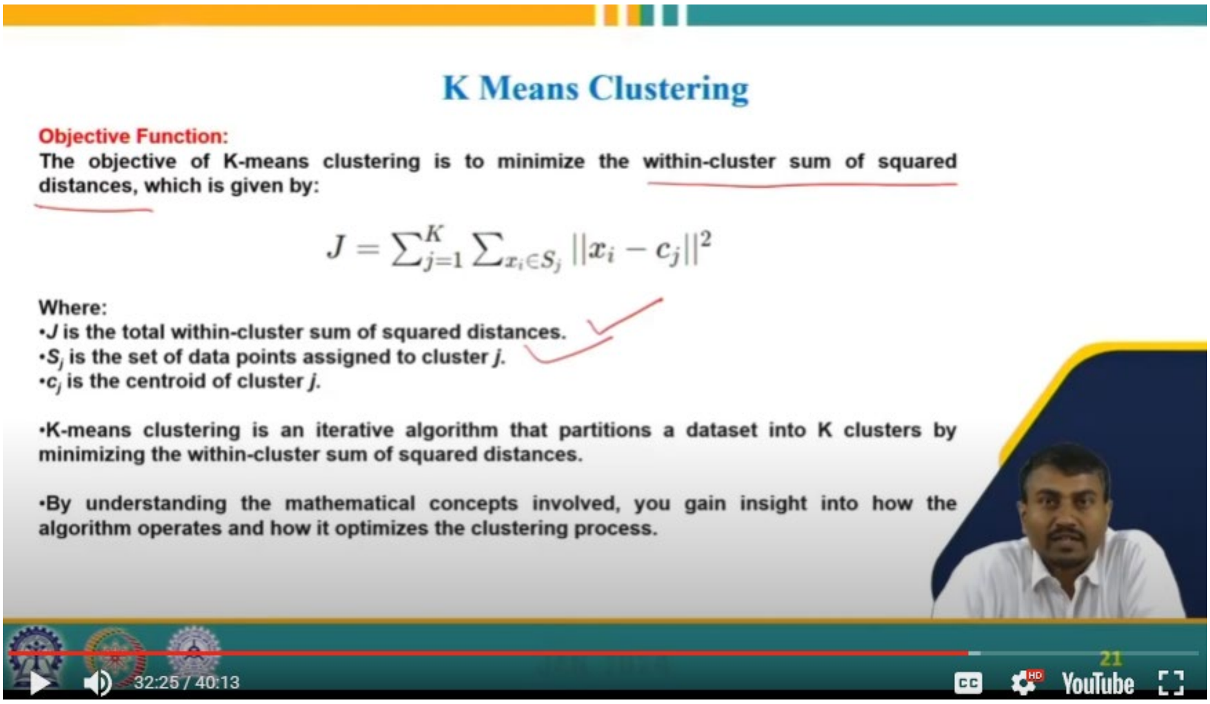
$|S_j|$ is the number of data points in cluster j.

Convergence:

Repeat steps 2 and 3 until convergence. Convergence occurs when the centroids no longer change significantly or when a predefined number of iterations is reached.

Objective function. So, the objective function of K-means clustering is to minimize the within-cluster sum of square distance, which is given by summation over j, which is equal to 1 to k, and summation over xi, which is an element within the Sj, xi minus cj mod whole square. The absolute difference between xi and cj and their whole square. So, j is the total within cluster sum of square distance, Sj is the set of data points assigned to the cluster, and cj is a centroid of the cluster j.

$$J = \sum_{j=1}^{K} \sum_{x_i \in S_j} ||x_i - c_j||^2$$



K Means Clustering

Objective Function:

The objective of K-means clustering is to minimize the within-cluster sum of squared distances, which is given by:

$$J = \sum_{j=1}^{K} \sum_{x_i \in S_j} ||x_i - c_j||^2$$

Where:

- *J* is the total within-cluster sum of squared distances.
- S_j is the set of data points assigned to cluster *j*.
- c_j is the centroid of cluster *j*.

- K-means clustering is an iterative algorithm that partitions a dataset into K clusters by minimizing the within-cluster sum of squared distances.

- By understanding the mathematical concepts involved, you gain insight into how the algorithm operates and how it optimizes the clustering process.

So, K-means clustering is an iterative algorithm we have seen, and we have to assign the k, and then we have to update and apply this concept. It is necessary that we understand this mathematics and the process. So, this is the original data set: data set 1, data set 2, data set 3, and data set 4. Now it has been subdivided into four clusters, and the centroids are exactly at the center point within that class. That is the final conclusion from the input to the output. So, what are the assumptions of K-means clustering? Assumptions are that the data are spherical or globular in shape, which means that in 2D space, more or less within a sphere or within a circle, they are located in a three-dimensional space. The data are within the sphere or the globular shape. That means the variance of the data points within each cluster is similar, and the and the variance within each cluster is similar because of this assumption across all dimensions. So, if the data or these clusters are non-spherical in shape or have varying variance, K-means clustering may perform suboptimally well.

Clusters are of similar size if the size of the clusters is more or less the same, or if the density of the clusters is the same as the number of points present. So, with more or less equal weightage, there will be no bias, but if there is a significant difference in density between those clusters, then higher weightage may be given to the larger clusters. So, that may lead to some biased results. Clusters are linearly separable, and we have assumed at the start of the relation that the data points are linearly classifiable. So, and separable, so if these assumptions are violated and the data are non-linearly separable in that kind of case, it would be very difficult to handle using K-means clustering. The centroid represents the cluster center. K-means assumes that the cluster centroid accurately represents the center of the cluster, and each data point is assigned to the cluster whose centroid is closest to it.

So, if the centroid is not accurately identified in the iterations, we have to go through a number of iterations to basically do away with this kind of error. Initial centroid positions matter, so that is why this can only be overcome by a number of iterative methods and initialization. So, the initialization of the centroid is very important. Continuous variable, K-means assume that the input variables are continuous, and it may not perform well with the categorical variables unless appropriate pre-processing is done to convert them into a suitable numerical representation. Now, what are the advantages? The major advantages are its simplicity and ease of implementation, which is why most people choose the K-means clustering method. It is very efficient and computationally efficient, too. Scalability is high, which is why we can use it for high-dimensional data as well. Versatility: it can be used for a wide range of applications, including segmentation, anomaly detection, image segmentation, etcetera.

Interpretability: cluster centroids produce K-means that can be easily interpreted, providing insight about the characteristics of the data and its clusters. Work well with well-separated clusters; this is what we have already seen. The disadvantages are sensitive to initial centroid positions; we have already discussed that. It requires a predefined number of clusters, so if this predefined number of clusters does not suit us, then sometimes we land on suboptimal solutions. Assume spherical clusters, so if it is a non-spherical one, then it will not perform well. Struggle with out layers. K-means clusters are very sensitive to out layers, and they can significantly affect the positions of cluster centroids, resulting in cluster assignments. It may converge to local optima, so K-means clustering is very prone to converging to local optima, and especially when the initial centroids are poorly chosen and cannot handle non-linear data, this is one of the major problems.

In the applications of exploratory data analysis, particularly identifying the natural grouping of the minerals based on the mineral data and mineralization data of the mining company, we can use this to analyze the possibility of mineral quantization in a particular area. Ore body segmentation: we can use it based on exploratory data, drill data, geochemical data, and grade data to basically attribute to which area of the ore body is located. Rock classifications: based on the different properties of the rock, we can basically find out the similarities and dissimilarities, and we can subdivide these rocks into several classes. Fault detection and monitoring and K-means clustering can also be detected for fault detection and monitoring in mining operations as well as in equipment, so this is a very useful method.

Market segmentation and demand forecasting are major areas where K-means clustering is nowadays used by different companies. So, based on the customer purchase behavior, geographical positions, where they are situated, whether in an in an urban or rural area, their marketing strategies, the products they are offering, and the pricing model based on that, So, these are the inputs based on which marketing by a company of a particular product is done, and this is a very helpful method. So, these are the references, so let me summarize in a few sentences what we have covered today.

So, we have discussed what unsupervised learning is, and in unsupervised learning, we have discussed its potential applications and challenges. Then, we introduce clustering, and K-means clustering is a common method that we use, its advantages, its mathematical assumptions, its advantages, and its applications in the mining industry we have discussed. Thank you.