The Ultimate Guide to K-Means Clustering

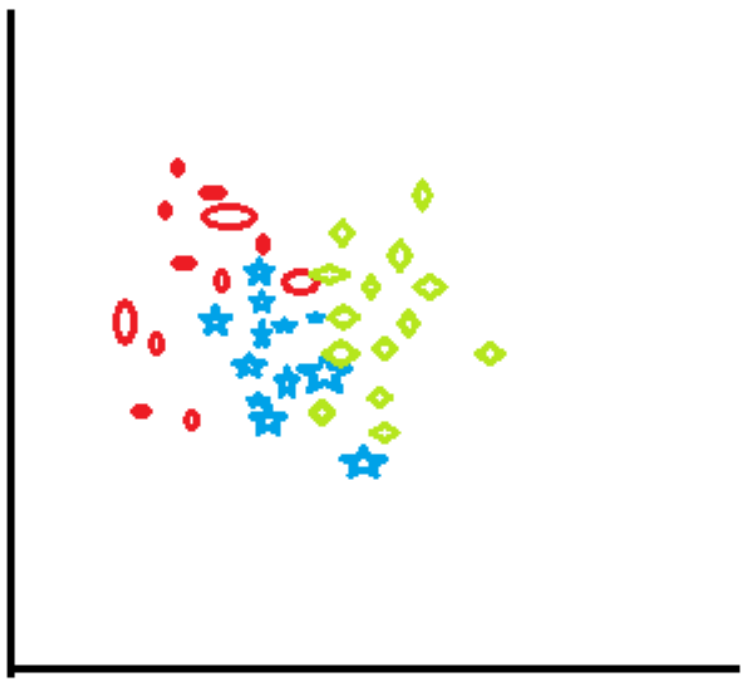


fig 1: before applying k-means clustering

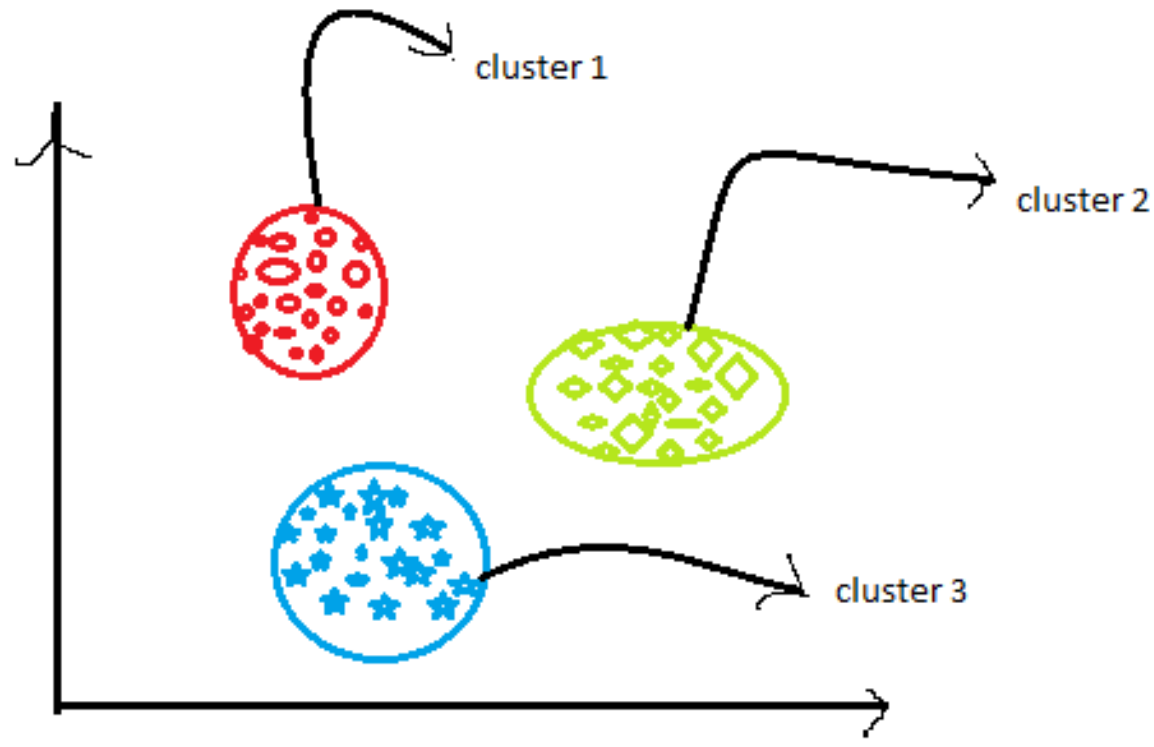


fig 2: After applying K-means clustering

Let's look at a simple example. A bank wants to give credit card offers to its customers. Based on the details of each customer, they decide which offer to give each customer.

The bank now has millions of customers. Making a decision based on each customer's details is too time-consuming.

Analytics Vidhya

To segment its customers into different groups, the bank can group the customers based on their income:

The bank can now make three different strategies/offers, one for each group. Instead of creating different strategies for each customer, they only have to create three. This will reduce the effort as well as the time.

What Is K-Means Clustering?

K-means clustering is a very famous and powerful unsupervised machine learning algorithm. It segregates the unlabeled data into various groups, called clusters, based on having similar features and common patterns. It uses vector quantization and aims to assign each observation to the cluster with the nearest mean or centroid, which serves as a prototype for the cluster.

Analytics Vidhya

… But How to Choose the Right Number of Clusters in K-Means Clustering?

Elbow Method

Plot the number of clusters against the sum of squared distances within each cluster. Look for the "elbow" point where adding more clusters doesn't significantly reduce the distance. This can indicate a good number of clusters.

Silhouette Coefficient

Calculate the average silhouette coefficient for different cluster numbers. The coefficient measures how well each data point fits within its cluster and ranges from -1 to 1. Choose the number of clusters with the highest average coefficient.

Analytics Vidhya

How to Apply K-Means Clustering Algorithm?

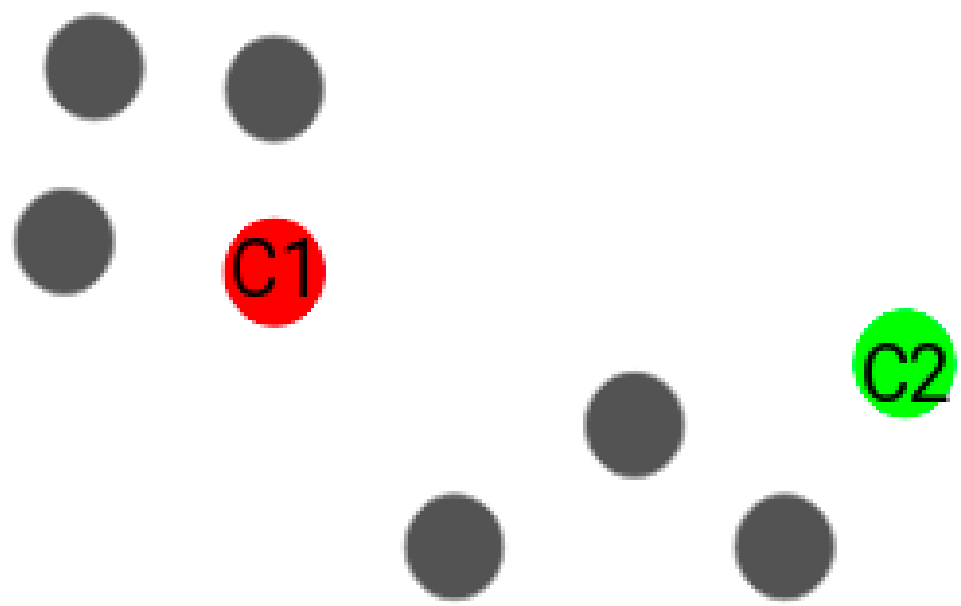
We have these 8 points, and we want to apply k-means to create clusters for these points. Here's how we can do it.

1. Choose the number of clusters k

The first step in k-means is to pick the number of clusters, k.

2. Select k random points from the data as centroids

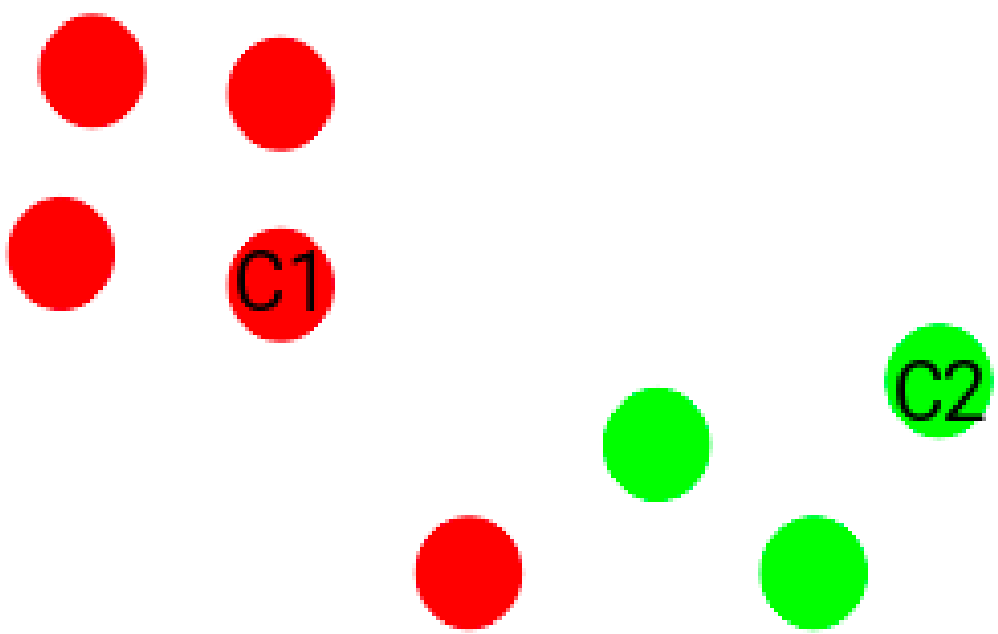
Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



Here, the red and green circles represent the centroid for these clusters.

3. Assign all the points to the closest cluster centroid

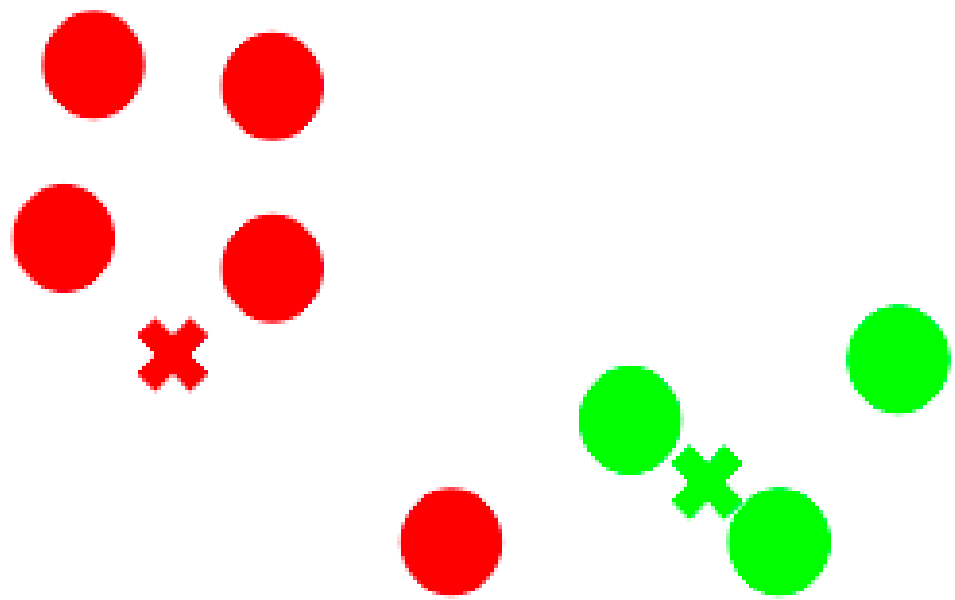
Once we have initialized the centroids, we assign each point to the closest cluster centroid:



Here you can see that the points closer to the red point are assigned to the red cluster, whereas the points closer to the green point are assigned to the green cluster.

4. Recompute the centroids of newly formed clusters

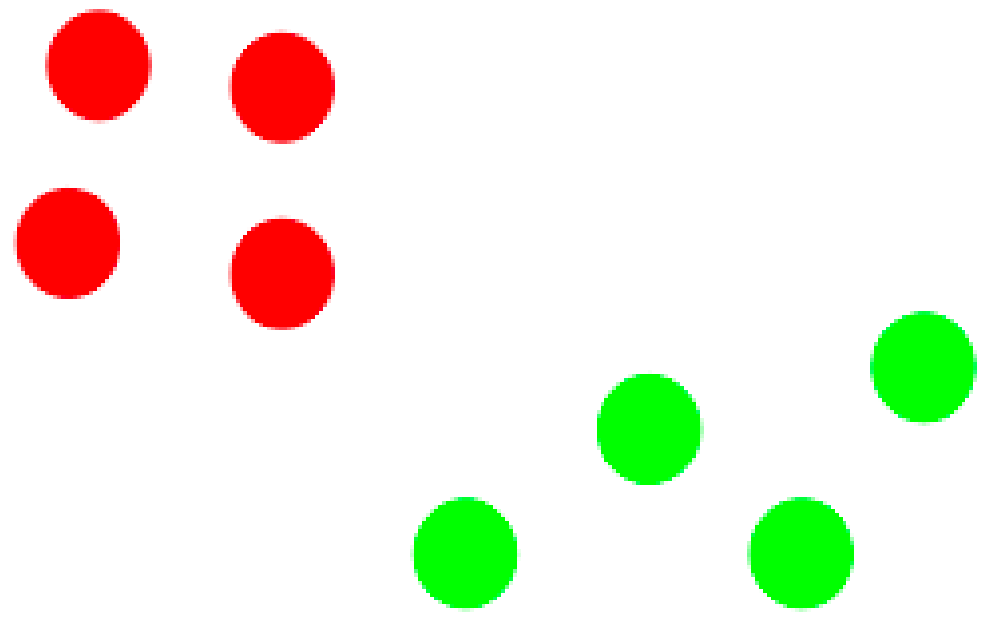
Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

5. Repeat steps 3 and 4

We then repeat steps 3 and 4:



The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration. But wait – when should we stop this process? It can’t run till eternity, right?

Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

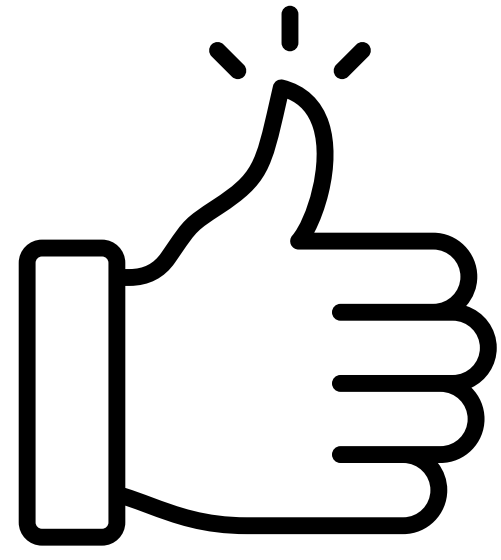
1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations is reached

Analytics Vidhya

Advantages & Disadvantages

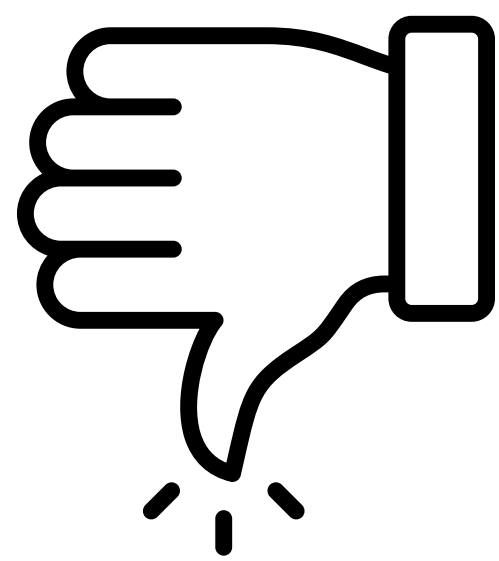
Advantages of K-means

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



Disadvantages of K-means

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

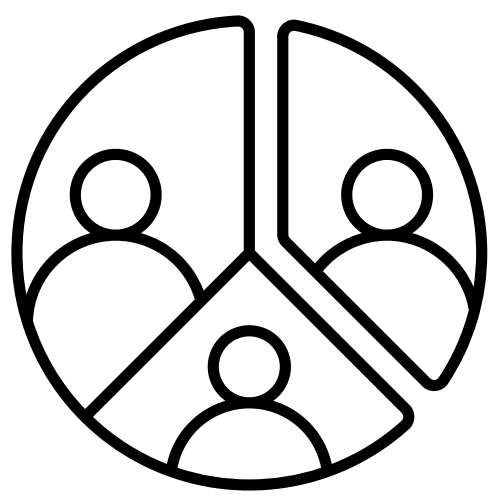


Analytics Vidhya

Applications of K -Means Clustering

Customer Segmentation

The most common applications of clustering is customer segmentation.This strategy is across functions, including telecom, e-commerce, sports, advertising, sales, etc.

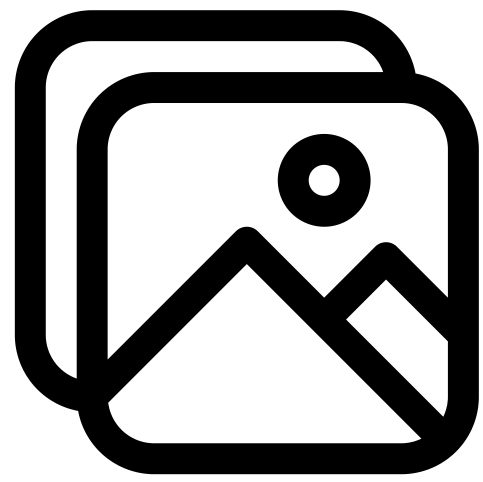


Document Clustering

Say you have multiple documents and you need to cluster similar documents together. Clustering helps us group these documents such that similar documents are in the same clusters.

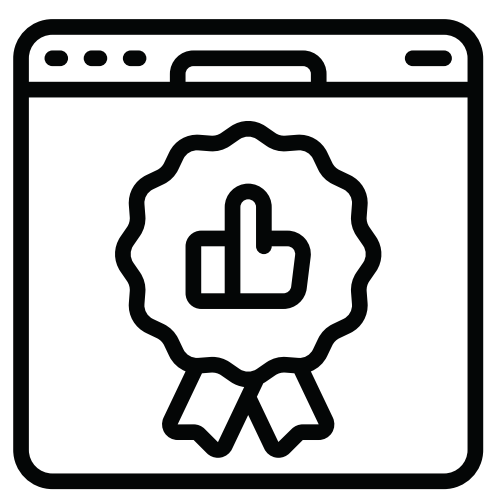
Image Segmentation

In this we try to club similar pixels in the image together. We can apply clustering to create clusters having similar pixels in the same group.



Recommendation Engines

Say you want to recommend songs to your friends. You can look at the songs liked by that person and then use clustering to find similar songs and finally recommend the most similar songs.



Analytics Vidhya

That's a wrap. Was this post Helpful?

Follow us for more!

Analytics Vidhya