Q1 [20 Marks]

Apply the agglomerative hierarchical clustering algorithm with the following distance matrix and the single linkage. Plot the cluster tree and mark out all the merging levels.

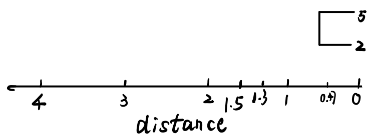
	1	2	3	4
2	2.33			
3	3.15	1.30		
4	1.90	1.50	3.70	
5	3.01	0.47	1.40	1.82

Table 1： distance matrix

$$\begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 2.33 & 0 & & & \\ 3.15 & 1.3 & 0 & & \\ 1.90 & 1.5 & 3.7 & 0 & \\ 3.01 & 0.47 & 1.4 & 1.82 & 0 \end{bmatrix} \end{array}$$

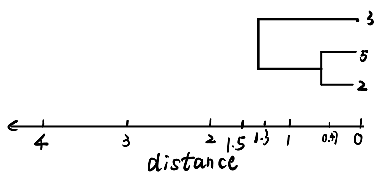
$$\begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & 2.33 & 3.15 & 1.9 & 3.01 \\ 2.33 & 0 & 1.3 & 1.5 & 0.47 \\ 3.15 & 1.3 & 0 & 3.7 & 1.4 \\ 1.90 & 1.5 & 3.7 & 0 & 1.82 \\ 3.01 & 0.47 & 1.4 & 1.82 & 0 \end{bmatrix} \end{array}$$

merge 2,5



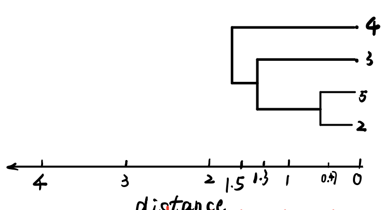
$$\begin{array}{cc} & \begin{array}{cccc} 1 & 3 & 4 & (2,5) \end{array} \\ \begin{array}{c} 1 \\ 3 \\ 4 \\ (2,5) \end{array} & \begin{bmatrix} 0 & 3.15 & 1.9 & 2.33 \\ 3.15 & 0 & 3.7 & 1.3 \\ 1.9 & 3.7 & 0 & 1.5 \\ 2.33 & 1.3 & 1.5 & 0 \end{bmatrix} \end{array}$$

merge 2,5,3



$$\begin{array}{cc} & \begin{array}{ccc} 1 & 4 & (2,3,5) \end{array} \\ \begin{array}{c} 1 \\ 4 \\ (2,3,5) \end{array} & \begin{bmatrix} 0 & 1.9 & 2.33 \\ 1.9 & 0 & 1.5 \\ 2.33 & 1.5 & 0 \end{bmatrix} \end{array}$$

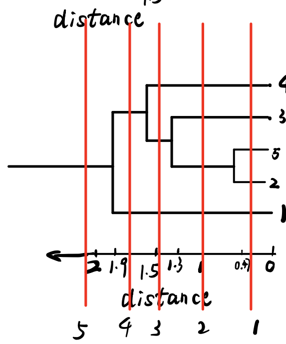
merge 2,3,5,4



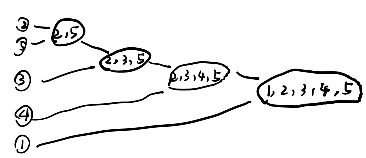
$$\begin{array}{cc} & \begin{array}{cc} 1 & (2,3,4,5) \end{array} \\ \begin{array}{c} 1 \\ (2,3,4,5) \end{array} & \begin{bmatrix} 0 & \\ 1.9 & 0 \end{bmatrix} \end{array}$$

merge 12345

Dendrogram.



merge level.



Q2 [20 Marks]

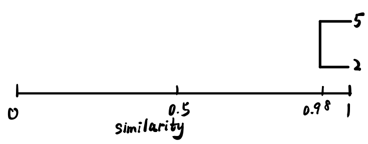
Use the similarity matrix in Table 2 to perform single-link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the clusters are merged.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

discover the maximum similarity

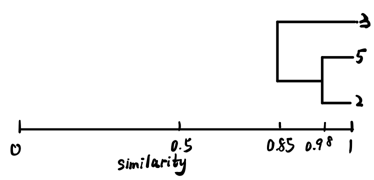
Table 2: Similarity matrix for Q2

$$\begin{array}{c|c} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 1 & 0.1 & 0.41 & 0.55 & 0.35 \\ 0.1 & 1 & 0.64 & 0.47 & 0.98 \\ 0.41 & 0.64 & 1 & 0.44 & 0.85 \\ 0.55 & 0.47 & 0.44 & 1 & 0.76 \\ 0.35 & 0.98 & 0.85 & 0.76 & 1 \end{bmatrix} \end{array}$$



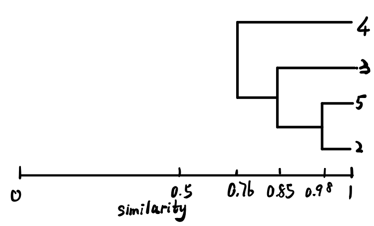
merge 2,5 ↓

$$\begin{array}{c|c} & \begin{matrix} 1 & 3 & 4 & (2,5) \end{matrix} \\ \begin{matrix} 1 \\ 3 \\ 4 \\ (2,5) \end{matrix} & \begin{bmatrix} 1 & & & \\ 0.41 & 1 & & \\ 0.55 & 0.44 & 1 & \\ 0.35 & 0.85 & 0.76 & 1 \end{bmatrix} \end{array}$$



merge 2,3,5 ↓

$$\begin{array}{c|c} & \begin{matrix} 1 & 4 & (2,3,5) \end{matrix} \\ \begin{matrix} 1 \\ 4 \\ (2,3,5) \end{matrix} & \begin{bmatrix} 1 & & \\ 0.55 & 1 & \\ 0.41 & 0.76 & 1 \end{bmatrix} \end{array}$$

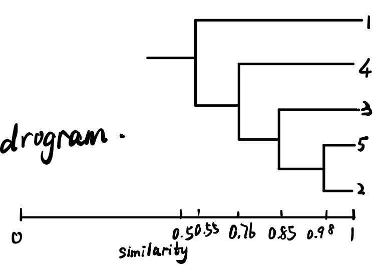


merge 1,2,3,5 ↓

$$\begin{array}{c|c} & \begin{matrix} 1 & (2,3,4,5) \end{matrix} \\ \begin{matrix} 1 \\ (2,3,4,5) \end{matrix} & \begin{bmatrix} 1 & \\ 0.55 & 1 \end{bmatrix} \end{array}$$

merge 12345

Dendrogram.



Q3 [30 Marks]

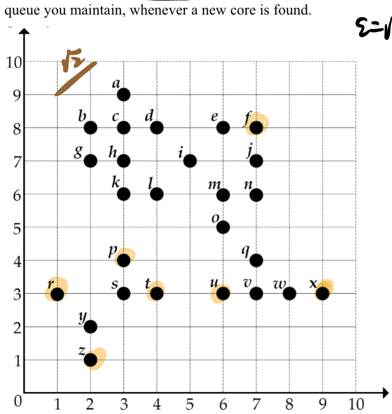
Apply DBSCAN with parameters MinPts=4 and Eps = $\sqrt{2}$ (ε) to get clustering results.

First, for every data point, answer if it is a core, a border, or an outlier.

Second, for data points that are not outliers, show the clusters detected.

Third, show your detailed steps of DBSCAN process, including the content of the queue you maintain, whenever a new core is found.

$\varepsilon = \sqrt{2}$



cancluate the $N(p)$, $\sqrt{2}$-neighborhood of point p

$N(a)=\{a,b,c,d\}$
$N(b)=\{b,a,c,g,h\}$
$N(c)=\{c,a,b,d,g,h\}$
$N(d)=\{d,a,c,h,i\}$
$N(e)=\{e,f,i,j\}$
$N(f)=\{f,e,j\}$
$N(g)=\{g,b,c,h,k\}$
$N(h)=\{h,b,c,d,g,k,l\}$
$N(i)=\{i,d,e,l,m\}$
$N(j)=\{j,e,f,m,n\}$
$N(k)=\{k,g,h,l\}$
$N(l)=\{l,h,i,k\}$
$N(m)=\{m,i,j,n,o\}$
$N(n)=\{n,j,m,o\}$
$N(o)=\{o,m,n,q\}$
$N(p)=\{p,s,t\}$
$N(q)=\{q,o,u,v,w\}$
$N(r)=\{r,y\}$
$N(s)=\{s,p,t,y\}$
$N(t)=\{t,p,s\}$
$N(u)=\{u,q,v\}$
$N(v)=\{v,q,u,w\}$
$N(w)=\{w,q,v,x\}$
$N(x)=\{x,w\}$
$N(y)=\{y,r,s,z\}$
$N(z)=\{z,y\}$

Here the given MinPts is 4, thus the size of $N(p)$ is at least 4.

we can find:
$N(a)=\{a,b,c,d\}$
$N(b)=\{b,a,c,g,h\}$
$N(c)=\{c,a,b,d,g,h\}$
$N(d)=\{d,a,c,h,i\}$
$N(e)=\{e,f,i,j\}$
$N(g)=\{g,b,c,h,k\}$
$N(h)=\{h,b,c,d,g,k,l\}$
$N(i)=\{i,d,e,l,m\}$
$N(j)=\{j,e,f,m,n\}$
$N(k)=\{k,g,h,l\}$
$N(l)=\{l,h,i,k\}$
$N(m)=\{m,i,j,n,o\}$
$N(n)=\{n,j,m,o\}$
$N(o)=\{o,m,n,q\}$
$N(q)=\{q,o,u,v,w\}$
$N(s)=\{s,p,t,y\}$
$N(v)=\{v,q,u,w\}$
$N(w)=\{w,q,v,x\}$
$N(y)=\{y,r,s,z\}$

Thus the core points are:
$\{a,b,c,d,e,g,h,i,j,k,l,m,n,o,q,s,v,w,y\}$

N(f)={f,e,j} e,j are core points
N(p)={p,s,t} s is a core point
N(r)={r,y} y is a core point
N(t)={t,p,s} s is a core point
N(u)={u,q,v} q,v are core points
N(x)={x,w} w is a core point
N(z)={z,y} y is a core point

Thus the border points are: {f,p,r,t,u,x,z}

noise points: none

• now we choose a. a is a core point. a cluster is formed

retrieve all points density-reachable from a: {a,b,c,d}. C_1={a,b,c,d}

all points density-reachable from b N(b)={b,a,c,g,h}, set b to be core point
C_1={a,b,c,d,g,h}

all points density-reachable from c N(c)={c,a,b,d,g,h}, set c to be core point
C_1={a,b,c,d,g,h}

all points density-reachable from d N(d)={d,a,c,h,i}, set d to be core point
C_1={a,b,c,d,g,h,i}

all points density-reachable from e N(e)={e,f,i,j}, set e to be core point
C_1={a,b,c,d,e,f,g,i,j}

all points density-reachable from f N(f)={f,e,j}, f to be border point.

all points density-reachable from g N(g)={g,b,c,h,k}, set g to be core point
C_1={a,b,c,d,e,g,h,i,j,k}

all points density-reachable from h N(h)={h,b,c,d,g,k,l} set h to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l}

all points density-reachable from i N(i)={i,d,e,l,m}, set i to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m}

all points density-reachable from j N(j)={j,e,f,m,n}, set j to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n}

all points density-reachable from k N(k)={k,g,h,l}, set k to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n}

all points density-reachable from l N(l)={l,h,i,k}, set l to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n}

all points density-reachable from m N(m)={m,i,j,n,o}, set m to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n,o}

all points density-reachable from N(n)={n,j,m,o}, set n to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n,o}

all points density-reachable from o N(o)={o,m,n,q}, set o to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n,o,q}

all points density-reachable from q N(q)={q,o,u,v,w}, set q to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n,o,q,u,v,w}

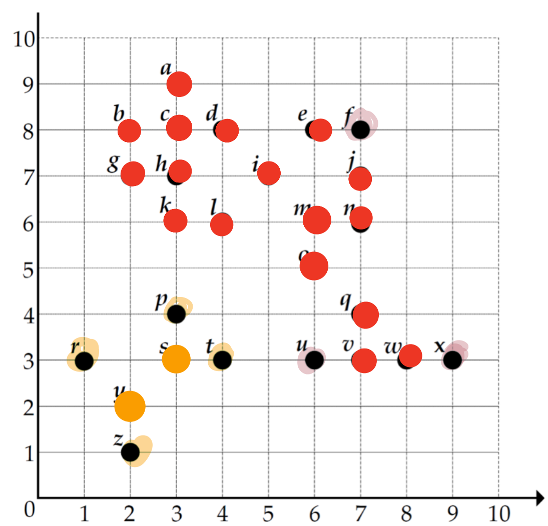
all points density-reachable from u, N(u)={u,q,v}, u to be border point.

all points density-reachable from v N(v)={v,q,u,w}, set v to be core point
C_1={a,b,c,d,e,g,h,i,j,k,l,m,n,o,q,u,v,w}

all points density-reachable from w $N(w)=\{w,q,v,x\}$, set w to be core point
$C\_1=\{a,b,c,d,e,g,h,i,j,k,l,m,n,o,q,u,v,w,x\}$
all points density-reachable from x $N(x)=\{x,w\}$, set x to be border point.
$C\_1=\{a,b,c,d,e,g,h,i,j,k,l,m,n,o,q,u,v,w,x\}$
so we have Cluster_1: {a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, q, u, v, w, x}

Next we choose s,
all points density-reachable from s $N(s)=\{s,p,t,y\}$, set s to be cord point
$C\_2=\{s,p,t,y\}$
all points density-reachable from p $N(p)=\{p,s,t\}$, set p to be border point
all points density-reachable from t $N(t)=\{t,p,s\}$, set t to be border point
all points density-reachable from y $N(y)=\{y,r,s,z\}$, set y to be cord point
$C\_2=\{s,p,r,t,y,z\}$
all points density-reachable from z $N(z)=\{z,y\}$, set z to be border point
all points density-reachable from t $N(t)=\{t,p,s\}$ set t to be border point

so we have Cluster_2: {s, p, r, t, y, z}



Q4 [20 Marks] Fuzzy Cluster

Assume there are 2 clusters in which the data is to be divided, initializing the data point randomly. Each data point lies in both clusters with some membership value which can be assumed anything in the initial state.

The table below represents the values of the data points along with their membership (gamma) in each cluster.

Cluster	(1,3)	(2,5)	(4,8)	(7,9)	(9,12)
1)	0.8	0.7	0.5	0.3	0.1
2)	0.2	0.3	0.5	0.7	0.9

Please work out the centroids, the distance of each point from centroid, and the cluster membership value.

Iteration 1:

$$M^T = \begin{bmatrix} 0.8 & 0.7 & 0.5 & 0.3 & 0.1 \\ 0.2 & 0.3 & 0.5 & 0.7 & 0.9 \end{bmatrix}$$

$$C_1 = \left(\frac{0.8^2\times1 + 0.7^2\times2 + 0.5^2\times4 + 0.3^2\times7 + 0.1^2\times9}{0.8^2+0.7^2+0.5^2+0.3^2+0.1^2}, \frac{0.8^2\times3 + 0.7^2\times5 + 0.5^2\times8 + 0.3^2\times9 + 0.1^2\times12}{0.8^2+0.7^2+0.5^2+0.3^2+0.1^2} \right)$$

$$= (2.2568, 4.9324)$$

$$C_2 = \left(\frac{0.2^2\times1 + 0.3^2\times2 + 0.5^2\times4 + 0.7^2\times7 + 0.9^2\times9}{0.2^2+0.3^2+0.5^2+0.7^2+0.9^2}, \frac{0.2^2\times3 + 0.3^2\times5 + 0.5^2\times8 + 0.7^2\times9 + 0.9^2\times12}{0.2^2+0.3^2+0.5^2+0.7^2+0.9^2} \right)$$

$$= (7.1071, 9.9405)$$

Iteration 2:

$$W_{01} = \frac{dist(O, C_2)^2}{dist(O, C_2)^2 + dist(O, C_1)^2} = \frac{(O_x - 7.1071)^2 + (O_y - 9.9405)^2}{(O_x - 7.1071)^2 + (O_y - 9.9405)^2 + (O_x - 2.2568)^2 + (O_y - 4.9324)^2}$$

$$W_{02} = 1 - W_{01}$$

```python
import numpy as np

# initialization
data_points = np.array([[1,3], [2,5], [4,8], [7,9], [9,12]])
membership_matrix = np.array([[0.8, 0.7, 0.5, 0.3, 0.1], [0.2, 0.3, 0.5, 0.7, 0.9]]).T

# define number of clusters and fuzziness parameter
c = 2
m = 2.0

for step in range(3):
    print(f"Step {step+1}")
    print("Membership Matrix:")
    print(membership_matrix)

    # calculate centroids
    centroids = np.dot(membership_matrix.T**m, data_points) / np.sum(membership_matrix.T**m, axis=1, keepdims=True)

    # calculate distance of each point from each centroid
    distances = np.linalg.norm(data_points[:, np.newaxis] - centroids, axis=2)

    # update membership values
    for i in range(c):
        membership_matrix[:, i] = 1.0 / np.sum((distances / distances[:, i:i+1])**(2/(m-1)), axis=1)

    print("Centroids:")
    print(centroids)
    print('#############################################')
```

Finally $M^T = \begin{bmatrix} 0.97 & 1 & 0.50 & 0.03 & 0.05 \\ 0.03 & 0 & 0.50 & 0.97 & 0.95 \end{bmatrix}$

$C_1 = (1.81, 4.49)$

$C_2 = (7.50, 10.17)$

$$D^T = \begin{pmatrix} 1.69 & 0.55 & 4.14 & 6.88 & 10.4 \\ 9.68 & 7.55 & 4.12 & 1.2 & 2.36 \end{pmatrix}$$

$\therefore$ $C_1 : \{O_1, O_2\}$

$C_2 : \{O_3, O_4, O_5\}$