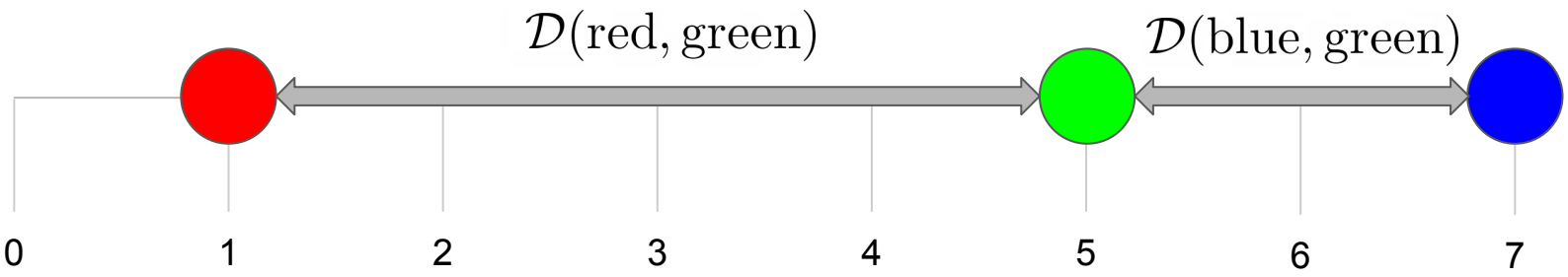
k-Means

MACHINE LEARNING

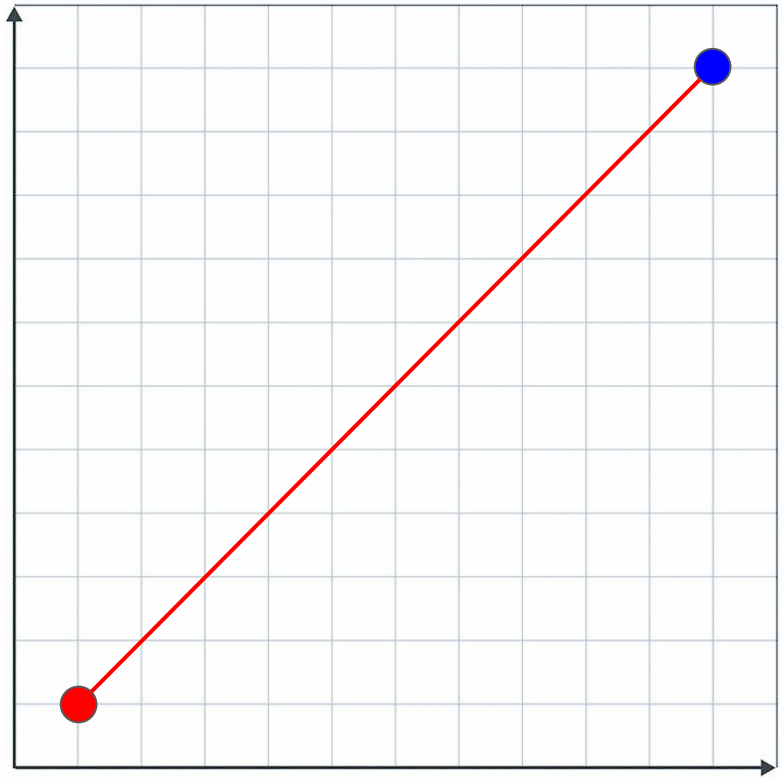
Pakarat Musikawan

Distance Metrics

$$\mathcal{D}(\text{red}, \text{green}) = |\text{red} - \text{green}|$$
$$\mathcal{D}(\text{blue}, \text{green}) = |\text{blue} - \text{green}|$$
$$\mathcal{D}(\text{red}, \text{green}) > \mathcal{D}(\text{blue}, \text{green})$$



Distance Metrics

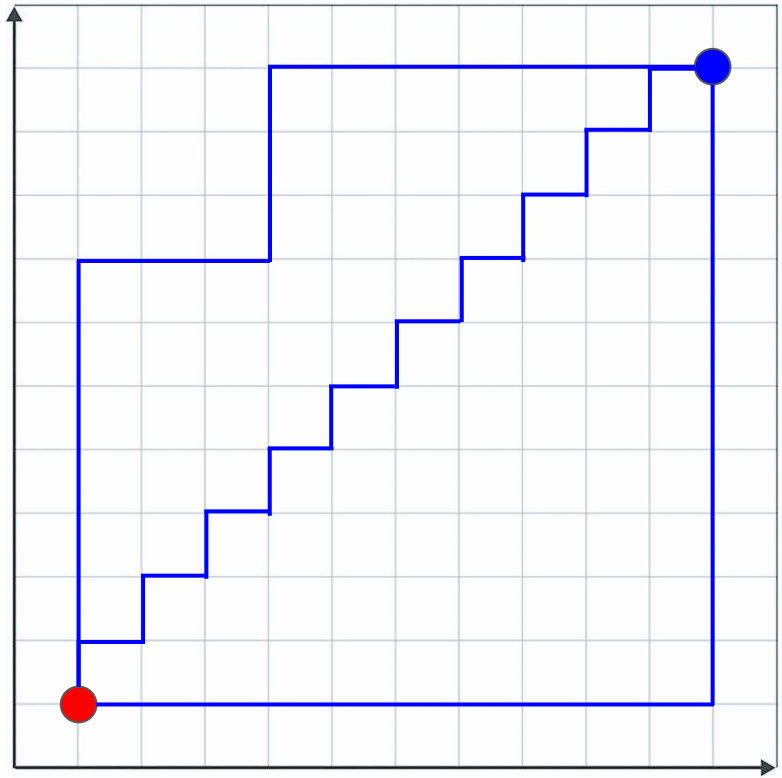


$$\mathcal{D}(\text{red}, \text{blue}) = d(\Delta x, \Delta y)$$
$$= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})$$

Euclidean

$$\mathcal{D}(\text{red}, \text{blue}) = d(\Delta x, \Delta y)$$
$$= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})$$
$$= \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2}$$

Distance Metrics

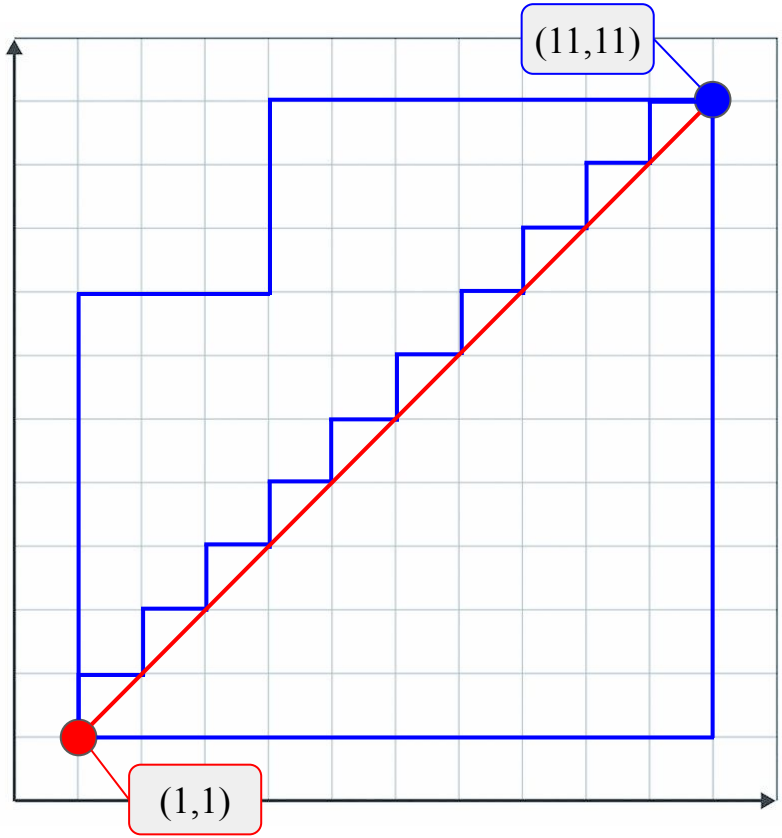


$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}})\end{aligned}$$

Manhattan

$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= |x_{\text{red}} - x_{\text{blue}}| + |y_{\text{red}} - y_{\text{blue}}|\end{aligned}$$

Distance Metrics



$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= d(1 - 11, 1 - 11)\end{aligned}$$

Euclidean

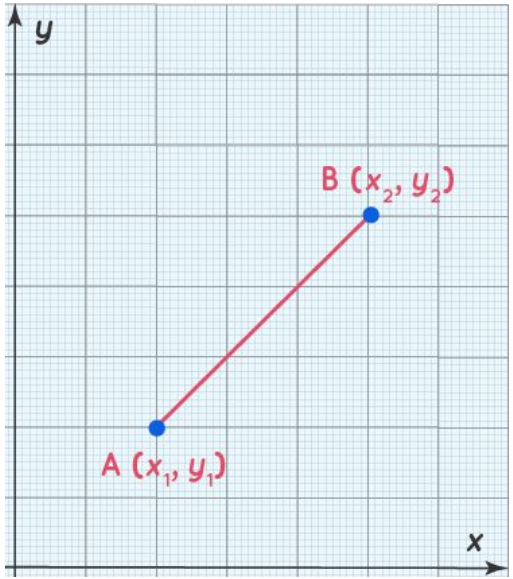
$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= \sqrt{(1 - 11)^2 + (1 - 11)^2} = 14.14\end{aligned}$$

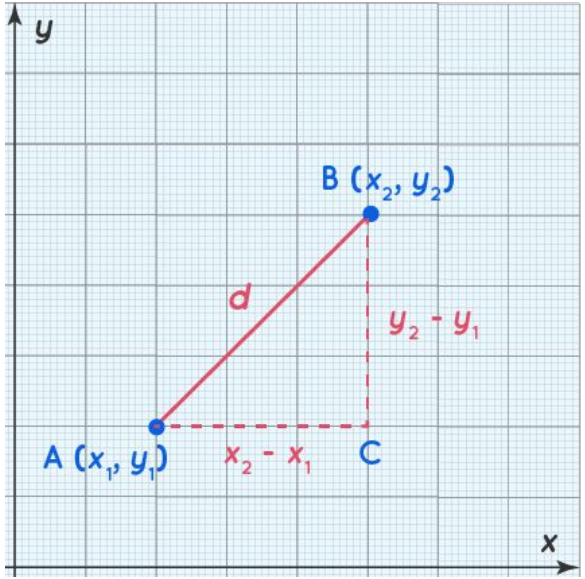
Manhattan

$$\begin{aligned}\mathcal{D}(\text{red}, \text{blue}) &= d(\Delta x, \Delta y) \\ &= d(x_{\text{red}} - x_{\text{blue}}, y_{\text{red}} - x_{\text{blue}}) \\ &= |1 - 11| + |1 - 11| = 20\end{aligned}$$

Euclidean Distance

- In coordinate geometry, Euclidean distance is the distance between two points.
- We derive the Euclidean distance formula using the Pythagoras theorem.





$$A = (x_1, y_1)$$

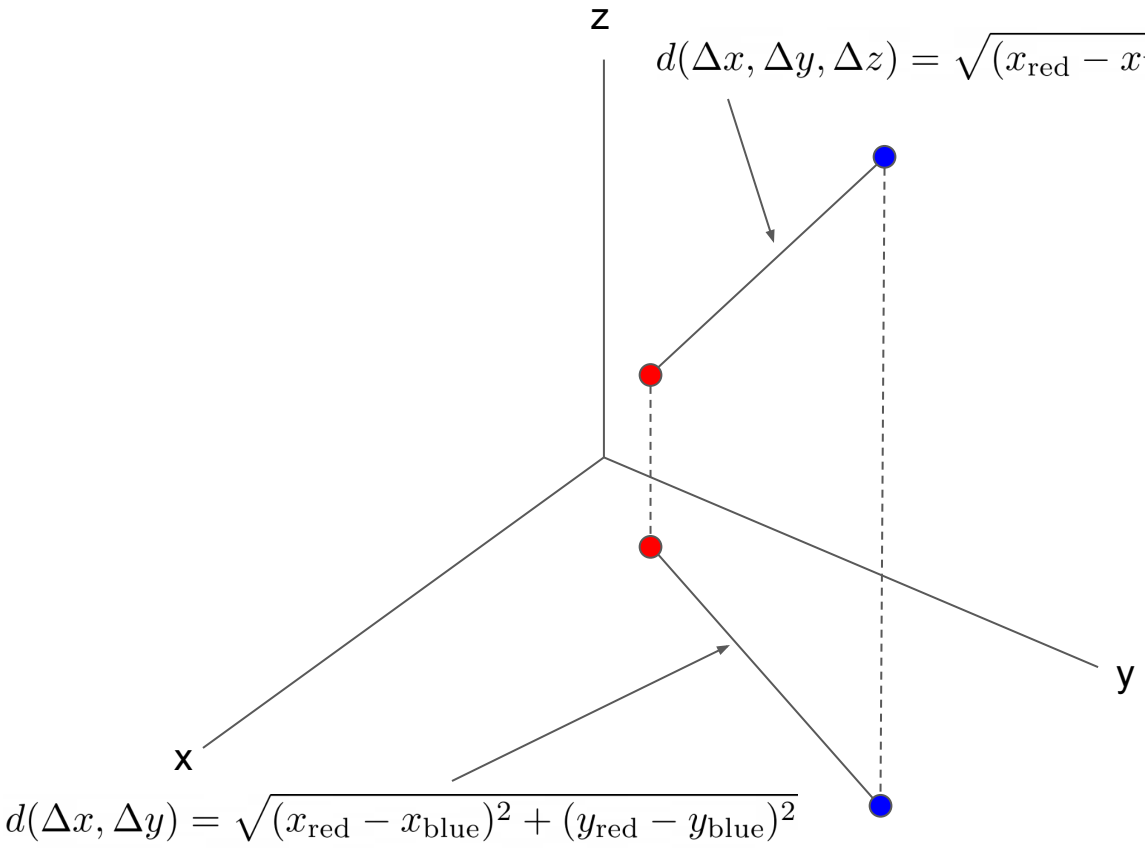
$$B = (x_2, y_2)$$

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

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

$$d = \|A - B\|_2$$

Euclidean Distance



$$d(\Delta x, \Delta y, \Delta z) = \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2 + (z_{\text{red}} - z_{\text{blue}})^2}$$

$$d(\Delta x, \Delta y) = \sqrt{(x_{\text{red}} - x_{\text{blue}})^2 + (y_{\text{red}} - y_{\text{blue}})^2}$$

Euclidean Distance

$$d(\Delta x) = \sqrt{(x_A - x_B)^2}$$

$$d(\Delta x, \Delta y) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

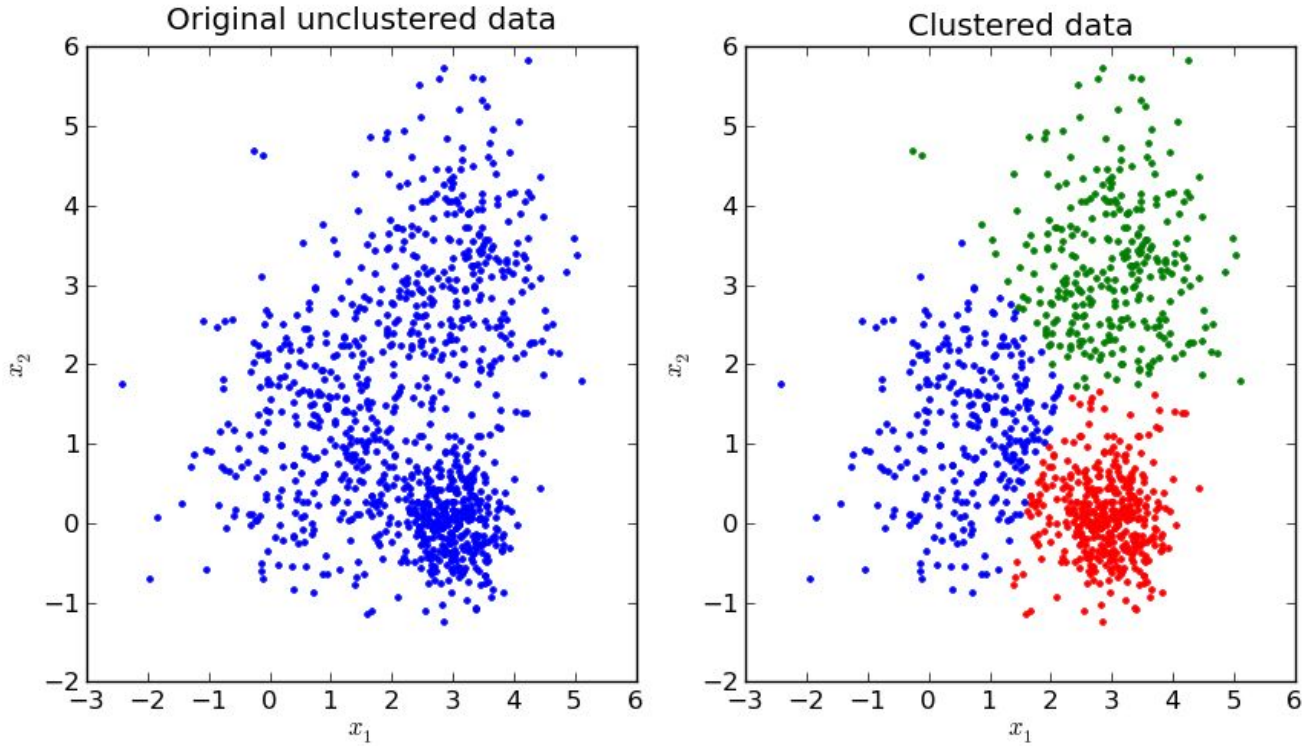
$$d(\Delta x, \Delta y, \Delta z) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$$

Euclidean Distance

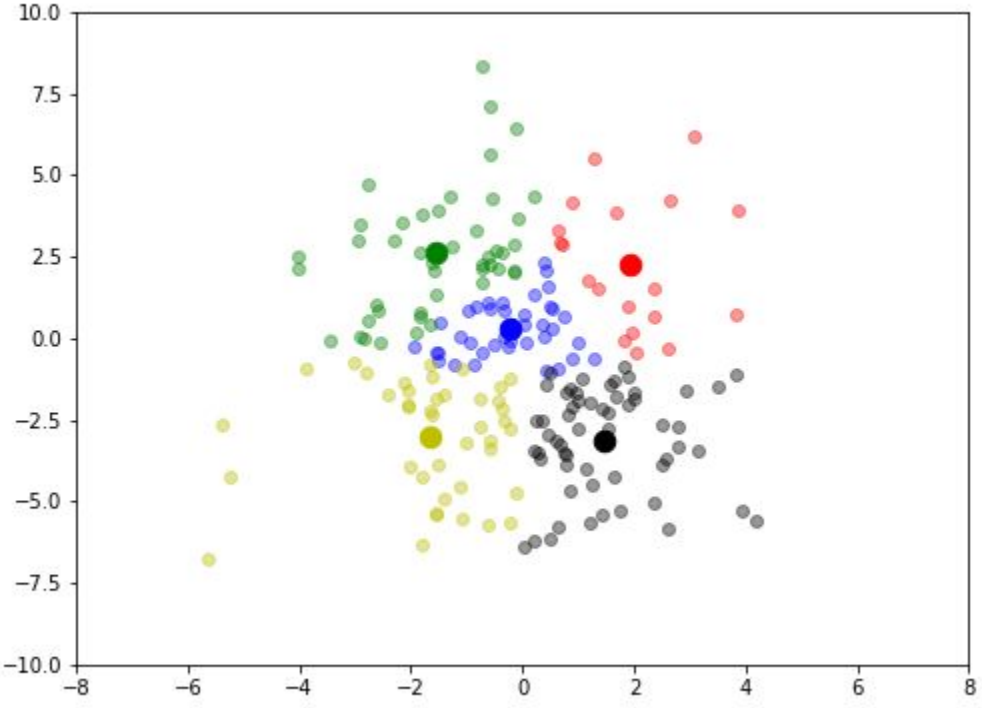
$$\boldsymbol{A} = [A_1, A_2, \ldots, A_n] \qquad \boldsymbol{B} = [B_1, B_2, \ldots, B_n]$$

$$\begin{aligned} \mathcal{D}(\boldsymbol{A}, \boldsymbol{B}) &= \sqrt{(A_1 - B_1)^2 + (A_2 - B_2)^2 + \ldots + (A_n - B_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} = \|\boldsymbol{A} - \boldsymbol{B}\| \end{aligned}$$

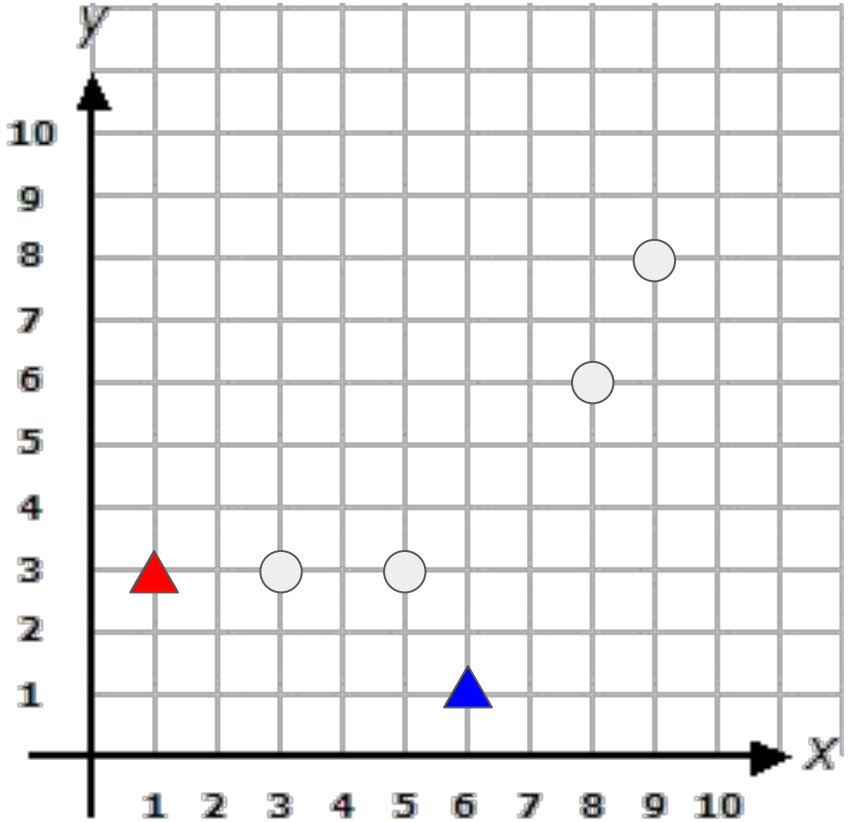
Clustering - 101



Clustering - 101



Example 1



$$x_1 = [3, 3]$$

$$x_2 = [5, 3]$$

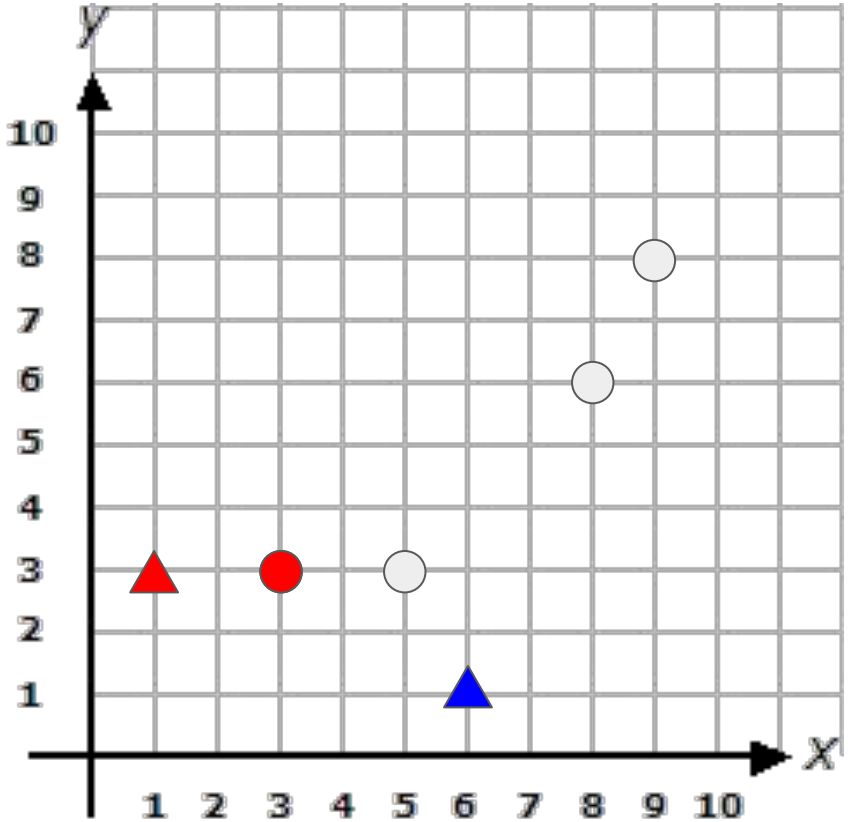
$$x_3 = [8, 6]$$

$$x_4 = [9, 8]$$

$$c_r = [1, 3]$$

$$c_b = [6, 1]$$

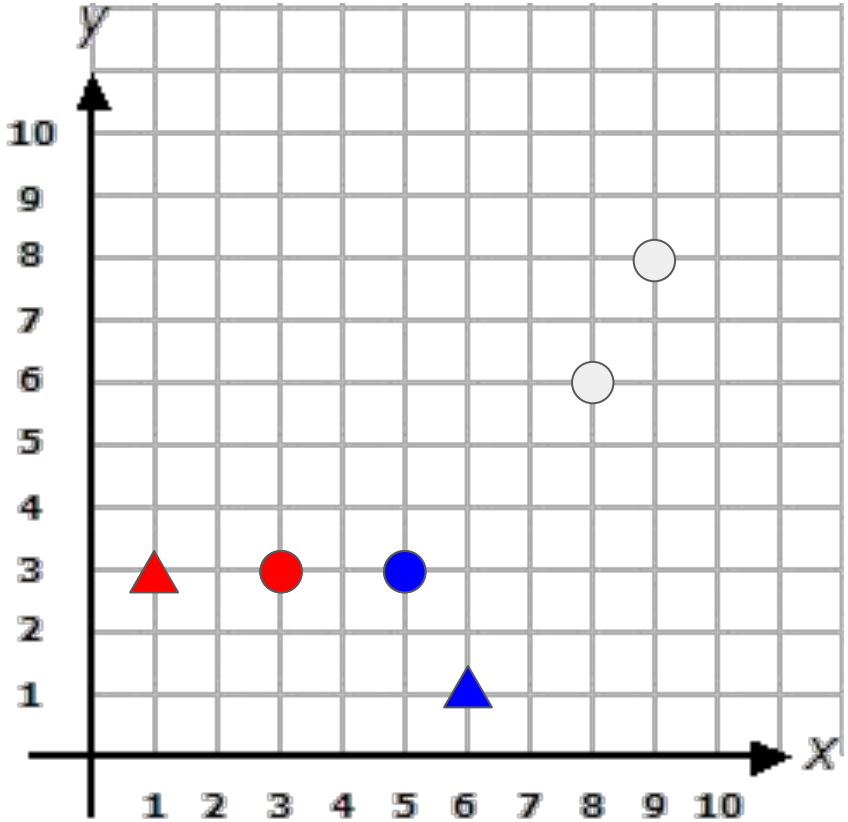
Example 1



$$d(x_1, c_r) = \sqrt{(1-3)^2 + (3-3)^2}$$

$$d(x_1, c_b) = \sqrt{(6-3)^2 + (1-3)^2}$$

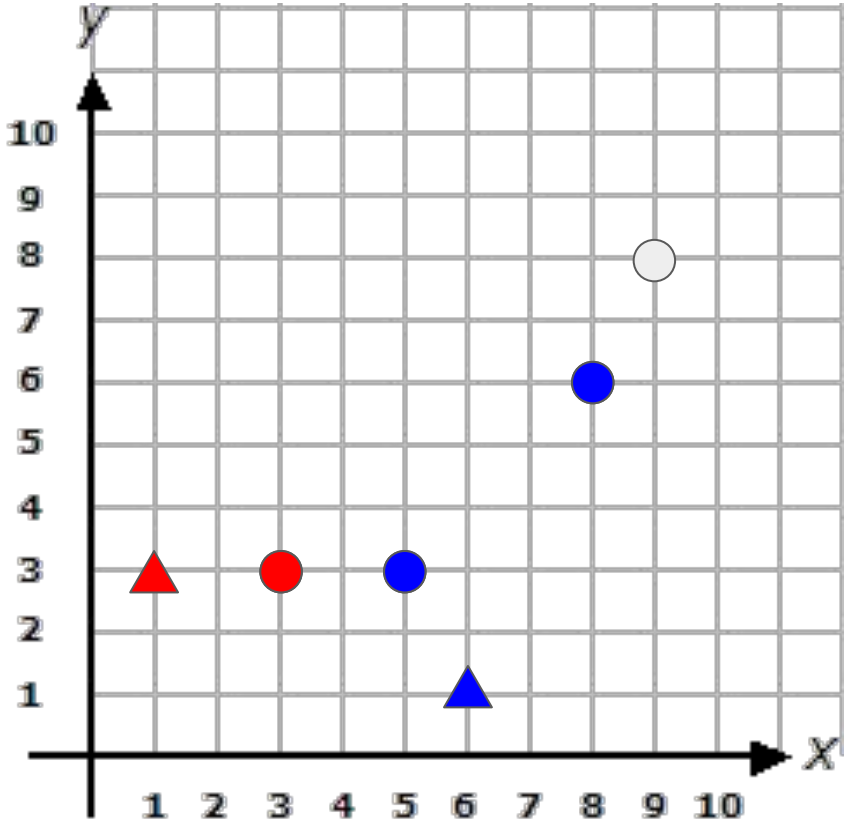
Example 1



$$d(x_2, c_r) = \sqrt{(1-5)^2 + (3-3)^2}$$

$$d(x_2, c_b) = \sqrt{(6-5)^2 + (1-3)^2}$$

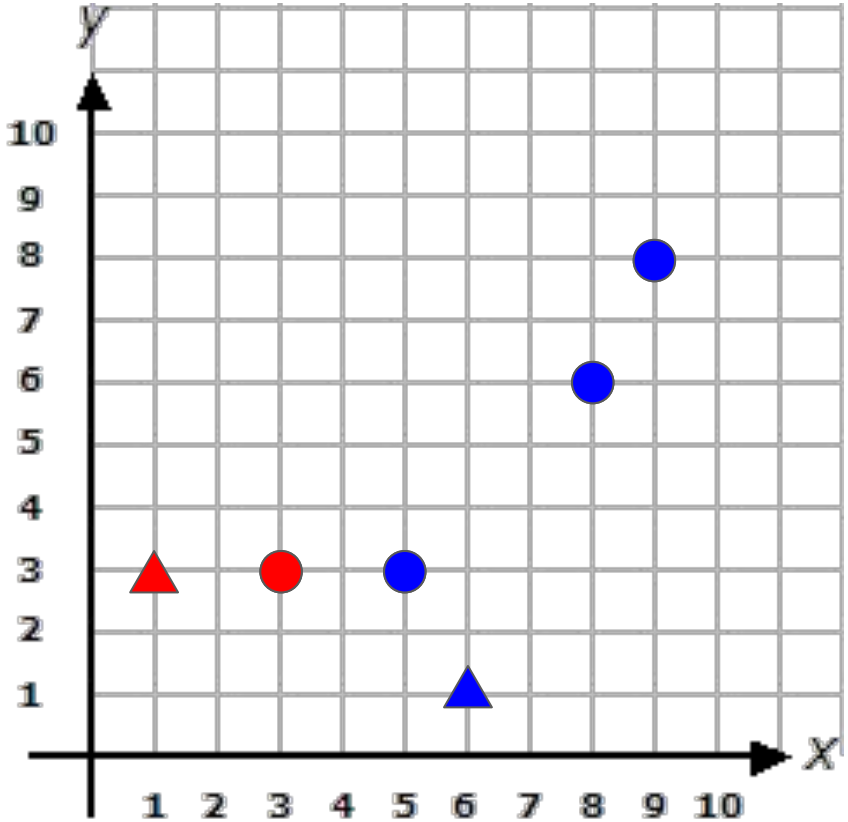
Example 1



$$d(x_3, c_r) = \sqrt{(1-8)^2 + (3-6)^2}$$

$$d(x_3, c_b) = \sqrt{(6-8)^2 + (1-6)^2}$$

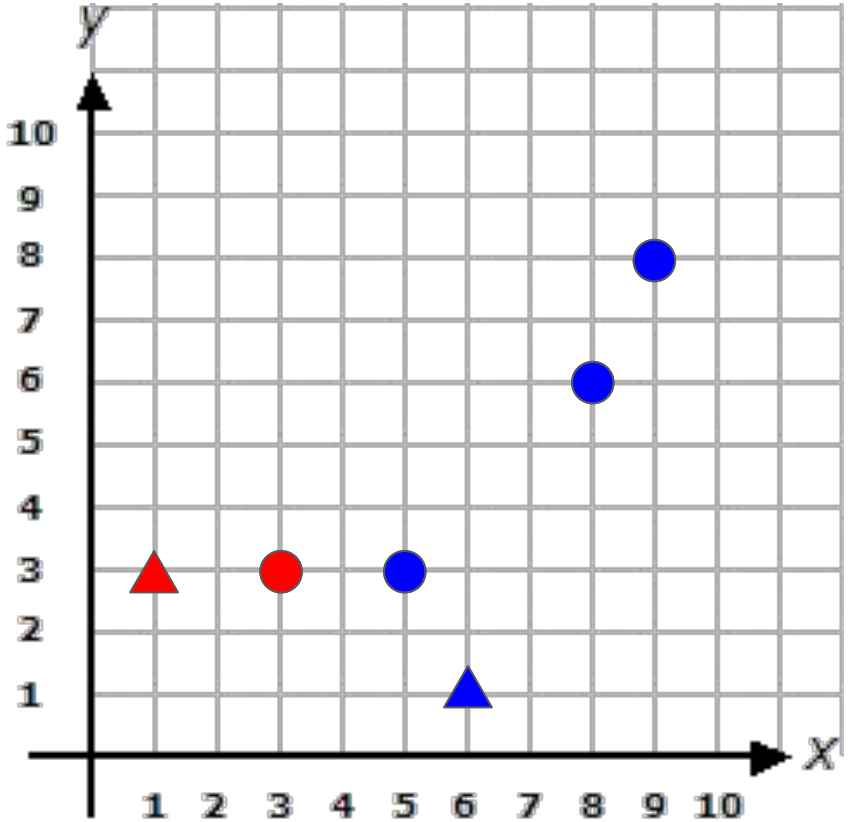
Example 1



$$d(x_4, c_r) = \sqrt{(1-9)^2 + (3-8)^2}$$

$$d(x_4, c_b) = \sqrt{(6-9)^2 + (1-8)^2}$$

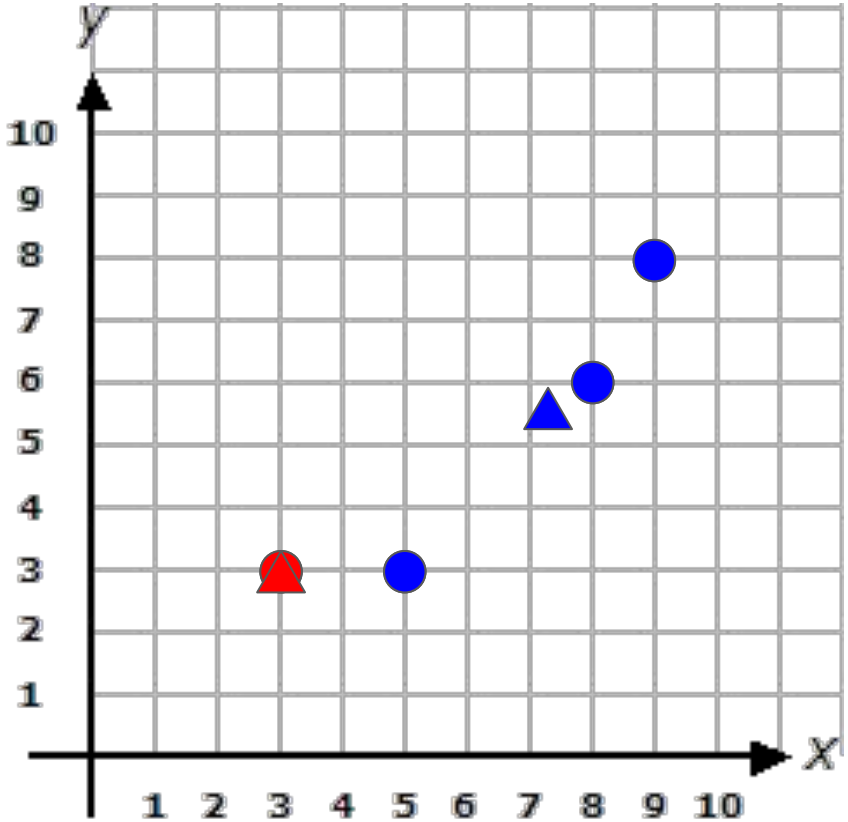
Example 1



$$c_r = \{x_1\}$$

$$c_b = \{x_2, x_3, x_4\}$$

Example 1



$$c_r = \{x_1\}$$

$$c_b = \{x_2, x_3, x_4\}$$

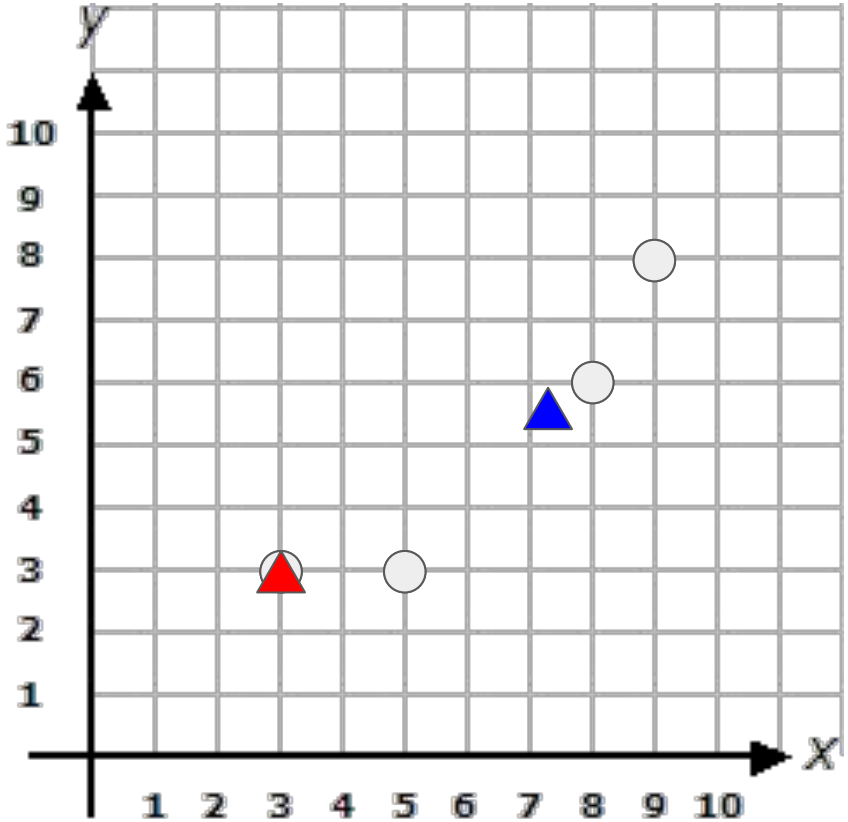
$$c_r = \left[\frac{3}{1}, \frac{3}{1}\right]$$

$$= [3, 3]$$

$$c_b = \left[\frac{5+8+9}{3}, \frac{3+6+8}{3}\right]$$

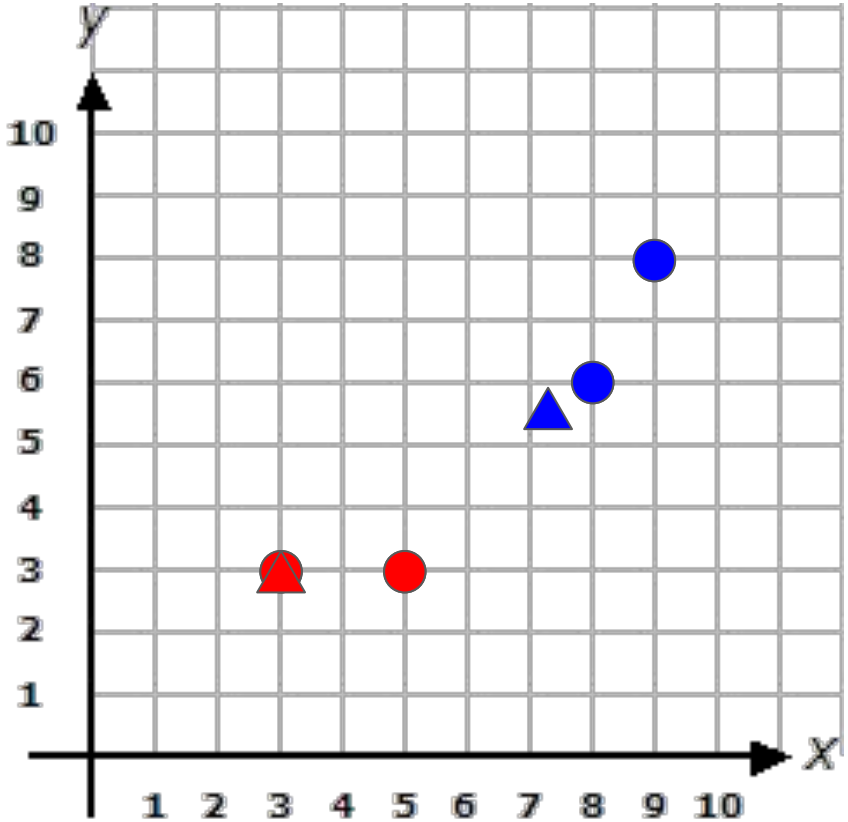
$$= [7.33, 5.66]$$

Example 1



$$d(x_1, c_r) = 0.00, d(x_1, c_b) = 5.08$$
$$d(x_2, c_r) = 2.00, d(x_2, c_b) = 3.54$$
$$d(x_3, c_r) = 5.83, d(x_3, c_b) = 0.75$$
$$d(x_4, c_r) = 7.81, d(x_4, c_b) = 2.87$$

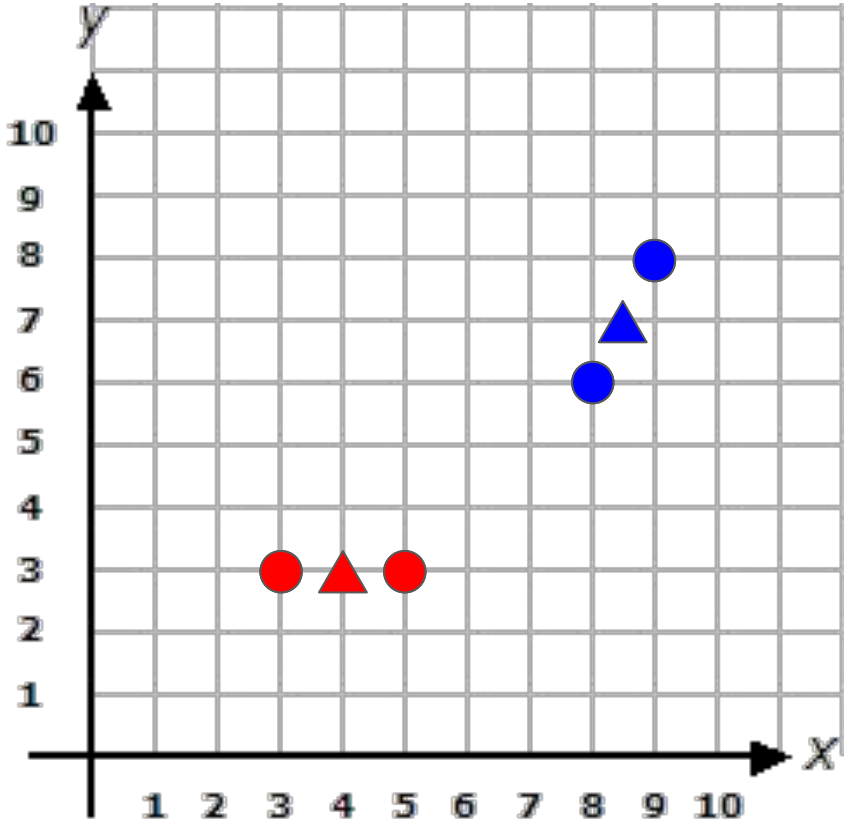
Example 1



$$c_r = \{x_1, x_2\}$$

$$c_b = \{x_3, x_4\}$$

Example 1



$$c_r = \{x_1, x_2\}$$

$$c_b = \{x_3, x_4\}$$

$$c_r = \left[\frac{3+5}{2}, \frac{3+3}{2}\right]$$

$$= [4, 3]$$

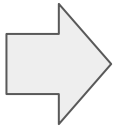
$$c_b = \left[\frac{8+9}{2}, \frac{6+8}{2}\right]$$

$$= [8.5, 7]$$

Example 2

$$x'_{j,i} = \frac{x_{j,i} - \min(X_i)}{\max(X_i) - \min(X_i)}$$

Customer	Age	Loan
A	25	40000
B	35	60000
C	45	80000
D	20	20000
E	52	18000
F	23	95000
G	40	62000
H	60	100000
I	48	220000
J	33	150000



Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

Age = [20-60]
Loan = [18000-220000]

Example 2

Centroid	Age	Loan
C1	0.255	0.225
C2	0.755	0.515

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

$$\sqrt{(0.125-0.255)^2+(0.109-0.225)^2}=0.174$$

$$\sqrt{(0.075-0.755)^2+(0.381-0.515)^2}=0.693$$

D(X,C1)	D(X,C2)
0.174	
	0.693

Example 2

Centroid	Age	Loan
C1	0.255	0.225
C2	0.755	0.515

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

D(X,C1)	D(X,C2)	Group
0.174	0.749	C1
0.121	0.489	C1
0.379	0.245	C2
0.334	0.908	C1
0.590	0.517	C2
0.238	0.693	C1
0.245	0.391	C1
0.767	0.268	C2
0.894	0.488	C2
0.434	0.452	C1

Example 2

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

Centroid	Age	Loan
C1	0.255	0.225
C2	0.755	0.515

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
D	0.000	0.010
F	0.075	0.381
G	0.500	0.218
J	0.325	0.653
new C1	0.233	0.263

Customer	Age	Loan
C	0.625	0.307
E	0.800	0.000
H	1.000	0.406
I	0.700	1.000
new C2	0.781	0.428

Example 2

Centroid	Age	Loan
C1	0.233	0.263
C2	0.781	0.428

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

D(X,C1)	D(X,C2)	Group
0.188	0.729	C1
0.152	0.462	C1
0.394	0.197	C2
0.344	0.886	C1
0.625	0.428	C2
0.197	0.708	C1
0.271	0.351	C1
0.780	0.220	C2
0.873	0.578	C2
0.401	0.508	C1

Example 2

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
C	0.625	0.307
D	0.000	0.010
E	0.800	0.000
F	0.075	0.381
G	0.500	0.218
H	1.000	0.406
I	0.700	1.000
J	0.325	0.653

Centroid	Age	Loan
C1	0.233	0.263
C2	0.781	0.428

Customer	Age	Loan
A	0.125	0.109
B	0.375	0.208
D	0.000	0.010
F	0.075	0.381
G	0.500	0.218
J	0.325	0.653
new C1	0.233	0.263

Customer	Age	Loan
C	0.625	0.307
E	0.800	0.000
H	1.000	0.406
I	0.700	1.000
new C2	0.781	0.428

Workshop

Given k is 3, iteration number is 2, and $D(.,.)$ is the Euclidean distance. Cluster the unlabeled dataset in the given table by using the k-means clustering algorithm. k initial centroids are chosen randomly from the input data.

sepal length	sepal width	petal length
5.1	3.5	1.4
4.9	3	1.4
4.7	3.2	1.3
4.6	3.1	1.5
7	3.2	4.7
6.4	3.2	4.5
6.9	3.1	4.9
5.5	2.3	4
6.5	2.8	4.6
6.3	3.3	6
5.8	2.7	5.1
7.1	3	5.9
6.3	2.9	5.6
6.5	3	5.8