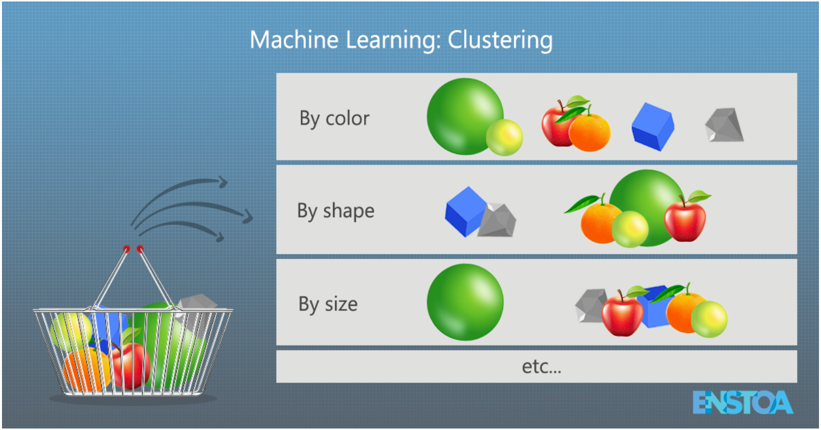
Clustering

What is clustering?

Clustering is grouping data points into groups where data points in one group are `similar` to each other.

What is clustering?



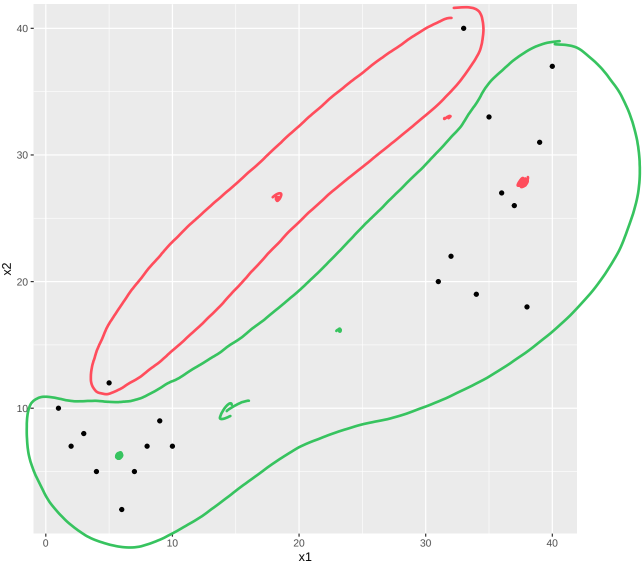
Methods of Clustering

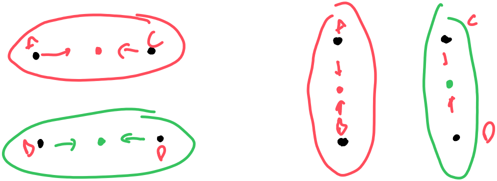
We will cover two clustering methods:

- K-means clustering and
- Hierarchical clustering

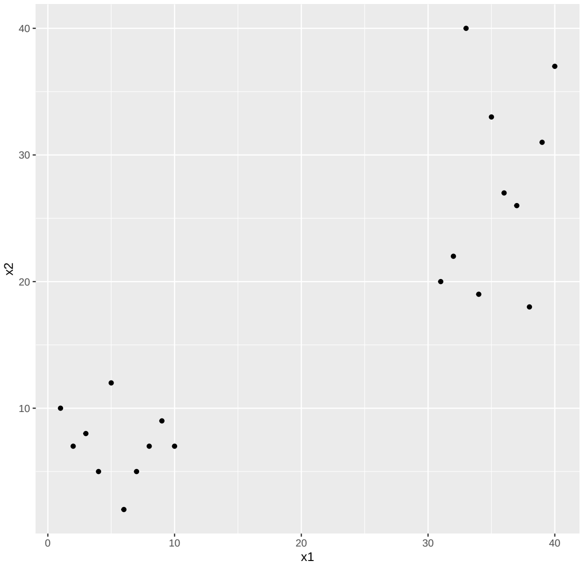
K-means clustering

Example - Data

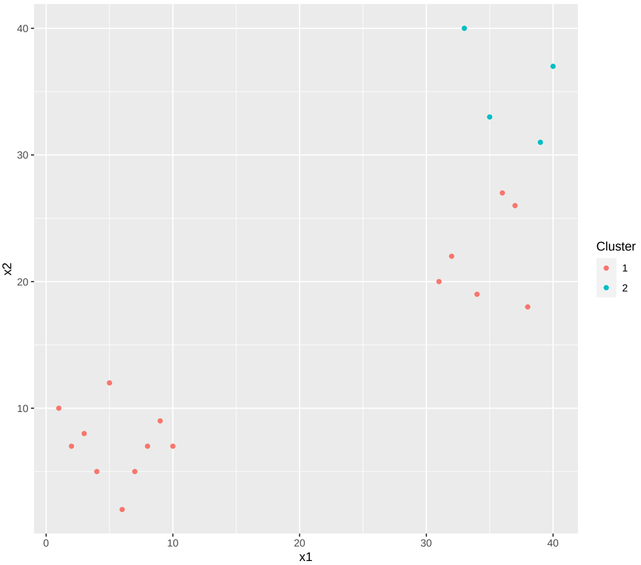



A
C
D
D
A
D
C
D

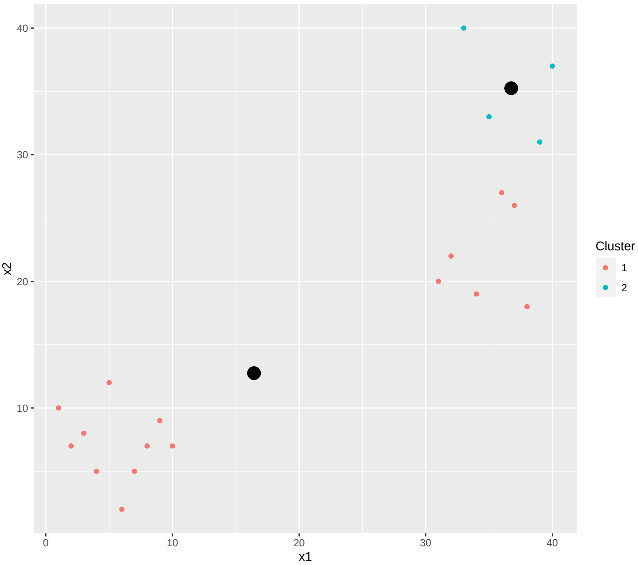
Step 1: Randomly Assign Points to Clusters



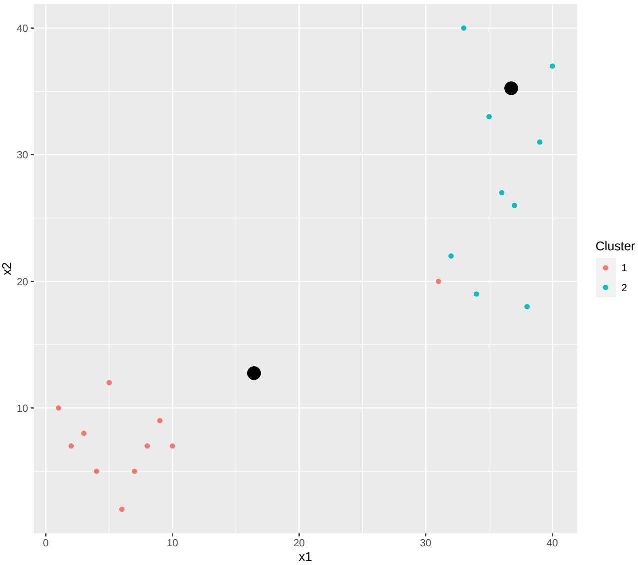
Step 1: Randomly Assign Points to Clusters



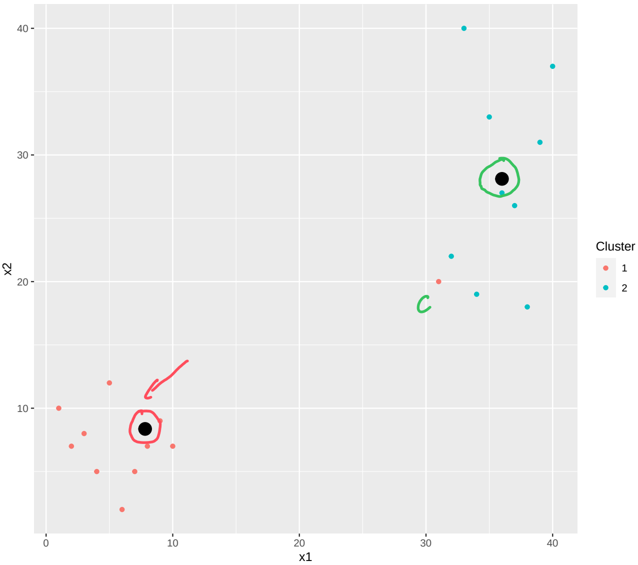
Locate centroids



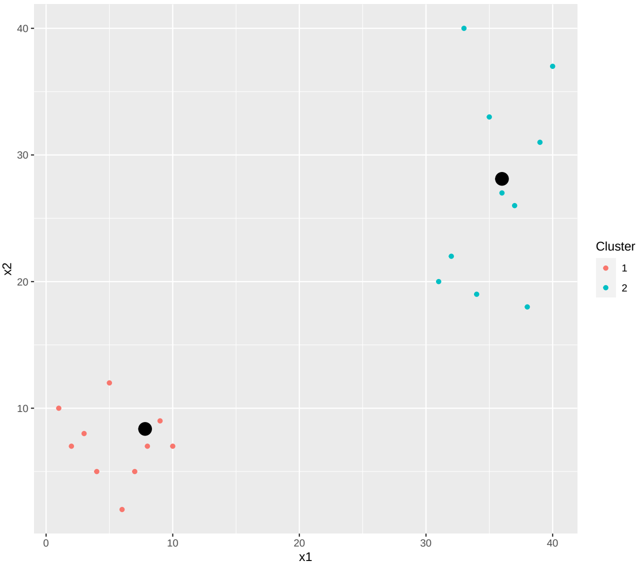
Reassign Points to clusters



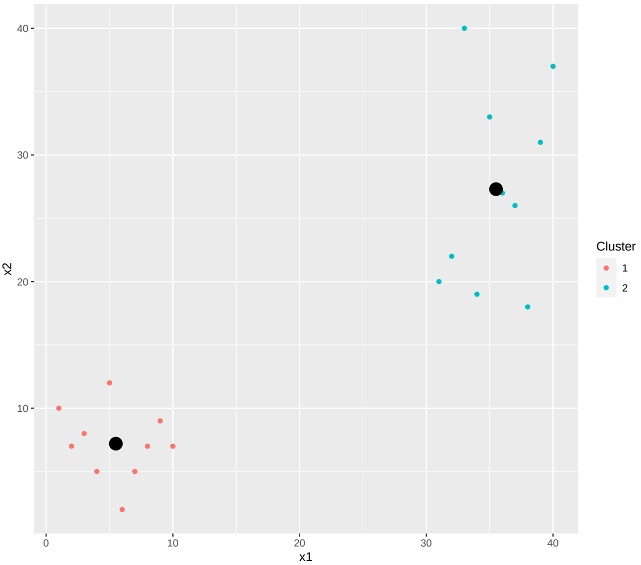
Relocate centroids



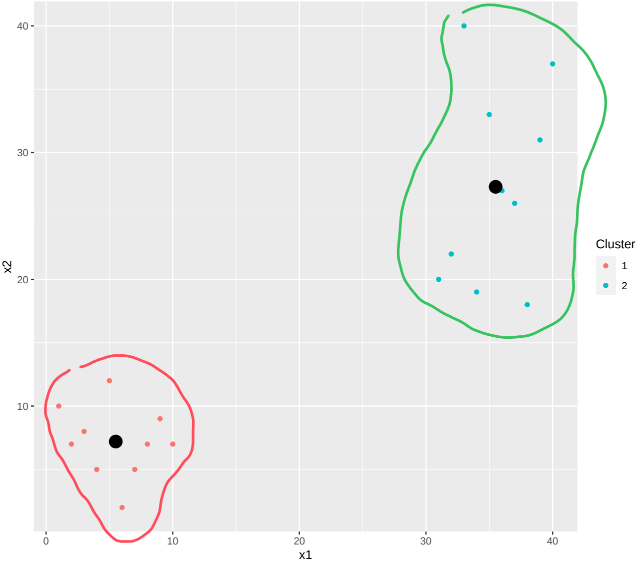
Reassign Points to clusters



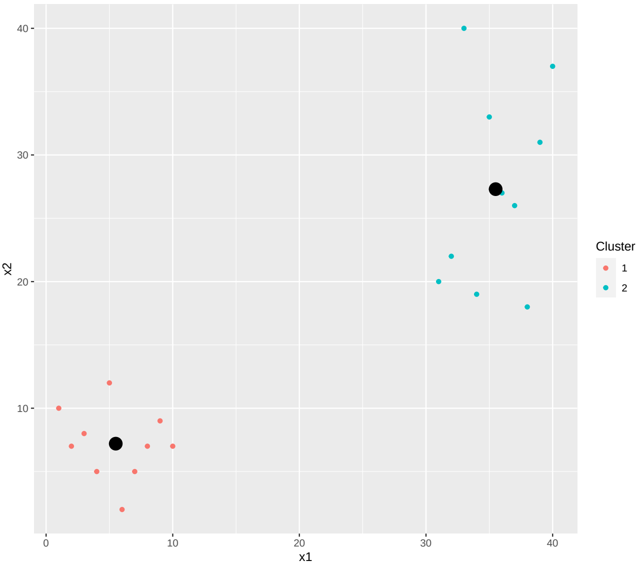
Relocate centroids



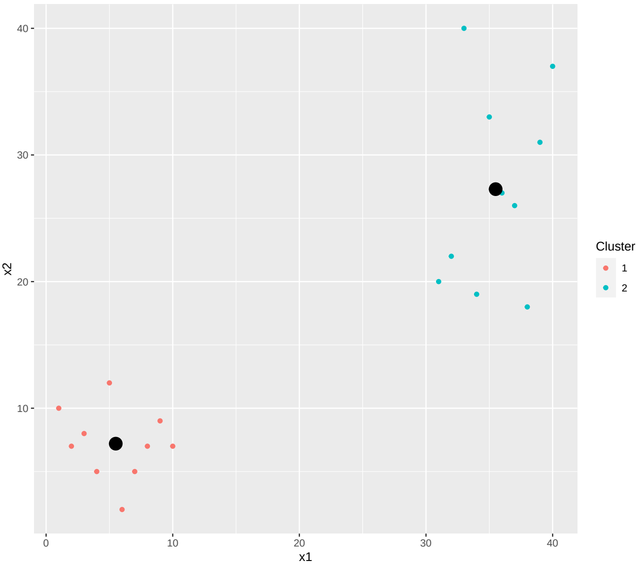
Reassign Points to clusters



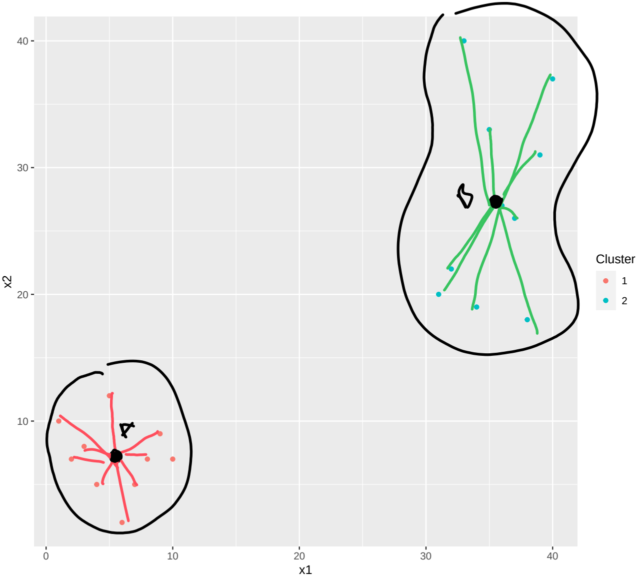
Relocate centroids



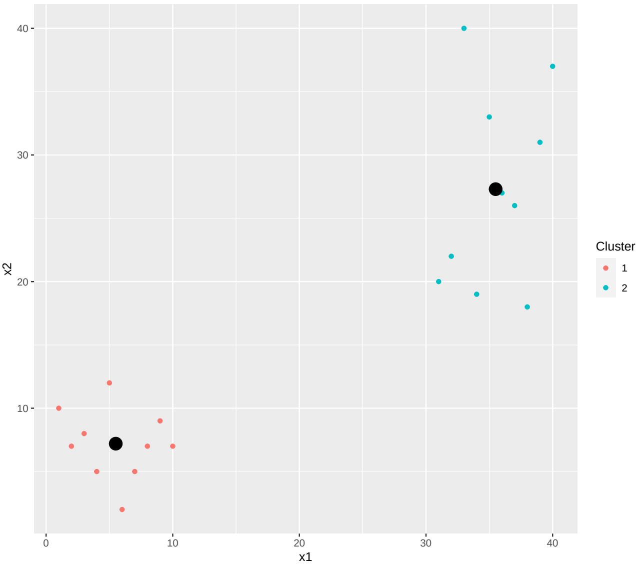
Step 2: Reassign Points to clusters



Step 2: Relocate centroids



Step 2: Reassign Points to clusters



Centroids

Cluster	x1	x2
1	5.5	7.2
2	35.5	27.3

K-means Algorithm

- 1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
- 2. Iterate until the cluster assignments stop changing:
    - (a) For each of the K clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    - (b) Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Dataset

Point	x	y
A	1	3
B	2	2
C	3	5
D	4	5
E	5	6

Randomly Assign Cluster to Points

Cluster	Point	x	y
1	A	1	3
2	B	2	2
1	C	3	5
1	D	4	5
2	E	5	6

Cluster	Point	x	y	C_1x	C_1y	C_2x	C_2y
1	A	1	3	2.67	4.33	3.5	4
2	B	2	2	2.67	4.33	3.5	4
1	C	3	5	2.67	4.33	3.5	4
1	D	4	5	2.67	4.33	3.5	4
2	E	5	6	2.67	4.33	3.5	4

Cluster	Point	x	y	C_1x	C_1y	C_2x	C_2y	dc1	dc2
1	A	1	3	2.67	4.33	3.5	4	2.13	2.69
2	B	2	2	2.67	4.33	3.5	4	2.42	2.50
1	C	3	5	2.67	4.33	3.5	4	0.75	1.12
1	D	4	5	2.67	4.33	3.5	4	1.49	1.12
2	E	5	6	2.67	4.33	3.5	4	2.87	2.50

Cluster	Point	x	y	dc1	dc2	min_distance
1	A	1	3	2.13	2.69	2.13
2	B	2	2	2.42	2.50	2.42
1	C	3	5	0.75	1.12	0.75
1	D	4	5	1.49	1.12	1.12
2	E	5	6	2.87	2.50	2.50

Cluster	Point	x	y	dc1	dc2	min_distance	New_Cluster
1	A	1	3	2.13	2.69	2.13	1
2	B	2	2	2.42	2.50	2.42	1
1	C	3	5	0.75	1.12	0.75	1
1	D	4	5	1.49	1.12	1.12	2
2	E	5	6	2.87	2.50	2.50	2

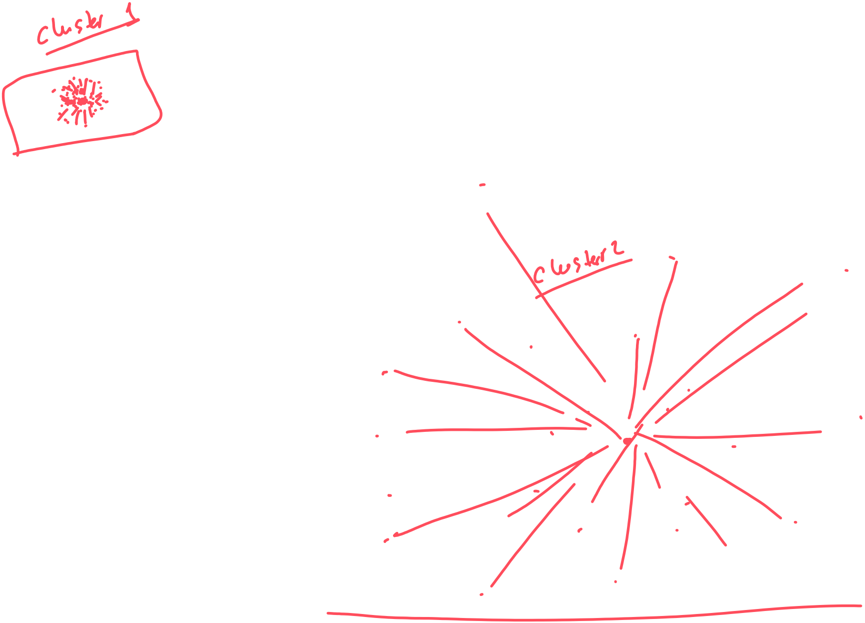
Total Variance within

- With the initial Clusters:
    - Cluster 1 = {A, C, D}
    - Cluster 2 = {B, E}
- Let M and N are the centroids of cluster 1 and 2 respectively

- Total Variance within

$$
\begin{aligned}
&= 2 \cdot (MA^2 + MC^2 + MD^2 + NB^2 + ND^2) \\
&= \frac{1}{3}(AB^2 + AC^2 + AD^2) + \frac{1}{2} \cdot BE^2
\end{aligned}
$$

- Total Variance within 19.83


cluster 1
cluster2

Total Variance within

- With the new Clusters:
  - Cluster 1 = $\{$A, B, C$\}$
  - Cluster 2 = $\{$D, E$\}$
- Let H and K are the centroids of cluster 1 and 2, respectively.

- Total Variance within

$$
\begin{aligned}
&= 2 \cdot (HA^2 + HB^2 + HC^2 + KD^2 + KE^2) \\
&= \frac{1}{3}(AB^2 + AC^2 + BC^2) + \frac{1}{2} \cdot DE^2
\end{aligned}
$$

- Total Variance within 7.67
- The process of k-means will minimize the total variance within

Example

You apply 2-means clustering to a set of five observations with two features. You are given the following initial cluster assignments:

Observation	X_1	X_2	Initial cluster
A	1	1	1
B	0	0	1
C	0	1	1
D	2	1	2
E	1	0	2

Calculate the total within-cluster variation (Total Variance within) of the initial cluster assignments, based on Euclidean distance measure.