*K- Means Clustering*

K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It is used to solve many complex unsupervised machine learning problems.

*A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.*

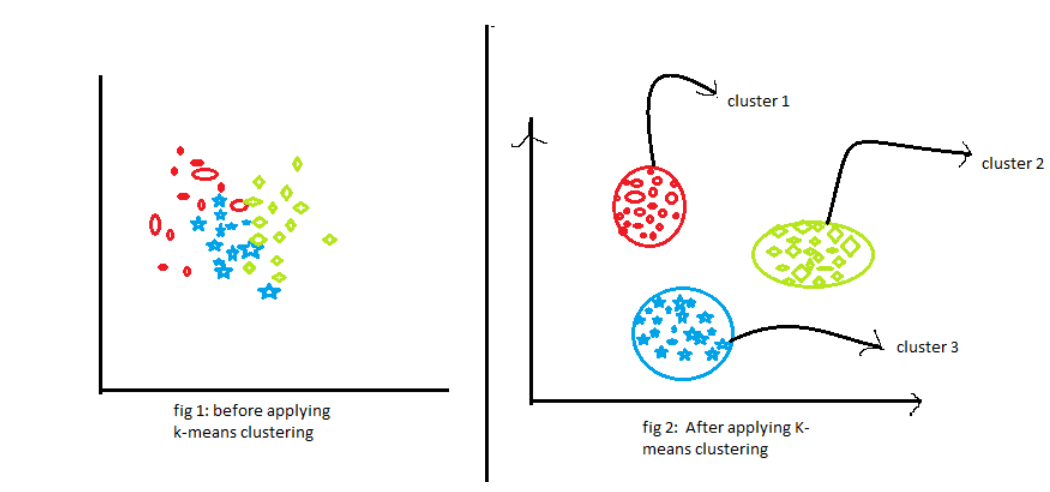
Let’s take an example. Suppose you went to a vegetable shop to buy some vegetables. There you will see different kinds of vegetables.

The one thing you will notice there that the vegetables will be arranged in a group of their types.

Like all the carrots will be kept in one place,

potatoes will be kept with their kinds and so on.

**If you will notice here then you will find that they are forming a group or cluster, where each of the vegetables is kept within their kind of group forming the clusters.**

******

Now, look at the above two figures. what did you observe? Let us talk about the first figure. The first figure shows the data before applying the k-means clustering algorithm. Here all three different categories are messed up. When you will see such data in the real world, you will not able to figure out the different categories.

Now, look at the second figure (fig 2). This shows the data after applying the K-means clustering algorithm. you can see that all three different items are classified into three different categories which are called clusters.

*How Does the K-means Clustering algorithm work?*

k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps.

Let’s see what are these three steps.

1. Select the k values.
2. Initialize the centroids.
3. Select the group and find the average.

*How to choose the value of K?*

One of the most challenging tasks in this clustering algorithm is to choose the right values of k. What should be the right k-value? How to choose the k-value? Let us find the answer to these questions. If you are choosing the k values randomly, it might be correct or may be wrong. If you will choose the wrong value then it will directly affect your model performance. So there are two methods by which you can select the right value of k.

1. Elbow Method.
2. Silhouette Method.

Advantages of K-means

1. It is very simple to implement.
2. It is scalable to a huge data set and also faster to large datasets.
3. it adapts the new examples very frequently.
4. Generalization of clusters for different shapes and sizes.

Disadvantages of K-means

1. It is sensitive to the outliers.
2. Choosing the k values manually is a tough job.
3. As the number of dimensions increases its scalability decreases.

*K mtlb kitne cluster humme chaye*

*Vo pehle 2 random point se start karega jishe hum centroid kahenge.*

*Fir vo jo centroid hai unko unke closest point se joddte jaao.*

*How to Determine Correct Number of Clusters ?*