* K-means clustering is one of the simplest and popular unsupervised machine learning algorithms (unsupervised learning algorithms make inferences from datasets using only input vectors without referring to known, or labeled.
* The objective of K-means is to group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroid as small as possible. The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

How the K-means algorithm works:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculation to optimize the positions of the centroids

**Algorithm summary**:

* Input: training data D, number K of clusters
* Output: M centroid and label vector for each data point in D

1. Step 1: for each instance, assign it to the cluster with the nearest centroid
2. Step 2: for each cluster, recompute its centroid form all the instances assigned to that cluster

**Advantages and disadvantages:**

* **Advantages:** Relatively simple to implement

Scales to large data sets

Guarantees convergence

Can warm-start the positions of centroids

Easily adapt to new examples

* **Disadvantages**: Choosing k manually

Being dependent on initial values

Clustering outliers

Clustering data of varying sizes and density