Data Mining HW4

202055364 황 성 윤

Exercises for Clustering

Problem 1)

① 최장연결법

	1	2	3	4	5
1	0				
2	7	0			
3	①	6	0		
4	9	3	8	0	
5	8	5	7	4	0

$d(1,3)=1 \Rightarrow G_1=\{1,3\}$

$d(G_1,2)=\max\{7,6\}=7$

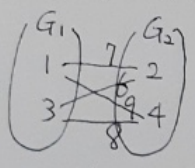
$d(G_1,4)=\max\{9,8\}=9$

$d(G_1,5)=\max\{8,7\}=8$

	{1,3}	2	4	5
⇒{1,3}	0			
2	7	0		
4	9	③	0	
5	8	5	4	0

$d(2,4)=3 \Rightarrow G_2=\{2,4\}$

$d(G_1,G_2)=\max\{7,9,6,8\}=9$



$d(G_2,5)=5$

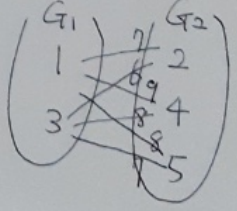
	{1,3}	{2,4}	5
⇒{1,3}	0		
{2,4}	9	0	
5	8	⑤	0

$d(G_2,5)=5 \Rightarrow G_3=\{2,4,5\}$

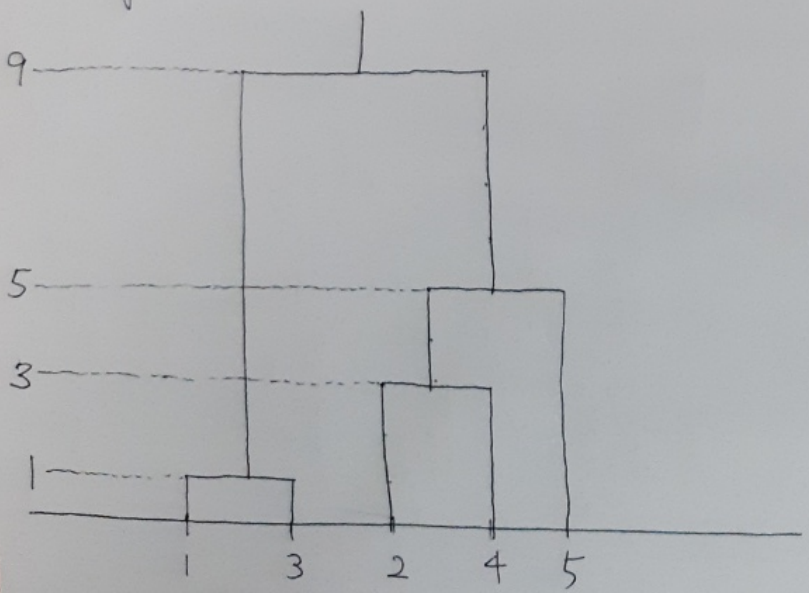
⇓

	{1,3}	{2,4,5}
{1,3}	0	
{2,4,5}	9	0

$d(G_1,G_3)=\max\{7,9,8,6,8,7\}=9$



(dendrogram)



② 평균연결법

	1	2	3	4	5
1	0				
2	7	0			
3	①	6	0		
4	9	3	8	0	
5	8	5	7	4	0

$d(1,3)=1 \Rightarrow G_1=\{1,3\}$

$d(G_1,2)=\frac{1}{2}(7+6)=6.5$

$d(G_1,4)=\frac{1}{2}(9+8)=8.5$

$d(G_1,5)=\frac{1}{2}(8+4)=6$

⇒

	{1,3}	2	4	5
{1,3}	0			
2	6.5	0		
4	8.5	3	0	
5	6	5	4	0

$d(2,4)=3 \Rightarrow G_2=\{2,4\}$

$d(G_1,G_2)=\frac{1}{4}(7+9+6+8)$
$=7.5$

$d(G_2,5)=\frac{1}{2}(5+4)=4.5$

⇒

	{1,3}	{2,4}	5
{1,3}	0		
{2,4}	7.5	0	
5	6	④.5	0

$d(G_2,5)=4.5 \Rightarrow G_3=\{2,4,5\}$

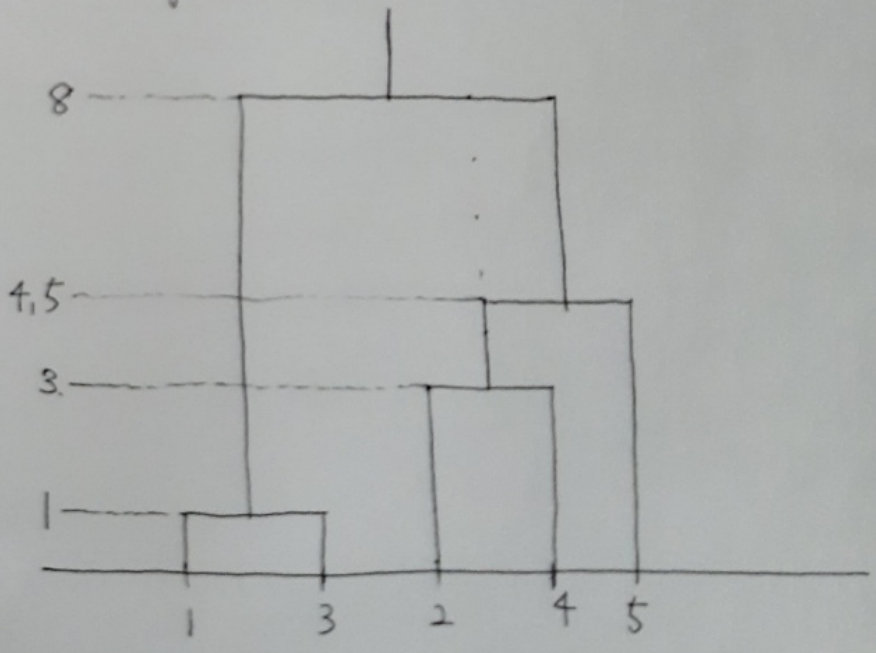
$d(G_1,G_3)=\frac{1}{6}(7+9+8+6+8+7)$
$=8$

⇓

	{1,3}	{2,4,5}
{1,3}	0	
{2,4,5}	8	0

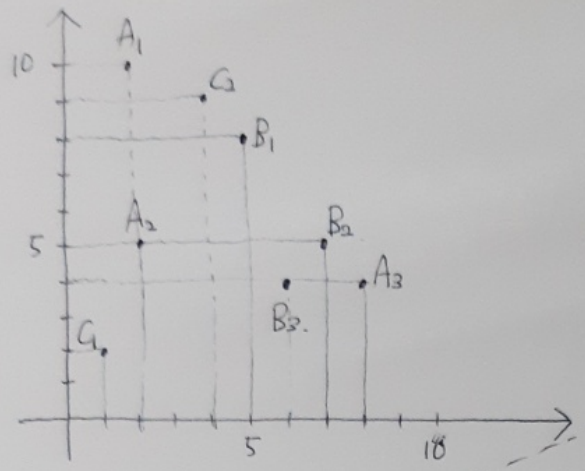
$G_4=\{1,2,3,4,5\}$

(dendrogram)



Problem 2) 거리는 Euclid distance 로 측정함.

* $A_1(2,10)$
$A_2(2,5)$
$A_3(8,4)$
$B_1(5,8)$
$B_2(7,5)$
$B_3(6,4)$
$C_1(1,2)$
$C_2(4,9)$



$k=3$

* 첫번째 결과

$G_1 = \{A_1\}$

$G_2 = \{B_1, A_2, A_3, B_2, B_3, C_1\}$

$G_3 = \{C_2\}$

new center point

$= \left(\frac{5+2+8+7+6+1}{6}, \frac{8+5+4+5+4+2}{6}\right)$

$= (4.83, 4.67) = G_2^*$

* 두번째 결과

$G_1 = \{A_1\} \rightarrow G_1^*$

$G_2 = \{A_2, A_3, B_2, B_3, C_1\} \rightarrow G_2^*$ → new center point $= \left(\frac{2+8+7+6+1}{5}, \frac{5+4+5+4+2}{5}\right) = (4.8, 4) = G_2^{**}$

$G_3 = \{C_2, B_1\} \rightarrow G_3^*$ → new center point $= \left(\frac{4+5}{2}, \frac{9+8}{2}\right) = (4.5, 8.5) = G_3^{**}$

* 세번째 결과

$G_1 = \{A_1\} \rightarrow G_1^{**}$

$G_2 = \{A_2, A_3, B_2, B_3, C_1\}$

$G_3 = \{B_1, C_2\}$

⇒ 이러한 과정을 통해 군집결과를 update 해 나가면 된다.
(중심의 위치에서 Euclid distance 가
가장 작은 군집으로 할당하는 것이 방법임.)

Problem 3) SOM

* 세번째 관측값 : $(0,1,0,1)$. , 초기 가중치 행렬 $W=(w_1, w_2)=\begin{bmatrix}0.5 & 0.1\\ 0.3 & 0.7\\ 0.6 & 0.8\\ 0.2 & 0.2\end{bmatrix}$

Step 2-1) 첫번째 입력 벡터 : $x=(0,1,0,1)$

Step 2-2) 유사도 거리 계산

$D_1 = \|x - w_1\|^2 = (0-0.5)^2+(1-0.3)^2+(0-0.6)^2+(1-0.2)^2 = 1.74$

$D_2 = \|x - w_2\|^2 = (0-0.1)^2+(1-0.7)^2+(0-0.8)^2+(1-0.2)^2 = 1.38$ ✓

Step 2-3) BMU : $C=2$

Step 2-4) 가중치 벡터 재조정 $w_c^{new} = w_c + \alpha(1)\cdot(x - w_c)$.

$$w_2^{new} = \begin{pmatrix}0.1\\0.7\\0.8\\0.2\end{pmatrix} + 0.6\cdot\begin{pmatrix}0-0.1\\1-0.7\\0-0.8\\1-0.2\end{pmatrix} = \begin{pmatrix}0.1-0.06\\0.7+0.18\\0.8-0.48\\0.2+0.48\end{pmatrix} = \begin{pmatrix}0.04\\0.88\\0.32\\0.68\end{pmatrix} \Rightarrow W=\begin{bmatrix}0.5 & 0.04\\ 0.3 & 0.88\\ 0.6 & 0.32\\ 0.2 & 0.68\end{bmatrix}$$

* 네 번째 관측값 : $(0,0,1,0)$

Step 2-1) 두번째 입력 벡터 : $x=(0,0,1,0)$

Step 2-2) 유사도 거리 계산

$D_1 = \|x - w_1\|^2 = (0-0.5)^2+(0-0.3)^2+(1-0.6)^2+(0-0.2)^2 = 0.54$ ✓

$D_2 = \|x - w_2\|^2 = (0-0.04)^2+(0-0.88)^2+(1-0.32)^2+(0-0.68)^2 = 1.7008$

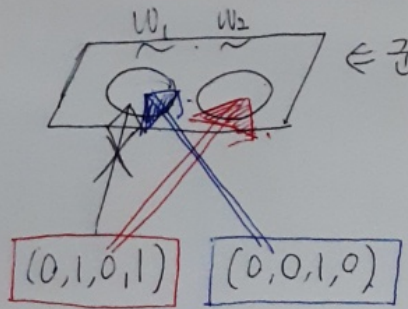
Step 2-3) BMU : $c=1$

Step 2-4) 가중치 벡터 재조정 $w_c^{new} = w_c + \alpha(1)\cdot(x - w_c)$

$$w_1^{new} = \begin{pmatrix}0.5\\0.3\\0.6\\0.2\end{pmatrix} + 0.6\begin{pmatrix}0-0.5\\0-0.3\\1-0.6\\0-0.2\end{pmatrix} = \begin{pmatrix}0.5-0.3\\0.3-0.18\\0.6+0.24\\0.2-0.12\end{pmatrix} = \begin{pmatrix}0.2\\0.12\\0.84\\0.08\end{pmatrix}$$

$$\Rightarrow W=\begin{bmatrix}0.2 & 0.04\\ 0.12 & 0.88\\ 0.84 & 0.32\\ 0.08 & 0.68\end{bmatrix}$$

최종 가중치 행렬



⇐ 군집결과