**Họ Tên:** Hà Huy Sơn

**MSSV:** 18574802010055

**Buổi thực hành 5: Phân cụm dữ liệu (K-Means)**

**Câu 1**: Cho tập dữ liệu gồm các mẫu có hai thuộc tính như sau:

S1[5.9, 3.2], S2[4.6, 2.9], S3[6.2, 2.8], S4[4.7, 3.2], S5[5.5, 4.2], S6[5.0, 3.0], S7[4.9 3.1], S8[6.7, 3.1], S9[5.1, 3.8], S10[6.0 3.0].

Phân cụm K‐means với K = 3 và độ đo khoảng cách giữa các điểm là khoảng cách Euclid. Các tâm cụm khởi tạo ban đầu C1(6.2,3.2); C2(6.6, 3.7); C3(6.5, 3.0).

Thực hiện các thao tác tính toán thủ công và trình bày kết quả tâm cụm sau mỗi lần lặp.

**Trả lời**:

Gọi C = {C1, C2, C3}

- Ta có cetroid:

+ C1(6.2, 3.2)

+ C2(6.6, 3.7)

+ C3(6.5, 3.0)

- **Lần lặp 1**:

+ S1[5.9, 3.2]

dist(S1, C1) = = 0.3

dist(S1, C2) = = 0.86

dist(S1, C3) = = 0.632

* S1 C1

+ S2[4.6, 2.9]

dist(S2, C1) = = 1.627

dist(S2, C2) = = 2.154

dist(S2, C3) = = 1.902

* S2 C1

+ S3[6.2, 2.8]

dist(S3, C1) = = 0.4

dist(S3, C2) = = 0.984

dist(S3, C3) = = 0.36

* S3 C3

+ S4[4.7, 3.2]

dist(S4, C1) = = 1.5

dist(S4, C2) = = 1.96

dist(S4, C3) = = 1.81

* S4 C1

+ S5[5.5, 4.2]

dist(S5, C1) = = 1.22

dist(S5, C2) = = 1.208

dist(S5, C3) = = 1.562

* S5 C2

+ S6[5.0, 3.0]

dist(S6, C1) = = 1.216

dist(S6, C2) = = 1.746

dist(S6, C3) = = 1.5

* S6 C1

+ S7[4.9, 3.1]

dist(S7, C1) = = 1.303

dist(S7, C2) = = 1.802

dist(S7, C3) = = 1.603

* S7 C1

+ S8[6.7, 3.1]

dist(S8, C1) = = 0.51

dist(S8, C2) = = 0.61

dist(S8, C3) = = 0.223

* S8 C3

+ S9[5.1, 3.8]

dist(S9, C1) = = 1.252

dist(S9, C2) = = 1.503

dist(S9, C3) = = 1.612

* S9 C1

+ S10[6.0, 3.0]

dist(S10, C1) = = 0.282

dist(S10, C2) = = 0.922

dist(S10, C3) = = 0.5

* S10 C1

**Vậy**: Ta thu được 3 cụm:

+ C1 = {S1, S2, S4, S6, S7, S9, S10}

+ C2 = {S5}

+ C3 = {S3, S8}

Cập nhật lại trọng tâm cụm:

C1 = (, ) = (5.17, 3.17)

C2 = (5.5, 4.2)

C3 = () = (6.45, 2.95)

- **Lần lặp 2**:

+ S1[5.9, 3.2]

dist(S1, C1) = = 0.730

dist(S1, C2) = = 1.077

dist(S1, C3) = = 0.604

* S1 C3

+ S2[4.6, 2.9]

dist(S2, C1) = = 0.63

dist(S2, C2) = = 1.58

dist(S2, C3) = = 1.85

* S2 C1

+ S3[6.2, 2.8]

dist(S3, C1) = = 1.09

dist(S3, C2) = = 1.56

dist(S3, C3) = = 0.29

* S3 C3

+ S4[4.7, 3.2]

dist(S4, C1) = = 0.47

dist(S4, C2) = = 1.28

dist(S4, C3) = = 1.76

* S4 C1

+ S5[5.5, 4.2]

dist(S4, C1) = = 1.08

dist(S4, C2) = = 0

dist(S4, C3) = = 1.57

* S5 C2

+ S6[5.0, 3.0]

dist(S6, C1) = = 0.24

dist(S6, C2) = = 1.3

dist(S6, C3) = = 1.45

* S6 C1

+ S7[4.9, 3.1]

dist(S7, C1) = = 0.27

dist(S7, C2) = = 1.25

dist(S7, C3) = = 1.55

* S7 C1

+ S8[6.7, 3.1]

dist(S8, C1) = = 1.53

dist(S8, C2) = = 1.62

dist(S8, C3) = = 0.29

* S8 C3

+ S9[5.1, 3.8]

dist(S9, C1) = = 0.63

dist(S9, C2) = = 0.56

dist(S9, C3) = = 1.59

* S9 C2

+ S10[6.0, 3.0]

dist(S10, C1) = = 0.84

dist(S10, C2) = = 1.3

dist(S10, C3) = = 0.45

* S10 C3

**Vậy**: Sau bước lặp thứ 2 ta thu được 3 cụm:

+ C1 = {S2, S4, S6, S7}

+ C2 = {S5, S9}

+ C3 = {S1, S3, S8, S10}

Cập nhật lại trọng tâm cụm:

C1 = (, ) = (4.8, 3.05)

C2 = () = (5.3, 4)

C3 = () = (6.2, 3.025)

- **Lần lặp 3**:

**+** S1[5.9, 3.2]

dist(S1,C1) = = 1.11

dist(S1,C2) = = 1

dist(S1,C3) = = 0.35

* S1 C3

+ S2[4.6, 2.9]

dist(S2,C1) = = 0.25

dist(S2,C2) = = 1.30

dist(S2,C3) = = 1.61

* S2 C1

+ S3[6.2, 2.8]

dist(S3,C1) = = 1.422

dist(S3,C2) = = 1.5

dist(S3,C3) = = 0.225

* S3 C3

+ S4[4.7, 3.2]

dist(S4,C1) = = 0.18

dist(S4,C2) = = 1

dist(S4,C3) = = 1.51

* S4 C1

+ S5[5.5, 4.2]

dist(S5,C1) = = 1.34

dist(S5,C2) = = 0.28

dist(S5,C3) = = 1.36

* S5 C2

+ S6[5.0, 3.0]

dist(S6,C1) = = 0.21

dist(S6,C2) = = 1.044

dist(S6,C3) = = 1.2

* S6 C1

+ S7[4.9 3.1]

dist(S7,C1) = = 0.111

dist(S7,C1) = = 0.98

dist(S7,C1) = = 1.302

* S7 C1

+ S8[6.7, 3.1]

dist(S8,C1) = = 1.9

dist(S8,C2) = = 1.66

dist(S8,C3) = = 0.505

* S8 C3

+ S9[5.1, 3.8]

dist(S9,C1) = = 0.807

dist(S9,C2) = = 0.28

dist(S9,C3) = = 1.345

* S9 C2

+ S10[6.0 3.0].

dist(S10,C1) = = 1.201

dist(S10,C1) = = 1.22

dist(S10,C1) = = 0.201

* S10 C3

Vậy: Sau bước lặp 3 ta thu được 3 cụm:

+ C1 = {S2, S4, S6, S7}

+ C2 = {S5, S9}

+ C3 = {S1, S3, S8, S10}

**Nhận xét:**

**-** Kết quả phân cụm giữ nguyên sau 3 lần lặp,giải thuật dừng và cho kết quả phân cụm:

* C1 = {S2, S4, S6, S7}
* C2 = {S5, S9}
* C3 = {S1, S3, S8, S10}

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cụm 1** | **Cụm 2** | **Cụm 3** |
| Khởi tạo | C1(6.2,3.2) | C2(6.6, 3.7) | C3(6.5, 3.0) |
| 1 | C1 = {S1, S2, S4, S6, S7, S9, S10} | C2 = {S5} | C3 = {S3, S8} |
| 2 | C1 = {S2, S4, S6, S7} | C2 = {S5, S9} | C3 = {S1, S3, S8, S10} |
| 3 | C3 = {S1, S3, S8, S10} | C3 = {S1, S3, S8, S10} | C3 = {S1, S3, S8, S10} |
| Kết thúc | C3 = {S1, S3, S8, S10} | C3 = {S1, S3, S8, S10} | C3 = {S1, S3, S8, S10} |