PCC-IT 307 ML
UNIT 2
Unsupervised Learning

Clustering

- Cluster analysis is part of the **unsupervised learning**.
- A cluster is a group of data that share similar features.
- Customer segmentation: Looks for similarity between groups of customers
- Stock Market clustering: Group stock based on performances

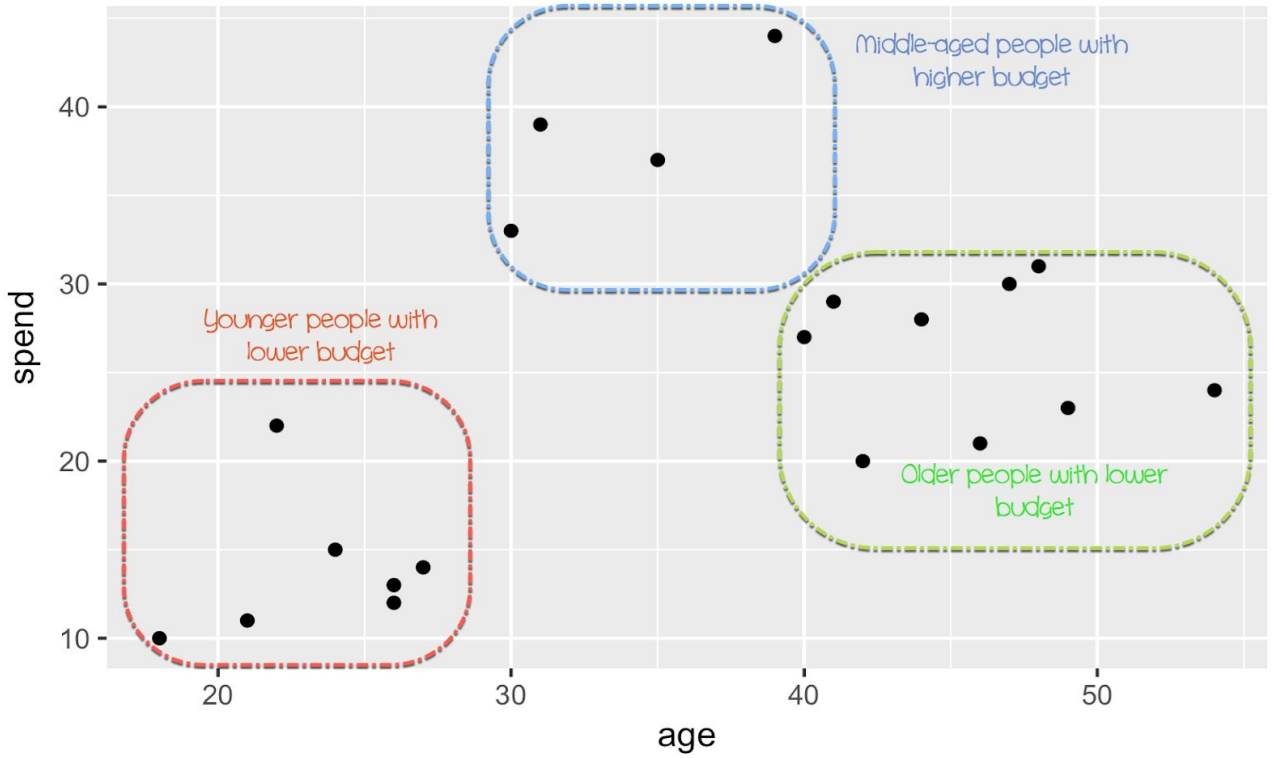
Clustering

- Based on Unsupervised Learning
- Unsupervised learning creates a new variable, the label.
- The machine helps the practitioner in the quest to label the data based on close relatedness.
- It is up to the analyst to make use of the groups and give a name to them.
- So, basically clustering partitions the dataset with similarities into different groups which can act as a base for further analysis. The result will be that objects in one group will be similar to one another but different from objects in another group.

Clustering Example

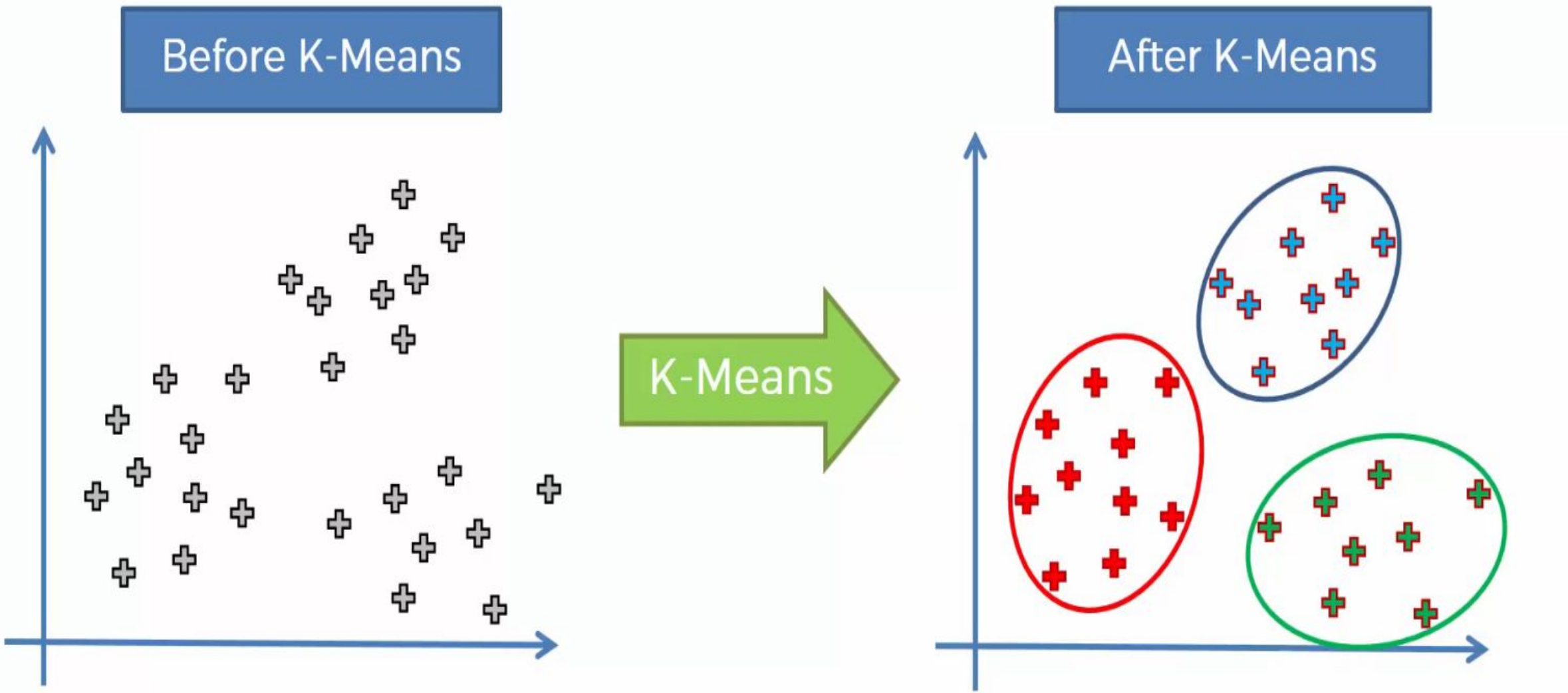
- You have data on the total spend of customers and their ages.
- To improve advertising, the marketing team wants to send more targeted emails to their customers.
- In the following graph, you plot the total spend and the age of the customers.

Clustering Example



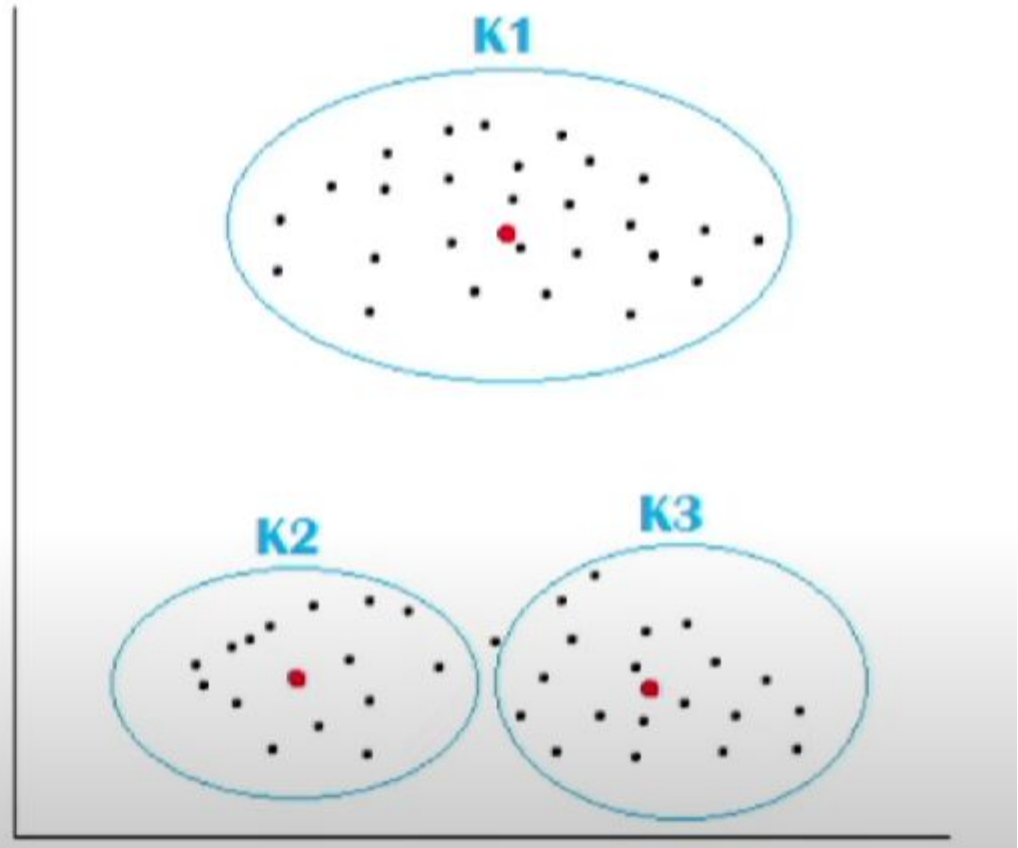
Clustering Example REF[2]

K Means Clustering


Before K-Means
K-Means
After K-Means

K-Means Clustering

- **K-means** clustering is a **distance-based unsupervised clustering algorithm** where data points that are close to each other are grouped in a given number of clusters/groups.



STEP 1: Choose the number K of clusters

STEP 2: Select at random K points, the centroids (not necessarily from your dataset)

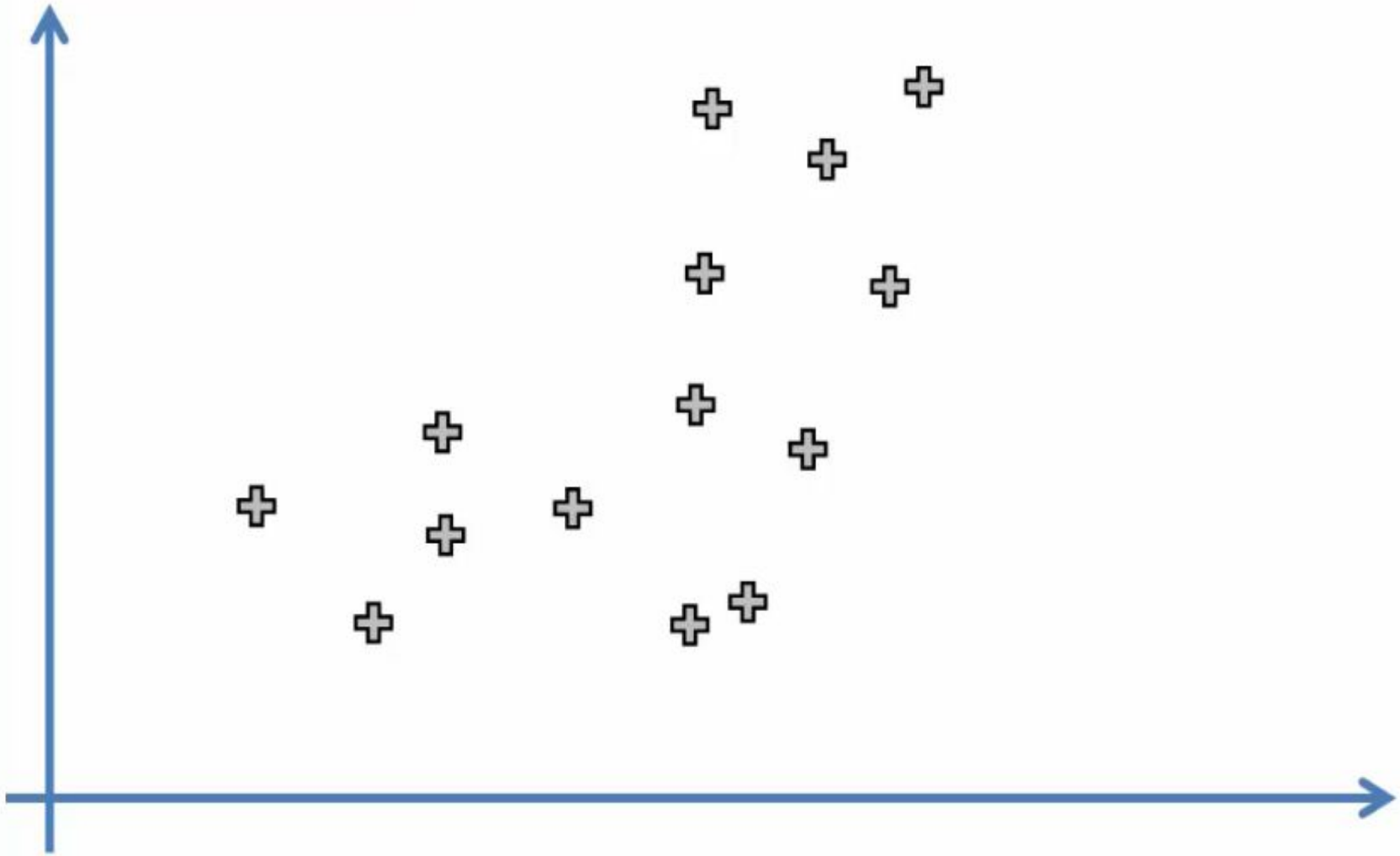
STEP 3: Assign each data point to the closest centroid → That forms K clusters

STEP 4: Compute and place the new centroid of each cluster

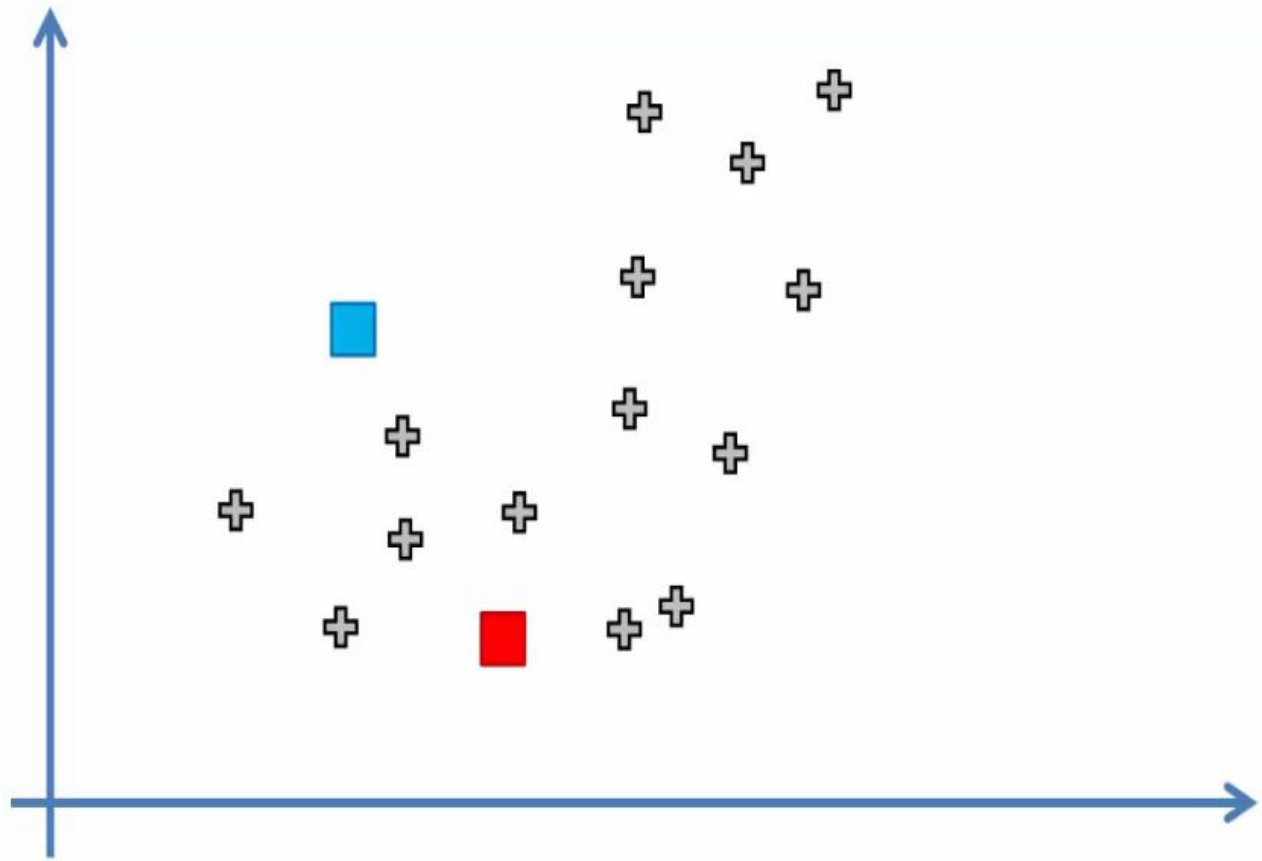
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

Your Model is Ready

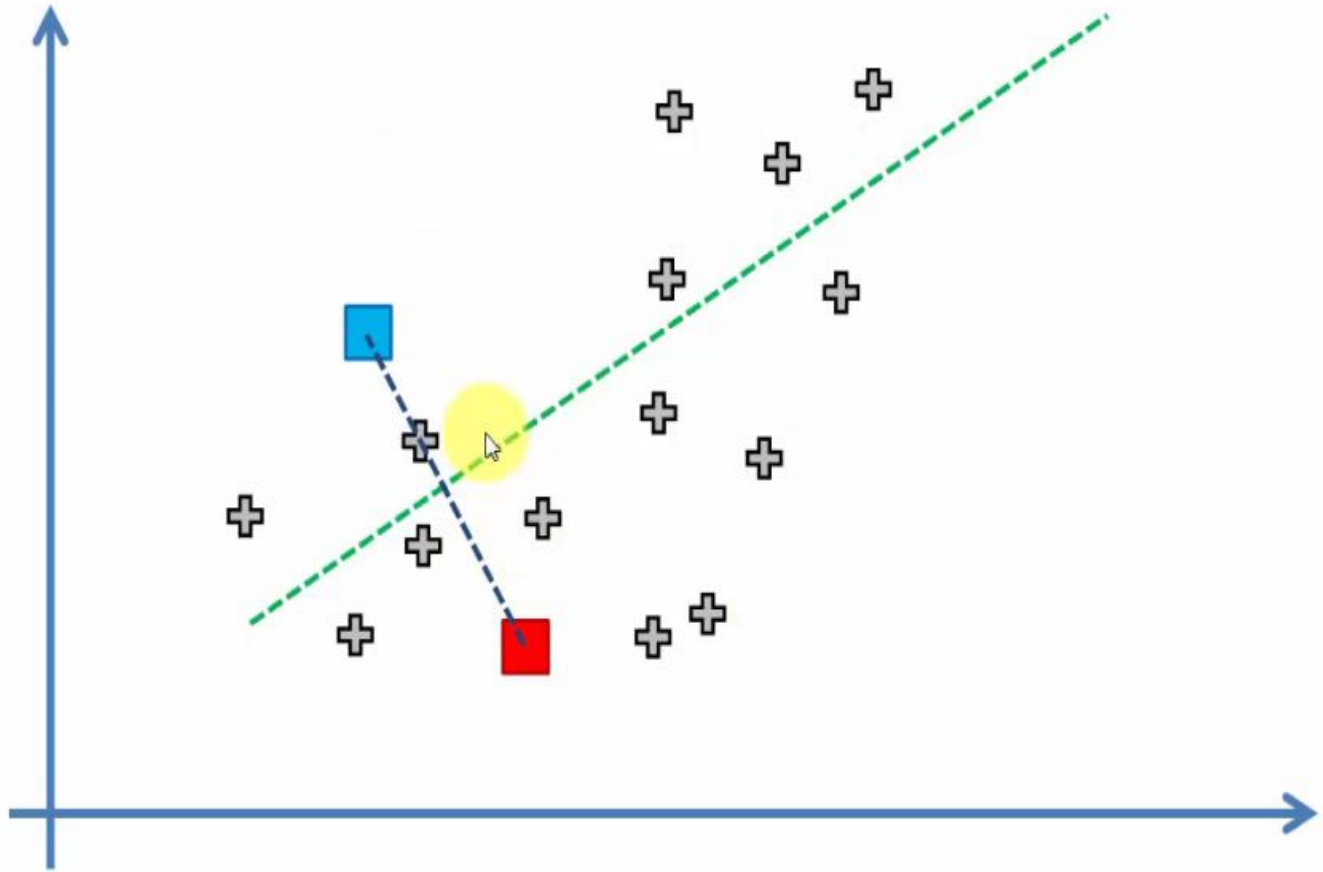
STEP 1: Choose the number K of clusters: K = 2



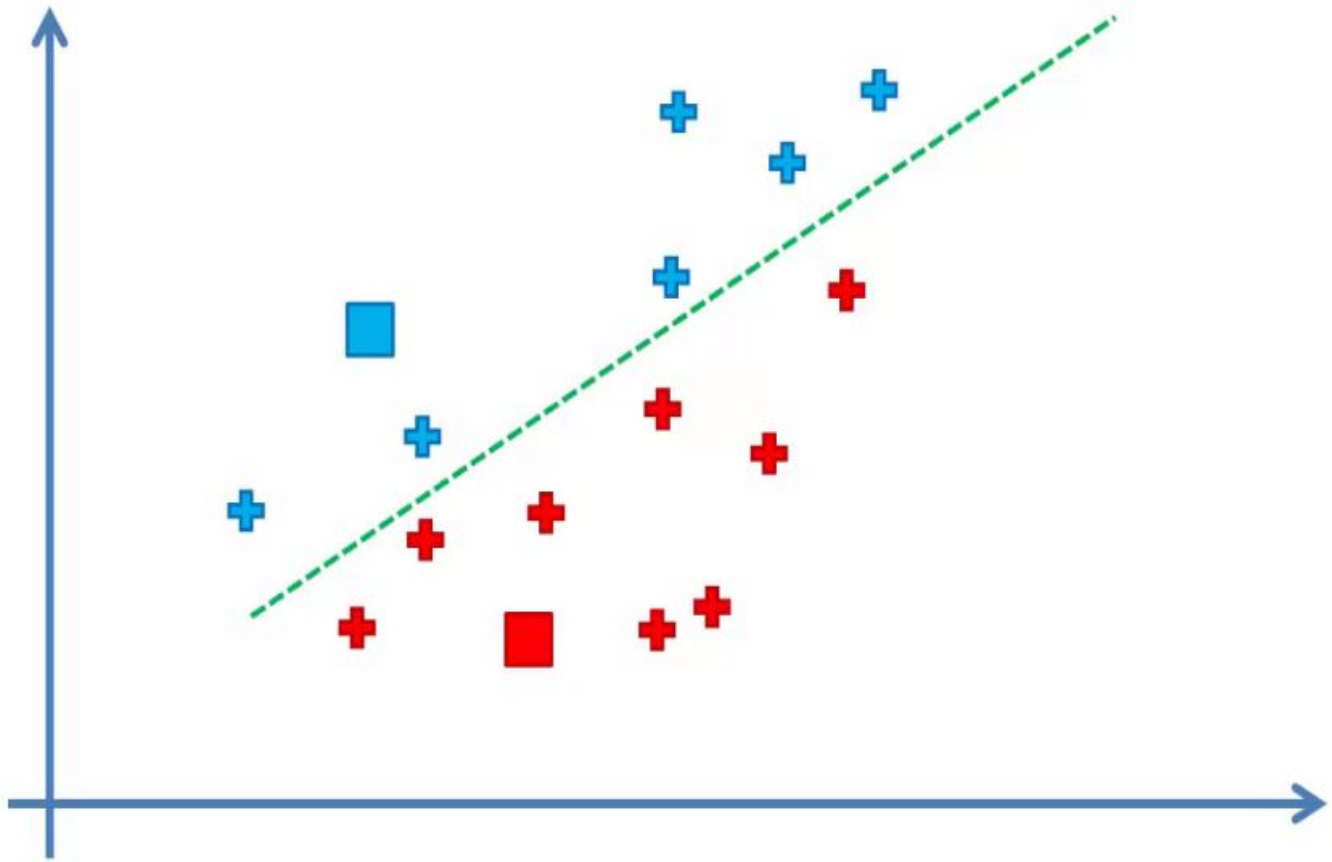
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)



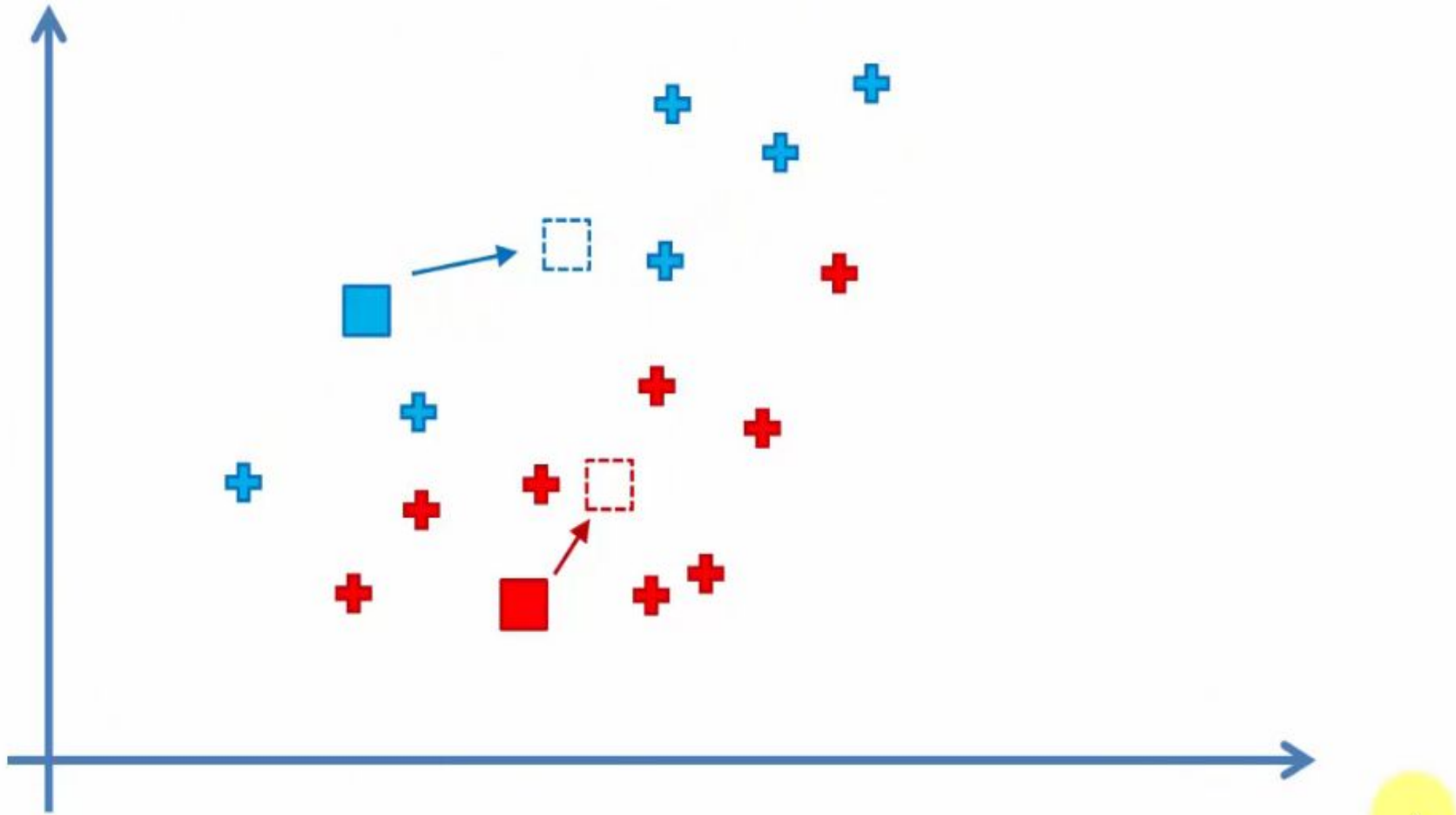
STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters



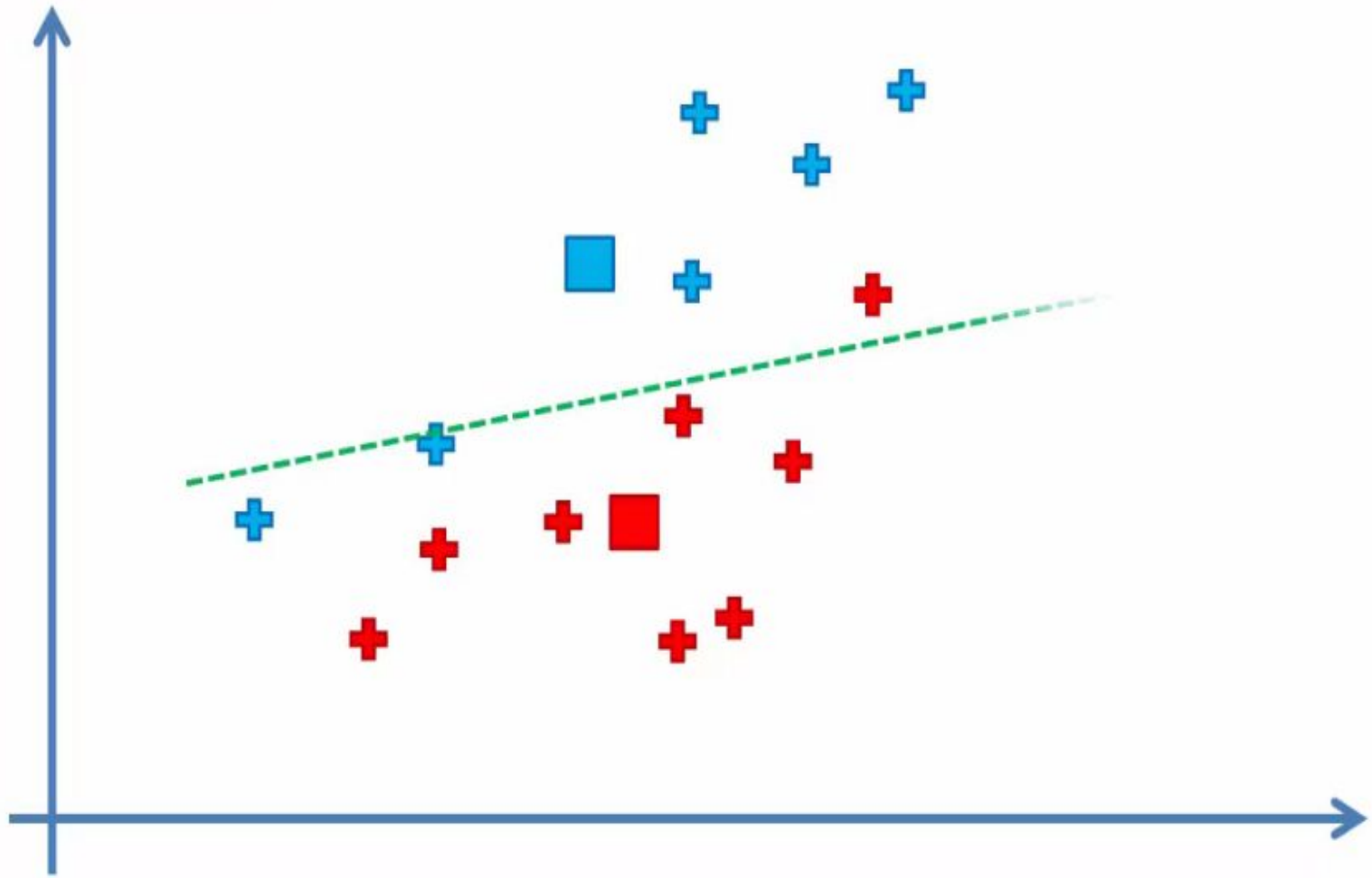
STEP 3: Assign each data point to the closest centroid ➡ That forms K clusters



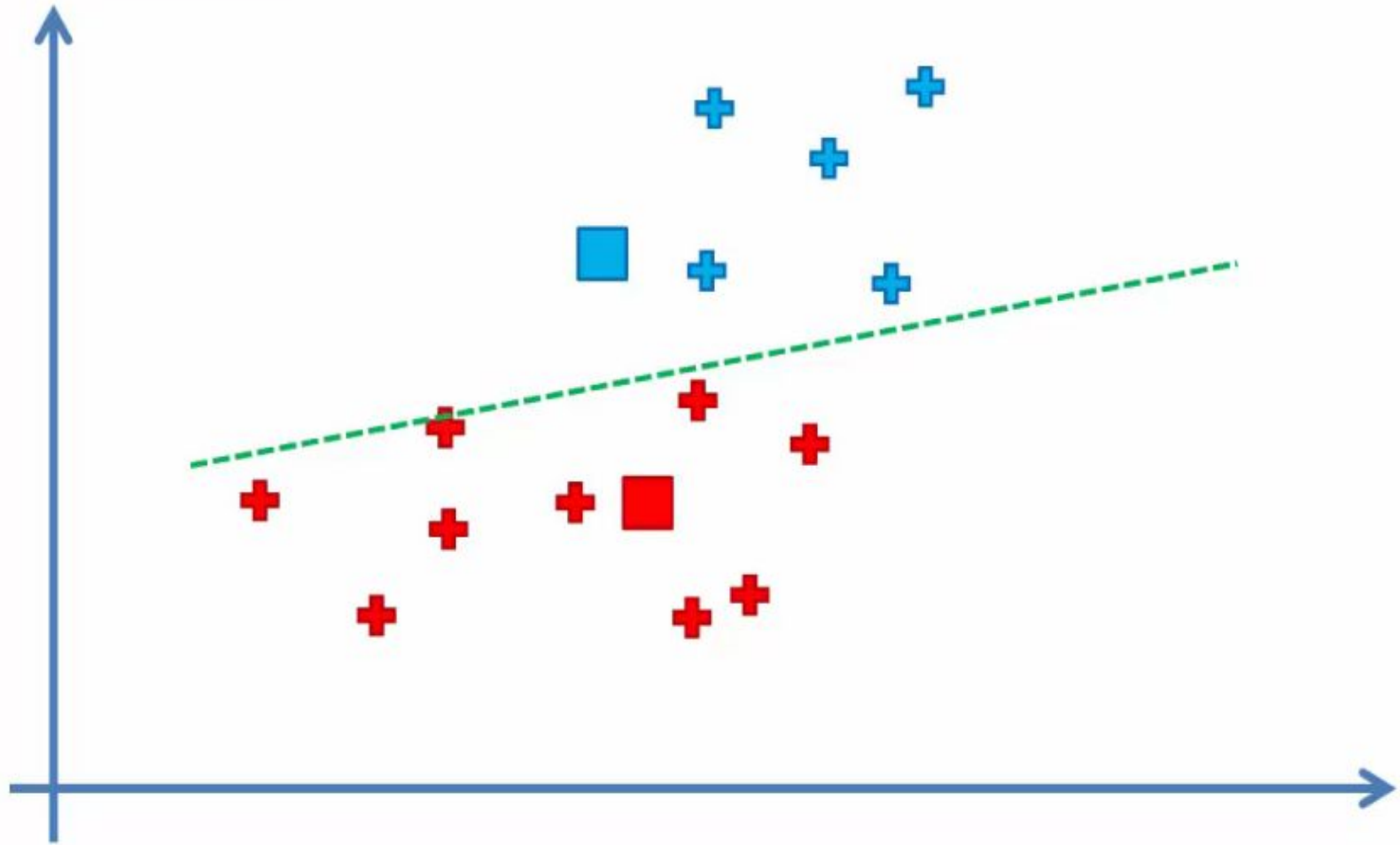
STEP 4: Compute and place the new centroid of each cluster



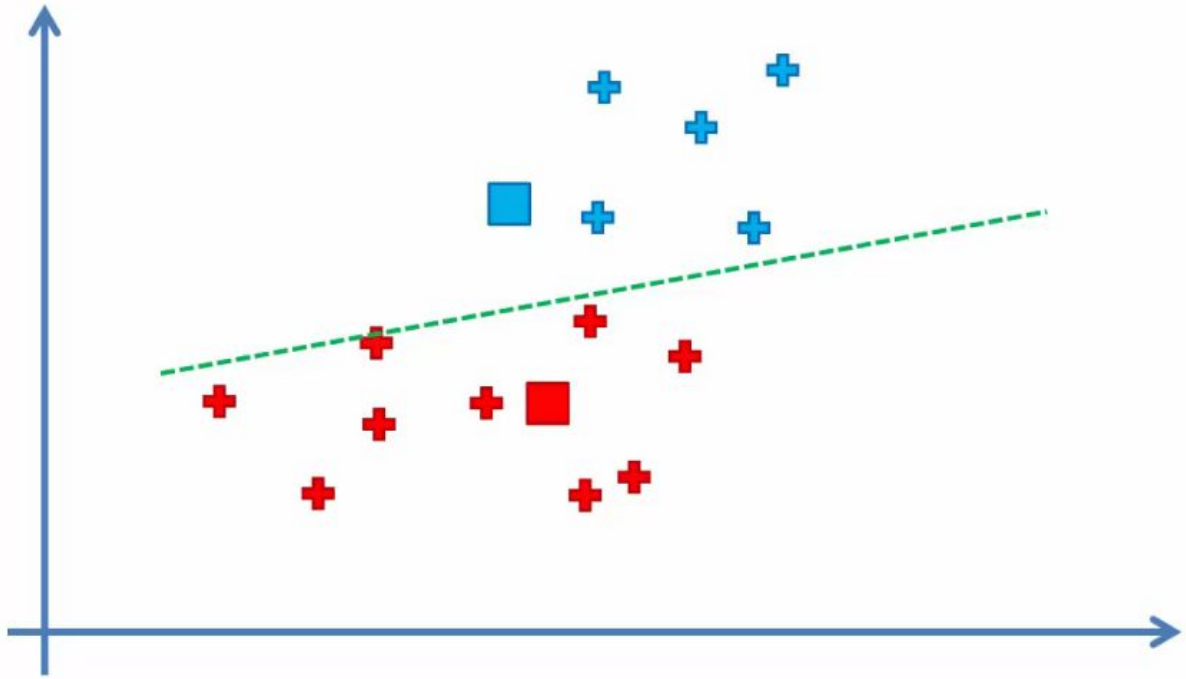
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



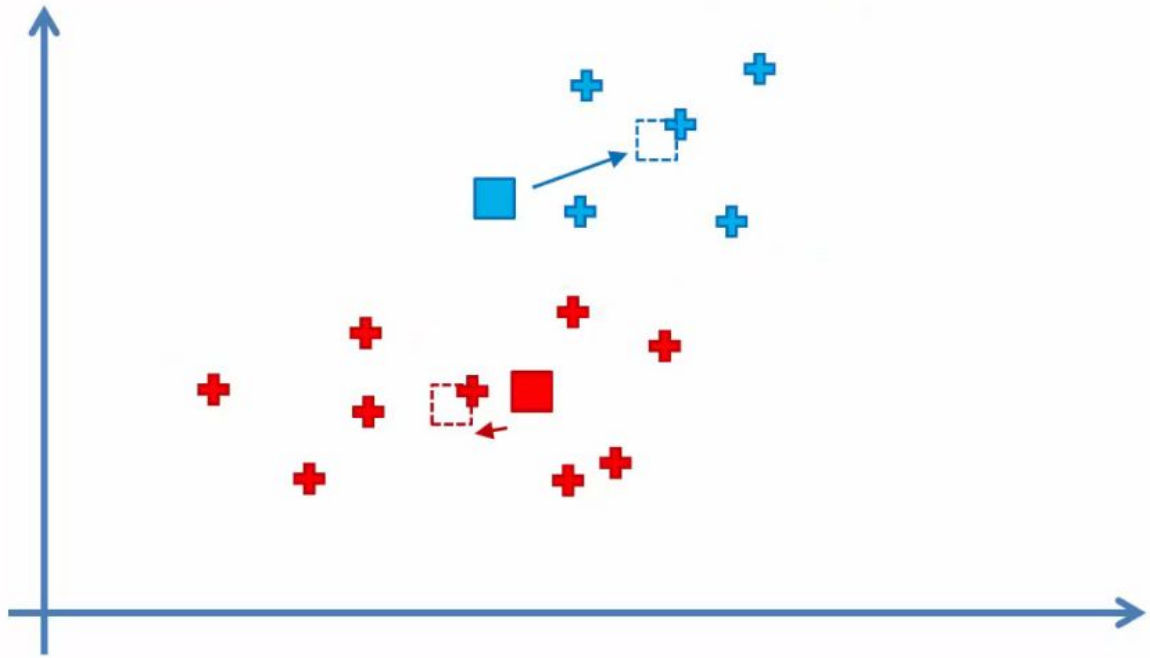
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



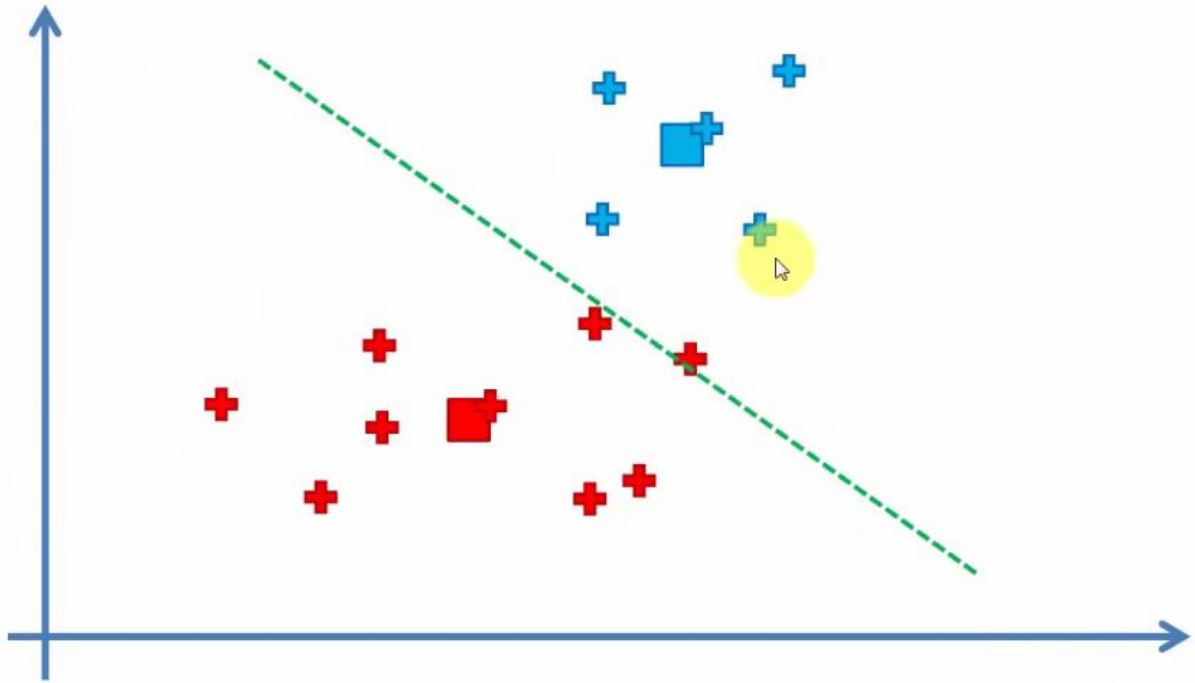
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



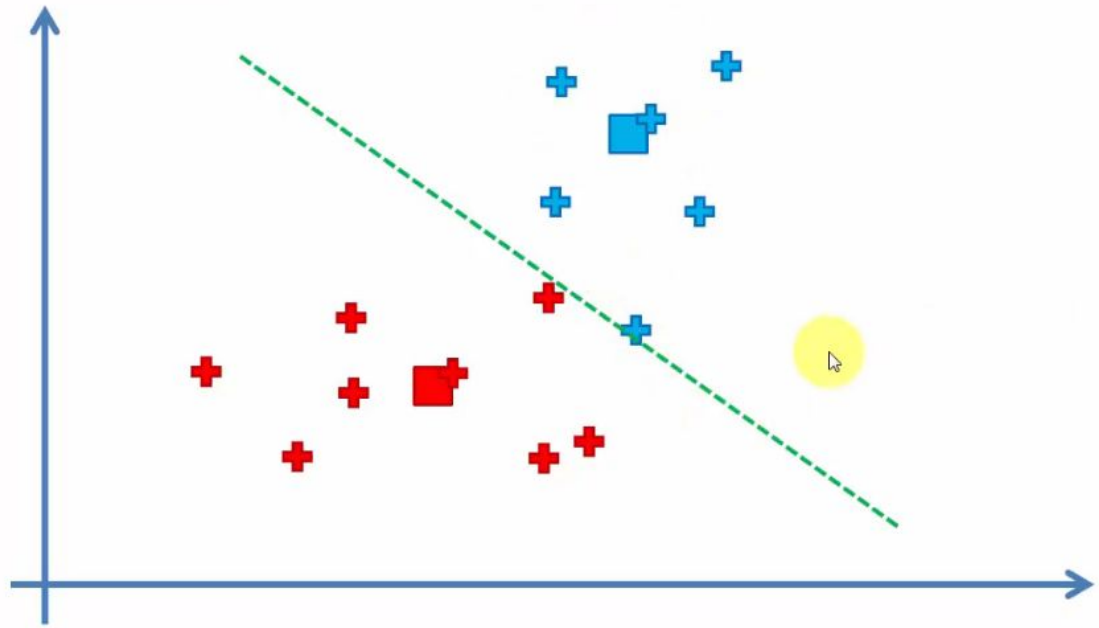
STEP 4: Compute and place the new centroid of each cluster



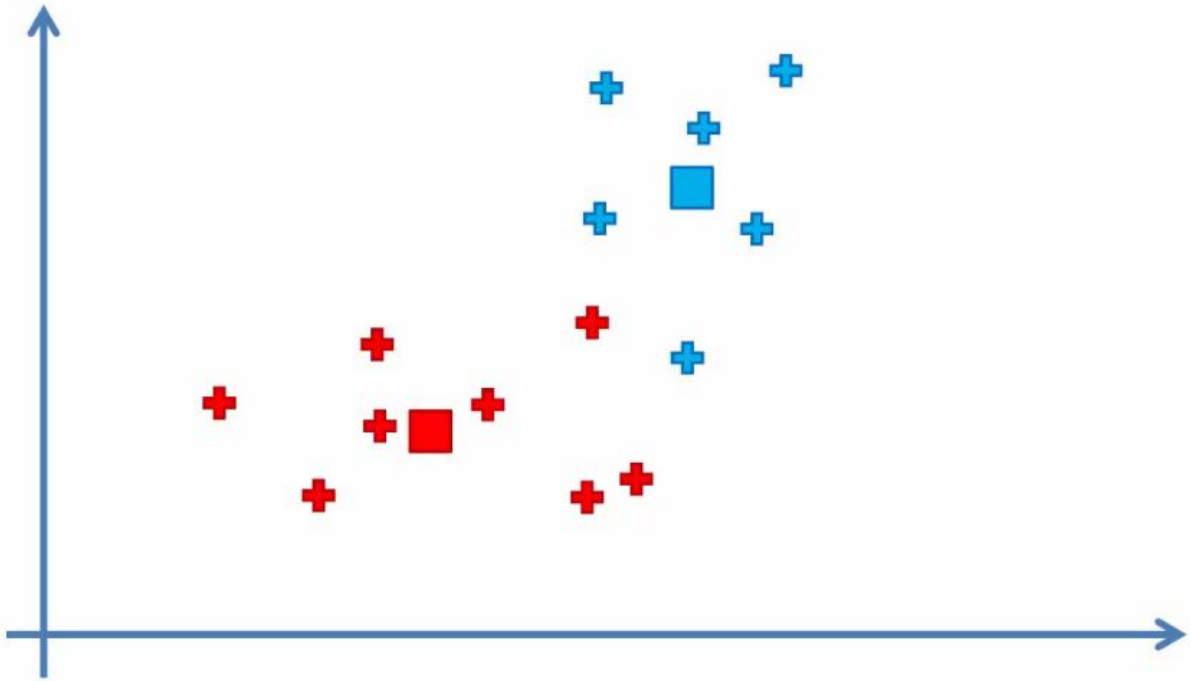
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.

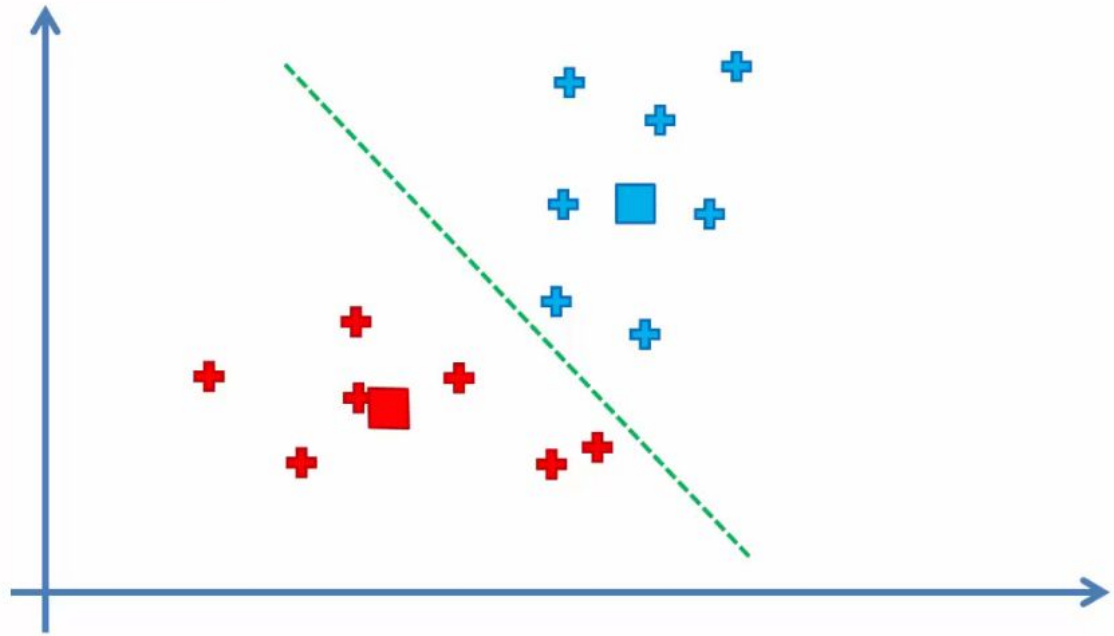


STEP 4: Compute and place the new centroid of each cluster

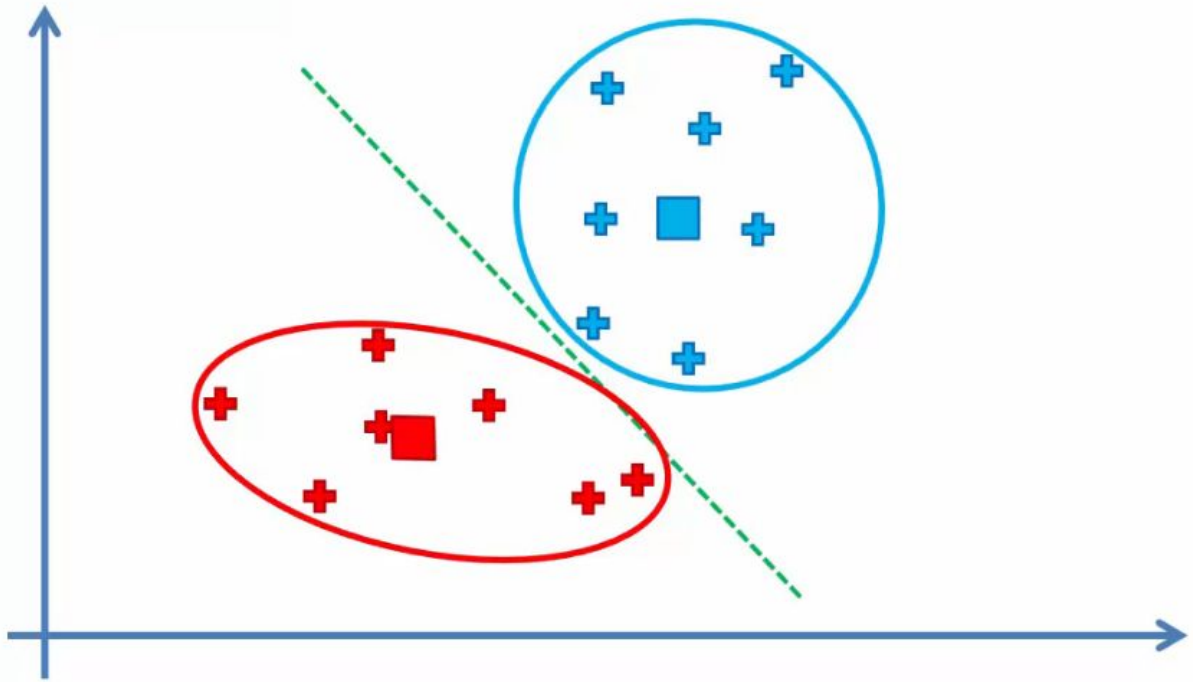


After doing iterations

STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.



FIN: Your Model Is Ready



K-Means Clustering – Solved Example

- Suppose that the data mining task is to cluster points into three clusters,
- where the points are
- $A1(2, 10)$, $A2(2, 5)$, $A3(8, 4)$, $B1(5, 8)$, $B2(7, 5)$, $B3(6, 4)$, $C1(1, 2)$, $C2(4, 9)$.
- The distance function is Euclidean distance.
- Suppose initially we assign $A1$, $B1$, and $C1$ as the center of each cluster, respectively.

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10								
A2	2	5								
A3	8	4								
B1	5	8								
B2	7	5								
B3	6	4								
C1	1	2								
C2	4	9								

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00							
A2	2	5	5.00							
A3	8	4	8.49							
B1	5	8	3.61							
B2	7	5	7.07							
B3	6	4	7.21							
C1	1	2	8.06							
C2	4	9	2.24							

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06			
A2	2	5	5.00		4.24		3.16			
A3	8	4	8.49		5.00		7.28			
B1	5	8	3.61		0.00		7.21			
B2	7	5	7.07		3.61		6.71			
B3	6	4	7.21		4.12		5.39			
C1	1	2	8.06		7.21		0.00			
C2	4	9	2.24		1.41		7.62			

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

New Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to			Cluster	New Cluster
A1	2	10				1	
A2	2	5				3	
A3	8	4				2	
B1	5	8				2	
B2	7	5				2	
B3	6	4				2	
C1	1	2				3	
C2	4	9				2	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2 →	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1

K-Means Clustering – Solved Example

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			2	10	6	6	1.5	1.5		
A1	2	10	0.00		5.66		6.52		1	
A2	2	5	5.00		4.12		1.58		3	
A3	8	4	8.49		2.83		6.52		2	
B1	5	8	3.61		2.24		5.70		2	
B2	7	5	7.07		1.41		5.70		2	
B3	6	4	7.21		2.00		4.53		2	
C1	1	2	8.06		6.40		1.58		3	
C2	4	9	2.24		3.61		6.04		1	

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	.6.5	5.25	1.5	3.5		
A1	2	10							1	
A2	2	5							3	
A3	8	4							2	
B1	5	8							2	
B2	7	5							2	
B3	6	4							2	
C1	1	2							3	
C2	4	9							1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	1
A2	2	5	4.61		4.51		1.58		3	3
A3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.50		3.13		5.70		2	1
B2	7	5	6.02		0.56		5.70		2	2
B3	6	4	6.26		1.35		4.53		2	2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.12		4.51		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	1
A2	2	5	4.61		4.51		1.58		3	3
A3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.50		3.13		5.70		2	1
B2	7	5	6.02		0.56		5.70		2	2
B3	6	4	6.26		1.35		4.53		2	2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.12		4.51		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

New Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	1
A2	2	5	4.61		4.51		1.58		3	3
A3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.50		3.13		5.70		2	1
B2	7	5	6.02		0.56		5.70		2	2
B3	6	4	6.26		1.35		4.53		2	2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.12		4.51		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
A1	2	10							1	
A2	2	5							3	
A3	8	4							2	
B1	5	8							1	
B2	7	5							2	
B3	6	4							2	
C1	1	2							3	
C2	4	9							1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	
A2	2	5	4.33		5.04		1.58		3	
A3	8	4	6.62		1.05		6.52		2	
B1	5	8	1.67		4.18		5.70		1	
B2	7	5	5.21		0.67		5.70		2	
B3	6	4	5.52		1.05		4.53		2	
C1	1	2	7.49		6.44		1.58		3	
C2	4	9	0.33		5.55		6.04		1	

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering – Solved Example

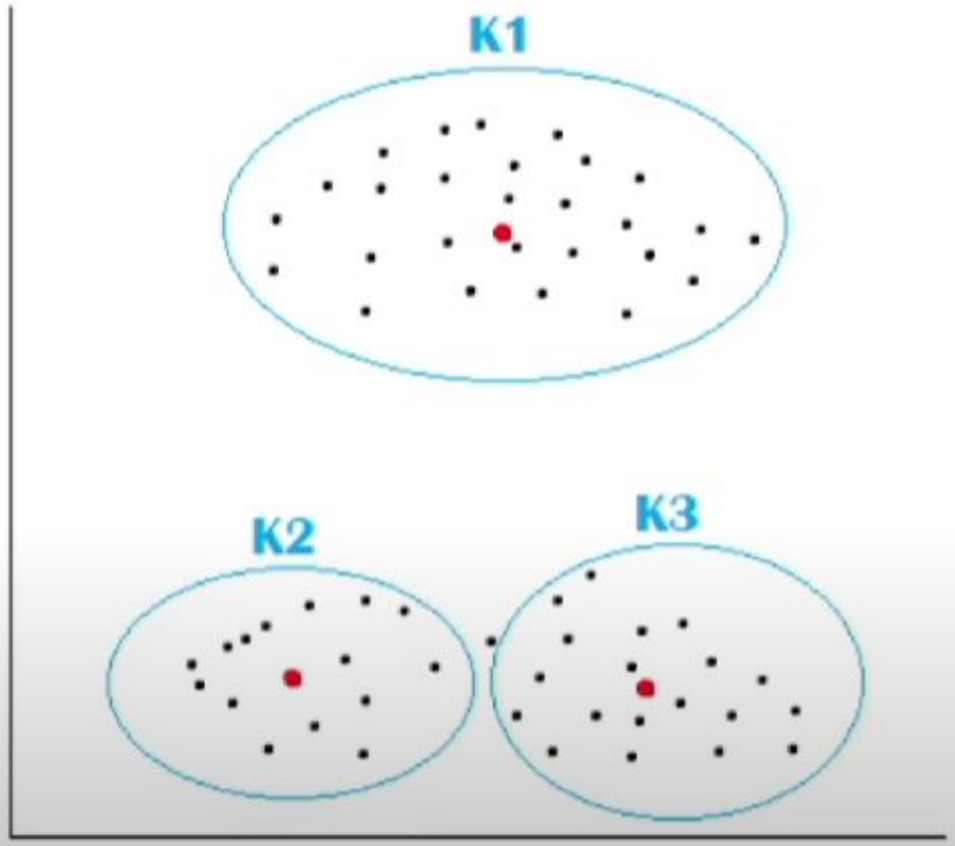
Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	1
A2	2	5	4.33		5.04		1.58		3	3
A3	8	4	6.62		1.05		6.52		2	2
B1	5	8	1.67		4.18		5.70		1	1
B2	7	5	5.21		0.67		5.70		2	2
B3	6	4	5.52		1.05		4.53		2	2
C1	1	2	7.49		6.44		1.58		3	3
C2	4	9	0.33		5.55		6.04		1	1

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K-Means Clustering - Disadvantages

- Need to select the initial centroids for each of the clusters.
- It requires to specify the **number of clusters (k) in advance**.
- Now, How to select optimal number of clusters (k) for the given data set.



K-Means Clustering

There are two methods to select optimal number of cluster

in K-Means algorithm

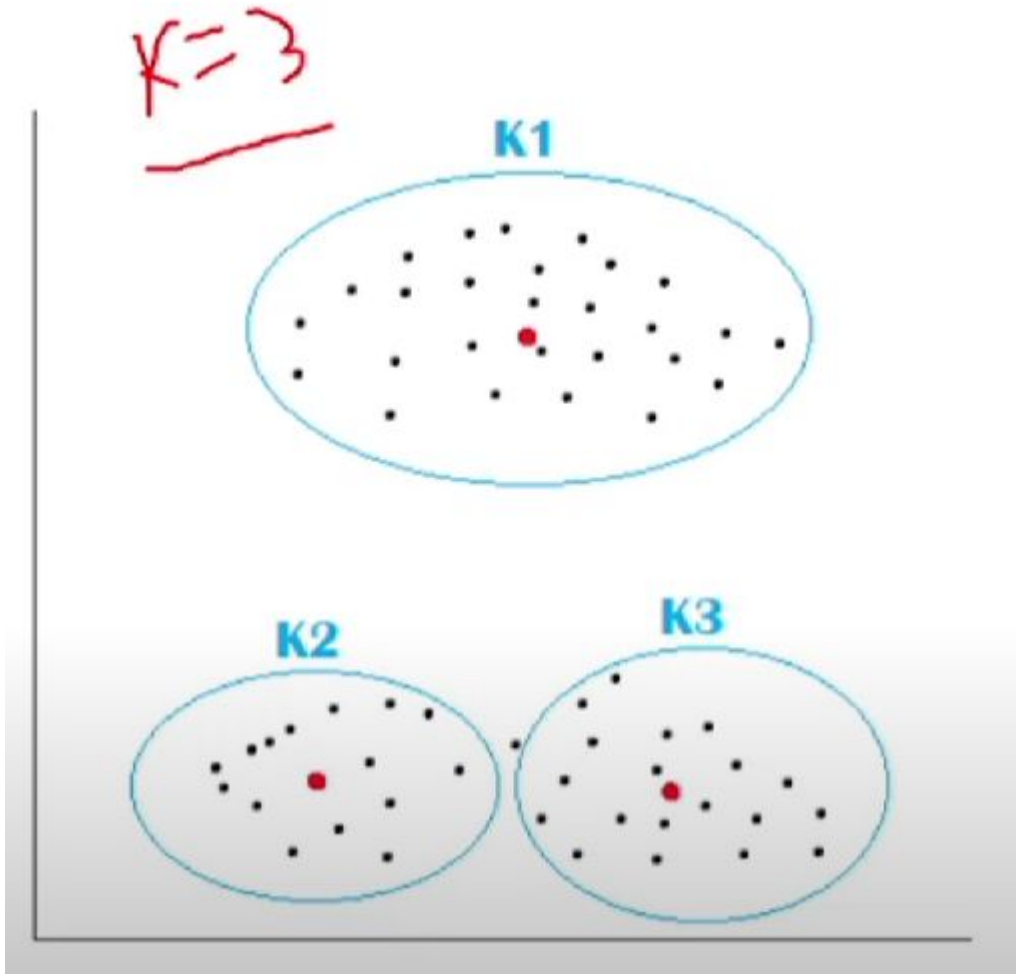
- Elbow Method
- Silhouette Method

K-Means Clustering – Elbow Method

- **Step 1:** Apply K-Means clustering algorithm and form the clusters for different values of k. For example, k=[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- **Step 2:** For each k calculate the within-cluster sum of squares (WCSS).

$$\text{WCSS} = \sum_{C_k}^{C_n} \left(\sum_{d_i \text{ in } C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

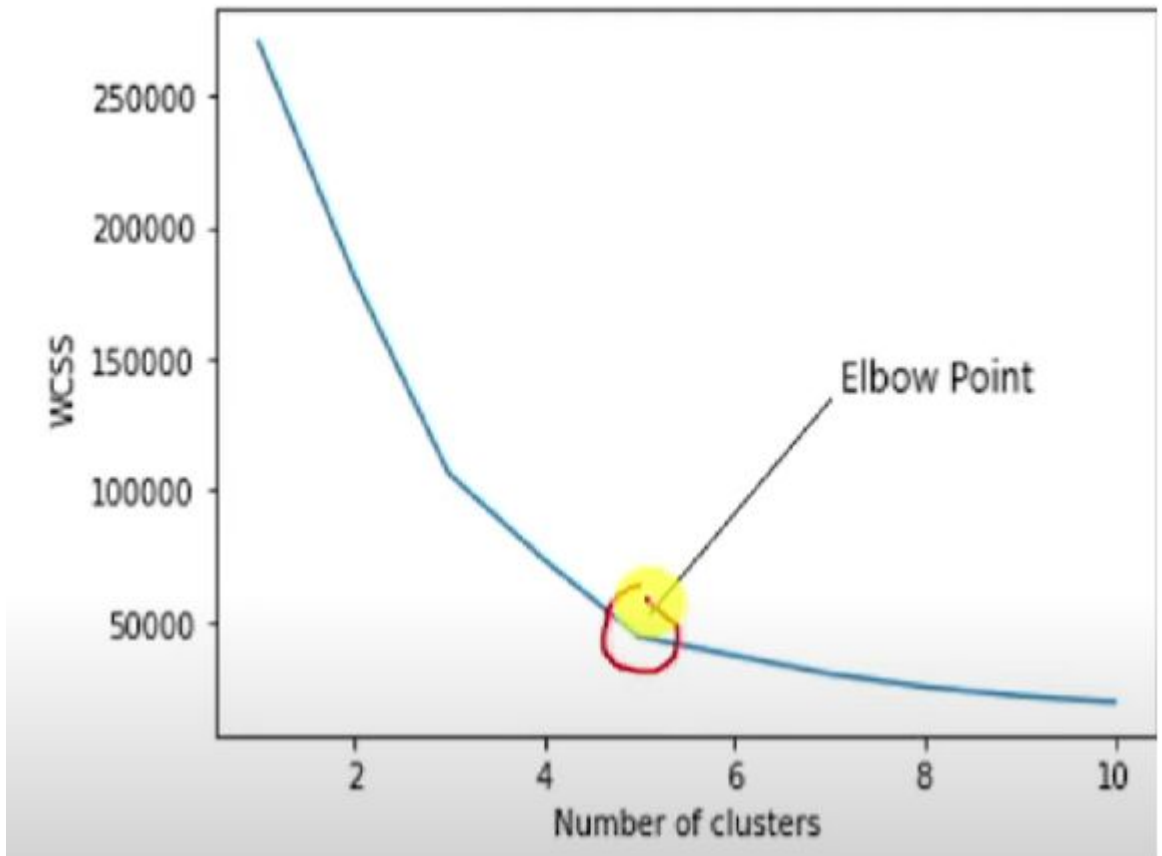
Where,
C is the cluster centroids and d is the data point in each Cluster.



K-Means Clustering – Elbow Method

- **Step 3:** Plot curve of WCSS according to the number of clusters.

- **Step 4:** The location of bend in the plot is generally considered an indicator of the approximate number of clusters.



K-Means Clustering – Silhouette Method

- The silhouette coefficient or silhouette score is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).
- For different values of k (say 1 to 10) calculate the silhouette coefficient.
- Plot Silhouette coefficient for each value of K.

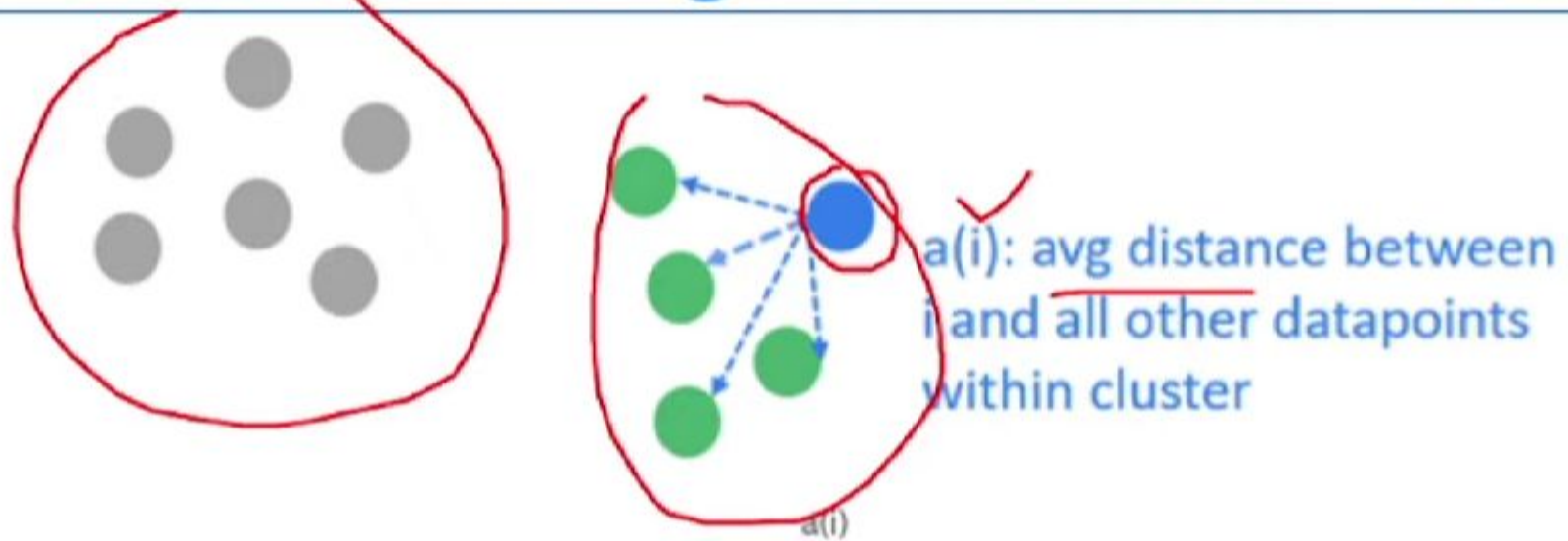
K-Means Clustering – Silhouette Method

- The equation for calculating the silhouette coefficient for a particular data point:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

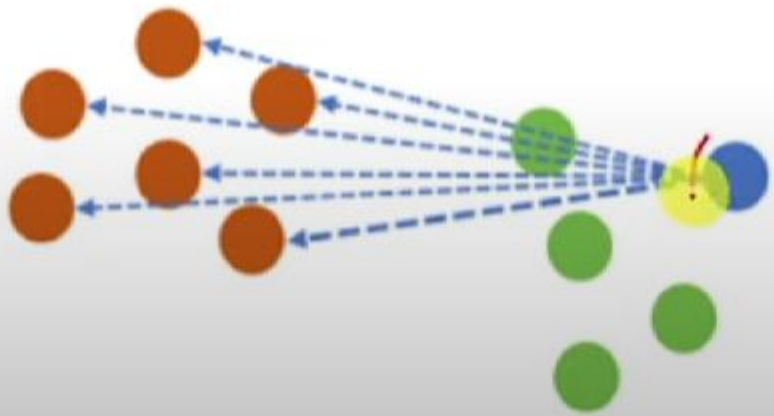
- $S(i)$ is the silhouette coefficient of the data point i.
- $a(i)$ is the average distance between i and all the other data points in the cluster to which i belongs.
- $b(i)$ is the average distance from i to all clusters to which i does not belong.

K-Means Clustering – Silhouette Method



a(i): avg distance between i and all other datapoints within cluster

a(i)



b(i): avg distance between i and all other datapoints outside/neighboring cluster

b(i)

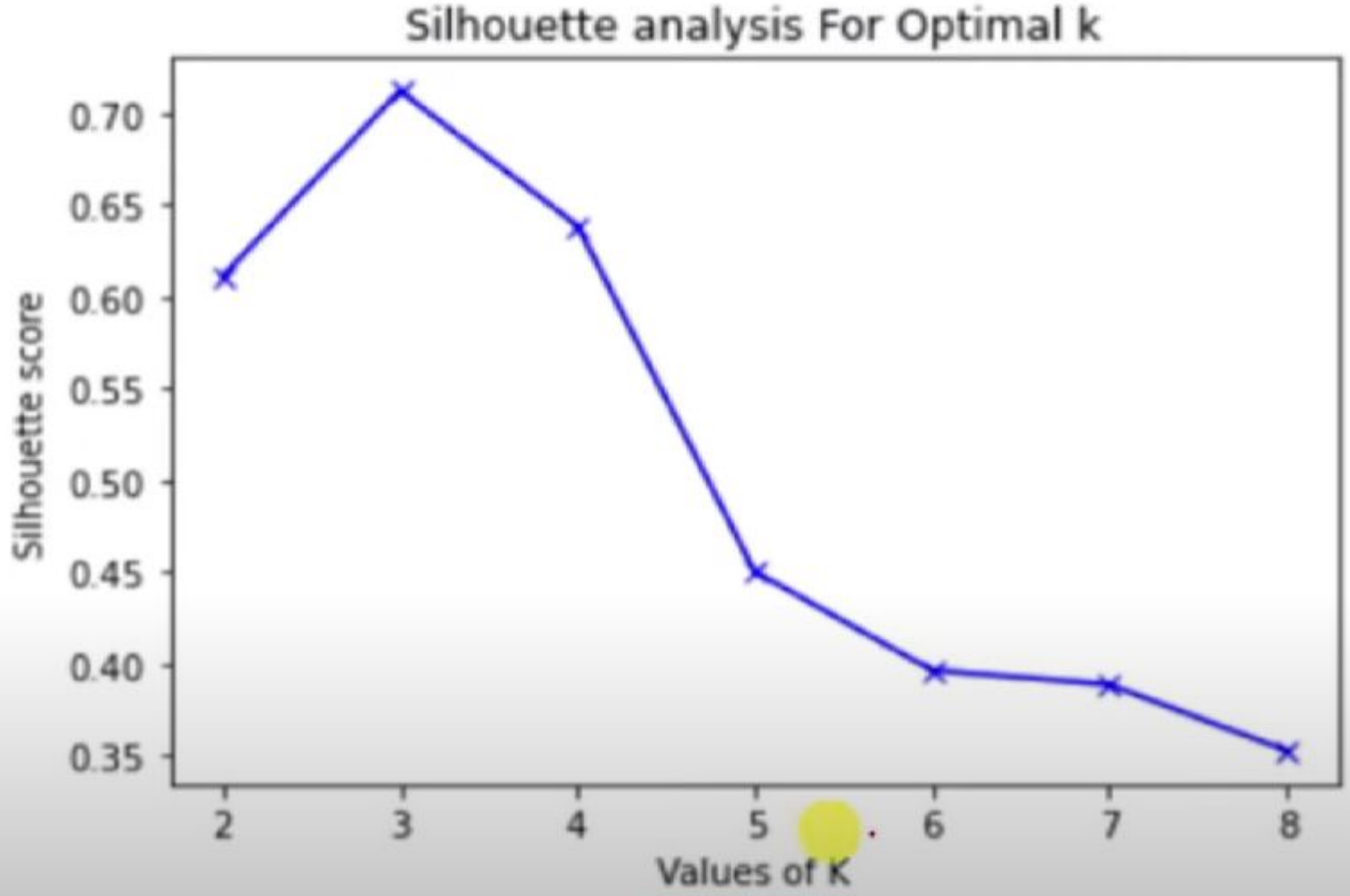
K-Means Clustering – Silhouette Method

- The equation for calculating the silhouette coefficient for a particular data point:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- $S(i)$ is the silhouette coefficient of the data point i.
- $a(i)$ is the average distance between i and all the other data points in the cluster to which i belongs.
- $b(i)$ is the average distance from i to all clusters to which i does not belong.
- Calculate the average_silhouette for every k. $\boldsymbol{Average - silhouette} = \frac{\boldsymbol{s(i)}}{\boldsymbol{n}}$

K-Means Clustering – Silhouette Method



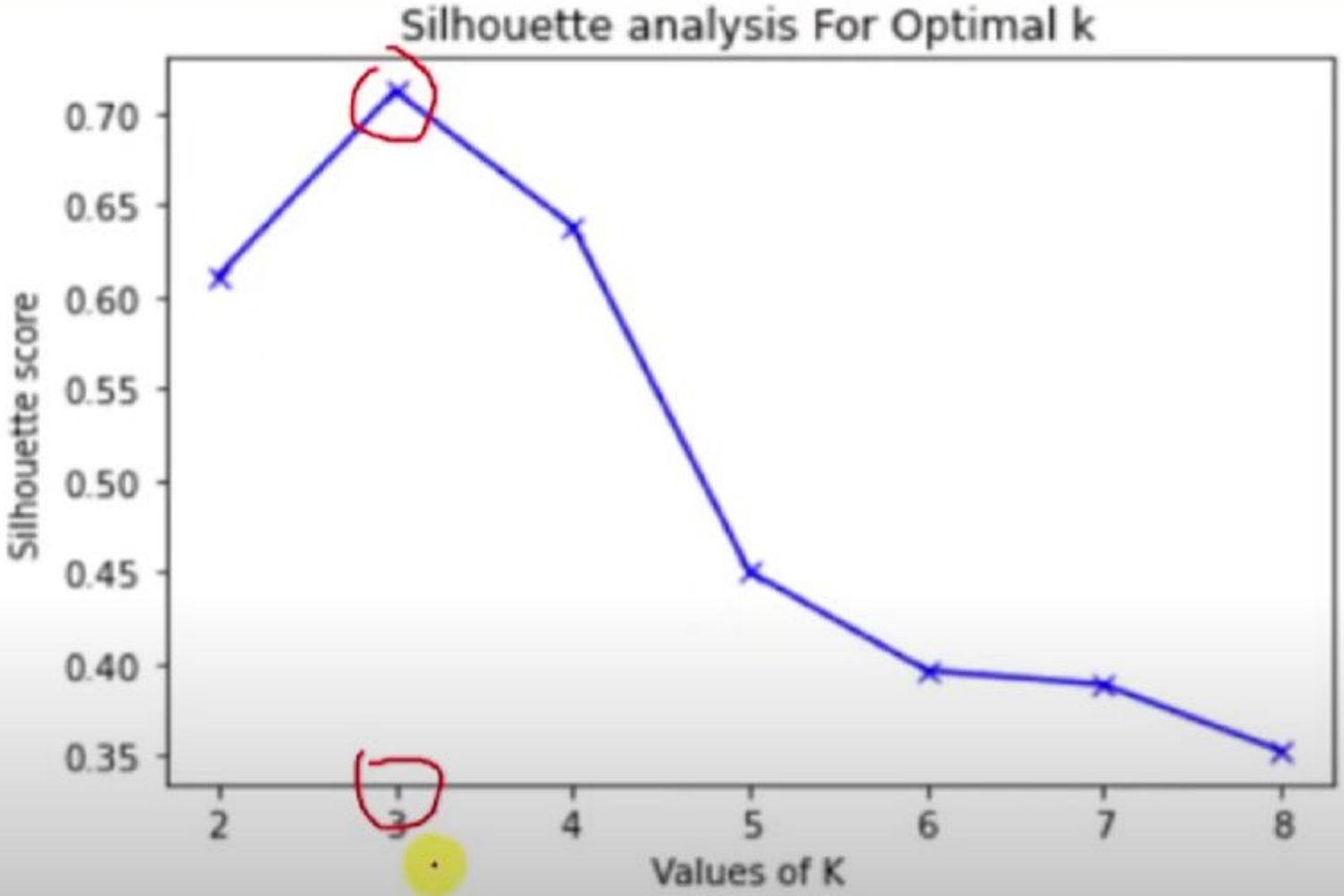
Line plot between K and Silhouette score

K-Means Clustering – Silhouette Method

Points to Remember While Calculating Silhouette Coefficient:

- The value of the silhouette coefficient is between [-1, 1].
- A score of 1 denotes the best, meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1.
- Values near 0 denote overlapping clusters.

K-Means Clustering – Silhouette Method



Line plot between K and Silhouette score

How to Compute Silhouette Coefficient – Clustering

Cluster Label

Point	Cluster Label
P1	1 ✓
P2	1 ✓
P3	2 ✓
P4	2 ✓

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

- Compute the Silhouette Coefficient for each point, each of the two clusters, and the overall clustering.

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

$$Silhouette\ Coefficient = 1 - \frac{a}{b}$$

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

$$Silhouette\ Coefficient = 1 - \frac{a}{b}$$

where

- a indicate the average distance of a point to other points in its cluster.
- b indicate the minimum of the average distance of a point to points in another cluster.

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Point P1:

- $a = \frac{0.1}{1} = 0.1 \quad and \quad b = \frac{0.65+0.55}{2} = 0.6$

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Point P1:

- $a = \frac{0.1}{1} = 0.1 \quad and \quad b = \frac{0.65+0.55}{2} = 0.6$
- $SC = 1 - \frac{a}{b} = 1 - \frac{0.1}{0.6} = \frac{5}{6} = 0.833$

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Point P1:

- $a = \frac{0.1}{1} = 0.1 \quad and \quad b = \frac{0.65+0.55}{2} = 0.6$
- $SC = 1 - \frac{a}{b} = 1 - \frac{0.1}{0.6} = \frac{5}{6} = 0.833$

Point P2:

- $SC = 1 - \frac{a}{b} = 1 - \frac{0.1}{\frac{0.7+0.6}{2}} = 0.846$

Point P3:

- $SC = 1 - \frac{a}{b} = 1 - \frac{0.3}{\frac{0.65+0.7}{2}} = 0.556$

Point P4:

- $SC = 1 - \frac{a}{b} = 1 - \frac{0.3}{\frac{0.55+0.6}{2}} = 0.478$

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Point P1: $SC = 0.833$

Point P2: $SC = 0.846$

Point P3: $SC = 0.556$

Point P4: $SC = 0.478$

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1

- $Average\ SC = \frac{0.833+0.846}{2} = 0.84$

Cluster 2

- $Average\ SC = \frac{0.556+0.478}{2} = 0.517$

How to Compute Silhouette Coefficient – Clustering

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Dissimilarity Matrix

Point	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Point P1: $SC = 0.833$

Point P2: $SC = 0.846$

Point P3: $SC = 0.556$

Point P4: $SC = 0.478$

Cluster 1 $Average\ SC = 0.84$

Cluster 2 $Average\ SC = 0.517$

Overall

- $Average\ SC = \frac{0.840+0.517}{2} = 0.68$