ÔN TẬP

Bài tập ôn

Cho tập dữ liệu gồm 8 đối tượng sau (biểu diễn thông qua tọa độ (x,y)):

Point	x	y
P1	0.40	0.53
P2	0.22	0.38
P3	0.353	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

- Sử dụng thuật toán phân cấp lần lượt với Single link, Complete link, và Average link để xác định 3 nhóm từ DL trên.
- Vẽ sơ đồ hình cây tương ứng tại mỗi bước.

Bài tập - Thuật toán AGNES

- Build distance matrix (Euclidean measure) between points

	P1	P2	P3	P4	P5	P6
P1	0.00	0.23	0.22	0.37	0.34	0.24
P2	0.23	0.00	0.15	0.19	0.14	0.24
P3	0.22	0.15	0.00	0.16	0.29	**0.10**
P4	0.37	0.19	0.16	0.00	0.28	0.22
P5	0.34	0.14	0.29	0.28	0.00	0.39
P6	0.24	0.24	**0.10**	0.22	0.39	0.00

Ví dụ - Thuật toán AGNES

Sử dụng Single Link :

1. Bước 1: mỗi điểm là một nhóm
2. Bước 2:
   - Trong số các nhóm gồm một điểm thì dist(3,6) – min nên gộp điểm P3 và P6 với nhau thành một nhóm
   - Thu được các nhóm : {1}, {4}, {2}, {5}, {3,6},
3. Quay lại bước 2 do chưa thu được nhóm “toàn bộ”
4. Tính khoảng cách giữa các nhóm .

Ví dụ: Dist({3,6},{1}) =min(dist(3,1),dist(6,1))

=min(0.22, 0.24) = 0.22

Thuật toán AGNES

Sử dụng Single Link :

5. dist(2,5) là nhỏ nhất nên gộp P2 và P5. Ta có các nhóm sau : {1}, {4}, {3,6}, {2,5}

6. Tính khoảng cách giữa các nhóm. Ví dụ :

- dist({3,6},{2,5})

= min(dist(3,2),dist(6,2),dist(3,5),dist(6,5))

= min(0.15, 0.24, 0.28, 0.39) = 0.15

….

- dist({3,6},{2,5}) nhỏ nhất nên gộp các nhóm {3,6}, {2,5} thành một nhóm.

Ta thu được các nhóm : {1},{4},{2,3,5,6}

Ví dụ - Thuật toán AGNES

Sử dụng Single Link :

7.Tiếp tục:

- Tính khoảng cách giữa các nhóm.
- Gộp {4} với {2,3,5,6} thu được các nhóm {1}, {2,3,4,5,6}

8.Gộp 2 nhóm này ta thu được nhóm “toàn bộ” và thuật toán dừng

Bài tập 1

Cho 8 đối tượng sau (biểu diễn thông qua tọa độ (x,y)) : A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9).
Giả sử gán A1, B1, C1 là các trung tâm của các nhóm tương ứng. Sử dụng thuật toán k-means (với k=3) để phân cụm các đối tượng trên:
–Tính độ đo SSE cho các nhóm sau vòng lặp thi hành đầu tiên.
–Tính độ đo SSE cho các nhóm sau vòng lặp thi hành cuối cùng.

Bài tập 2

Cho tập DL gồm 5 điểm trong không gian 2 chiều với ma trận khoảng cách đã cho.

a. Sử dụng thuật toán AGNES lần lượt với Single Link và Complete link để gom nhóm. Vẽ sơ đồ hình cây.

b. Xác định 3 nhóm thu được từ sơ đồ hình cây

	P1	**P2**	**P3**	**P4**	**P5**
P1	0.00	0.10	0.41	0.55	0.35
P2	0.10	0.00	0.64	0.47	0.98
P3	0.41	0.64	0.00	0.44	0.85
P4	0.55	0.47	0.44	0.00	0.76
P5	0.35	0.98	0.85	0.76	0.00

Bài tập 3

Cho tập dữ liệu huấn luyện sau:

Quang cảnh	Nhiệt độ	Độ ẩm	Sức gió	Chơi tennis
Nắng	Nóng	Cao	Yếu	Không
Nắng	Nóng	Cao	Mạnh	Không
Mây	Nóng	Cao	Yếu	Có
Mưa	TB	Cao	Yếu	Có
Mưa	Lạnh	BT	Yếu	Có
Mưa	Lạnh	BT	Mạnh	Không
Mây	Lạnh	BT	Mạnh	Có
Nắng	TB	Cao	Yếu	Không
Nắng	Lạnh	BT	Yếu	Có
Mưa	TB	BT	Yếu	Có
Nắng	TB	BT	Mạnh	Có
Mây	TB	Cao	Mạnh	Có
Mây	Nóng	BT	Yếu	Có
Mưa	TB	Cao	Mạnh	Không

Bài tập 3

a) Sử dụng lần lượt độ đo Gain, chỉ mục gini để xây dựng cây quyết định. So sánh kết quả của 2 độ đo.

b) Từ cây quyết định, biến đổi cây thành luật.

c) Sử dụng phương pháp ILA để xác định luật. So sánh với các luật thu được ở câu b). Nhận xét ?

d) Sử dụng lần lượt các tập luật thu được từ câu c) hoặc để xác định lớp cho mẫu mới. So sánh kết quả.

Quang cảnh	Nhiệt độ	Độ ẩm	Sức gió	Chơi Tennis
Mưa	TB	BT	Mạnh	?
Nắng	TB	Cao	Mạnh	?