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) = $\sqrt{(6.2-5.9)^2 + (3.2-3.2)^2}$ = 0.3

dist(S1, C2) = $\sqrt{(6.6-5.9)^2 + (3.7-3.2)^2}$ = 0.86

dist(S1, C3) = $\sqrt{(6.5-5.9)^2 + (3.0-3.2)^2}$ = 0.632

⇨ S1 $\in$ C1

+ S2[4.6, 2.9]

dist(S2, C1) = $\sqrt{(6.2-4.6)^2 + (3.2-2.9)^2}$ = 1.627

dist(S2, C2) = $\sqrt{(6.6-4.6)^2 + (3.7-2.9)^2}$ = 2.154

dist(S2, C3) = $\sqrt{(6.5-4.6)^2 + (3.0-2.9)^2}$ = 1.902

⇨ S2 $\in$ C1

+ S3[6.2, 2.8]

dist(S3, C1) = $\sqrt{(6.2-6.2)^2 + (3.2-2.8)^2}$ = 0.4

dist(S3, C2) = $\sqrt{(6.6-6.2)^2 + (3.7-2.8)^2}$ = 0.984

dist(S3, C3) = $\sqrt{(6.5-6.2)^2 + (3.0-2.8)^2}$ = 0.36

⇨ S3 $\in$ C3

+ S4[4.7, 3.2]

dist(S4, C1) = $\sqrt{(6.2-4.7)^2+\ (3.2-3.2)^2}$ = 1.5

dist(S4, C2) = $\sqrt{(6.6-4.7)^2+\ (3.7-3.2)^2}$ = 1.96

dist(S4, C3) = $\sqrt{(6.5-4.7)^2+\ (3.0-3.2)^2}$ = 1.81

- ⇨ S4 ∈ C1

+ S5[5.5, 4.2]

dist(S5, C1) = $\sqrt{(6.2-5.5)^2+\ (3.2-4.2)^2}$ = 1.22

dist(S5, C2) = $\sqrt{(6.6-5.5)^2+\ (3.7-4.2)^2}$ = 1.208

dist(S5, C3) = $\sqrt{(6.5-5.5)^2+\ (3.0-4.2)^2}$ = 1.562

- ⇨ S5 ∈ C2

+ S6[5.0, 3.0]

dist(S6, C1) = $\sqrt{(6.2-5.0)^2+\ (3.2-3.0)^2}$ = 1.216

dist(S6, C2) = $\sqrt{(6.6-5.0)^2+\ (3.7-3.0)^2}$ = 1.746

dist(S6, C3) = $\sqrt{(6.5-5.0)^2+\ (3.0-3.0)^2}$ = 1.5

- ⇨ S6 ∈ C1

+ S7[4.9, 3.1]

dist(S7, C1) = $\sqrt{(6.2-4.9)^2+\ (3.2-3.1)^2}$ = 1.303

dist(S7, C2) = $\sqrt{(6.6-4.9)^2+\ (3.7-3.1)^2}$ = 1.802

dist(S7, C3) = $\sqrt{(6.5-4.9)^2+\ (3.0-3.1)^2}$ = 1.603

- ⇨ S7 ∈ C1

+ S8[6.7, 3.1]

dist(S8, C1) = $\sqrt{(6.2-6.7)^2+\ (3.2-3.1)^2}$ = 0.51

dist(S8, C2) = $\sqrt{(6.6-6.7)^2+\ (3.7-3.1)^2}$ = 0.61

dist(S8, C3) = $\sqrt{(6.5-6.7)^2+\ (3.0-3.1)^2}$ = 0.223

- ⇨ S8 ∈ C3

+ S9[5.1, 3.8]

dist(S9, C1) = $\sqrt{(6.2-5.1)^2+\ (3.2-3.8)^2}$ = 1.252

dist(S9, C2) = $\sqrt{(6.6-5.1)^2 + (3.7-3.8)^2}$ = 1.503

dist(S9, C3) = $\sqrt{(6.5-5.1)^2 + (3.0-3.8)^2}$ = 1.612

⇨ S9 ∈ C1

+ S10[6.0, 3.0]

dist(S10, C1) = $\sqrt{(6.2-6.0\,)^2 + (3.2-3.0)^2}$ = 0.282

dist(S10, C2) = $\sqrt{(6.6-6.0\,)^2 + (3.7-3.0)^2}$ = 0.922

dist(S10, C3) = $\sqrt{(6.5-6.0\,)^2 + (3.0-3.0)^2}$ = 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 = ($\frac{5.9+4.6+4.7+5.0+4.9+5.1+6.0}{7}$, $\frac{3.2+2.9+3.2+3.0+3.1+3.8+3.0}{7}$) = (5.17, 3.17)

C2 = (5.5, 4.2)

C3 = ($\frac{6.2+6.7}{2}$, $\frac{2.8+3.1}{2}$) = (6.45, 2.95)

- Lần lặp 2:

+ S1[5.9, 3.2]

dist(S1, C1) = $\sqrt{(5.17-5.9)^2 + (3.17-3.2)^2}$ = 0.730

dist(S1, C2) = $\sqrt{(5.5-5.9)^2 + (4.2-3.2)^2}$ = 1.077

dist(S1, C3) = $\sqrt{(6.45-5.9)^2 + (2.95-3.2)^2}$ = 0.604

⇨ S1 ∈ C3

+ S2[4.6, 2.9]

dist(S2, C1) = $\sqrt{(5.17-4.6)^2 + (3.17-2.9)^2}$ = 0.63

dist(S2, C2) = $\sqrt{(5.5-4.6)^2 + (4.2-2.9)^2}$ = 1.58

dist(S2, C3) = $\sqrt{(6.45-4.6)^2 + (2.95-2.9)^2}$ = 1.85

⇨ S2 ∈ C1

+ S3[6.2, 2.8]

dist(S3, C1) = $\sqrt{(5.17-6.2)^2 + \ (3.17-2.8)^2}$ = 1.09

dist(S3, C2) = $\sqrt{(5.5-6.2)^2 + \ (4.2-2.8)^2}$ = 1.56

dist(S3, C3) = $\sqrt{(6.45-6.2)^2 + \ (2.95-2.8)^2}$ = 0.29

⇨ S3 ∈ C3

+ S4[4.7, 3.2]

dist(S4, C1) = $\sqrt{(5.17-4.7)^2 + \ (3.17-3.2)^2}$ = 0.47

dist(S4, C2) = $\sqrt{(5.5-4.7)^2 + \ (4.2-3.2)^2}$ = 1.28

dist(S4, C3) = $\sqrt{(6.45-4.7)^2 + \ (2.95-3.2)^2}$ = 1.76

⇨ S4 ∈ C1

+ S5[5.5, 4.2]

dist(S4, C1) = $\sqrt{(5.17-5.5)^2 + \ (3.17-4.2)^2}$ = 1.08

dist(S4, C2) = $\sqrt{(5.5-5.5)^2 + \ (4.2-4.2)^2}$ = 0

dist(S4, C3) = $\sqrt{(6.45-5.5)^2 + \ (2.95-4.2)^2}$ = 1.57

⇨ S5 ∈ C2

+ S6[5.0, 3.0]

dist(S6, C1) = $\sqrt{(5.17-5.0)^2 + \ (3.17-3.0)^2}$ = 0.24

dist(S6, C2) = $\sqrt{(5.5-5.0)^2 + \ (4.2-3.0)^2}$ = 1.3

dist(S6, C3) = $\sqrt{(6.45-5.0)^2 + \ (2.95-3.0)^2}$ = 1.45

⇨ S6 ∈ C1

+ S7[4.9, 3.1]

dist(S7, C1) = $\sqrt{(5.17-4.9)^2 + \ (3.17-3.1)^2}$ = 0.27

dist(S7, C2) = $\sqrt{(5.5-4.9)^2 + \ (4.2-3.1)^2}$ = 1.25

dist(S7, C3) = $\sqrt{(6.45-4.9)^2 + \ (2.95-3.1)^2}$ = 1.55

⇨ S7 ∈ C1

+ S8[6.7, 3.1]

dist(S8, C1) = $\sqrt{(5.17-6.7)^2+\ (3.17-3.1)^2}$ = 1.53

dist(S8, C2) = $\sqrt{(5.5-6.7)^2+\ (4.2-3.1)^2}$ = 1.62

dist(S8, C3) = $\sqrt{(6.45-6.7)^2+\ (2.95-3.1)^2}$ = 0.29

⇨ S8 ∈ C3

+ S9[5.1, 3.8]

dist(S9, C1) = $\sqrt{(5.17-5.1)^2+\ (3.17-3.8)^2}$ = 0.63

dist(S9, C2) = $\sqrt{(5.5-5.1)^2+\ (4.2-3.8)^2}$ = 0.56

dist(S9, C3) = $\sqrt{(6.45-5.1)^2+\ (2.95-3.8)^2}$ = 1.59

⇨ S9 ∈ C2

+ S10[6.0, 3.0]

dist(S10, C1) = $\sqrt{(5.17-6.0)^2+\ (3.17-3.0)^2}$ = 0.84

dist(S10, C2) = $\sqrt{(5.5-6.0)^2+\ (4.2-3.0)^2}$ = 1.3

dist(S10, C3) = $\sqrt{(6.45-6.0)^2+\ (2.95-3.0)^2}$ = 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 = ($\frac{4.6+4.7+5.0+4.9}{4}$, $\frac{2.9+3.2+3.0+3.1}{4}$) = (4.8, 3.05)

C2 = ($\frac{5.5+5.1}{2}$, $\frac{4.2+3.8}{2}$) = (5.3, 4)

C3 = ($\frac{5.9+6.2+6.7+6.0}{4}$, $\frac{3.2+2.8+3.1+3.0}{4}$) = (6.2, 3.025)

- Lần lặp 3:

+ S1[5.9, 3.2]

dist(S1,C1) = $\sqrt{(4.8-5.9)^2+\ (3.05-3.2)^2}$ = 1.11

dist(S1,C2) = $\sqrt{(5.3-5.9)^2+\ (4-3.2)^2}$ = 1 dist(S1,C3) = $\sqrt{(6.2-5.9)^2+\ (3.025-3.2)^2}$ = 0.35

⇨ S1 ∈ C3

+ S2[4.6, 2.9]

dist(S2,C1) = $\sqrt{(4.8-4.6)^2+\ (3.05-2.9)^2}$ = 0.25

dist(S2,C2) = $\sqrt{(5.3-4.6)^2+\ (4-2.9)^2}$ = 1.30

dist(S2,C3) = $\sqrt{(6.2-4.6)^2+\ (3.025-2.9)^2}$ = 1.61

⇨ S2 ∈ C1

+ S3[6.2, 2.8]

dist(S3,C1) = $\sqrt{(4.8-6.2)^2+\ (3.05-2.8)^2}$ = 1.422

dist(S3,C2) = $\sqrt{(5.3-6.2)^2+\ (4-2.8)^2}$ = 1.5

dist(S3,C3) = $\sqrt{(6.2-6.2)^2+\ (3.025-2.8)^2}$ = 0.225

⇨ S3 ∈ C3

+ S4[4.7, 3.2]

dist(S4,C1) = $\sqrt{(4.8-4.7)^2+\ (3.05-3.2)^2}$ = 0.18

dist(S4,C2) = $\sqrt{(5.3-4.7)^2+\ (4-3.2)^2}$ = 1

dist(S4,C3) = $\sqrt{(6.2-4.7)^2+\ (3.025-3.2)^2}$ = 1.51

⇨ S4 ∈ C1

+ S5[5.5, 4.2]

dist(S5,C1) = $\sqrt{(4.8-5.5)^2+\ (3.05-4.2)^2}$ = 1.34

dist(S5,C2) = $\sqrt{(5.3-5.5)^2+\ (4-4.2)^2}$ = 0.28

dist(S5,C3) = $\sqrt{(6.2-5.5)^2+\ (3.025-4.2)^2}$ = 1.36

⇨ S5 ∈ C2

+ S6[5.0, 3.0]

dist(S6,C1) = $\sqrt{(4.8-5.0)^2+\ (3.05-3.0)^2}$ = 0.21

dist(S6,C2) = $\sqrt{(5.3-5.0)^2+\ (4-3.0)^2}$ = 1.044

dist(S6,C3) = $\sqrt{(6.2-5.0)^2+\ (3.025-3.0)^2}$ = 1.2

⇨ S6 ∈ C1

+ S7[4.9 3.1]

dist(S7,C1) = $\sqrt{(4.8-4.9)^2 + (3.05-3.1)^2}$ = 0.111

dist(S7,C1) = $\sqrt{(5.3-4.9)^2 + (4-3.1)^2}$ = 0.98

dist(S7,C1) = $\sqrt{(6.2-4.9)^2 + (3.025-3.1)^2}$ = 1.302

- ⇨ S7 ∈ C1

+ S8[6.7, 3.1]

dist(S8,C1) = $\sqrt{(4.8-6.7)^2 + (3.05-3.1)^2}$ = 1.9

dist(S8,C2) = $\sqrt{(5.3-6.7)^2 + (4-3.1)^2}$ = 1.66

dist(S8,C3) = $\sqrt{(6.2-6.7)^2 + (3.025-3.1)^2}$ = 0.505

- ⇨ S8 ∈ C3

+ S9[5.1, 3.8]

dist(S9,C1) = $\sqrt{(4.8-5.1)^2 + (3.05-3.8)^2}$ = 0.807

dist(S9,C2) = $\sqrt{(5.3-5.1)^2 + (4-3.8)^2}$ = 0.28

dist(S9,C3) = $\sqrt{(6.2-5.1)^2 + (3.025-3.8)^2}$ = 1.345

- ⇨ S9 ∈ C2

+ S10[6.0 3.0].

dist(S10,C1) = $\sqrt{(4.8-6.0)^2 + (3.05-3.0)^2}$ = 1.201

dist(S10,C1) = $\sqrt{(5.3-6.0)^2 + (4-3.0)^2}$ = 1.22

dist(S10,C1) = $\sqrt{(6.2-6.0)^2 + (3.025-3.0)^2}$ = 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}