INFS 4203 / 7203 Data Mining

Tutorial 4: Clustering

T4-Q1: Calculating distance

Calculate the L1, L2, L-infinite norms of two points: (1, 0 ,5) and (2, 4, 9)

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}|^p \right)^{1/p}$$

Minkowski Distance

- d is the number of attributes
- x_{iu} is the u-th element of vector x_i
- x_{ju} is the u-th element of vector x_j

L_1 distance (Manhattan distance)

- $p = 1$

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{u=1}^{d} |x_{iu} - x_{ju}|$$

- For (1, 0 ,5) and (2, 4, 9)
  - d = 3
  - $x_{i1} = 1, x_{i2} = 0, x_{i3} = 5$
  - $x_{j1} = 2, x_{j2} = 4, x_{j3} = 9$
  - $dist(x_i, x_j) = |1 - 2| + |0 - 4| + |5 - 9| = 9$

L_2 distance (Euclidean distance)

0 ,5) and (2, 4, 9)

3

$= 1, x_{i2} = 0, x_{i3} = 5$

$= 2, x_{j2} = 4, x_{j3} = 9$

$\left(x_i, x_j\right) = (|1-2|^2 + |0-4|^2 + |5-9|^2)^{1/2} = \sqrt{33}$

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{d} |x_{iu} - x_{ju}|^2\right)^{1/2}$$

L_∞ distance (Chebyshev distance)

- $p = \infty$

$$dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \max_{u=\{1,2,\ldots,d\}} |x_{iu} - x_{ju}|$$

- For (1, 0 ,5) and (2, 4, 9)
  - d = 3
  - $x_{i1} = 1, x_{i2} = 0, x_{i3} = 5$
  - $x_{j1} = 2, x_{j2} = 4, x_{j3} = 9$
  - $dist(x_i, x_j) = \max(|1-2|, |0-4|, |5-9|) = 4$

T4-Q2: AGNES algorithm

Suppose the data mining task is to use agglomerative clustering to group measurements of the variable age.

- Age = {18, 28, 22, 33, 40, 48}

Commute the step-by-step agglomerative grouping using:

a. Single linkage

b. Complete linkage

c. Average linkage

with **Manhattan distance**

Algorithm

1. Compute the *distance matrix*
2. Let *each* data point be a cluster

Repeat

3. *merge* the two closest clusters
4. *update* the distance matrix

Until only a *single* cluster remains

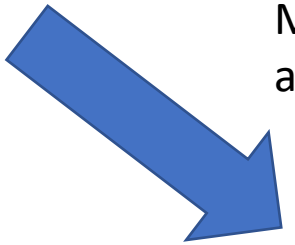
Step 1: Compute the distance matrix

	18	28	22	33	40	48
18						
28						
22						
33						
40						
48						

	18	28	22	33	40	48
18	0	10	4	15	22	30
28	10	0	6	5	12	20
22	4	6	0	11	18	26
33	15	5	11	0	7	15
40	22	12	18	7	0	8
48	30	20	26	15	8	0

Single linkage: Round 1

	18	28	22	33	40	48
18	0	10	4	15	22	30
28	10	0	6	5	12	20
22	4	6	0	11	18	26
33	15	5	11	0	7	15
40	22	12	18	7	0	8
48	30	20	26	15	8	0



Merge 18 and 22

	18, 22	28	33	40	48
18, 22	0	6	11	18	26
28	6	0	5	12	20
33	11	5	0	7	15
40	18	12	7	0	8
48	26	20	15	8	0

Single linkage: Round 2

	18, 22	28	33	40	48
18, 22	0	6	11	18	26
28	6	0	5	12	20
33	11	5	0	7	15
40	18	12	7	0	8
48	26	20	15	8	0

Merge 33 and 28

	18, 22	28, 33	40	48
18, 22	0	6	18	26
28, 33	6	0	7	15
40	18	7	0	8
48	26	15	8	0

Single linkage: Round 3

	18, 22	28, 33	40	48
18, 22	0	6	18	26
28, 33	6	0	7	15
40	18	7	0	8
48	26	15	8	0

Merge {18, 22} and {28, 33}

	18, 22, 28, 33	40	48
18, 22, 28, 33	0	7	15
40	7	0	8
48	15	8	0

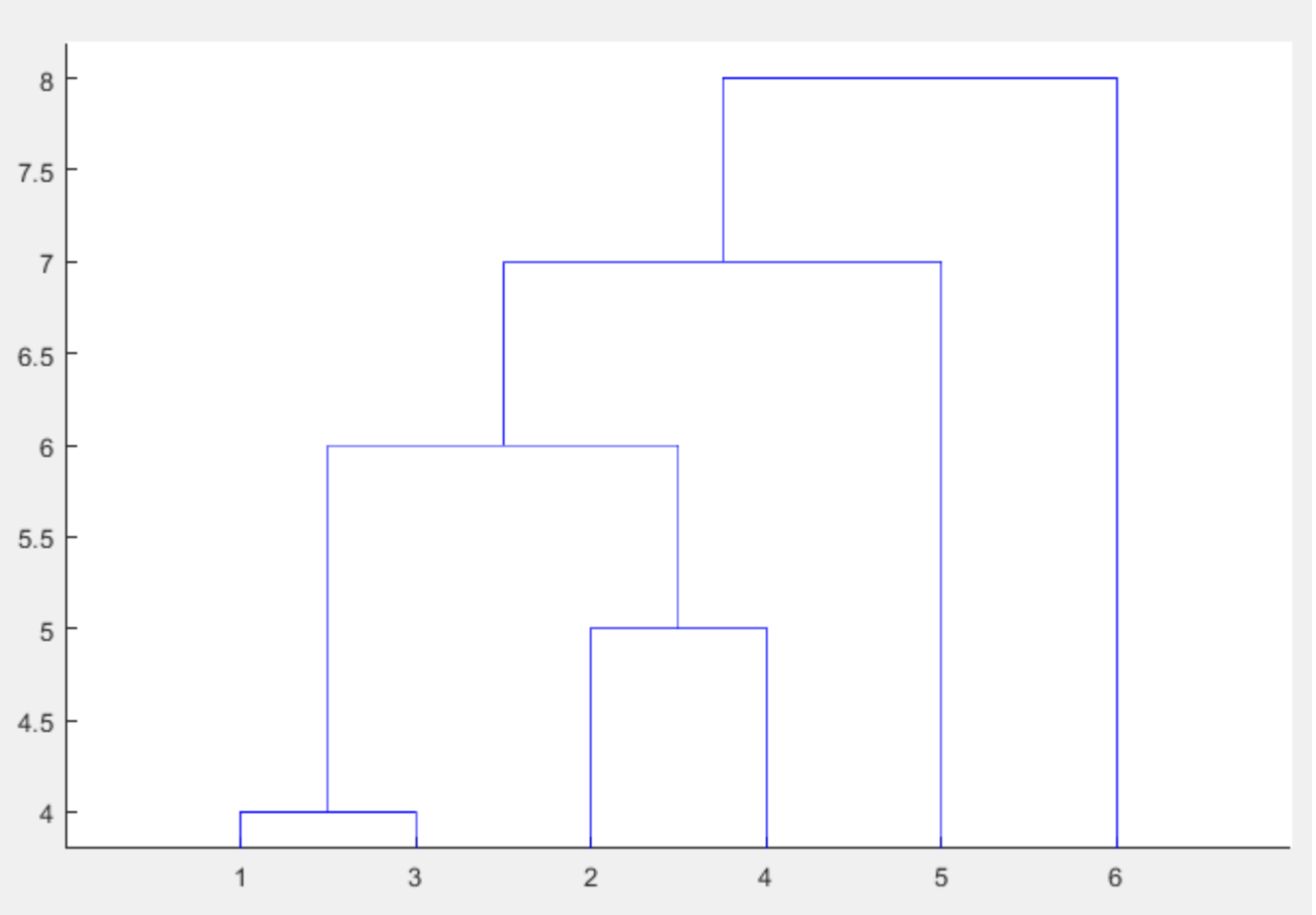
Single linkage: Round 4

	18, 22, 28, 33	40	48
18, 22, 28, 33	0	7	15
40	7	0	8
48	15	8	0

Merge {18, 22, 28, 33} and 40

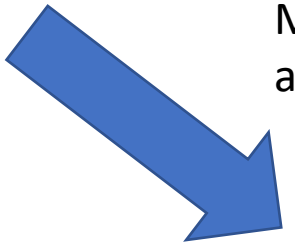
	18, 22, 28, 33, 40	48
18, 22, 28, 33, 40	0	8
48	8	0

Single linkage: Dendrogram



Complete linkage: Round 1

	18	28	22	33	40	48
18	0	10	4	15	22	30
28	10	0	6	5	12	20
22	4	6	0	11	18	26
33	15	5	11	0	7	15
40	22	12	18	7	0	8
48	30	20	26	15	8	0



Merge 18 and 22

	18, 22	28	33	40	48
18, 22	0	10	15	22	30
28	10	0	5	12	20
33	15	5	0	7	15
40	22	12	7	0	8
48	30	20	15	8	0

Complete linkage: Round 2

	18, 22	28	33	40	48
18, 22	0	10	15	22	30
28	10	0	5	12	20
33	15	5	0	7	15
40	22	12	7	0	8
48	30	20	15	8	0

Merge 33 and 28

	18, 22	28, 33	40	48
18, 22	0	15	22	30
28, 33	15	0	12	20
40	22	12	0	8
48	30	20	8	0

Complete linkage: Round 3

	18, 22	28, 33	40	48
18, 22	0	15	22	30
28, 33	15	0	12	20
40	22	12	0	8
48	30	20	8	0

Merge 48 and 40

	18, 22	28, 33	40, 48
18, 22	0	15	30
28, 33	15	0	20
40, 48	30	20	0

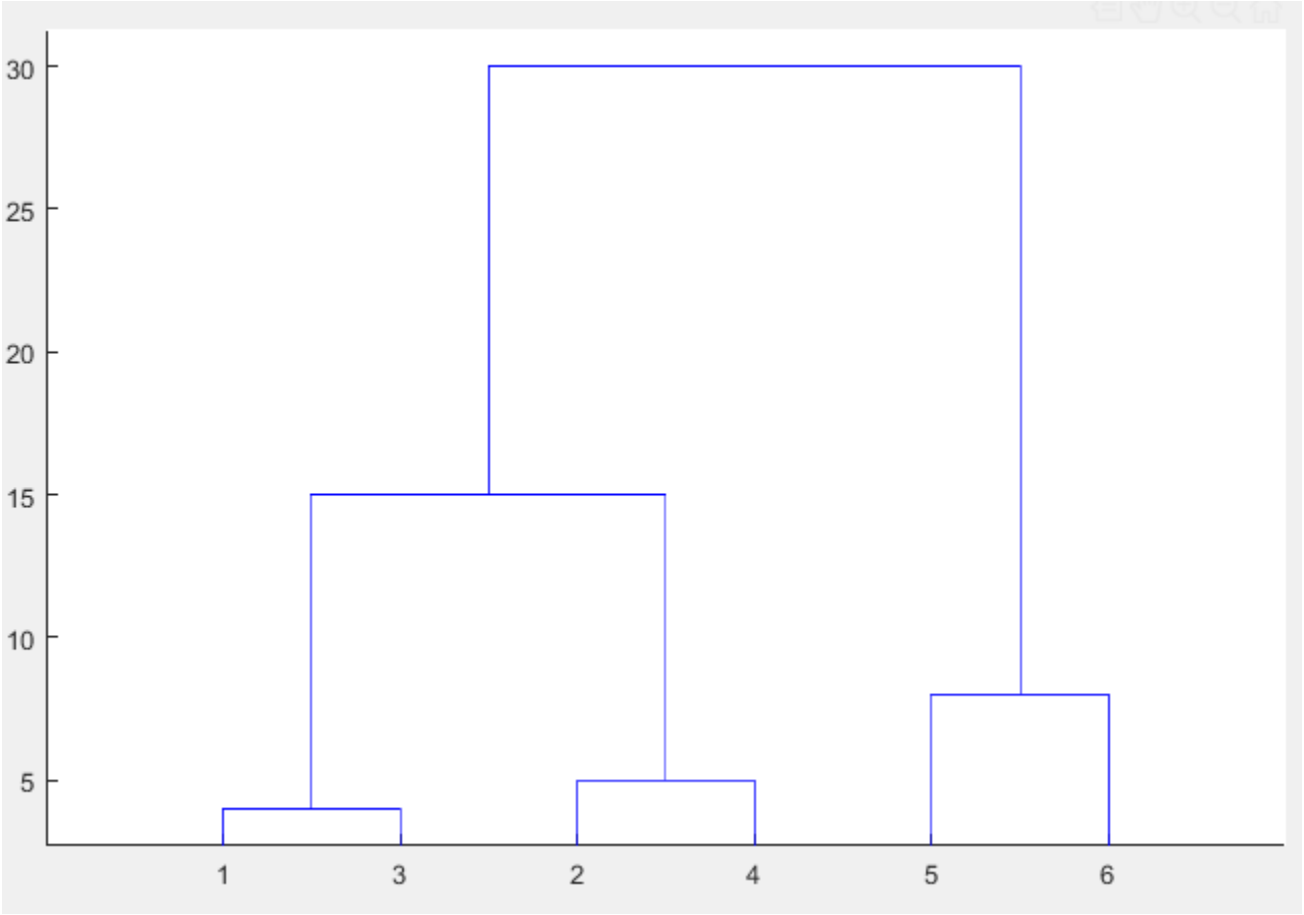
Complete linkage: Round 4

	18, 22	28, 33	40, 48
18, 22	0	15	30
28, 33	15	0	20
40, 48	30	20	0

Merge {18, 22} and {28, 33}

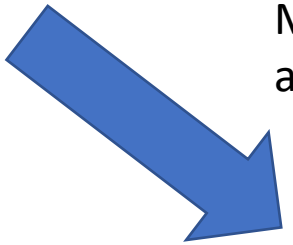
	18, 22, 28, 33	40, 48
18, 22, 28, 33	0	30
40, 48	30	0

Complete linkage: Dentrogram



Average linkage: Round 1

	18	28	22	33	40	48
18	0	10	4	15	22	30
28	10	0	6	5	12	20
22	4	6	0	11	18	26
33	15	5	11	0	7	15
40	22	12	18	7	0	8
48	30	20	26	15	8	0



Merge 18 and 22

	18, 22	28	33	40	48
18, 22	0	8	13	20	28
28	8	0	5	12	20
33	13	5	0	7	15
40	20	12	7	0	8
48	28	20	15	8	0

Average linkage: Round 2

	18, 22	28	33	40	48
18, 22	0	8	13	20	28
28	8	0	5	12	20
33	13	5	0	7	15
40	20	12	7	0	8
48	28	20	15	8	0

Merge 33 and 28

	18, 22	28, 33	40	48
18, 22	0	10.5	20	28
28, 33	10.5	0	9.5	17.5
40	20	9.5	0	8
48	28	17.5	8	0

Average linkage: Round 3

	18, 22	28, 33	40	48
18, 22	0	10.5	20	28
28, 33	10.5	0	9.5	17.5
40	20	9.5	0	8
48	28	17.5	8	0

Merge 48 and 40

	18, 22	28, 33	40, 48
18, 22	0	10.5	24
28, 33	10.5	0	13.5
40, 48	24	13.5	0

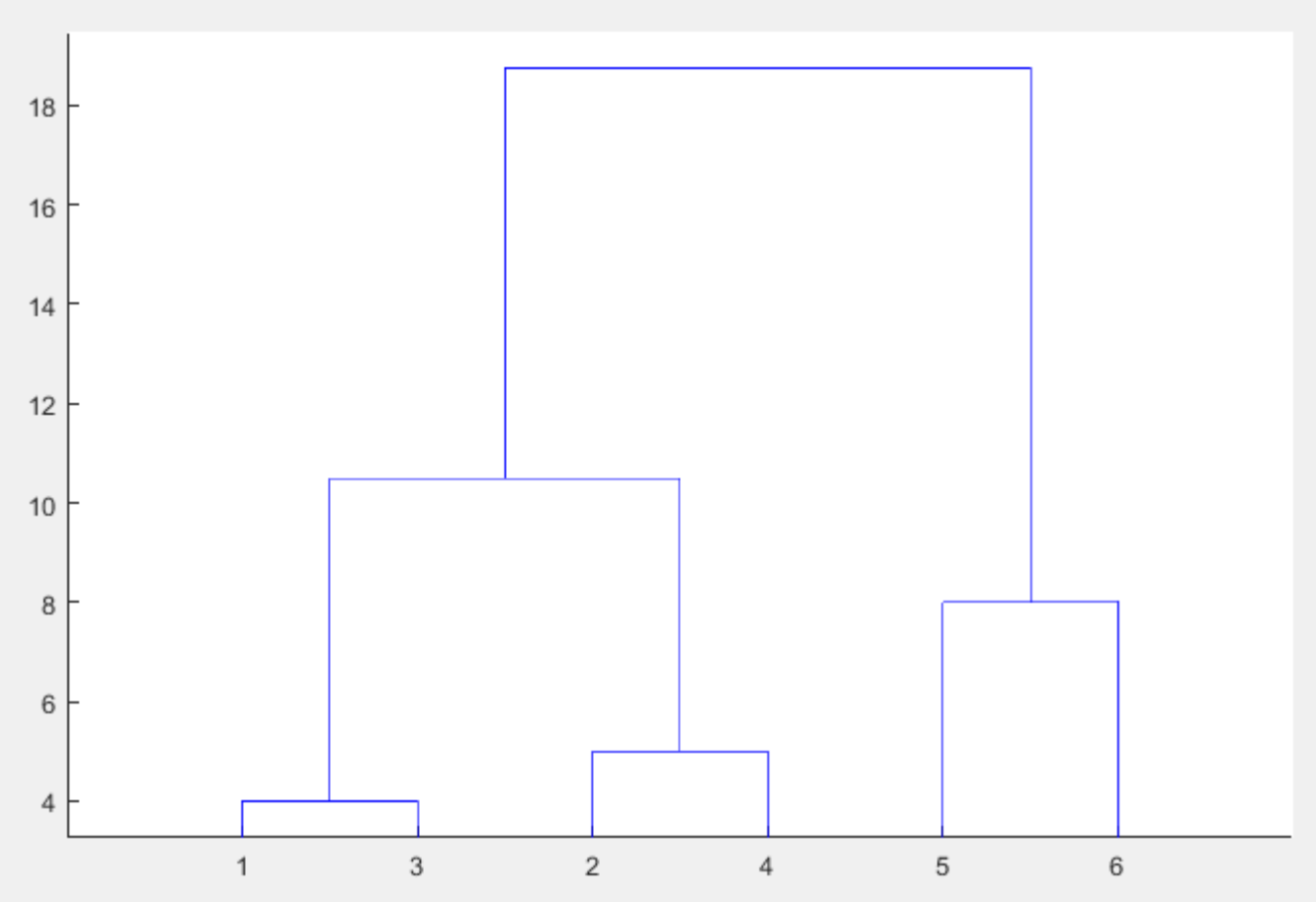
Average linkage: Round 4

	18, 22	28, 33	40, 48
18, 22	0	10.5	24
28, 33	10.5	0	13.5
40, 48	24	13.5	0

Merge {18, 22} and {28, 33}

	18, 22, 28, 33	40, 48
18, 22, 28, 33	0	18.75
40, 48	18.75	0

Average linkage: Dentrogram



T4-Q3: *k*-means and SSE calculation

Suppose the data mining task is to cluster the following measurements

Age = {18, 22, 25, 42, 27, 43, 33, 35, 56, 28} into **three** groups. For initial centroid of {18, 27, 35}:

a. Use the *k*-means algorithm to show the clustering procedures step by step;

b. Calculate the final SSE using **Manhattan distance**

K-means algorithm

1. --Select *k* points as the initial centroids
2. Repeat
3. --Form *k* clusters by assigning points to the closest centroid
4. --Update the centroid of each cluster
5. Until --The cluster assignment don’t change

k-means: Round 1

Cluster	Centroid	Cluster Elements	Updated Centroid
1	18	18, 22	20
2	27	25, 27, 28	26.7
3	35	33, 35, 42, 43, 56	41.8

k-means: Round 2

Cluster	Centroid	Cluster Elements	Updated Centroid
1	20	18, 22	20
2	26.7	25, 27, 28, 33	28.25
3	41.8	35, 42, 43, 56	44

k-means: Round 3

Cluster	Centroid	Cluster Elements	Updated Centroid
1	20	18, 22	20
2	28.25	25, 27, 28, 33, 35	29.6
3	44	42, 43, 56	47

k-means: Round 4

Cluster	Centroid	Cluster Elements	Updated Centroid
1	20	18, 22	20
2	29.6	25, 27, 28, 33, 35	29.6
3	47	42, 43, 56	47

Calculate the SSE

Cluster	Centroid	Cluster Elements	Updated Centroid
1	20	18, 22	20
2	29.6	25, 27, 28, 33, 35	29.6
3	47	42, 43, 56	47

$$SSE = \sum_{i=1}^{3} \sum_{x \in C_i} dist(x, c_i)^2$$

$$= (20 - 18)^2 + (22 - 20)^2$$
$$+(25 - 29.6)^2 + (27 - 29.6)^2 + (28 - 29.6)^2 + (33 - 29.6)^2 + (35 - 29.6)^2$$
$$+(42 - 47)^2 + (43 - 47)^2 + (56 - 47)^2 = 201.2$$

Thanks for your attention