

BÀI TẬP VỀ PHÂN CỤM DỮ LIỆU

Bài 1. Cho 08 điểm dữ liệu:

A(1,2), B(0,3), C(3,1), D(4,2), E(4,0), F(2,2), G(3,3), H(1,4). Số cụm $k = 2$.

Áp dụng giải thuật phân cụm (k-mean, pam, clara) để phân cụm 8 điểm trên với tâm khởi tạo là A, B (thực hiện với 2 lần lặp).

Bài 2. Cho 08 điểm dữ liệu:

A(1,2), B(0,3), C(3,1), D(4,2), E(4,0), F(2,2), G(3,3), H(1,4). Số cụm $k = 2$.

Áp dụng giải thuật phân cụm (k-mean, pam, clara) để phân cụm 8 điểm trên với tâm khởi tạo ngẫu nhiên (làm sau 2 lần lặp).

Bài 3. Cho 05 điểm dữ liệu:

A(1, 0.1, 0.41, 0.55, 0.35), B(0.1, 1, 0.64, 0.47, 0.98), C(0.41, 0.64, 1, 0.44, 0.85), D(0.55, 0.47, 0.44, 1, 0.76), E(0.35, 0.98, 0.85, 0.76, 1). Số cụm $k = 2$.

Áp dụng giải thuật phân cụm (k-mean, pam, clara) phân cụm 5 điểm trên với tâm khởi tạo ngẫu nhiên (làm sau 2 lần lặp).

Bài 4. Cho bảng dữ liệu sau:

Masv	Toan1	Toan2	Toan3	Toan4
1	7	9	5	9
2	5	4	7	8
3	9	6	6	8
4	6	5	8	6
5	2	7	7	8
6	4	6	4	5
7	6	9	3	4
8	6	4	7	3

Áp dụng giải thuật phân cụm (k-mean, pam, clara) phân cụm với 8 sinh viên trên với tâm khởi tạo ngẫu nhiên (làm sau 2 lần lặp).

Bài 5. Một xe đón khách về bến xe Mỹ Đình của công ty Hà Lan muốn đón n khách hàng. Do thời gian đón khách ít nên công ty muốn gom khách về k địa điểm để tiện việc đón. Giả sử $n = 5$ và $k = 2$. Tọa độ của 5 khách hàng đang ở các vị trí A(1,1), B(3,1), C(3,3), D(4,2),

E(1,3). Anh/chị hãy cho biết nên hẹn khách nào tại địa điểm nào để việc đưa đón là thuận tiện nhất. Cho biết tọa độ của 2 địa điểm cần đón khách? Giả sử độ đo khoảng cách được sử dụng là độ đo Euclidean.

Bài 6. Sinh viên làm thực hành trên các bộ dữ liệu lớn lấy từ UCI:

“<https://archive.ics.uci.edu/ml/datasets.html>”