**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 | 2 |
| y4 | 1 | 8 |
| y5 | 6 | 6 |
| y6 | 1 | 4  |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

Follow:

1. K-means [minimize squared errorE = ∑ ∑ | *p - mi* |2

*i*=1 *p*∈*Ci*

1. 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:**

Distance matrix

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

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 |
|  | 5.25 | 3.25 |
| y3 | 6 | 2 |
| y5 | 6 | 6 |
| y7 | 4 | 2 |
| y8 | 5 | 3 |

C(Seed2)

|  |  |  |
| --- | --- | --- |
| Y | M | N |
|  | 1.25 | 4.5 |
| y1 | 1 | 2 |
| y2 | 2 | 4 |
| y4 | 1 | 8 |
| y6 | 1 | 4 |

**Solution to Problem 2**

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} | 3 | 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},{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 | 3.75 | x |  |  |
| {Y4} | 6 | 4 | 8.5 | x |  |
| {Y5} | 11 | 6 | ? | 7 | x |

d({y1}, {{Y3,Y7},Y8})= 1/2D(Y1,{Y3,Y7}) + 1/2d(Y1,Y8) – 1/2d({Y3,Y7},Y8) =

½\*3 + ½\*5 -1/2\*1 = 3/2 + 5/2 – ½ = 7/2

d({Y2,Y6},{{Y3.Y7},Y8})= 1/2d({Y2,Y6},{Y3,Y7}) + 1/2d({Y2,Y6},Y8} – 1/2d({Y3,Y7},Y8)=

½\*4 + ½\*4 – ½\*1 = 3.75

d(Y4, {{Y3,Y7},Y8}) = ½\* d(Y4, {Y3,Y7}) + ½\* d(Y4,Y8) – 1/2d{{Y3,Y7},Y8})

= ½\*9 + ½\*9 - ½\* 1 = 8.5