Đại học ABC từ năm 1976 đến nay đã sưu tập và lưu giữ hầu hết các giống lúa mùa cổ truyền của vùng ĐBSCL. Số lượng cụ thể là hơn 1.900 mẫu. Mỗi một giống lúa trong ngân hàng thông tin được mô tả bởi 69 đặc điểm hình thái, đặc tính nông sinh học (ví dụ: góc lá đòng, độ cứng cây, chiều cao cây, thời gian sinh trưởng, số bông hữu hiệu/khóm, số hạt chắc, lép/bông, khối lượng nghìn hạt, năng suất lý thuyết - theo thang điểm của International Rice Research Institute). Chúng ta cần tạo ra các công cụ tin học giúp cho các nhà nghiên cứu có thể phát hiện ra các mẫu lúa cùng nhóm với nhau. Giải thuật máy học nào sau đây phù hợp cho bài toán này?

Với dữ liệu là điểm môn học của 5 sinh viên như hình bên, anh/chị hãy sử dụng phương pháp Hierarchical agglomerative clustering để xây dựng biểu đồ Dendrogram, khoảng cách Euclid được sử dụng để đo khoảng cách giữa các đối tượng và phương Single Link để đo khoảng cách giữa các cluster

Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

Bước 1: Xây dựng ma trận khoảng cách

ID	1	2	3	4	5
1					
2					
3					
4					
5					

Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

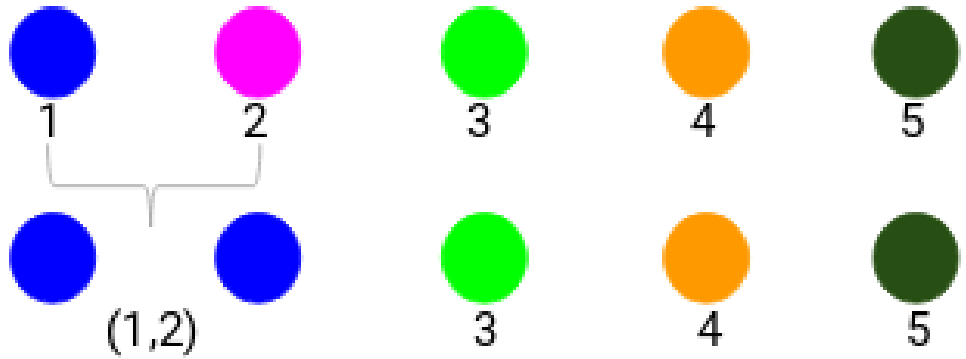
Bước 1: Xây dựng ma trận khoảng cách

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Bước 2: Gom nhóm đầu tiên. Tính lại ma trận khoảng cách

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0



Thực hiện tương tự bước 2 cho đến khi gom được tất cả các nhóm về thành 1 nhóm lớn

Cho 5 điểm dữ liệu sau, anh/chị hãy sử dụng phương pháp Hierarchical agglomerative clustering để xây dựng biểu đồ Dendrogram, khoảng cách Euclid được sử dụng để đo khoảng cách giữa các đối tượng và phương Single Link để đo khoảng cách giữa các cluster

Point	a	b
P1	0.07	0.83
P2	0.85	0.14
P3	0.66	0.89
P4	0.49	0.64
P5	0.80	0.46

	P1	P2	P3	P4	P5
P1	0				
P2	1.04139	0			
P3	0.59304	0.77369	0		
P4	0.46098	0.61612	0.30232	0	
P5	0.81841	0.32388	0.45222	0.35847	

Point	a	b
P1	0.07	0.83
P2	0.85	0.14
P3	0.66	0.89
P4	0.49	0.64
P5	0.80	0.46

	P1	P2	P3	P4	P5
P1	0				
P2	1.04139	0			
P3	0.59304	0.77369	0		
P4	0.46098	0.61612	0.30232	0	
P5	0.81841	0.32388	0.45222	0.35847	0

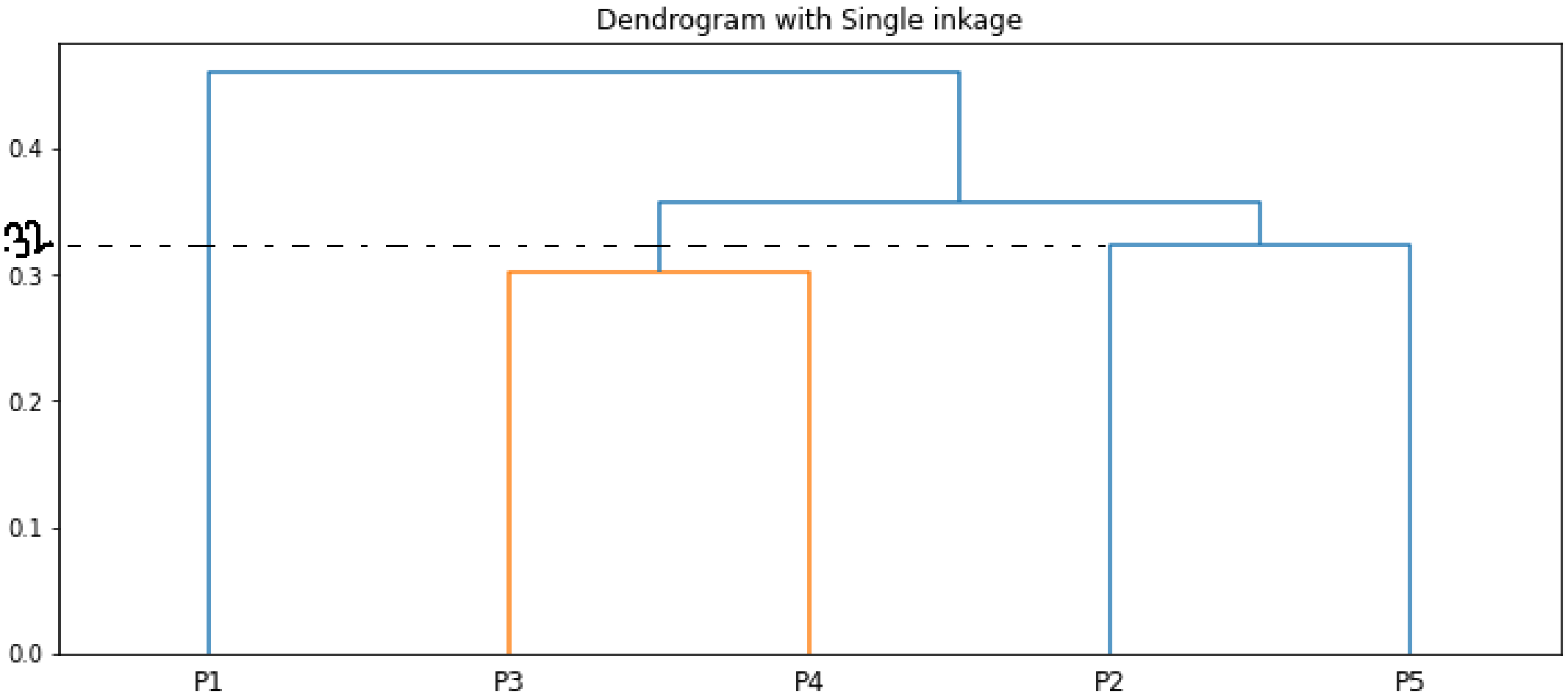
	P1	P2	P3,P4	P5
P1	0			
P2	1.04139	0		
P3,P4	0.46098	0.61612	0	
P5	0.81841	0.32388	0.35847	0

	P1	P2	P3,P4	P5
P1	0			
P2	1.04139	0		
P3,P4	0.46098	0.61612	0	
P5	0.81841	0.32388	0.35847	0

	P1	P2,P5	P3,P4
P1	0		
P2,P5	0.81841	0	
P3,P4	0.46098	0.35847	0

	P1	P2,P5	P3,P4
P1	0		
P2,P5	0.81841	0	
P3,P4	0.46098	0.35847	0

	P1	P2,P5,P3,P4
P1	0	
P2,P5,P3,P4	0.46098	0


Dendrogram with Single inkage
0.4
0.3
0.2
0.1
0.0
P1
P3
P4
P2
P5

Các anh chị làm bài tập sau trong 10 phút

Cho tập dữ liệu O gồm có 06 đối tượng. Mỗi đối tượng có 2 thuộc tính x1, x2 như sau

	O1	**O2**	**O3**	**O4**	**O5**	**O6**
X1	70	40	20	0	90	60
X2	30	50	40	10	70	80

Xét 2 cụm dữ liệu với 2 điểm khởi tạo là C1 = (90, 70) và C2 = (60, 80). Áp dụng giải thuật K-means, anh / chị hãy cho biết các đối tượng và tâm của 2 cụm dữ liệu khi giải thuật kết thúc. Khoảng cách Euclidea được sử dụng để đo khoảng cách giữa các đối tượng.

C1 = (90, 70) và C2 = (60, 80).

	X1	X2	Bình phương khoảng cách đến C1	Bình phương khoảng cách đến C2	Cluster? (1 hay 2)
O1	70	30			
O2	40	50			
O3	20	40			
O4	0	10			
O5	90	70			
O6	60	80			

Khoảng cách Euclidea

$$d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$$

Vòng lặp 1 C1 = (90, 70) và C2 = (60, 80).

	X1	X2	Bình phương khoảng cách đến C1	Bình phương khoảng cách đến C2	Cluster? (1 hay 2)
O1	70	30	2,000	2,600	1
O2	40	50	2,900	1,300	2
O3	20	40	5,800	3,200	2
O4	0	10	11,700	8,500	2
O5	90	70	-	1,000	1
O6	60	80	1,000	-	2

Tâm mới

	X1	X2
C1	80	50
C2	30	45

Vòng lặp 2			C1 = (80, 50)	C2 = (30, 45)	
	X1	X2	Bình phương khoảng cách đến C1	Bình phương khoảng cách đến C2	Cluster? (1 hay 2)
O1	70	30	500	1,825	1
O2	40	50	1,600	125	2
O3	20	40	3,700	125	2
O4	0	10	8,000	2,125	2
O5	90	70	500	4,225	1
O6	60	80	1,300	2,125	1

Tâm mới

	X1	X2
C1	73.3	60
C2	20	33.33

Vòng lặp 3			C1 = (73.3, 60)	C2 = (20, 33.33)	
	X1	X2	Bình phương khoảng cách đến C1	Bình phương khoảng cách đến C2	Cluster? (1 hay 2)
O1	70	30	911	2,511	1
O2	40	50	1,209	678	2
O3	20	40	3,241	44	2
O4	0	10	7,873	944	2
O5	90	70	379	6,245	1
O6	60	80	577	3,778	1

Tâm không thay đổi, các phần tử trong cluster không thay đổi.
Giải thuật dừng