**Clustering Sample Problems.**

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

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

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.

**Problem 1:**

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

Distance matrix

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

Single Link Technique (connected components):

Level 1 (1) - {y2,y6}, {y1}, {y3}, {y4}, {y5}, {y7}

Level 2 (2) - {y1,y6}-> {y1,y2,y6}, {y3,y5}, {y4}, {y7}

Level 3 (3) – {y1,y2,y6,y7}, {y3,y5}, {y4}

Level 4 (4) - {y1,y2,y6,y7,y3,y5,y4}

**Problem 2**

Complete Link Technique (cliques):

Level 1 - {y2,y6}, {y1}, {y3}, {y4}, {y5}, {y7}

Level 2 (2) - {y1,y6}-> {y1,y2,y6}, {y3,y5}, {y4}, {y7}

3

Level 3 (3) – {y1,y2,y6,y7}, {y3,y5}, {y4}

5 1

Level 4 (4) - {y1,y2,y6,y7,y3,y5,y4} 7\*6/2 = 21

20

Level 8 (9) -

**Problem 3**

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

Solution.

Y1 - random attribute

AD(Y1)=33/6 =5.5 AD(Y1)= 24/5 = 4.2

AD(Y3)= 7+4+6+2+5+4= 28/6 =4+1/3

**Problem 4 (K - Means)**

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

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)=7 , d(Y1,Y6)=9

Take Y2, d(Y2,Y3)=4, d(Y2,Y6)=1

Take Y4, d(Y4,Y3)=6, d(Y4,Y6)=7

Take Y5, d(Y5,Y3)=2, d(Y5,Y6)=4

Take Y7, d(Y7,Y3)=4, d(Y7,Y6)=5

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

C(Y3)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y3 | 6 | 4 |
| y1 | 1 | 2 |
| y4 | 4 | 8 |
| y3 | 6 | 6 |
| y7 | 4 | 2 |

21/5 22/5 Center=(4.2, 4.4)

C(Y6)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
| y6 | 1 | 4 |
| y2 | 2 | 4 |
| y |  |  |

3/2 8/2 Center=(1.5, 4) (4,2)

C((4.2,4.4))

|  |  |  |
| --- | --- | --- |
| Y | M | N |
|  | 4.2 | 4.4 |
| y3 | 6 | 4 |
| y4 | 4 | 8 |
| y5 | 6 | 6 |
| y7 | 4 | 2 |

20/4 20/4

C((1.5,4))

|  |  |  |
| --- | --- | --- |
| Y | M | N |
|  | 1.5 | 4 |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y6 | 1 | 4 |

4/3 10/3

**Problem 5** (Top-down: TV Tree) solution /data are slightly changed/:

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

Dom(M): 1 2 4 6

p1 p2 p3

Dom(N): 2 4 6 8

q1 q2 q3

**Problem 6**

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

**Solution**

{y2,y6} - cluster

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

d({y2,y6},y1)= ½ d(y1,y2) + ½ d(y1,y6) -1/2d(y2,y6)= 3/2 + 2/2 – ½ = 2

d({y2,y6}, y3) = ½ d(y3,y2) + ½ d(y3,y6) – ½ d(y2,y6)= 4/2 + 5/2 – ½ = 4

d({y2,y6},y4)= ½ d(y4,y2) + ½ d(y4,y6) – ½ d(y2,y6)= 6/2 + 7/2 – ½ = 6

d({y2,y6},y5) = ½ d(y5,y2) + ½ d(y5,y6) – ½ d(y2,y6) = 6/2 + 7/2 – ½ = 6

d({y2,y6},y7) = ½ d(y7,y2) + ½ d(y7,y6) – ½ d(y2,y6)=4/2 + 5/2 – ½ = 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | {Y2,Y6,Y1} | Y3 | Y4 | Y5 | Y7 |
| {Y2,Y6,Y1} | x |  |  |  |  |
| Y3 | 4.5 | x |  |  |  |
| Y4 | 6.5 | 6 | x |  |  |
| Y5 | 6.5 | 2 | 4 | x |  |
| Y7 | 2.5 | 4 | 6 | 6 | x |

d({Y2,Y6,Y1},Y3)= ½ d({Y2,Y6},Y3) + 1/2d(Y1,Y3) – ½ d({Y2,Y6},Y1)= 4/2 + 7/2 – 1= 4.5

d({Y2,Y6,Y1},Y4) = 1/2d({Y2,Y6},Y4) + 1/2d(Y1,Y4) – 1/2d({Y2,Y6},Y4) = ½\*6 + ½\*9 -1/2\*2=6.5

d({Y2,Y6,Y1},Y5) = 1/2d({Y2,Y6},Y5) + 1/2d(Y1,Y5) – 1/2d({Y2,Y6},Y4)=1/2\*6 + ½\*9-1= 6.5

d({Y2,Y6,Y1},Y7) = 1/2d({Y2,Y6},Y7) + 1/2d(Y1,Y7) – 1/2d({Y2,Y6},Y4)= ½\*4+1/2\*3 – 1 =2.5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | {Y2,Y6,Y1} | {Y3,Y5} | Y4 | Y7 |
| {Y2,Y6,Y1} | x |  |  |  |
| {Y3,Y5} | 4.5 | x |  |  |
| Y4 | 6.5 | ? | x |  |
| Y7 | 2.5 | ? | 6 | x |

d({Y3,Y5}, {Y2,Y6,Y1})= ½\* d(Y3,{Y2,Y6,Y1}) +1/2\* d(Y5,{Y2,Y6,Y1})- ½\*d(Y3,Y5)=

1/2\*4.5 + ½\*6.5 – ½\*2 = 5.5-1 =4.5