K Means Clustering is a widely used unsupervised machine learning algorithm with a fundamental goal of dividing a given dataset into 'K' distinct clusters, where each cluster is characterized by a centroid, representing the mean of all the data points within that cluster. The algorithm iteratively assigns data points to the cluster whose centroid is closest in terms of Euclidean distance, followed by the update of centroids based on the newly formed clusters. This iterative process continues until the centroids stabilize, i.e., when data points are no longer reassigned to different clusters or after a predefined number of iterations. It is an iterative algorithm that divides the unlabeled dataset into K different clusters in such a way that each dataset belongs only one group that has similar properties.

Underlying the K Means Clustering model are the principles of similarity and distance. The algorithm seeks to group data points with similar characteristics into the same clusters, while maximizing the dissimilarity between different clusters. It accomplishes this by utilizing the concept of centroids as representative centers for each cluster, and the distance metric, typically Euclidean distance, to measure the dissimilarity between data points.

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.



1. **Underlying Principles:**
   * K Means Clustering is based on the concept of partitioning data points into 'K' clusters, where each cluster is represented by its centroid, which is the mean of all the points in that cluster.
   * The algorithm iteratively assigns each data point to the cluster with the nearest centroid and then updates the centroids based on the new assignment.
   * The process continues until the centroids stabilize, and data points are no longer reassigned or until a predefined number of iterations is reached.
2. **Assumptions:**
   * K Means Clustering assumes that the data points can be represented as vectors in a multidimensional space, and the distance between data points can be calculated using a distance metric such as Euclidean distance.
   * The algorithm assumes that the number of clusters 'K' is known in advance or can be specified by the user.
3. **Equations Involved:**
   * Euclidean Distance: Given two data points 'X' and 'Y' with 'n' features (dimensions), the Euclidean distance between them can be calculated as follows:

d(X, Y) = sqrt( (X1 - Y1)^2 + (X2 - Y2)^2 + ... + (Xn - Yn)^2 )

1. **Model Learning Process**:
   * Initialization: The K Means algorithm starts by randomly selecting 'K' data points from the dataset as the initial centroids. These initial centroids act as the starting points for the clustering process.
   * Assignment Step: In this step, each data point in the dataset is assigned to the nearest centroid based on the Euclidean distance. The distance between a data point 'X' and a centroid 'C' is calculated, and 'X' is assigned to the cluster represented by the centroid 'C' with the minimum distance.
   * Update Step: After all data points are assigned to clusters, the centroids are recalculated. Each new centroid is set to be the mean of all data points assigned to the corresponding cluster. This step aims to update the centroid positions to better represent the clusters.
   * Convergence Check: The algorithm checks if the centroids have changed significantly after the update step. If the change is below a certain tolerance level or after a predefined number of iterations, the algorithm terminates, and the final clusters are obtained.
2. **Model Predictions**:
   * After the algorithm converges, and the centroids are fixed, new data points can be assigned to the existing clusters by calculating their distances to the cluster centroids and assigning them to the nearest cluster.
   * The prediction for a new data point 'X' is done by finding the centroid 'C' with the minimum Euclidean distance to 'X' and then assigning 'X' to the cluster represented by centroid 'C'.

## How to choose the value of "K number of clusters" in K-means Clustering?

The performance of the K-means clustering algorithm depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal number of clusters, but here we are discussing the most appropriate method to find the number of clusters or value of K. The method is given below:

# Elbow Method

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

WCSS= ∑Pi in Cluster1 distance(Pi C1)2 +∑Pi in Cluster2distance(Pi C2)2+∑Pi in CLuster3 distance(Pi C3)2

In the above formula of WCSS,

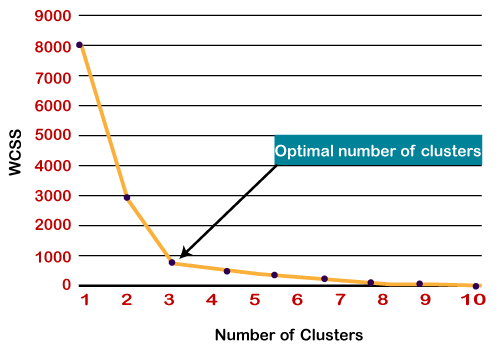
∑Pi in Cluster1 distance (Pi C1)2: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

* It executes the K-means clustering on a given dataset for different K values.
* For each value of K, calculates the WCSS value.
* Plots a curve between calculated WCSS values and the number of clusters K.
* The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow; hence it is known as the elbow method. The graph for the elbow method looks like the below image:



It is vital to note that K Means Clustering can be sensitive to the initial positions of centroids and may converge to local optima. To mitigate this issue, it is advisable to run the algorithm multiple times with different random initializations to improve the likelihood of finding the global optimum clustering.

In summary, K Means Clustering is an unsupervised algorithm that partitions data into 'K' clusters based on similarity. It uses the concept of centroids as representative points for each cluster and iteratively updates the centroids until convergence. The algorithm makes predictions for new data points by assigning them to the nearest cluster centroid based on their distances. It is important to note that K Means Clustering can be sensitive to the initial centroid positions and may converge to local optima, which is why it is often recommended to run the algorithm multiple times with different initializations to improve the chances of finding the global optimum.