2021년도 2학기 바이오빅데이터와데이터마이닝
기말고사 보고서

이화여자대학교 컴퓨터공학과
1871056 한지수

1번 문제 풀이

K-Means Clustering 풀이 방법 (k:3, Distance metric: Manhattan distance)

1. 중심점을 바탕으로 cluster한다.
2. Cluster된 점들을 바탕으로 중심점을 다시 조정한다.
3. 중심점을 바탕으로 cluster 한 경우,
    1. 수렴할 경우 해당 중심점으로 채택
    2. 수렴하지 않을 경우 2로 다시 돌아감. 이를 수렴할 때까지 반복

*문제 계산과정이 매우 많아 엑셀로 Distance matrix를 계산하였습니다.(엑셀 문제 1.xlsx, 3쪽 참조)

문제에서 제시한 15개의 레코드와 초기 중심점은 다음과 같다.

Instance (point)	X	Y
1	4	19
2	1	9
3	16	1
4	14	18
5	15	13
6	13	7
7	8	20
8	8	4
9	17	13
10	6	12
11	17	8
12	16	2
13	17	0
14	10	8
15	13	3

〔 전체 데이터셋 〕

초기 중심점	X	Y
Cluster 1	15	15
Cluster 2	7	11
Cluster 3	17	3

〔 초기 중심점 〕

1번 문제 풀이

초기. 앞에서 제시된 초기중심점을 바탕으로 Distance Metric에 맞게 cluster 처리를 한 결과는 다음과 같다.

1. 계산된 Distance 결과

다음 표는 instance 별로 cluster 중심점과의 거리를 구하고(밑 example 참조),
최소 거리를 바탕으로 Clustering한 결과를 볼 수 있다.

instance	(Distance with) cluster 1	cluster 2	cluster 3	Cluster Num
1	15	11	29	2
2	20	8	22	2
3	15	19	3	3
4	4	14	18	1
5	2	10	12	1
6	10	10	8	3
7	12	10	26	1
8	18	8	10	2
9	4	12	10	1
10	12	2	20	2
11	9	13	5	3
12	14	18	2	3
13	17	21	3	3
14	12	6	12	2
15	14	14	4	3

계산 Example. Instance 1 의 Cluster별 Distance (Excel 표현: ABS(C10-C3)+ABS(D10-D3))

- $cluster\ 1 = |4 - 15| + |19 - 15| = 11 + 4 = 15$
- $cluster\ 2 = |4 - 7| + |19 - 11| = 3 + 8 = 11$
- $cluster\ 3 = |4 - 17| + |19 - 3| = 13 + 16 = 29$

2. 업데이트된 Cluster 결과

Cluster Num에 맞게 나눈 Cluster 정보는 다음과 같으며, Cluster 정보에 따라 중심점도 옮겨졌다.

- **Cluster 1. 중심점 (13.5, 16)**

instance	X	Y
4	14	18
5	15	13
7	8	20
9	17	13

- **Cluster 2. 중심점 (5.8, 10.4)**

instance	X	Y
1	4	19
2	1	9
8	8	4
10	6	12
14	10	8

- **Cluster 3. 중심점 (15.333, 3.5)**

instance	X	Y
3	16	1
6	13	7
11	17	8
12	16	2
13	17	0
15	13	3

1번 문제 풀이

1차 시도. 첫번째 시도를 통해 얻은 중심점을 바탕으로 Distance Metric에 맞게 cluster 처리를 한 결과는 다음과 같다.

1. 계산된 Distance 결과

다음 표는 instance 별로 중심점과의 거리를 구하고, 최소 거리를 바탕으로 Cluster를 연결한 결과를 볼 수 있다. 첫번째 시도에 분류된 Cluster와 현재 분류된 Cluster를 비교해보았을 때 8번 instance이 수렴하지 않은 상황으로 Cluster를 다시 조정하고자 한다.

instance	cluster 1	cluster 2	cluster 3	NOW CLUSTER	PREVIOUS CLUSTER
1	12.5	10.4	26.833	2	2
2	19.5	6.2	19.833	2	2
3	17.5	19.6	3.167	3	3
4	2.5	15.8	15.833	1	1
5	4.5	11.8	9.833	1	1
6	9.5	10.6	5.833	3	3
7	9.5	11.8	23.833	1	1
8	17.5	8.6	7.833	3	2
9	6.5	13.8	11.167	1	1
10	11.5	1.8	17.833	2	2
11	11.5	13.6	6.167	3	3
12	16.5	18.6	2.167	3	3
13	19.5	21.6	5.167	3	3
14	11.5	6.6	9.833	2	2
15	13.5	14.6	2.833	3	3

2. 업데이트된 Cluster 결과

Cluster Num에 맞게 나눈 Cluster 정보는 다음과 같으며,Cluster 정보에 따라 중심점도 업데이트 됨을 알 수 있다.

- **Cluster 1. 중심점 (13.5, 16)**

instance	X	Y
4	14	18
5	15	13
7	8	20
9	17	13

- **Cluster 2. 중심점 (5.25, 12)**

instance	X	Y
1	4	19
2	1	9
10	6	12
14	10	8

- **Cluster 3. 중심점 (14.286, 3.571)**

instance	X	Y
3	16	1
6	13	7
11	17	8
12	16	2
13	17	0
15	13	3
8	8	4

1번 문제 풀이

2차 시도. 첫번째 시도를 통해 얻은 중심점을 바탕으로 Distance Metric에 맞게 cluster 처리를 한 결과는 다음과 같다.

1. 계산된 Distance 결과

다음 표는 instance 별로 중심점과의 거리를 구하고, 최소 거리를 바탕으로 Cluster를 연결한 결과를 볼 수 있다. 첫번째 시도에 분류된 Cluster와 현재 분류된 Cluster를 비교해보았을 때 14번 instance이 수렴하지 않은 상황으로 Cluster를 다시 조정하고자 한다.

instance	cluster 1	cluster 2	cluster 3	NOW CLUSTER	PREVIOUS CLUSTER
1	12.5	8.25	25.715	2	2
2	19.5	7.25	18.715	2	2
3	17.5	21.75	4.285	3	3
4	2.5	14.75	14.715	1	1
5	4.5	10.75	10.143	1	1
6	9.5	12.75	4.715	3	3
7	9.5	10.75	22.715	1	1
8	17.5	10.75	6.715	3	3
9	6.5	12.75	12.143	1	1
10	11.5	0.75	16.715	2	2
11	11.5	15.75	7.143	3	3
12	16.5	20.75	3.285	3	3
13	19.5	23.75	6.285	3	3
14	11.5	8.75	8.715	3	2
15	13.5	16.75	1.857	3	3

2. 업데이트된 Cluster 결과

Cluster Num에 맞게 나눈 Cluster 정보는 다음과 같으며, Cluster 정보에 따라 중심점도 업데이트 됨을 알 수 있다.

- **Cluster 1. 중심점 (13.5, 16)**

instance	X	Y
4	14	18
5	15	13
7	8	20
9	17	13

- **Cluster 2. 중심점 (3.667, 13.333)**

instance	X	Y
1	4	19
2	1	9
10	6	12

- **Cluster 3. 중심점 (13.75, 4.125)**

instance	X	Y
3	16	1
6	13	7
11	17	8
12	16	2
13	17	0
15	13	3
8	8	4
14	10	8

1번 문제 풀이

3차 시도. 두번째 시도를 통해 얻은 중심점을 바탕으로 Distance Metric에 맞게 cluster 처리를 한 결과는 다음과 같다.

1. 계산된 Distance 결과

다음 표는 instance 별로 초기 중심점과의 거리를 구하고, 최소 거리를 바탕으로 Cluster를 연결한 결과를 볼 수 있다.

두번째 시도에 분류된 Cluster와 현재 분류된 Cluster를 비교해보았을 때 **모든 instance들이 수렴하고 있는 것으로 보아 최종 클러스터가 결정되었음**을 알 수 있다.

instance	cluster 1	cluster 2	cluster 3	NOW CLUSTER	PREVIOUS CLUSTER
1	12.500	6.010	24.630	2	2
2	19.500	6.990	17.630	2	2
3	17.500	24.670	5.370	3	3
4	2.500	15.010	14.130	1	1
5	4.500	11.670	10.130	1	1
6	9.500	15.670	3.630	3	3
7	9.500	11.010	21.630	1	1
8	17.500	13.670	5.870	3	3
9	6.500	13.670	12.130	1	1
10	11.500	3.670	15.630	2	2
11	11.500	18.670	7.130	3	3
12	16.500	23.670	4.370	3	3
13	19.500	26.670	7.370	3	3
14	11.500	11.670	7.630	3	3
15	13.500	19.670	1.870	3	3
CONVERGE					

1번 문제 풀이

최종 Clustering 결과

- **Cluster 1. 중심점 (13.5, 16)**

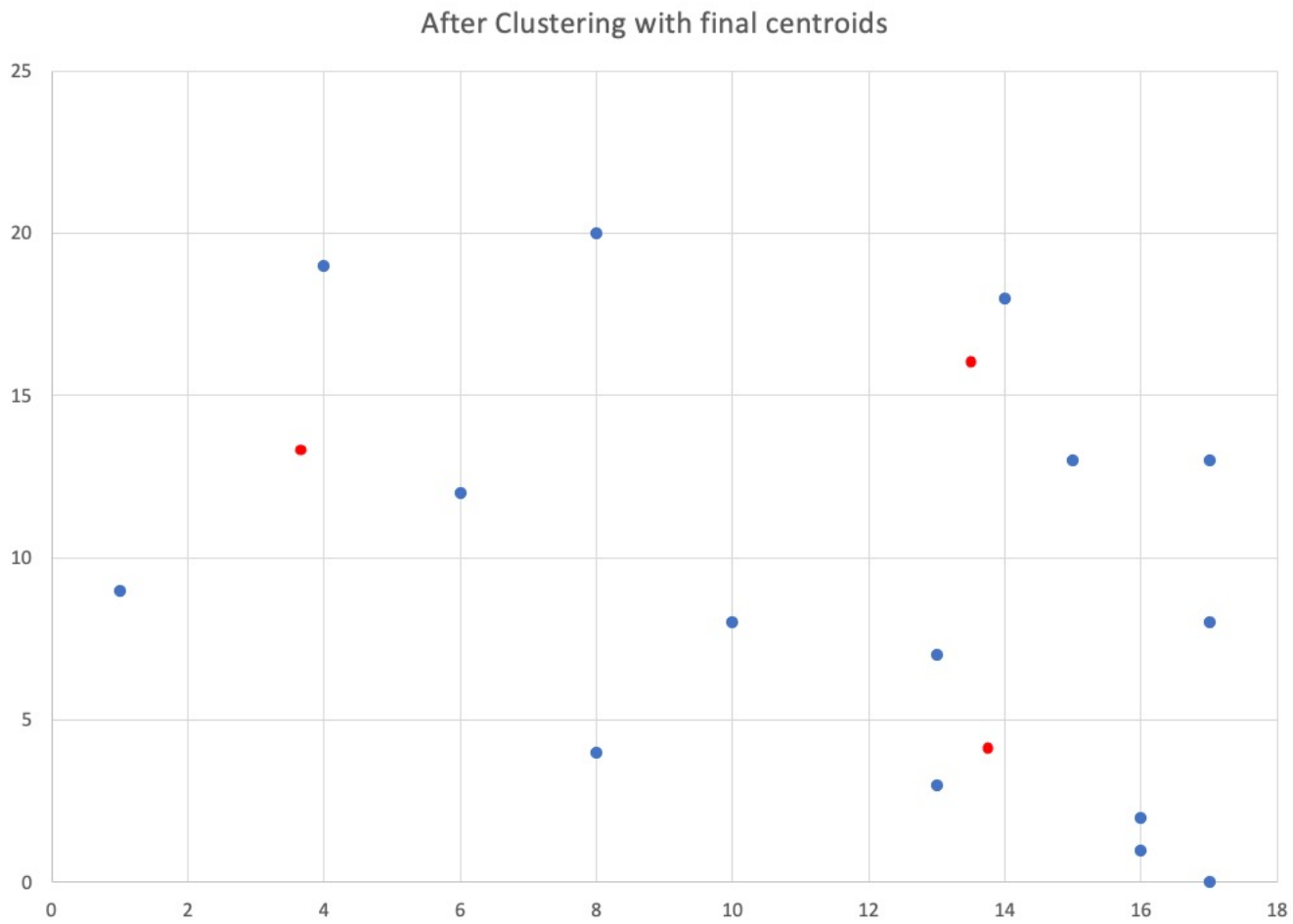
instance	X	Y
4	14	18
5	15	13
7	8	20
9	17	13

- **Cluster 2. 중심점 (3.667, 13.333)**

instance	X	Y
1	4	19
2	1	9
10	6	12

- **Cluster 3. 중심점 (13.75, 4.125)**

instance	X	Y
3	16	1
6	13	7
11	17	8
12	16	2
13	17	0
15	13	3
8	8	4
14	10	8



2번 문제 풀이

Hierarchical Clustering 풀이 방법 (Distance metric: Euclidean distance, Linkage method: MIN or Single)
3개의 클러스터가 발생될 때까지 반복{

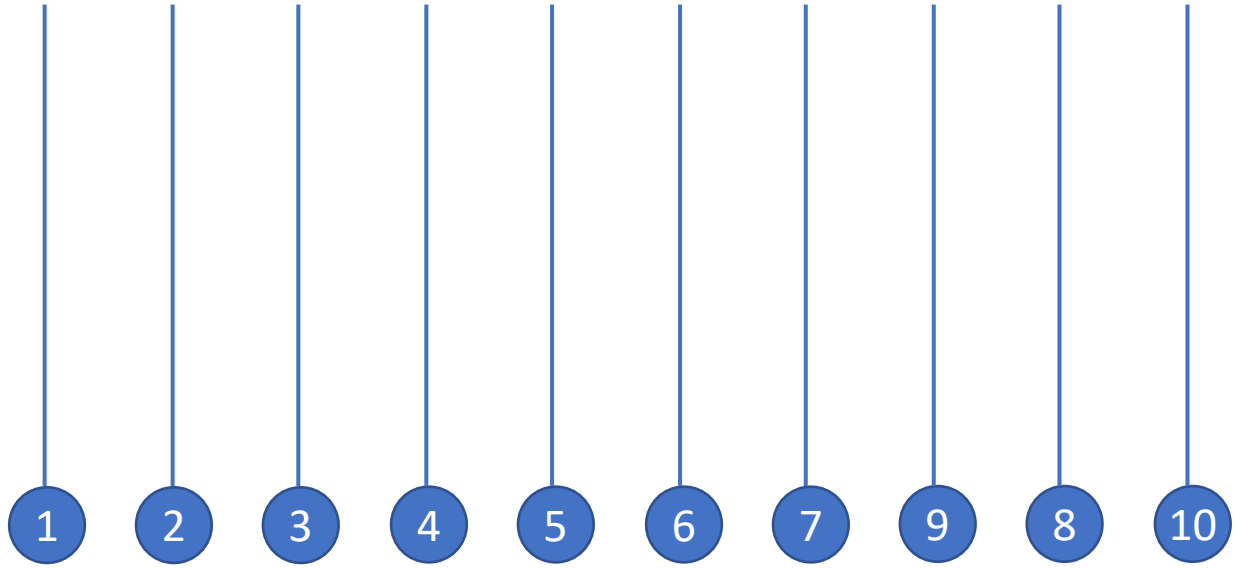
1. Proximity matrix(Instance(point)간의 distance)를 구한다.
2. Linkage method에 따라 cluster 처리하여 cluster 내용(dendrogram), Proximity matrix을 업데이트 한다

}
*Proximity matrix 계산과정이 매우 많아 엑셀로 Distance matrix를 계산하였습니다.(엑셀 문제 2.xlsx, 9쪽 참조)

문제에서 제시한 10개의 레코드와 초기 dendrogram은 다음과 같으며 현재 10개의 클러스터로 이루어져있다.

instance (point)	X	Y
1	139	49
2	52	-64
3	-115	-65
4	28	-91
5	-6	-13
6	-44	-21
7	3	19
8	19	-9
9	13	29
10	26	1

〔 전체 데이터셋 〕



〔 초기 dendrogram: 총 10개 cluster〕

2번 문제 풀이

- Proximity Matrix 계산

Cluster	1	2	3	4	5	6	7	8	9	10
1	0.000									
2	142.611	0.000								
3	278.410	167.003	0.000							
4	178.664	36.125	145.344	0.000						
5	157.699	77.233	120.768	85.088	0.000					
6	195.931	105.190	83.528	100.419	38.833	0.000				
7	139.270	96.385	144.845	112.805	33.242	61.717	0.000			
8	133.282	64.140	145.231	82.492	25.318	64.133	32.249	0.000		
9	127.577	100.846	158.808	120.934	46.098	75.822	14.142	38.471	0.000	
10	122.772	70.007	155.682	92.022	34.928	73.376	29.206	12.207	30.871	0.000

*행렬의 왼쪽위-오른쪽아래로 대각선을 그으면 나타나는 상부와 하부 부분의 계산 결과가 동일하기에 하부 부분만을 표시하였습니다.

*엑셀로 계산하였으며 소수 셋째자리에서 반올림 처리되었습니다.

*행렬에서의 클러스터 표현은 instance 1로 이루어진 cluster 면 Cluster 1로, instance 2와 3으로 이루어져 있다면 Cluster (2,3) 형식으로 표기하였습니다.

(참고. (5,8,6)과 ((5,8),6) 클러스터는 hierarchical proximity에 따라 다른 표현이므로 구분하고자 후자의 표현방법을 사용합니다. Hierarchical Clustering의 표기인 ((5,8),6)이 6보다 (5,8)이 더 근접함을 나타내므로 다음과 같이 표기합니다.)

Cluster 1 의 Cluster별 Distance (Excel 표현: =SQRT(POWER(C5-C4,2)+POWER(D5-D4,2)))

- $Dist(cluster\ 1\ -cluster\ 2) = \sqrt{(139-52)^2+(49-(-64))^2} = 142.611$
- $Dist(cluster\ 1\ -cluster\ 3) = \sqrt{\left(139-(-115)\right)^2+\left(49-(-65)\right)^2} = 278.410$
- $Dist(cluster\ 1\ -cluster\ 4) = \sqrt{(139-28)^2+(49-(-91))^2} = 178.664$
- $Dist(cluster\ 1\ -cluster\ 5) = \sqrt{\left(139-(-6)\right)^2+\left(49-(-13)\right)^2} = 157.699$
- $Dist(cluster\ 1\ -cluster\ 6) = \sqrt{\left(139-(-44)\right)^2+\left(49-(-21)\right)^2} = 195.931$
- $Dist(cluster\ 1\ -cluster\ 7) = \sqrt{(139-3)^2+(49-19)^2} = 138.270$
- $Dist(cluster\ 1\ -cluster\ 8) = \sqrt{(139-19)^2+\left(49-(-9)\right)^2} = 133.282$
- $Dist(cluster\ 1\ -cluster\ 9) = \sqrt{(139-13)^2+(49-29)^2} = 127.577$
- $Dist(cluster\ 1\ -cluster\ 10) = \sqrt{(139-26)^2+(49-1)^2} = 122.772$

2번 문제 풀이

- Proximity Matrix 계산

Cluster 5의 Cluster별 Distance

- $Dist(cluster\ 2\ -cluster\ 3) = \sqrt{\left(52-(-115)\right)^2+\left(64-(-65)\right)^2} = 167.003$
- $Dist(cluster\ 2\ -cluster\ 4) = \sqrt{(52-28)^2+(64-(-91))^2} = 36.125$
- $Dist(cluster\ 2\ -cluster\ 5) = \sqrt{\left(52-(-6)\right)^2+\left(64-(-13)\right)^2} = 77.233$
- $Dist(cluster\ 2\ -cluster\ 6) = \sqrt{\left(52-(-44)\right)^2+\left(64-(-21)\right)^2} = 105.190$
- $Dist(cluster\ 2\ -cluster\ 7) = \sqrt{(52-3)^2+(64-19)^2} = 96.385$
- $Dist(cluster\ 2\ -cluster\ 8) = \sqrt{(52-19)^2+\left(64-(-9)\right)^2} = 64.140$
- $Dist(cluster\ 2\ -cluster\ 9) = \sqrt{(52-13)^2+(64-29)^2} = 100.846$
- $Dist(cluster\ 2\ -cluster\ 10) = \sqrt{(52-26)^2+(64-1)^2} = 70.007$

Cluster 3의 Cluster별 Distance

- $Dist(cluster\ 3\ -cluster\ 4) = \sqrt{(-115-28)^2+(-65-(-91))^2} = 145.344$
- $Dist(cluster\ 3\ -cluster\ 5) = \sqrt{\left(-115-(-6)\right)^2+\left(-65-(-13)\right)^2} = 120.768$
- $Dist(cluster\ 3\ -cluster\ 6) = \sqrt{\left(-115-(-44)\right)^2+\left(-65-(-21)\right)^2} = 83.528$
- $Dist(cluster\ 3\ -cluster\ 7) = \sqrt{(-115-3)^2+(-65-19)^2} = 144.845$
- $Dist(cluster\ 3\ -cluster\ 8) = \sqrt{(-115-19)^2+\left(-65-(-9)\right)^2} = 145.231$
- $Dist(cluster\ 3\ -cluster\ 9) = \sqrt{(-115-13)^2+(-65-29)^2} = 158.808$
- $Dist(cluster\ 3\ -cluster\ 10) = \sqrt{(-115-26)^2+(-65-1)^2} = 155.682$

Cluster 4의 Cluster별 Distance

- $Dist(cluster\ 4\ -cluster\ 5) = \sqrt{\left(28-(-6)\right)^2+\left(-91-(-13)\right)^2} = 85.088$
- $Dist(cluster\ 4\ -cluster\ 6) = \sqrt{\left(28-(-44)\right)^2+\left(-91-(-21)\right)^2} = 100.419$
- $Dist(cluster\ 4\ -cluster\ 7) = \sqrt{(28-3)^2+(-91-19)^2} = 122.805$
- $Dist(cluster\ 4\ -cluster\ 8) = \sqrt{(28-19)^2+\left(-91-(-9)\right)^2} = 82.492$
- $Dist(cluster\ 4\ -cluster\ 9) = \sqrt{(28-13)^2+(-91-29)^2} = 120.934$
- $Dist(cluster\ 4\ -cluster\ 10) = \sqrt{(28-26)^2+(-91-1)^2} = 92.022$

2번 문제 풀이

- Proximity Matrix 계산

Cluster 5의 Cluster별 Distance

- $Dist(cluster\ 5\ -cluster\ 6) = \sqrt{\left(-6-(-44)\right)^2+\left(-13-(-21)\right)^2} = 38.833$
- $Dist(cluster\ 5\ -cluster\ 7) = \sqrt{(-6-3)^2+(-13-19)^2} = 33.242$
- $Dist(cluster\ 5\ -cluster\ 8) = \sqrt{(-6-19)^2+\left(-13-(-9)\right)^2} = 25.318$
- $Dist(cluster\ 5\ -cluster\ 9) = \sqrt{(-6-13)^2+(-13-29)^2} = 46.098$
- $Dist(cluster\ 5\ -cluster\ 10) = \sqrt{(-6-26)^2+(-13-1)^2} = 34.928$

Cluster 6의 Cluster별 Distance

- $Dist(cluster\ 6\ -cluster\ 7) = \sqrt{(-44-3)^2+(-21-19)^2} = 61.717$
- $Dist(cluster\ 6\ -cluster\ 8) = \sqrt{(-44-19)^2+\left(-21-(-9)\right)^2} = 64.133$
- $Dist(cluster\ 6\ -cluster\ 9) = \sqrt{(-44-13)^2+(-21-29)^2} = 75.822$
- $Dist(cluster\ 6\ -cluster\ 10) = \sqrt{(-44-26)^2+(-21-1)^2} = 73.376$

Cluster 7의 Cluster별 Distance

- $Dist(cluster\ 7\ -cluster\ 8) = \sqrt{(3-19)^2+\left(19-(-9)\right)^2} = 32.249$
- $Dist(cluster\ 7\ -cluster\ 9) = \sqrt{(3-13)^2+(19-29)^2} = 14.142$
- $Dist(cluster\ 7\ -cluster\ 10) = \sqrt{(3-26)^2+(19-1)^2} = 29.206$

Cluster 8의 Cluster별 Distance

- $Dist(cluster\ 8\ -cluster\ 9) = \sqrt{(19-13)^2+(-9-29)^2} = 38.471$
- $Dist(cluster\ 8\ -cluster\ 10) = \sqrt{(19-26)^2+(-9-1)^2} = 12.207$

Cluster 9의 Cluster별 Distance

- $Dist(cluster\ 9\ -cluster\ 10) = \sqrt{(13-26)^2+(29-1)^2} = 30.871$

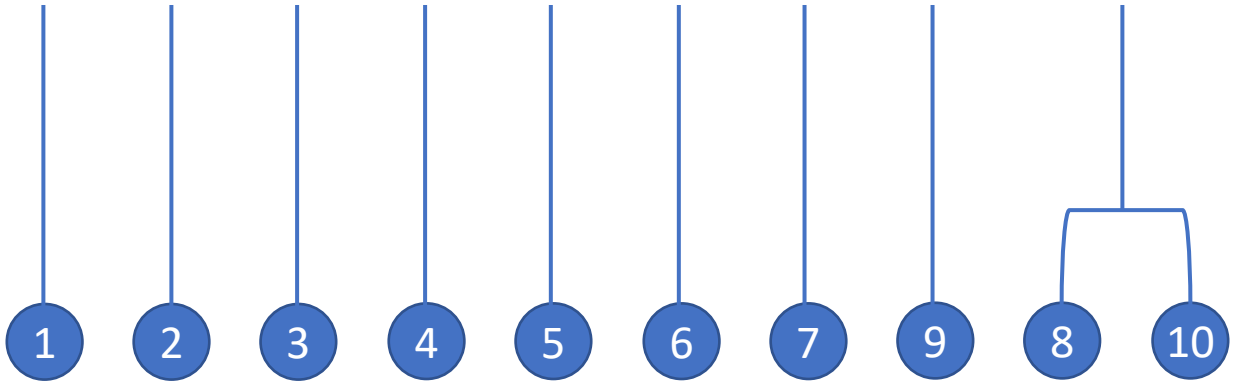
2번 문제 풀이

1차 시도.

- Proximity Matrix

Cluster	1	2	3	4	5	6	7	8	9	10
1	0.000									
2	142.611	0.000								
3	278.410	167.003	0.000							
4	178.664	36.125	145.344	0.000						
5	157.699	77.233	120.768	85.088	0.000					
6	195.931	105.190	83.528	100.419	38.833	0.000				
7	139.270	96.385	144.845	112.805	33.242	61.717	0.000			
8	133.282	64.140	145.231	82.492	25.318	64.133	32.249	0.000		
9	127.577	100.846	158.808	120.934	46.098	75.822	14.142	38.471	0.000	
10	122.772	70.007	155.682	92.022	34.928	73.376	29.206	12.207	30.871	0.000

Distance가 가장 작은 8-10을 합치게 된다. 업데이트된 Dendrogram은 다음과 같다.
총 9개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 9개 cluster 〕

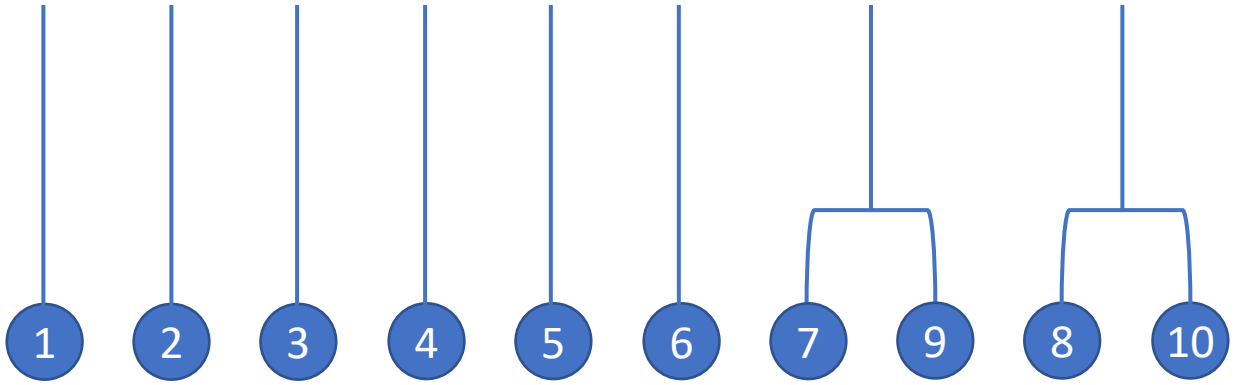
2번 문제 풀이

2차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	2	3	4	5	6	7	(8,10)	9
1	0.000								
2	142.611	0.000							
3	278.410	167.003	0.000						
4	178.664	36.125	145.344	0.000					
5	157.699	77.233	120.768	85.088	0.000				
6	195.931	105.190	83.528	100.419	38.833	0.000			
7	139.270	96.385	144.845	112.805	33.242	61.717	0.000		
(8,10)	122.772	64.140	145.231	82.492	25.318	64.133	29.206	0.000	
9	127.577	100.846	158.808	120.934	46.098	75.822	14.142	30.871	0.000

Distance가 가장 작은 7-9가 합쳐진다. 업데이트된 Dendrogram은 다음과 같다. 클러스터는 (1),(2),(3),(4),(5),(6),(7,9),(8,10)이 존재한다.
총 8개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 8개 cluster 〕

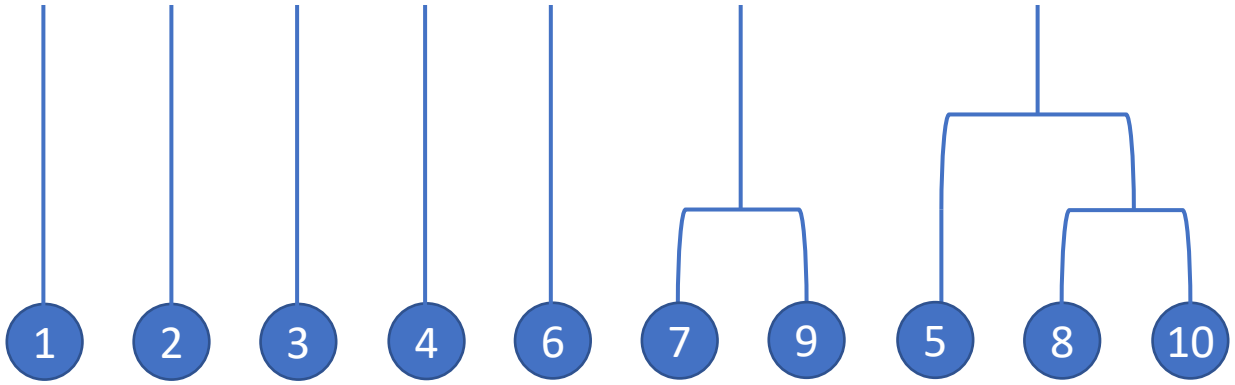
2번 문제 풀이

3차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	2	3	4	5	6	(7,9)	(8,10)
1	0.000							
2	142.611	0.000						
3	278.410	167.003	0.000					
4	178.664	36.125	145.344	0.000				
5	157.699	77.233	120.768	85.088	0.000			
6	195.931	105.190	83.528	100.419	38.833	0.000		
(7,9)	127.577	96.385	144.845	112.805	33.242	61.717	0.000	
(8,10)	122.772	64.140	145.231	82.492	25.318	64.133	29.206	0.000

Distance가 가장 작은 5와 (8, 10)가 합쳐진다. 업데이트된 Dendrogram은 다음과 같다. 클러스터는 (1), (2), (3), (4), (6), (7,9), (5,(8,10))이 존재한다.
총 7개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 7개 cluster 〕

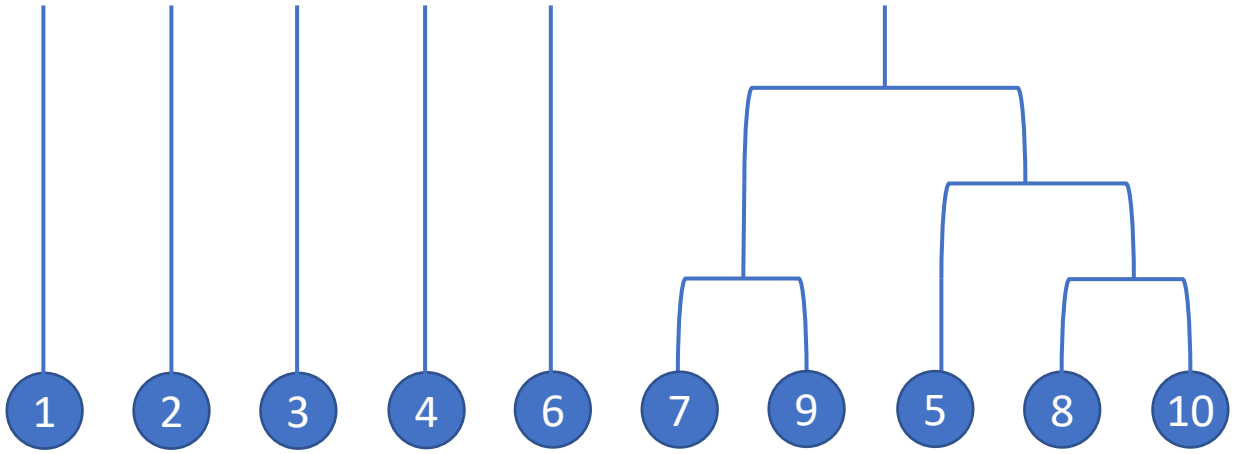
2번 문제 풀이

4차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	2	3	4	(5,(8,10))	6	(7,9)
1	0.000						
2	142.611	0.000					
3	278.410	167.003	0.000				
4	178.664	36.125	145.344	0.000			
(5,(8,10))	122.772	64.140	120.768	82.492	0.000		
6	195.931	105.190	83.528	100.419	38.833	0.000	
(7,9)	127.577	96.385	144.845	112.805	29.206	61.717	0.000

Distance가 가장 작은 (5,(8,10))과 (7,9) 를 합치게 된다. 업데이트 된 Dendrogram은 다음과 같다. 클러스터는 (1), (2), (3), (4), (6), ((7,9), (5,(8,10)))이 존재한다.
총 6개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 6개 cluster 〕

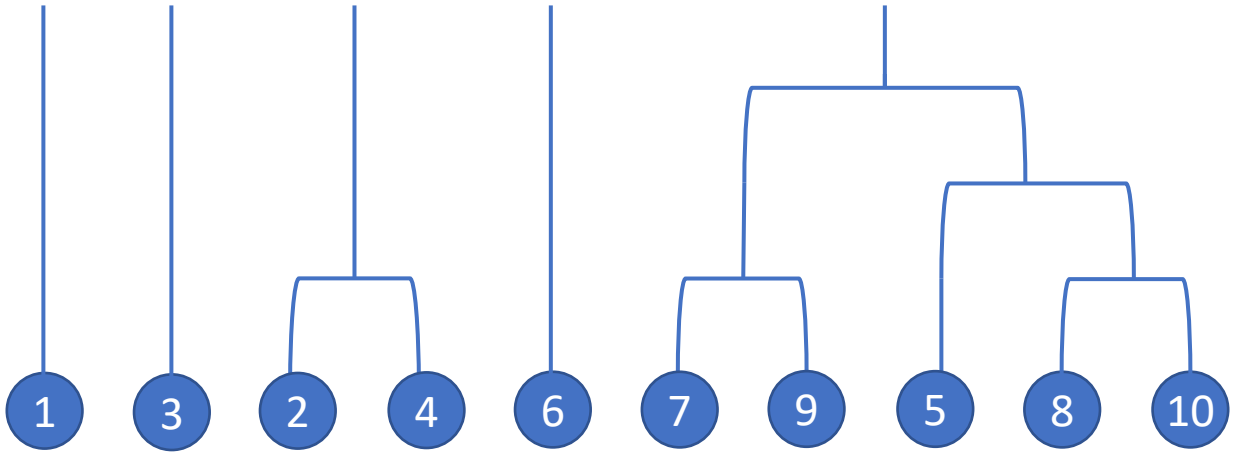
2번 문제 풀이

5차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	2	3	4	((7,9),(5,(8,10)))	6
1	0.000					
2	142.611	0.000				
3	278.410	167.003	0.000			
4	178.664	36.125	145.344	0.000		
((7,9),(5,(8,10)))	122.772	64.140	120.768	82.492	0.000	
6	195.931	105.190	83.528	100.419	38.833	0.000

Distance가 가장 작은 2와 4를 합치게 된다. 업데이트된 Dendrogram은 다음과 같다. 클러스터는 (1), (3), (2, 4), (6), ((7,9), (5,(8,10)))이 존재한다.
총 5개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 5개 cluster 〕

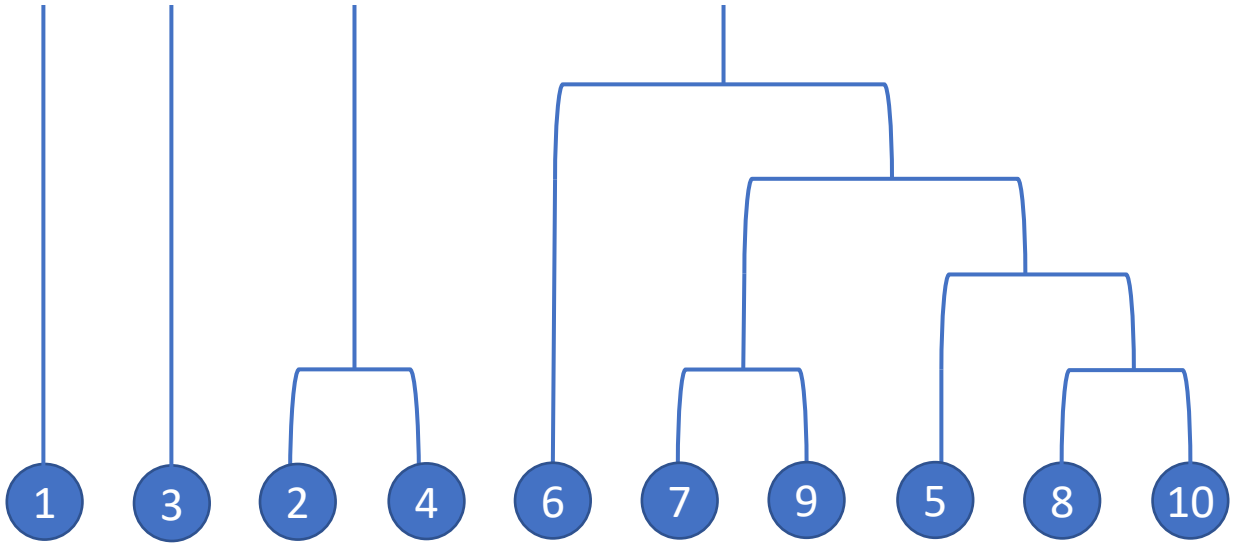
2번 문제 풀이

6차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	(2,4)	3	((7,9),(5,(8,10)))	6
1	0.000				
(2,4)	142.611	0.000			
3	278.410	145.344	0.000		
((7,9),(5,(8,10)))	122.772	64.140	120.768	0.000	
6	195.931	100.419	83.528	38.833	0.000

Distance가 가장 작은 6와 ((7,9),(5,(8,10)))를 합치게 된다. 업데이트된 Dendrogram은 다음과 같다. 클러스터는 (1), (3), (2, 4), (6, ((7,9), (5,(8,10))))이 존재한다.
총 4개의 클러스터가 발생되므로 hierarchical clustering을 계속 진행한다.



〔 업데이트된 dendrogram: 총 4개 cluster 〕

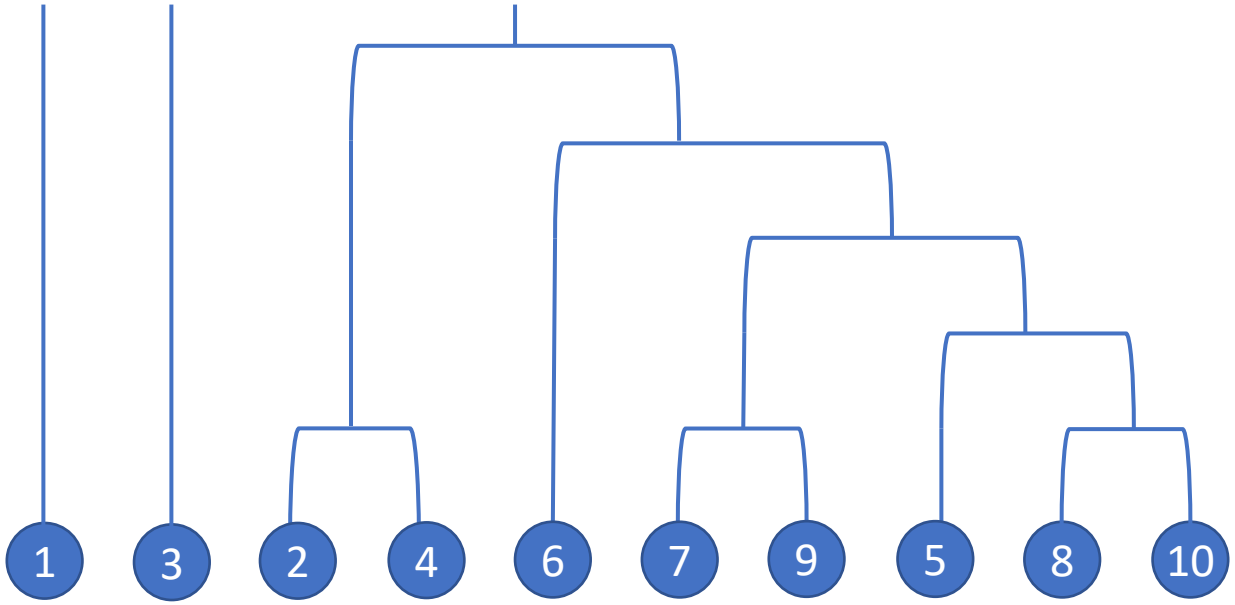
2번 문제 풀이

7차 시도.

- Proximity Matrix: MIN값을 기준으로 업데이트된 Proximity Matrix는 다음과 같다.

Cluster	1	2,4	3	((7,9),(5,(8,10)))
1	0.000			
(2,4)	142.611	0.000		
3	278.410	145.344	0.000	
((7,9),(5,(8,10)))	122.772	64.140	83.528	0.000

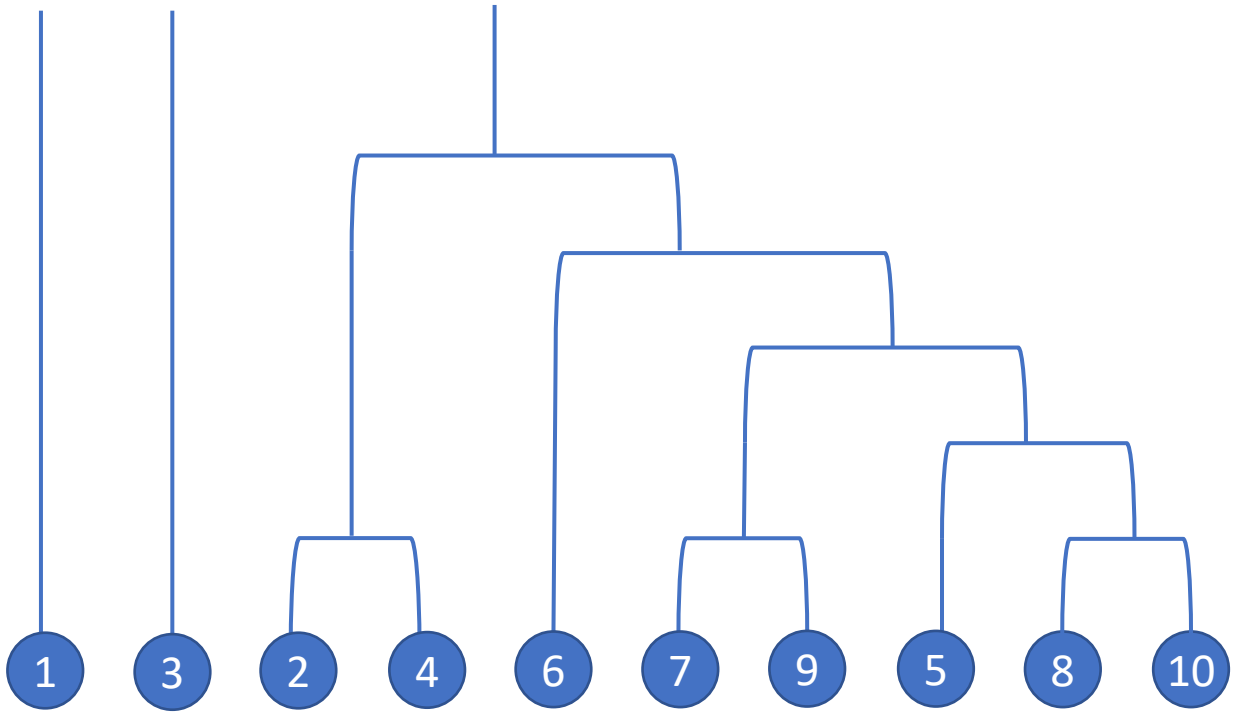
Distance가 가장 작은 (2,4)와 ((7,9),(5,(8,10)))를 합치게 된다. 업데이트된 Dendrogram은 다음과 같다. 클러스터는 (1), (3), ((2, 4), (6, ((7,9), (5,(8,10)))))이 존재한다.
총 3개의 클러스터가 발생되므로 hierarchical clustering을 마친다.



〔 업데이트된 dendrogram: 총 3개 cluster 〕

2번 문제 풀이

최종 Dendrogram결과



3번 문제 - 이론

두 개의 데이터 시퀀스 시퀀스 **a=[3, 7, 19],** 시퀀스 **b=[2, 4, 6, 15, 22]** 에 대한 Dynamic Time Warping Distance를 구하고, 두 시퀀스의 warping path 구현하기

Dynamic Time Warping Distance matrix를 계산한 결과는 다음과 같다.

Distance		2	4	6	15	22
	0	inf	inf	inf	inf	inf
3	inf	1	2	5	17	36
7	inf	6	4	3	11	26
19	inf	23	19	16	7	10

위에서 아래로 한 열이 완성되면 다음 열의 위에서 아래로 계산되는 방식으로 다음 순서와 같이 계산되었습니다.
*Cost는 강의교재와 동일하게 difference로 정의합니다.

1. $DTW(3,2) = cost(3,2) + minmum\ (0, inf, inf) = 1 + 0 = 1$
2. $DTW(7,2) = cost(7,2) + minmum\ (inf, 1, inf) = 5 + 1 = 6$
3. $DTW(19,2) = cost(19,2) + minmum(inf, 6, inf) = 17 + 6 = 23$

4. $DTW(3,4) = cost(4,3) + minmum(inf, inf, 1) = 1 + 1 = 2$
5. $DTW(7,4) = cost(7,4) + minimum(1,2,6) = 3 + 1 = 4$
6. $DTW(19,4) = cost(19,4) + minimum(6,4,23) = 15 + 4 = 19$

7. $DTW(3,6) = cost(3,6) + minimium(inf, inf, 2) = 3 + 2 = 5$
8. $DTW(7,6) = cost(7,6) + minimum(2,5,4) = 1 + 2 = 3$
9. $DTW(19,6) = cost(19,6) + minimum(4,3,19) = 13 + 3 = 16$
10. $DTW(3,15) = cost(3,15) + minimum(inf, inf, 5) = 12 + 5 = 17$
11. $DTW(7,15) = cost(7,15) + minimum(5,17,3) = 8 + 3 = 11$
12. $DTW(19, 15) = cost(19, 15) + minimum(3,11,16) = 4 + 3 = 7$

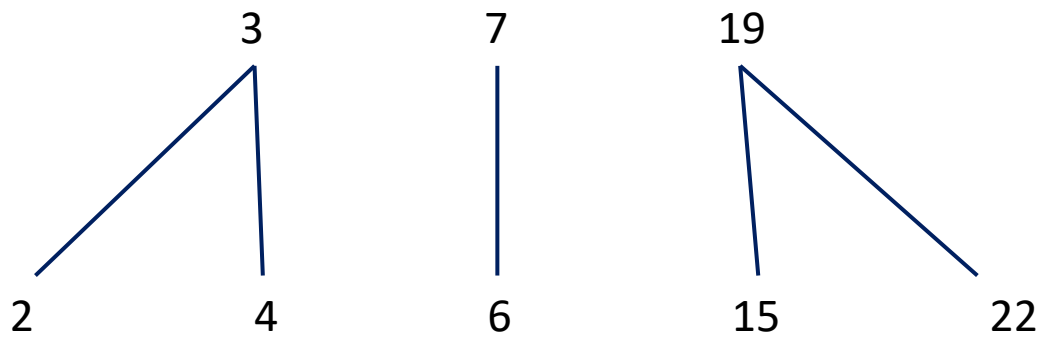
13. $DTW(3,22) = cost(3,22) + minimum(inf, inf, 17) = 19 + 17 = 36$
14. $DTW(7,22) = cost(7,22) + minimum(17, 36, 11) = 15! + 11 = 26$
15. $DTW(19, 22) = cost(19, 22) + minimum(11,26,7) = 3 + 7 = 10$

3번 문제 - 이론

Dynamic Time Warping Distance Matrix를 바탕으로 Warping path를 만들고자 합니다.
최소 warping path를 Matrix상에 표현해보면 다음과 같으며, 이를 바탕으로 warping path를 정의하면 다음과 같습니다.

Distance		2	4	6	15	22
	0	inf	inf	inf	inf	inf
3	inf	1	2	5	17	36
7	inf	6	4	3	11	26
19	inf	23	19	16	7	10

〔Dynamic Time Warping Distance Matrix〕



〔이론상 최종 Warping Path〕

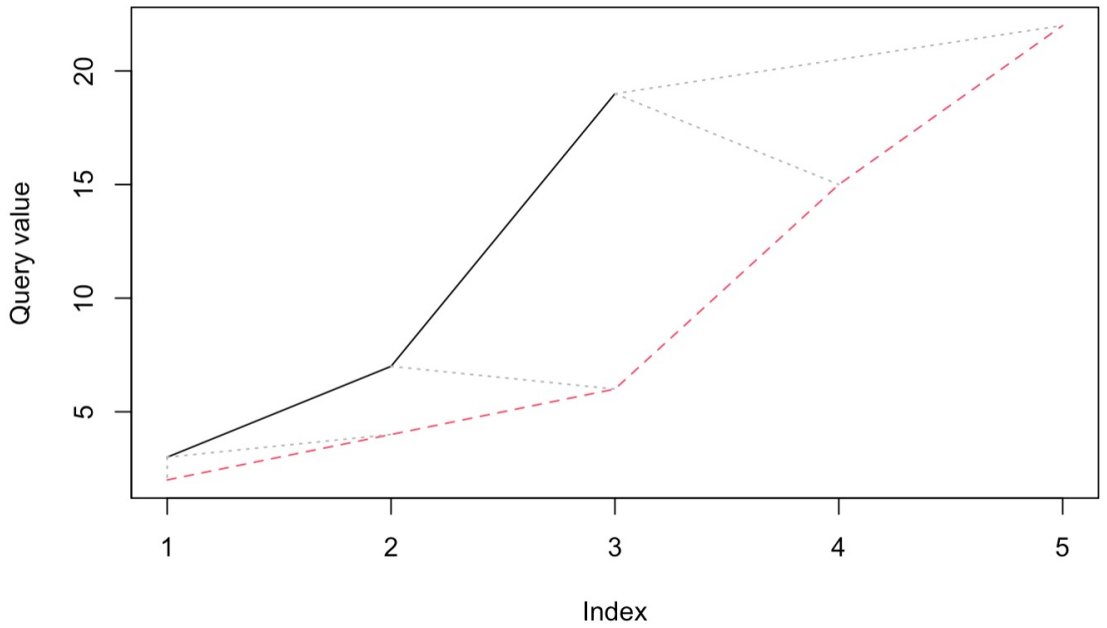
3번 문제 - 실습 코드

강의노트에서 구현한 Matrix의 DTW를 구하는게 아닌 시퀀스 2개만의 DTW를 계산하는 방법이므로 다음과 같이 구현하였습니다.
왼쪽은 구현된 코드(문제3.R)이며 오른쪽은 구현시 나타나는 DTW warping path 입니다.

문제 3.R

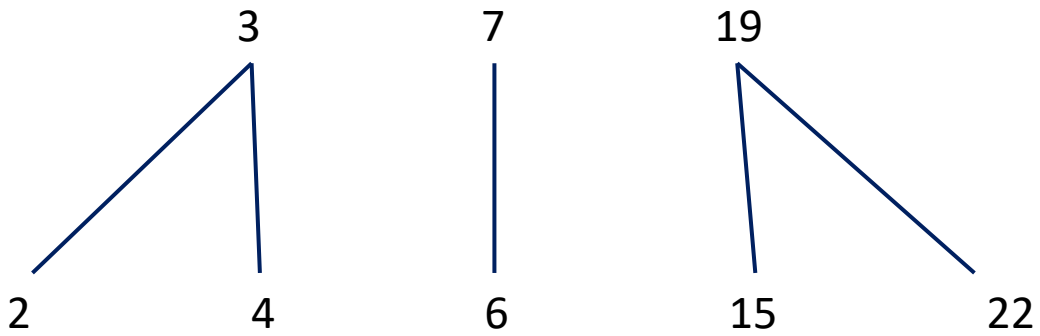
```
rm(list = ls())                 #초기 세팅
#install.packages("dtw")
library(dtw)                    #DTW 라이브러리 불러오기
library(proxy)                  #계산에 도움이 되는 추가 라이브러리
a = c(3,7,19)                   #시퀀스 a 생성
b = c(2,4,6,15,22)              #시퀀스 b 생성
dtww = dtw(a,b,keep=T)          #둘 만의 DTW 생성
summary(dtww)
dtwPlotTwoWay(dtww)             #오른쪽 plot 결과
dtwPlotThreeWay(dtww)
```

추가.
정수형 난수 생성 코드: `sample(x=범위, size=개수, replace=FALSE)`
예시코드. `sample(x=1:45, size=6, replace=FALSE)`
*본 문제에서는 시퀀스 a=[3, 7, 19], 시퀀스 b=[2, 4, 6, 15, 22]로 설정하였습니다.

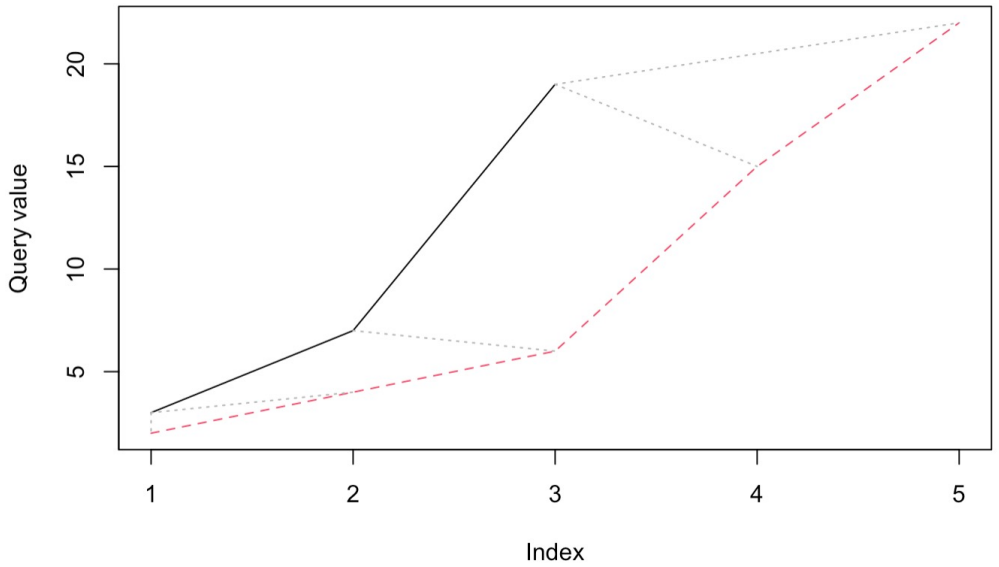


(실습 상 Warping Path, dtwPlotTwoWay 결과)

3번 문제 - 이론 & 실습 결과 비교



(이론 상 Warping Path)



(실습 상 Warping Path, dtwPlotTwoWay 결과)

실습 상 warping path를 보면, 검은 선이 시퀀스 a, 빨간 선이 시퀀스 b를 가르킨다.

이론 상의 warping path와 실습 상 warping path를 비교해보면 **동일한 결과를 보여줌**을 알 수 있다.