**Clustering Sample Problems.**

Assume that the matrix below shows the distances between points in a two dimensional space.

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y3 | 6 | 2 |
| y4 | 1 | 8 |
| y5 | 6 | 6 |
| y6 | 1 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

Follow:

1. Single Link Technique (maximal connected components in a graph) to find the clusters. Show the resulting dendogram.
2. Complete Link Technique (looks for cliques) to find the clusters. Show the resulting dendogram.
3. Minimum Spanning Tree (starts with complete graph, removes largest inconsistent edge) to find the clusters
4. K-means [minimize squared errorE = ∑ ∑ | *p - mi* |2

*i*=1 *p*∈*Ci*

1. K-medoids method to find the clusters
2. TV-tree method to find the clusters (threshold to be active = 2)
3. Agglomerative strategy to find the clusters. Use Manhattan distance

/ d(yi, yj) = |Mi – Mj | + |Ni – Nj | / for objects yi, yjand the distance d(R,Q)= 1/2**⋅**d(A,Q) + 1/2**⋅**d(B,Q) - 1/2**⋅**d(A,B) between clusters R and Q, where R is formed by merging clusters A and B.

**Solution – Problem 1:**

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y3 | 6 | 2 |
| y4 | 1 | 8 |
| y5 | 6 | 6 |
| y6 | 1 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

Distance matrix (we use Manhattan distance)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
| Y1 | x |  |  |  |  |  |  |  |
| Y2 | 3 | x |  |  |  |  |  |  |
| Y3 | 5 | 6 | x |  |  |  |  |  |
| Y4 | 6 | 5 | 11 | x |  |  |  |  |
| Y5 | 11 | 6 | 4 | 7 | X |  |  |  |
| Y6 | 2 | 1 | 7 | 4 | 7 | x |  |  |
| Y7 | 3 | 4 | 2 | 9 | 6 | 5 | X |  |
| Y8 | 5 | 4 | 2 | 9 | 4 | 5 | 2 | x |

Single Link Technique (connected components):

Level 1 - {y6, y2}

Level 2 - {y1,y6}, {{y3, y7}, y8} -> {{y1,y6}, y2}, {y3, y7, y8}->

{y1, y2, y6}, {y3, y7, y8}

Level 3 - {y1, y2, y6, y3, y7, y8} Remark - clique

2 nodes – 1 link

Level 4 - {y1, y2, y6, y3, y7, y8, y5} 3 nodes – 3 links

4 nodes – 12/2=6

**Solution - Problem 2**

Double Link Technique (cliques): 5 nodes – 20/2=10

Level 1 - {y6, y2}- clique 6 nodes – 30/2=15

Level 2 – {y6,y2,y1}- 2 edges, {y3,y7} - >{y3,y7,y8}- 3 edges (clique)

Level 3 - {y6,y2,y1,y3,y7,y8}- 7 edges

Level 4 - {y6,y2,y1,y3,y7,y8,y5,y4}- 11 edges

Level 5 -

**Solution - Problem 3**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
| Y1 | x |  |  |  |  |  |  |  |
| Y2 | 3 | x |  |  |  |  |  |  |
| Y3 | 5 | 6 | x |  |  |  |  |  |
| Y4 | 6 | 5 | 11 | x |  |  |  |  |
| Y5 | 11 | 6 | 4 | 7 | x |  |  |  |
| Y6 | 2 | 1 | 7 | 4 | 7 | x |  |  |
| Y7 | 3 | 4 | 2 | 9 | 6 | 5 | X |  |
| Y8 | 5 | 4 | 2 | 9 | 4 | 5 | 2 | x |

Heuristic Algorithm (chose any node): Let’s take Y1.

D(Y1,Y2)=3, D(Y1,Y3)=5,….; We add these numbers – 35; 35/7= 5 ;

Edge (Y1,Y5) is removed

**Solution – Problem 4 (K - Means)**

Let’s take K=2 and assume that Y3, Y6 are the seeds.

We build clusters C(Y3), C(Y6).

Take Y1, d(Y1,Y3)=5, d(Y1,Y6)=2. So, C(Y3)={Y3}, C(Y6)={Y6,Y1}

Take Y2, d(Y2,Y3)=6, d(Y2,Y6)=1. So, C(Y3)={Y3}, C(Y6)={Y6,Y1,Y2}

Take Y4, d(Y4,Y3)=11, d(Y4,Y6)=4. So, C(Y3)={Y3}, C(Y6)={Y6,Y1,Y2,Y4}

Take Y5, d(Y5,Y3)=4, d(Y5,Y6)=7. So, C(Y3)={Y3,Y5}, C(Y6)={Y6,Y1,Y2,Y4}

Take Y7, d(Y7,Y3)=2, d(Y7,Y6)=5. So, C(Y3)={Y3,Y5,Y7}, C(Y6)={Y6,Y1,Y2,Y4}

Take Y8, d(Y8,Y3)=2, d(Y8,Y6)=5. So, C(Y3)={Y3,Y5,Y7,Y8}, C(Y6)={Y6,Y1,Y2,Y4}

Now we have to find center for C(Y3) and C(Y6)

C(Y3)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y3 | 6 | 2 |
| y5 | 6 | 6 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

21/4 13/4 -> Seed1 = (5.25, 3.25)

C(Y6)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 8 |
| y6 | 1 | 4 |

5/4 18/4 -> Seed2= (1.25, 4.5)

Now, we have to build clusters around these two seeds.

C(Seed1)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| Seed1 | 5.25 | 3.25 |
| Y3 | 6 | 2 |
| Y5 | 6 | 6 |
| Y7 | 4 | 2 |
| Y8 | 5 | 3 |

C(Seed2)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| Y1 | 1 | 2 |
| Y2 | 2 | 4 |
| Y4 | 1 | 8 |
| Y6 | 1 | 4 |
| Seed2 | 1.25 | 4.5 |

**Problem 6** (top-down: TV Tree) solution /data are slightly changed/:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y3 | 6 | 2 |
| y4 | 1 | 6 |
| y5 | 5 | 4 |
| y6 | 1 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

Dom(M): 1 2 4 5 6

p1 p2 p3 p4

Dom(N): 2 3 4 6

q1 q2 q3

**Checking p1**: Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |

M-active

Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y2 | 2 | 4 |
| y3 | 6 | 2 |
| y5 | 5 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

N-active

(Total- 2 active dimensions)

**Checking p2**: Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |

M - active

**Checking p2**: Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y3 | 6 | 2 |
| y5 | 5 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

M-active, N-active (Total- 3 active dimensions)

**Checking p3**: Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |
| y7 | 4 | 2 |

No Active dimensions

**Checking q1**: Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y3 | 6 | 2 |
| y7 | 4 | 2 |

N- active

**Checking q1**: Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y5 | 5 | 4 |
| y6 | 1 | 4 |
| y8 | 5 | 3 |

No active dimensions (Total=1)

**Checking q2**: Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y3 | 6 | 2 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

N-active

**Checking q2**: Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y5 | 5 | 4 |
| y6 | 1 | 4 |

N- active (Total=2)

Winner – p2

T1: Pointer to Left sub-table

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |

M - active (Center=1.5; Radius=0.5)

T2: Pointer to Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y3 | 6 | 2 |
| y5 | 5 | 4 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

M-active (Center=5; Radius=1), N-active (Center=3, Radius=1)

A picture containing clock

Description automatically generated

Table T1 has to be split.

Dom(N): 2 4 6

p5 p6 (give the same split)

Let’s take p5.

Table T1

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |

**Checking p5**: T3 – Pointer to Left sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y1 | 1 | 2 |

N- active (C=2, R=0)

**Checking p5**: T4- Pointer to Right sub-table:

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y2 | 2 | 4 |
| y4 | 1 | 6 |
| y6 | 1 | 4 |

N- active (C=5, R=1)

A close up of a device

Description automatically generated

**Solution to Problem 7**

Agglomerative strategy to find the clusters. Use Manhattan distance

/ d(yi, yj) = |Mi – Mj | + |Ni – Nj | / for objects yi, yjand the distance d(R,Q)= 1/2**⋅**d(A,Q) + 1/2**⋅**d(B,Q) - 1/2**⋅**d(A,B) between clusters R and Q, where R is formed by merging clusters A and B.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Y1 | Y2 | Y3 | Y4 | Y5 | Y6 | Y7 | Y8 |
| Y1 | x |  |  |  |  |  |  |  |
| Y2 | 3 | x |  |  |  |  |  |  |
| Y3 | 5 | 6 | x |  |  |  |  |  |
| Y4 | 6 | 5 | 11 | x |  |  |  |  |
| Y5 | 11 | 6 | 4 | 7 | x |  |  |  |
| Y6 | 2 | 1 | 7 | 4 | 7 | x |  |  |
| Y7 | 3 | 4 | 2 | 9 | 6 | 5 | X |  |
| Y8 | 5 | 4 | 2 | 9 | 4 | 5 | 2 | x |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | {Y1} | {Y2, Y6} | {Y3} | {Y4} | {Y5} | {Y7} | {Y8} |
| {Y1} | x |  |  |  |  |  |  |
| {Y2, Y6} | 2 | x |  |  |  |  |  |
| {Y3} | 5 | ? | x |  |  |  |  |
| {Y4} | 6 | ? | 11 | x |  |  |  |
| {Y5} | 11 | ? | 4 | 7 | x |  |  |
| {Y7} | 3 | ? | 2 | 9 | 6 | X |  |
| {Y8} | 5 | ? | 2 | 9 | 4 | 2 | x |

d(R,Q)= 1/2**⋅**d(A,Q) + 1/2**⋅**d(B,Q) - 1/2**⋅**d(A,B), where R = A∪B

d({Y2,Y6},Y1)= 1/2d(T2,Y1) +1/2d(Y6,Y1) – 1/2d(Y2,Y7)=1/2\*3 + ½\*2 -1/2\*1 = 2

d({Y2,Y6},{Y3})= 1/2⋅d(Y2,Y3)+1/2⋅d(Y6,Y3)-1/2⋅d({Y2},{Y6})=

(1/2)⋅6 + (1/2)⋅7 – (1/2)⋅1 = 3+3 = 6

d({Y2,Y6},{Y4})= 1/2⋅d(Y2,Y4)+1/2⋅d(Y6,Y4)-1/2⋅d({Y2},{Y6})=

(1/2)⋅5 + (1/2)⋅4 – (1/2)⋅1 = 2+2 = 4

d({Y2,Y6},{Y5})= 1/2⋅d(Y2,Y5)+1/2⋅d(Y6,Y5)-1/2⋅d({Y2},{Y6})=

(1/2)⋅6 + (1/2)⋅7 – (1/2)⋅1 = 3 + 3 = 6

d({Y2,Y6},{Y7})= 1/2⋅d(Y2,Y7)+1/2⋅d(Y6,Y7)-1/2⋅d({Y2},{Y6})=

(1/2)⋅4 + (1/2)⋅5 – (1/2)⋅1 = 2 + 2 = 4

d({Y2,Y6},{Y8})= 1/2⋅d(Y2,Y8)+1/2⋅d(Y6,Y8)-1/2⋅d({Y2},{Y6})=

(1/2)⋅4 + (1/2)⋅5 – (1/2)⋅1 = 2 + 2 = 4

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | {Y1} | {Y2, Y6} | {Y3} | {Y4} | {Y5} | {Y7} | {Y8} |
| {Y1} | x |  |  |  |  |  |  |
| {Y2, Y6} | 3 | x |  |  |  |  |  |
| {Y3} | 5 | 6 | x |  |  |  |  |
| {Y4} | 6 | 4 | 11 | x |  |  |  |
| {Y5} | 11 | 6 | 4 | 7 | x |  |  |
| {Y7} | 3 | 4 | 2 | 9 | 6 | x |  |
| {Y8} | 5 | 4 | 2 | 9 | 4 | 2 | x |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | {Y1} | {Y2, Y6} | {Y3, Y7} | {Y4} | {Y5} | {Y8} |
| {Y1} | x |  |  |  |  |  |
| {Y2, Y6} | 3 | x |  |  |  |  |
| {Y3, Y7} | ? | ? | x |  |  |  |
| {Y4} | 6 | 4 | ? | x |  |  |
| {Y5} | 11 | 6 | ? | 7 | x |  |
| {Y8} | 5 | 4 | ? | 9 | 4 | x |

d(R,Q)= 1/2**⋅**d(A,Q) + 1/2**⋅**d(B,Q) - 1/2**⋅**d(A,B), where R = A∪B

d({Y1},{Y3,Y7})= 1/2⋅d({Y1},{Y3}) + 1/2⋅d({Y1},{Y7}) – 1/2⋅d({Y3,Y7}) =

= (1/2)⋅5 + (1/2)⋅3 – (1/2)⋅2 = 4 – 1 = 3

d({Y2,Y6},{Y3,Y7})= 1/2⋅d({Y2,Y6},{Y3}) + 1/2⋅d({Y2,Y6},{Y7}) – 1/2⋅d({Y3,Y7}) =

= (1/2)⋅6 + (1/2)⋅4 – (1/2)⋅2 = 5 – 1 = 4

d({Y4},{Y3,Y7})= 1/2⋅d({Y4},{Y3}) + 1/2⋅d({Y4},{Y7}) – 1/2⋅d({Y3,Y7}) =

= (1/2)⋅11 + (1/2)⋅9 - (1/2)⋅2 = 10 – 1 = 9

d({Y5},{Y3,Y7})= 1/2⋅d({Y5},{Y3}) + 1/2⋅d({Y5},{Y7}) – 1/2⋅d({Y3,Y7}) =

= (1/2)⋅4 + (1/2)⋅6 - (1/2)⋅2 = 5 – 1 = 4

d({Y8},{Y3,Y7})= 1/2⋅d({Y8},{Y3}) + 1/2⋅d({Y8},{Y7}) – 1/2⋅d({Y3,Y7}) =

= (1/2)⋅2 + (1/2)⋅2 - (1/2)⋅2 = 1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | {Y1} | {Y2, Y6} | {Y3, Y7} | {Y4} | {Y5} | {Y8} |
| {Y1} | x |  |  |  |  |  |
| {Y2, Y6} | 3 | x |  |  |  |  |
| {Y3, Y7} | 3 | 4 | x |  |  |  |
| {Y4} | 6 | 4 | 9 | x |  |  |
| {Y5} | 11 | 6 | 4 | 7 | x |  |
| {Y8} | 5 | 4 | 1 | 9 | 4 | x |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | {Y1} | {Y2, Y6} | {Y3, Y7},Y8 | {Y4} | {Y5} |
| {Y1} | x |  |  |  |  |
| {Y2, Y6} | 3 | x |  |  |  |
| {Y3, Y7},Y8 | 3.5 | 3.5 | x |  |  |
| {Y4} | 6 | 4 | ? | x |  |
| {Y5} | 11 | 6 | ? | 7 | x |

d(Y1,{ {Y3, Y7},Y8}) = 1/2 d(Y1,{Y3, Y7}) +  1/2d(Y1,Y8) – 1/2d({Y3, Y7},Y8) = ½\*3 + ½\*5 -1/2\*1 = 3.5

d({Y2,Y6},{ {Y3, Y7},Y8}) = ½\* d({Y2,Y6}, {Y3, Y7}) +1/2 d({Y2,Y6},Y8) -1/2 d({Y3, Y7},Y8)= ½\*4 + ½\*4 – ½\*1