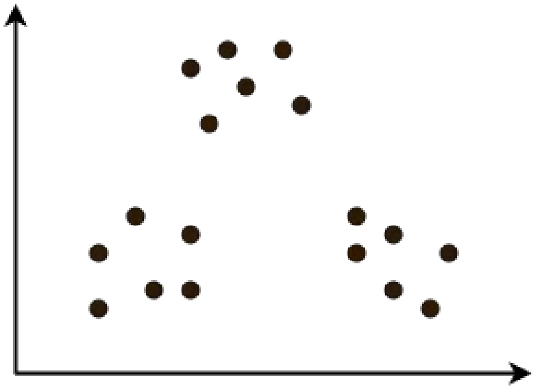
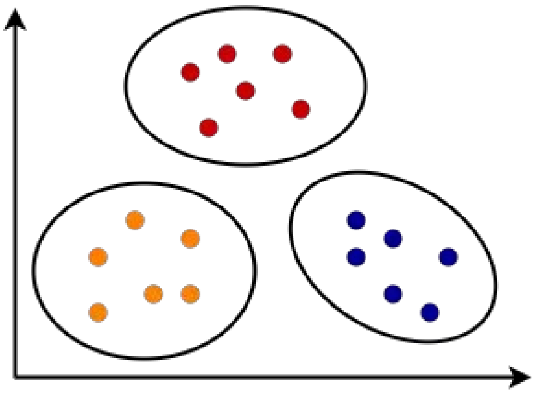
# K-Means Clustering

- K-Means clustering is an unsupervised iterative clustering technique.
- It partitions the given data set into **K** predefined distinct clusters.
- A cluster is defined as a collection of data points exhibiting certain similarities.



Before K-Means



After K-Means

It partitions the data set such that-

- Each data point belongs to a cluster with the nearest mean.
- Data points belonging to one cluster have high degree of similarity.
- Data points belonging to different clusters have high degree of dissimilarity.

### Example Question:

Cluster the following five points into two clusters:

A1(2, 1), A2(3 2), A3(1, 2), A4(2, 2), A5(3, 3)

### Solution:

Here, K = 2;

So we need to identify two random points as initial clusters.

Lets the initial clusters are A1(2,1) and A5(3,3)

Step 1:

Now we calculate the centroids via “Euclidean Distance” formula.

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

$d(C1,A1) = \sqrt{((x_1-x_2)^2 + (y_1-y_2)^2)}$

$= \sqrt{((2-2)^2 + (1-1)^2)}$

$= 0$

$d(C1,A2) = \sqrt{((x_1-x_2)^2 + (y_1-y_2)^2)}$

$= \sqrt{((2-3)^2 + (1-2)^2)}$

$= \sqrt{2}$

…

…

$d(C2,A1) = \sqrt{((x_1-x_2)^2 + (y_1-y_2)^2)}$

$= \sqrt{((3-2)^2 + (3-1)^2)}$

$= \sqrt{5}$

…

…

| | **A1(2,1)** | **A2(3,2)** | **A3(1,2)** | **A4(2,2)** | **A(3,3)** |
|---|---|---|---|---|---|
| **C1: (2,1)** | 0 | √2 | √2 | √1 | √5 |
| **C2: (3,3)** | √5 | √1 | √5 | √2 | 0 |
| | **C1** | **C2** | **C1** | **C1** | **C2** |

New C1 (x,y),

x = (2+1+2) ÷ 3

= 1.67

y = (1+2+2) ÷ 3

= 1.67

So, C1 => (1.67, 1.67)

New C2 (x,y),

x = (3+3) ÷ 2

= 3

x = (2+3) ÷ 2

= 2.5

So, C2 => (3, 2.5)

Step 2:

| | **A1(2,1)** | **A2(3,2)** | **A3(1,2)** | **A4(2,2)** | **A(3,3)** |
|---|---|---|---|---|---|
| **C1: (1.67, 1.67)** | 0.71 | 1.3 | 0.74 | 0.46 | 1.8 |
| **C2: (3, 2.5)** | 1.8 | 0.5 | 2.0 | 1.1 | 0.5 |
| | **C1** | **C2** | **C1** | **C1** | **C2** |

Here we see that the group/cluster of step 1 and step 2 are same. So we can stop the iteration.

So the final clusters/groups are:

Group 1 : (2,1), (1,2) and (2,2)

Group 2 : (3,2) and (3,3)