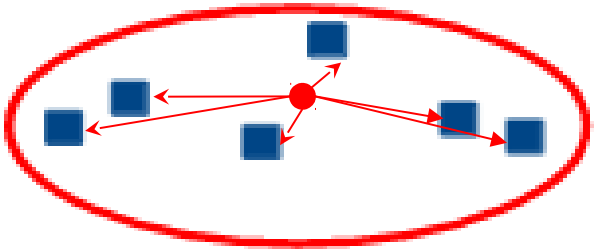
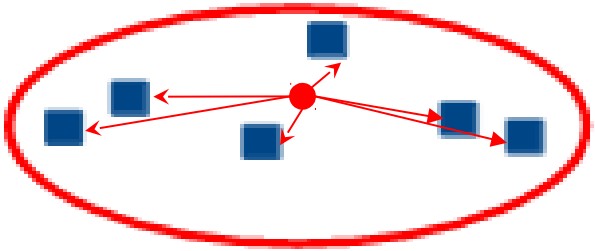
Clustering Techniques

Clustering Techniques

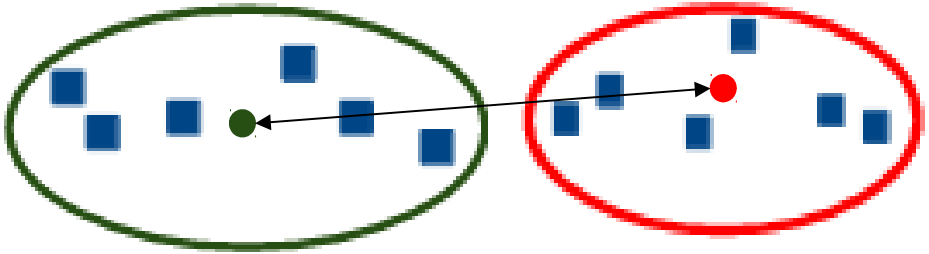


Intra cluster distance

Clustering Techniques



Intra cluster distance



Inter cluster distance

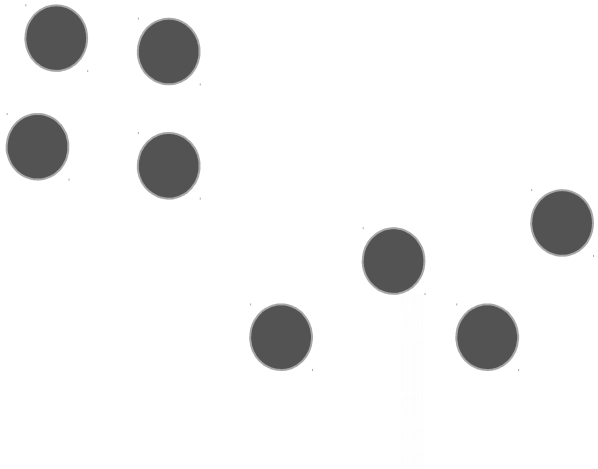
K-means

- Centroid based algorithm
- Each cluster has a centroid

K-means

Objective : To minimize the sum of distances between the points and their respective cluster centroid.

Steps to perform K-means



Steps to perform K-means

1. Choose the number of clusters (k)

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.

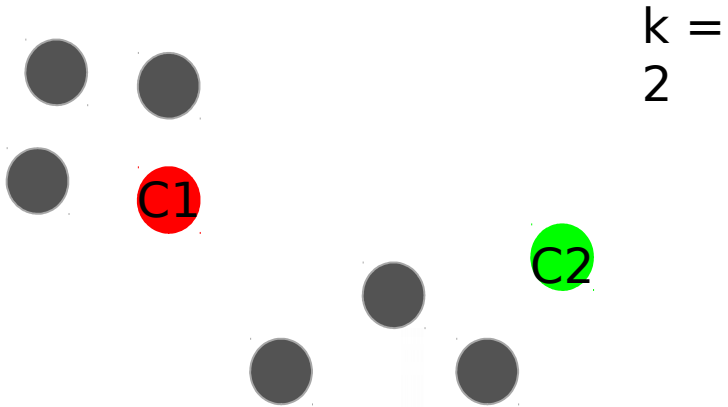
Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.

k = 2

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.

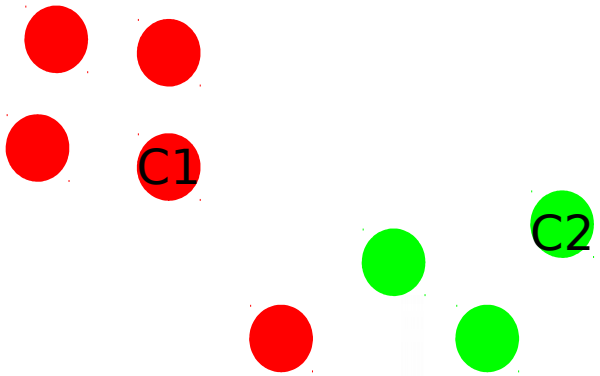


Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid

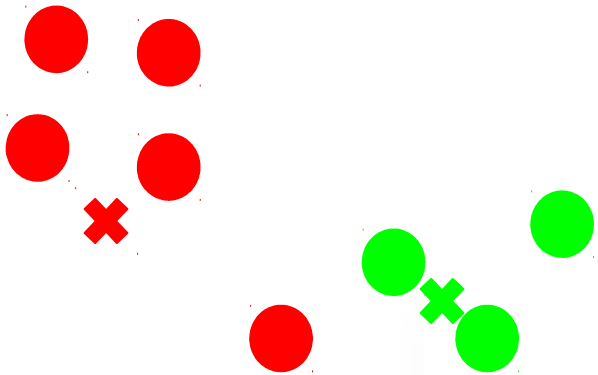


Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid
4. Recompute centroids of newly formed clusters

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid
4. Recompute centroids of newly formed clusters



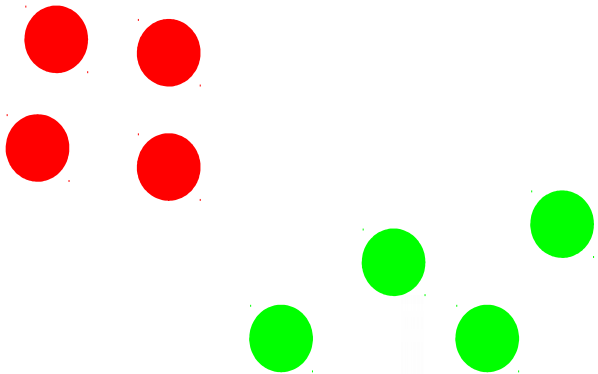
Analytics Vidhya
Learn everything about analytics

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid
4. Recompute centroids of newly formed clusters
5. Repeat step 3 and 4.

Steps to perform K-means

1. Choose the number of clusters (k)
2. Select k random points from the data as centroids.
3. Assign all the points to the closest cluster centroid
4. Recompute centroids of newly formed clusters
5. Repeat step 3 and 4.



Stopping Criteria for K-means

1. Centroids of newly formed clusters do not change

2. Points remains in the same cluster

3. Maximum number of iterations are reached

How to measure similarity between points?

Euclidean Distance

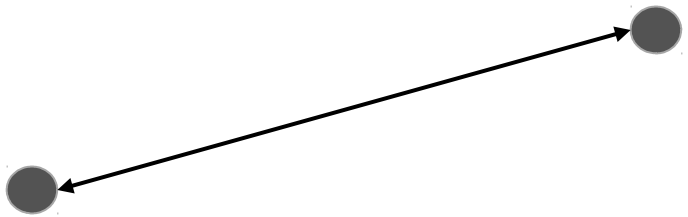
How to measure similarity between points?

Euclidean Distance

Similar

How to measure similarity between points?

Euclidean Distance



Not
Similar

How to measure similarity between points?

Euclidean Distance

$$D = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}$$

n = number of variables

$p_1, p_2, p_3, \ldots$ = features of first point

$q_1, q_2, q_3, \ldots$ = features of second point

Issue with distance based algorithms

- Age
- Income (rupees)

Issue with distance based algorithms

ID	Age	Income(rupees)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Issue with distance based algorithms

ID	Age	Income(rupees)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Euclidean Distance =

$[(100000 - 80000)^2 + (30 - 25)^2]^{1/2}$

Issue with distance based algorithms

ID	Age	Income(rupees)
1	25	80,000
2	30	100,000
3	40	90,000
4	30	50,000
5	40	110,000

Euclidean Distance =

$[(100000 - 80000)^2 + (30 - 25)^2]^{1/2}$

$= 20000.000625$

Issue with distance based algorithms

Solution : Scaling or

Normalizing

Need to scale the data

ID	Age	Income(rupees)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Need to scale the data

ID	Age	Income(rupees)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Euclidean Distance =

$[(0.608+0.260)^2 + (-0.447+1.192)^2]^{1/2}$

Need to scale the data

ID	Age	Income(rupees)
1	-1.192	-0.260
2	-0.447	0.608
3	1.043	0.173
4	-0.447	-1.563
5	1.043	1.042

Euclidean Distance =

$[(0.608+0.260)^2 + (-0.447+1.192)^2]^{1/2}$

= 1.1438

Thank
You!

Analytics Vidhya
Learn everything about analytics