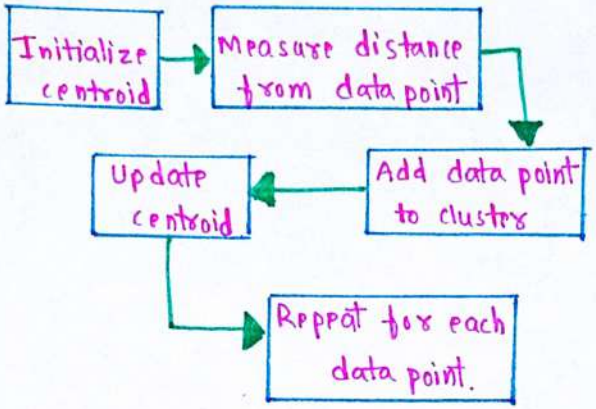
* K-means Clustering:

- Unsupervised algorithm
- Used to capture meaningful structure, underlying processes and grouping inherent in dataset.
- Aim in clustering is to divide the population into several groups in such way that data points in each group more similar to each other than to data points in other groups.
- With this we can discover intrinsic groups in unlabelled data.
- K-means is a centroid based clustering algorithm where we calculate distance between each data point and a centroid to assign it to a cluster.
- The goal is to identify 'K' number of clusters.
- Centroid is center of cluster, which corresponds to arithmetic mean of data points assigned to cluster.

A. Working

1. Random selection of number of clusters or K.
2. Initialize centroids by selecting random data points as centroid. No. of centroid is equal to number of clusters.
3. Calculate euclidian distance from each data point to each cluster centroid.
4. Assign the data point to the cluster, whose centroid is closer to that data point. i.e minimum distance.
5. Update the centroid after adding each new data point by calculating new mean.
6. Iterate over all data points with same approach.

- In short,



- Eucledian distance

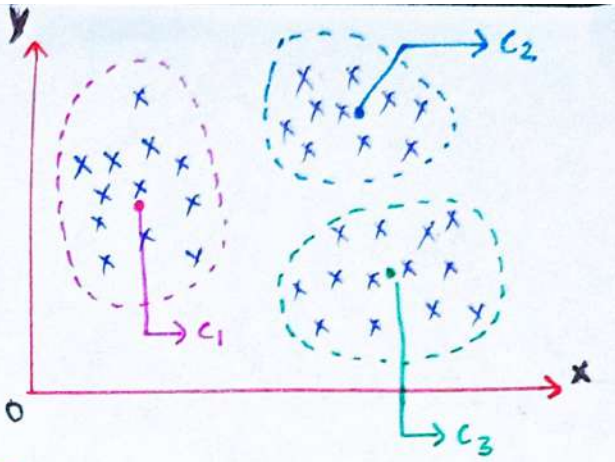
$$= \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

where $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ are two data points.

- $$\text{Centroid} = \frac{\text{Value of all data points of cluster}}{\text{Total no. of data points in cluster}}$$

or

$$\text{Centroid}(c_1) = \frac{1}{n}\sum_{i=1}^{n} x_i$$



B. Choosing value of K.

- Can be done using elbow method
- Intra cluster distance: distance between data point and centroid of cluster to which it belongs.
- Inter cluster distance: distance between data points of two clusters.
- Within cluster sum of squares (WCSS): defines total variations within clusters

$$WCSS = \sum_{i=1}^{n} d(x_i, c_1)^2 + \ldots\ldots + \text{other clusters}$$

where d is distance of centroid(c) from data point (x_i)

- When we plot WCSS vs K while experimenting, the point where curve bent looks like elbow, is best value of K.