Tut 11: Clustering with Minkowski Distance

Feb 2025

Minkowski Distance

1. (May 2020 Final Q3(b)) Given an appropriate example to explain why the Minkowski distance

$$M(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^r \right)^{\frac{1}{r}}, \quad \boldsymbol{x},\ \boldsymbol{y} \in \mathbb{R}^p$$

will no longer be a distance function when $r = \frac{1}{2}$. (2 marks)

Solution. Note that when $0 < r = \frac{1}{2} < 1$, the nonnegativity and symmetric property will be true. .... [0.5 mark]

So, we need to show that it violates the triangle inequality. .... [0.5 mark]

Let $p = 2$ and consider three points $(0, 0)$, $(1, 0)$, $(5, 4)$, therefore,

$$M((0, 0), (1, 0)) = (|0 - 0|^{1/2} + |0 - 1|^{1/2})^2 = 1$$

$$M((1, 0), (5, 4)) = (|1 - 5|^{1/2} + |0 - 4|^{1/2})^2 = (2 + 2)^2 = 16$$

However,

$$\begin{aligned} M((0, 0), (5, 4)) &= (|0 - 5|^{1/2} + |0 - 4|^{1/2})^2 \\ &= 9 + 4 \times \sqrt{5} > M((0, 0), (1, 0)) + M((1, 0), (5, 4)). \end{aligned}$$ [1 mark]

□

K-means Floyd Algorithm with Minkowski Distance

Note that K-means is theoretically founded on Euclidean distance but not Minkowski distance.

2. (Jan 2022 Final Q5(b)) Given the three-dimensional points in Table 5.2,

Table 5.2: Three-dimensional points.

Label	x_1	x_2	x_3
P_1	3.3	4.4	2.5
P_2	2.4	3.1	2.1
P_3	0.1	1.9	1.1
P_4	0.3	2.4	1.5
P_5	-0.6	1.1	1.1
P_6	-2.9	-0.1	0.1
P_7	4.3	6.4	5.5
P_8	3.4	5.1	5.1
P_9	1.1	3.9	4.1

Use the k-means clustering method with **Manhattan distance** to cluster the given points into $k = 3$ clusters by using P_5, P_4, P_7 as the initial clusters, find the **stable cluster centres**.

(8 marks)

Solution. Step 1 : Update the distance table based on the distance of each point to the initial cluster centres.

Point	x_1	x_2	x_3	Centre 1	Centre 2	Centre 3	Cluster label
P_1	3.3	4.4	2.5	8.6	6	6	2
P_2	2.4	3.1	2.1	6	3.4	8.6	2
P_3	0.1	1.9	1.1	1.5	1.1	13.1	2
P_4	0.3	2.4	1.5	2.6	0	12	2
P_5	-0.6	1.1	1.1	0	2.6	14.6	1
P_6	-2.9	-0.1	0.1	4.5	7.1	19.1	1
P_7	4.3	6.4	5.5	14.6	12	0	3
P_8	3.4	5.1	5.1	12	9.4	2.6	3
P_9	1.1	3.9	4.1	7.5	4.9	7.1	2

. . . . [3 marks]

The new cluster centres are

$$C_1 = (-1.75, 0.5, 0.6), \quad C_2 = (1.44, 3.14, 2.26), \quad C_3 = (3.85, 5.75, 5.3)$$

. . . . [1 mark]

Step 2: Update the distance table based on the distance of each point to the updated cluster centres.

Point	x_1	x_2	x_3	Centre 1	Centre 2	Centre 3	Cluster label
P_1	3.3	4.4	2.5	10.85	3.36	4.7	2
P_2	2.4	3.1	2.1	8.25	1.16	7.3	2
P_3	0.1	1.9	1.1	3.75	3.74	11.8	2
P_4	0.3	2.4	1.5	4.85	2.64	10.7	2
P_5	-0.6	1.1	1.1	2.25	5.24	13.3	1
P_6	-2.9	-0.1	0.1	2.25	9.74	17.8	1
P_7	4.3	6.4	5.5	16.85	9.36	1.3	3
P_8	3.4	5.1	5.1	14.25	6.76	1.3	3
P_9	1.1	3.9	4.1	9.75	2.94	5.8	2

. . . . [3 marks]

The **stable cluster centres** are

$$C_1 = (-1.75, 0.5, 0.6), \quad C_2 = (1.44, 3.14, 2.26), \quad C_3 = (3.85, 5.75, 5.3)$$

. . . . [1 mark] □

3. (Final Exam Jan 2023, Q4(b)) Given the four-dimensional points in Table 4.2.

Obs.	x_1	x_2	x_3	x_4
P_1	3.77	2.09	4.88	4.58
P_2	1.37	1.75	1.80	2.22
P_3	2.31	3.13	2.50	1.34
P_4	0.17	1.29	1.54	3.57
P_5	4.75	3.27	6.36	3.00
P_6	3.46	4.42	4.08	5.43
P_7	0.21	1.93	0.78	2.72

Table 4.2: Four-dimensional points.

Use the k-means clustering method with **Manhattan distance** to cluster the given points into **three clusters** by using P_6, P_4, P_2 as the initial centres, find the **stable cluster centres**.

(9 marks)

Solution. Given the initial centres:

$$Centre_1 = P_6(3.46, 4.42, 4.08, 5.43),$$
$$Centre_2 = P_4(0.17, 1.29, 1.54, 3.57),$$
$$Centre_3 = P_2(1.37, 1.75, 1.80, 2.22)$$

Step 1 : Update table based on distance to cluster centres

x_1	x_2	x_3	x_3	dist.1	dist.2	dist.3	label
3.77	2.09	4.88	4.58	4.29	8.75	8.18	1
1.37	1.75	1.80	2.22	10.25	3.27	0	3
2.31	3.13	2.50	1.34	8.11	7.17	3.9	3
0.17	1.29	1.54	3.57	10.82	0	3.27	2
4.75	3.27	6.36	3.00	7.15	11.95	10.24	1
3.46	4.42	4.08	5.43	0	10.82	10.25	1
0.21	1.93	0.78	2.72	11.75	2.29	2.86	2

. . . . [4 marks]

The new cluster centres are . . . . [1 mark]

$$Centre_1 = (3.9933, 3.2600, 5.1067, 4.3367),$$
$$Centre_2 = (0.19, 1.61, 1.16, 3.145),$$
$$Centre_3 = (1.84, 2.44, 2.15, 1.78)$$

Step 2 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1^2	dist.2^2	dist.3^2	label^2	label^3
3.77	2.09	4.88	4.58	1.8633	9.215	7.81	1
1.37	1.75	1.80	2.22	9.5567	2.885	1.95	3
2.31	3.13	2.50	1.34	7.4167	6.785	1.95	3
0.17	1.29	1.54	3.57	10.1267	1.145	5.22	2
4.75	3.27	6.36	3.00	3.3567	11.565	9.17	1
3.46	4.42	4.08	5.43	3.8133	11.285	9.18	1
0.21	1.93	0.78	2.72	11.0567	1.145	4.45	2

. . . . [3 marks]

The cluster centres stabilises to the stable cluster centres

$$Centre_1 = (3.9933, 3.2600, 5.1067, 4.3367),$$
$$Centre_2 = (0.19, 1.61, 1.16, 3.145),$$
$$Centre_3 = (1.84, 2.44, 2.15, 1.78)$$

[1 mark]

Average: 3.47 / 9 marks in Jan 2023; 18% below 4.5 marks.

□

Agglomerative Hierarchical Clustering with Minkowski Distance

4. (May 2020 Final Q3(c)) Group the observations in Table 3.1 using hierarchical clustering and the **Minkowski distance** with $r = 3$ (refer to part (b) for the definition of Minkowski distance) and **complete linkage** and draw the dendrogram formed by the hierarchical clustering.

Table 3.1: Unlabelled data.

Obs	x_1	x_2	x_3
A	1	3	2
B	5	7	9
C	6	9	8
D	7	8	9
E	2	3	5
F	1	4	3

(4 marks)

Solution. First, we construct the distance matrix using the Minkowski distance with $r = 3$:

	A	B	C	D	E	F
A	0					
B	7.7805	0				
C	8.2278	2.1544	0			
D	8.8109	2.0801	1.4422	0		
E	3.0366	5.3717	6.7460	6.7969	0	
F	1.2599	6.7460	7.2112	7.9158	2.1544	0

. . . . [1 mark]

Height = 1.2599; Cluster: A, F

	A,F	B	C	D	E
A,F	0				
B	7.7805	0			
C	8.2278	2.1544	0		
D	8.8109	2.0801	1.4422	0	
E	3.0366	5.3717	6.7460	6.7969	0

. . . . [0.5 mark]

Height = 1.4422; Cluster: C, D

	A,F	B	C,D	E
A,F	0			
B	7.7805	0		
C,D	8.8109	2.1544	0	
E	3.0366	5.3717	6.7969	0

. . . . [0.5 mark]

Height = 2.1544; Cluster: B, (C, D)

	A,F	B,C,D	E
A,F	0		
B,C,D	8.8109	0	
E	3.0366	6.7969	0

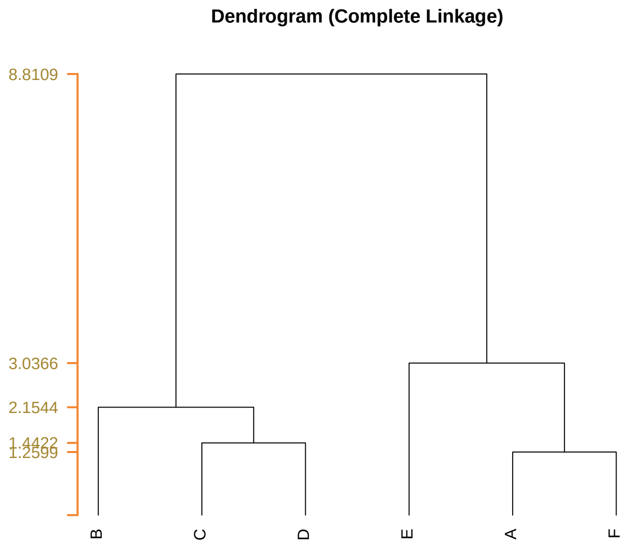
. . . . [0.5 mark]

Height = 3.0366; Cluster: A, (F, E)

	A,F	B,C,D	E
A,F,E	0		
B,C,D	8.8109		

. . . . [0.5 mark]

With the above information, we can construct a nice dendrogram (marks will be deducted without appropriate labels).



. . . . [1 mark] □

5. (Jan 2021 Final Q4(a). Hand calculation is possible but Excel/R is recommended) Group the observations in Table 4.1 using hierarchical clustering and the **Manhattan distance** and **single linkage** and draw the dendrogram formed by the hierarchical clustering.

Table 4.1: Unlabelled data.

Obs	x_1	x_2
A	-2.68	-2.02
B	3.06	-0.83
C	1.91	1.57
D	-1.06	-0.88
E	0.49	2.42
F	0.83	1.75
G	-0.71	-0.84
H	-2.01	-1.92

(5 marks)

Solution. The first step is to construct the distance matrix using the Manhattan distance:

	A	B	C	D	E	F	G	H
A	0							
B	6.93	0						
C	8.18	3.55	0					
D	2.76	4.17	5.42	0				
E	7.61	5.82	2.27	4.85	0			
F	7.28	4.81	1.26	4.52	1.01	0		
G	3.15	3.78	5.03	0.39	4.46	4.13	0	
H	0.77	6.16	7.41	1.99	6.84	6.51	2.38	0

. . . . [1.5 marks]

The height is $\boxed{0.39}$. Cluster: D, G.

	A	B	C	D,G	E	F	H
A	0						
B	6.93	0					
C	8.18	3.55	0				
D,G	2.76	3.78	5.03	0			
E	7.61	5.82	2.27	4.46	0		
F	7.28	4.81	1.26	4.13	1.01	0	
H	0.77	6.16	7.41	1.99	6.84	6.51	0

. . . . [0.5 mark]

The height is $\boxed{0.77}$. Cluster: A, H.

	A,H	B	C	D,G	E	F
A,H	0					
B	6.16	0				
C	7.41	3.55	0			
D,G	1.99	3.78	5.03	0		
E	6.84	5.82	2.27	4.46	0	
F	6.51	4.81	1.26	4.13	1.01	0

. . . . [0.5 mark]

The height is $\boxed{1.01}$. Cluster: E, F.

	A,H	B	C	D,G	E,F
A,H	0				
B	6.16	0			
C	7.41	3.55	0		
D,G	1.99	3.78	5.03	0	
E,F	6.51	4.81	1.26	4.13	0

. . . . [0.5 mark]

The height is $\boxed{1.26}$. Cluster: C, (E,F).

	A,H	B	C,(E,F)	D,G
A,H	0			
B	6.16	0		
C,(E,F)	6.51	3.55	0	
D,G	1.99	3.78	4.13	0

. . . . [0.5 mark]

The height is $\boxed{1.99}$. Cluster: C, (E,F).

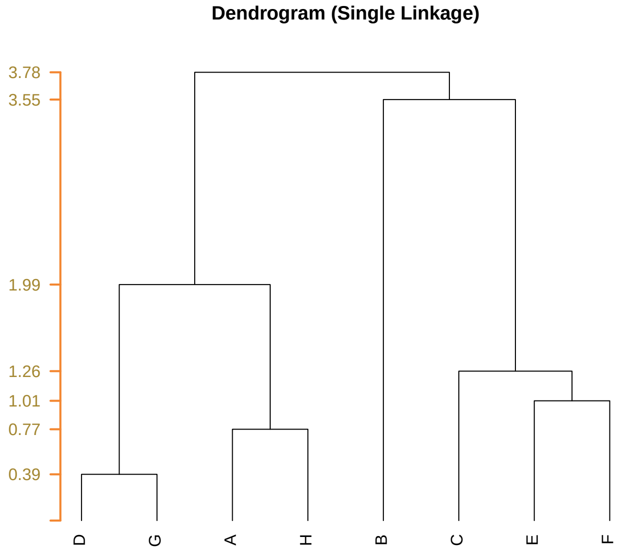
	A,H,D,G	B	C,(E,F)
A,H,D,G	0		
B	3.78	0	
C,(E,F)	4.13	3.55	0

. . . . [0.3 mark]

The height is $\boxed{3.55}$. Cluster: C, (E,F).

	A,H,D,G	B,C,(E,F)
A,H,D,G	0	
B,C,(E,F)	3.78	0

. . . . [0.2 mark]



Marks are deducted for the lack of proper labels. . . . . [1 mark] □

6. (Final Exam May 2024 Sem, Q5(b)) Given a three-dimensional mixture of Gaussian data X in Table 5.2.

Table 5.2: three-dimensional mixture of Gaussian data X

Obs.	x_1	x_2	x_3
A	3.19	2.25	1.73
B	4.90	4.17	5.35
C	0.64	2.35	0.52
D	-0.79	0.56	1.30
E	0.38	3.72	-0.49
F	1.09	1.87	0.96

Suppose the computer output of R command `dist(`X`, method="minkowski", p=3))` is listed below.

```
      A     B     C     D     E
B 3.904
C 2.638 5.809
D 4.081 6.677 2.090
E 3.311 6.631 1.535 3.388
F 2.138 5.338 0.660 2.075 2.135
```

Use the computer output to find the hierarchical clustering with **average linkage** for the data X in Table 5.2 and then **draw the dendrogram** with appropriate labels. (13 marks)

Solution. The minimum distance is 0.660, so C and F should be grouped. . . . . [1 mark]

```
          A      B      CF     D      E
A
B      3.904
CF     2.388  5.5735
D      4.081  6.677   2.0825
E      3.311  6.631   1.835   3.388
```

. . . . [2 marks]

The minimum distance is 1.835, so CF and E should be grouped. . . . . [1 mark]

```
          A        B     CFE       D
A
B      3.904
```

```
CFE    2.695667 5.926
D      4.081    6.677   2.517667
```

. . . . [2 marks]

The minimum distance is 2.517667$\approx$ 2.518, so CFE and D should be grouped. . . . [1 mark]

```
            A         B       CFED
A
B      3.904
CFED   3.042    6.11375
```

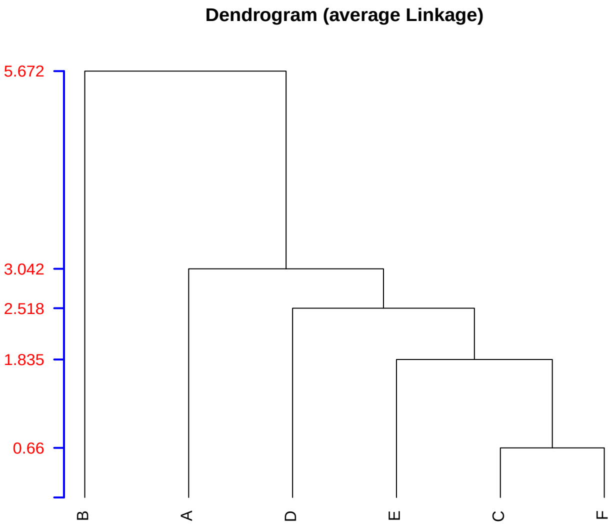
. . . . [2 marks]

The minimum distance is 3.042, so A and CFED should be grouped. . . . . [1 mark]

```
       ACFED         B
ACFED
B      5.6718
```

. . . . [1 mark]

We can now sketch the dendrogram:



. . . . [2 marks] □