**Ví dụ K-means:**

Giả sử ta có tập hợp các điểm sau trong không gian 2 chiều:

D = { (5,3), (10,15), (15,12), (24,10), (30,45), (85,70), (71,80), (60,78), (55,52), (80,91) }

Ta muốn chia bộ dữ liệu này thành 2 phân cụm, C1 và C2 dựa vào độ giống nhau giữa các điểm.

Đầu tiên ta chọn ngẫu nhiên các trọng tâm cho 2 phân cụm. Gọi c1, c2 lần lượt là trọng tâm của phân cụm C1 và C2. Chọn c1 = (5,3) và c2 = (10,15)

Bây giờ ta sẽ bắt đầu vòng lặp.

* **VÒNG LẶP 1:**

| **Điểm** | **Khoảng cách Euclid từ trọng tâm c1 = (5,3)** | **Khoảng cách Euclid từ trọng tâm c2 = (10,15)** | **Phân cụm** |
| --- | --- | --- | --- |
| (5,3) | 0 | 13 | C1 |
| (10,15) | 13 | 0 | C2 |
| (15,12) | 13.45 | 5.83 | C2 |
| (24,10) | 20.24 | 14.86 | C2 |
| (30,45) | 48.87 | 36 | C2 |
| (85,70) | 104.35 | 93 | C2 |
| (71,80) | 101.41 | 89 | C2 |
| (60,78) | 93 | 80 | C2 |
| (55,52) | 70 | 58 | C2 |
| (80,91) | 115.52 | 103.32 | C2 |

Dựa vào khoảng cách Euclid từ mỗi điểm đến 2 trọng tâm mà ta sẽ quyết định đưa điểm đó vào phân cụm nào. Ví dụ, khoảng cách từ điểm (15,12) đến c1 là 13.45 và khoảng cách từ điểm (15,12) đến c2 là 5.83 nên ta đưa điểm (15,12) vào phân cụm C2.

Sau đó, ta tìm các trọng tâm mới của 2 phân cụm bằng cách tính trung bình từng thành phần tọa độ của các điểm trong mỗi phân cụm.

C1 chỉ có 1 điểm (5,3) nên trọng tâm mới của C1 vẫn là (5,3).

C2 hiện có 9 điểm. Ta gọi các thành phần tọa độ của mỗi điểm là x và y.

Giá trị tọa độ x của trọng tâm c2 mới là trung bình các giá trị tọa độ x của các điểm thuộc C2 như sau:

c2(x) = (10 + 15 + 24 + 30 + 85 + 71 + 60 + 55 + 80) / 9 = 47.77

Giá trị tọa độ y của trọng tâm c2 mới là trung bình các giá trị tọa độ y của các điểm thuộc C2 như sau:

c2(y) = (15 + 12 + 10 + 45 + 70 + 80 + 78 + 52 + 91) / 9 = 50.33

Vậy trọng tâm mới của C2 là:  c2 = (47.77, 50.33).

* **VÒNG LẶP 2:**

| **Điểm** | **Khoảng cách Euclid từ trọng tâm c1 = (5,3)** | **Khoảng cách Euclid từ trọng tâm c2 = (47.77,50.33)** | **Phân cụm** |
| --- | --- | --- | --- |
| (5,3) | 0 | 63.79 | C1 |
| (10,15) | 13 | 51.71 | C1 |
| (15,12) | 13.45 | 50.42 | C1 |
| (24,10) | 20.24 | 46.81 | C1 |
| (30,45) | 48.87 | 18.55 | C2 |
| (85,70) | 104.35 | 42.10 | C2 |
| (71,80) | 101.41 | 37.68 | C2 |
| (60,78) | 93 | 30.25 | C2 |
| (55,52) | 70 | 7.42 | C2 |
| (80,91) | 115.52 | 51.89 | C2 |

Tương tự ở vòng lặp 1, ta tính được các trọng tâm mới như sau:

c1(x) = (5, 10, 15, 24) / 4 = 13.5

c1(y) = (3, 15, 12, 10) / 4 = 10.0

Vậy trọng tâm mới của C1 là c1 = (13.5, 10.0).

c2(x) = (30 + 85 + 71 + 60 + 55 + 80) / 6 = 63.5

c2(y) = (45 + 70 + 80 + 78 + 52 +91) / 6 = 69.33

Vậy trọng tâm mới của C2 là c2 = (63.5, 69.33).

* **VÒNG LẶP 3:**

| **Điểm** | **Khoảng cách Euclid từ trọng tâm c1= (13.5,10)** | **Khoảng cách Euclid từ trọng tâm c2= (63.5,69.33)** | **Phân cụm** |
| --- | --- | --- | --- |
| (5,3) | 11.01 | 88.44 | C1 |
| (10,15) | 6.10 | 76.24 | C1 |
| (15,12) | 2.5 | 75.09 | C1 |
| (24,10) | 10.5 | 71.27 | C1 |
| (30,45) | 38.69 | 41.40 | C1 |
| (85,70) | 93.33 | 21.51 | C2 |
| (71,80) | 90.58 | 13.04 | C2 |
| (60,78) | 82.37 | 9.34 | C2 |
| (55,52) | 59.04 | 19.30 | C2 |
| (80,91) | 104.80 | 27.23 | C2 |

Ta tính được các trọng tâm mới như sau:

c1(x) = (5, 10, 15, 24, 30) / 5 = 16.8

c1(y) = (3, 15, 12, 10, 45) / 5 = 17.0

Vậy trọng tâm mới của C1 là c1 = (16.8, 17.0).

c2(x) = (85 + 71 + 60 + 55 + 80) / 5 = 70.2

c2(y) = (70 + 80 + 78 + 52 + 91) / 5 = 74.2

Vậy trọng tâm mới của C2 là c2 = (70.2, 74.2).

* **VÒNG LẶP 4:**

| **Điểm** | **Khoảng cách Euclid từ trọng tâm c1 = (16.8,17)** | **Khoảng cách Euclid từ trọng tâm c2 = (70.2,74.2)** | **Phân cụm** |
| --- | --- | --- | --- |
| (5,3) | 18.30 | 96.54 | C1 |
| (10,15) | 7.08 | 84.43 | C1 |
| (15,12) | 5.31 | 83.16 | C1 |
| (24,10) | 10.04 | 79.09 | C1 |
| (30,45) | 30.95 | 49.68 | C1 |
| (85,70) | 86.37 | 15.38 | C2 |
| (71,80) | 83.10 | 5.85 | C2 |
| (60,78) | 74.74 | 10.88 | C2 |
| (55,52) | 51.80 | 26.90 | C2 |
| (80,91) | 97.31 | 19.44 | C2 |

Ta tính được các trọng tâm mới như sau:

c1(x) = (5, 10, 15, 24, 30) / 5 = 16.8

c1(y) = (3, 15, 12, 10, 45) / 5 = 17.0

Vậy trọng tâm mới của C1 là c1 = (16.8, 17.0).

c2(x) = (85 + 71 + 60 + 55 + 80) / 5 = 70.2

c2(y) = (70 + 80 + 78 + 52 + 91) / 5 = 74.2

Vậy trọng tâm mới của C2 là c2 = (70.2, 74.2).

Sau vòng lặp 4, ta thấy giá trị mới của c1, c2 không đổi so với sau vòng lặp 3. Điều này cho thấy bộ dữ liệu đã phân cụm xong.

**Kết luận**: c1 = (16.8, 17.0) và c2 = (70.2, 74.2) lần lượt là trọng tâm cho phân cụm C1 và C2.

: C1 gồm các điểm (5,3), (10,15), (15,12), (24,10), (30,45)

: C2 gồm các điểm (85,70), (71,80), (60,78), (55,52), (80,91)