DSAA 5002 - HW3

50015976 Ruiming ZHANG

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

Solution:

According to the single-link algorithm, we merge the two points with the smallest distance. And the new distance is the smallest distance between the two points and the other points. Then we find that the smallest distance is 0.47 between 2 and 5. Merge them and we get new distance matrix:

	1	(2&5)	3
(2&5)	2.33		
3	3.15	1.30	
4	1.90	1.50	3.70

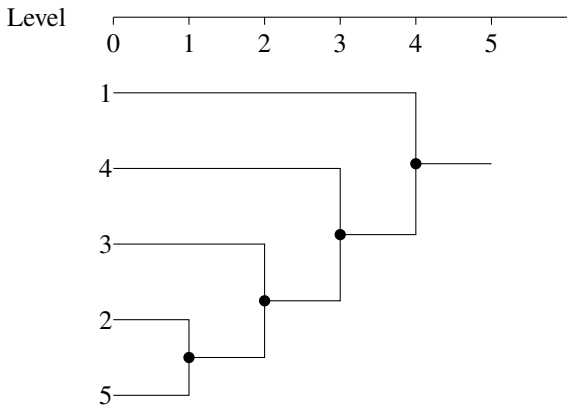
Then the distance 1.30 between 3 and (2&5) is the smallest one.

	1	(2&3&5)
(2&3&5)	2.33	
4	1.90	1.50

Then the distance 1.50 between (2&3&5) and 4 is the smallest one.

	1
(2&3&4&5)	1.90

Hence we get the cluster tree:



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

Table 2: Similarity matrix for Q2

Solution:

Regenerate the matrix by eliminating the diagonal and the upper triangle:

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

According to the single-link algorithm, we merge the two points with the max similarity. And the new similarity is the max similarity between the two points and the other points. Then we find that the max similarity is 0.98 between p2 and p5.

	p1	p2&5	p3
p2&5	0.35		
p3	0.41	0.85	
p4	0.55	0.76	0.44

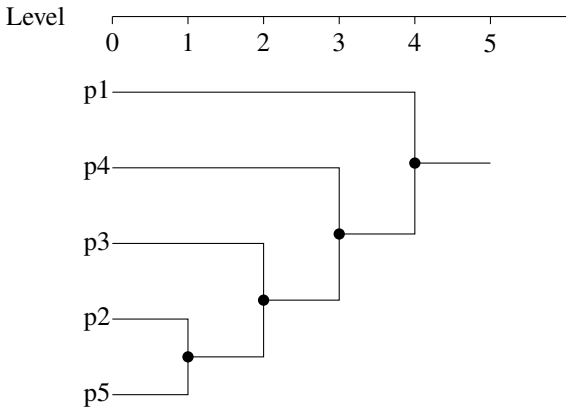
In the new similarity matrix, we find that the max similarity is 0.85 between p3 and p2&5.

	p1	p2&3&5
p2&3&5	0.41	
p4	0.55	0.76

In the new similarity matrix, we find that the max similarity is 0.76 between p4 and p2&3&5.

	p1
p2&3&4&5	0.55

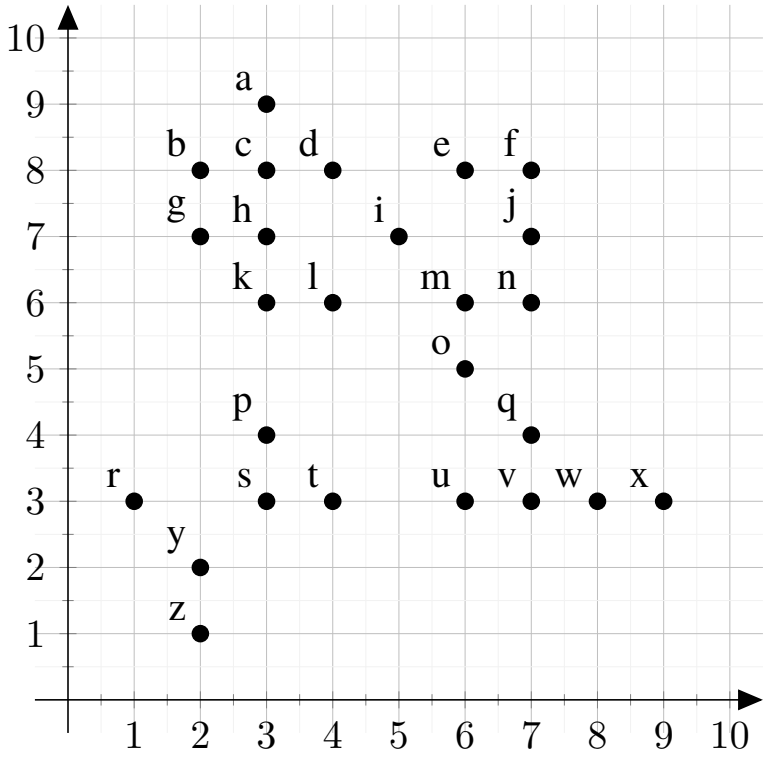
Hence we get the cluster tree:



Q3 [30 Marks]

Apply DBSCAN with parameters MinPts = 4 and Eps = $\sqrt{2}$ 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.

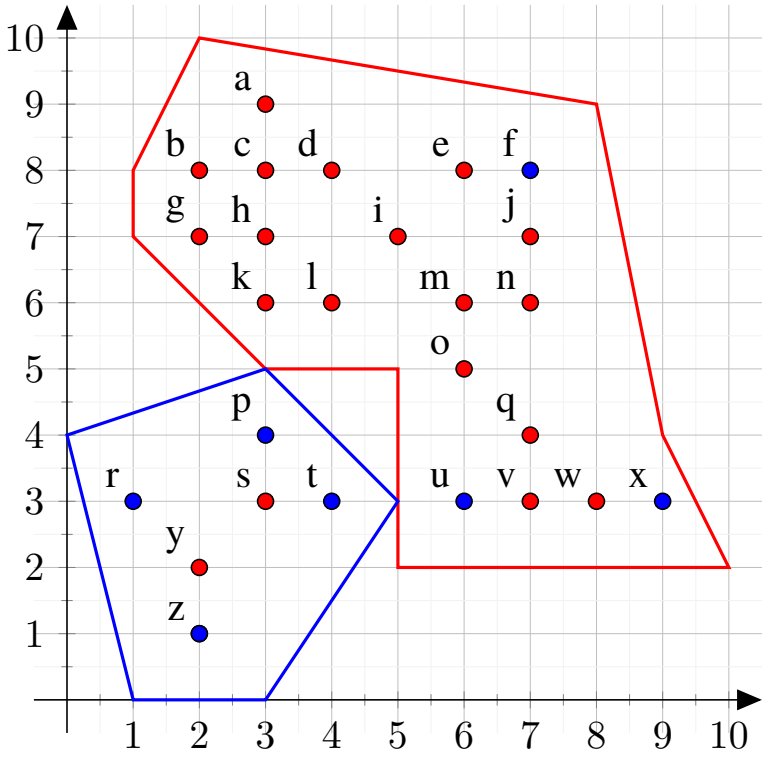


Solution:

If the size of N(p) is at least 4, then p is a core point.
We can find:

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

We find that points a, b, c, d, e, g, h, i, j, k, l, m, n, o, q, s, v, w, y are core points.
The other points are border points.
And there are no outliers.



In the scatter plot below, the red points are core points, the blue points are border points, and the green points are outliers (which actually do not exist).
And we detect 2 clusters which are circled in red and blue frames.

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.

Solution:

About the Fuzzy Clustering Using the EM Algorithm, we have the following formula:
E step:

$$W_{i1} = \frac{\frac{1}{\text{dist}(o_i,c_1)^2}}{\left(\frac{1}{\text{dist}(o_i,c_1)^2} + \frac{1}{\text{dist}(o_i,c_2)^2}\right)} = \frac{\text{dist}(o_i,c_2)^2}{\text{dist}(o_i,c_2)^2 + \text{dist}(o_i,c_1)^2}$$

$$W_{i2} = \frac{\frac{1}{\text{dist}(o_i,c_2)^2}}{\left(\frac{1}{\text{dist}(o_i,c_1)^2} + \frac{1}{\text{dist}(o_i,c_2)^2}\right)} = \frac{\text{dist}(o_i,c_1)^2}{\text{dist}(o_i,c_2)^2 + \text{dist}(o_i,c_1)^2} = 1 - W_{i1}$$

M step:

$$c_1 = \left(\frac{\sum_{\text{each point o}}(w_{o,c_1}^2 * o_x)}{\sum_{\text{each point o}} w_{o,c_1}^2}, \frac{\sum_{\text{each point o}}(w_{o,c_1}^2 * o_y)}{\sum_{\text{each point o}} w_{o,c_1}^2}\right)$$

$$c_2 = \left(\frac{\sum_{\text{each point o}}(w_{o,c_2}^2 * o_x)}{\sum_{\text{each point o}} w_{o,c_2}^2}, \frac{\sum_{\text{each point o}}(w_{o,c_2}^2 * o_y)}{\sum_{\text{each point o}} w_{o,c_2}^2}\right)$$

Iteration 1:

According to the table below, we already have the result of E step:

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

Then we have the M step:

$$c_1 = \left(\frac{0.8^2 * 1 + 0.7^2 * 2 + 0.5^2 * 4 + 0.3^2 * 7 + 0.1^2 * 9}{0.8^2 + 0.7^2 + 0.5^2 + 0.3^2 + 0.1^2}, \frac{0.8^2 * 3 + 0.7^2 * 5 + 0.5^2 * 8 + 0.3^2 * 9 + 0.1^2 * 12}{0.8^2 + 0.7^2 + 0.5^2 + 0.3^2 + 0.1^2}\right)$$

$$c_2 = \left(\frac{0.2^2 * 1 + 0.3^2 * 2 + 0.5^2 * 4 + 0.7^2 * 7 + 0.9^2 * 9}{0.2^2 + 0.3^2 + 0.5^2 + 0.7^2 + 0.9^2}, \frac{0.2^2 * 3 + 0.3^2 * 5 + 0.5^2 * 8 + 0.7^2 * 9 + 0.9^2 * 12}{0.2^2 + 0.3^2 + 0.5^2 + 0.7^2 + 0.9^2}\right)$$

Then we have $c_1 = (2.2568, 4.9324)$ and $c_2 = (7.1071, 9.9405)$.

Iteration 2:

E step:

Here we just calculate point d = (1,3) as an example:

First calculate the distance between the point and the centroids.

To simplify the calculation, we just calculate the square of the distance.

$$dist_sq((1,3), c_1) = (1 - 2.2568)^2 + (3 - 4.9324)^2 = 5.3137$$

$$dist_sq((1,3), c_2) = (1 - 7.1071)^2 + (3 - 9.9405)^2 = 85.4672$$

Then we have the result of E step:

$$w_{d,c_1} = \frac{85.4672}{85.4672 + 5.3137} = 0.9415$$

$$w_{d,c_2} = 1 - w_{d,c_1} = 0.0585$$

The rest of the calculation is similar to the first iteration. Then we have the result of M step:

Cluster	(1,3)	(2,5)	(4,8)	(7,9)	(9,12)
1)	0.9415	0.9986	0.5188	0.0224	0.0758
2)	0.0585	0.0014	0.4812	0.9776	0.9242

M step:

Similar to the first iteration, we have the result of M step: we have $c_1 = (1.8585, 4.5724)$ and $c_2 = (7.4856, 10.1299)$.

Iteration 3:
Similar to the second iteration, just show the full result of E step and M step:

Cluster	(1,3)	(2,5)	(4,8)	(7,9)	(9,12)	c_x	c_y
1)	0.9666	0.9964	0.5053	0.0318	0.0517	1.8171	4.5061
2)	0.0334	0.0036	0.4947	0.9682	0.9483	7.5079	10.1747

I have put the entire calculation process in the attachment of this homework, a .ipynb file.