CAP5510 Fall 2019, HW#3

Question no 1:

(1) The formula of Pearson's correlation coefficient is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- Calculating Pearson's correlation coefficient r for gene P53 and Mdm2:

Condition	P53 (X)	Mdm2 (Y)	X_iY_i	X_i^2	Y_i^2
1	10	9	90	100	81
2	4	1	4	16	1
3	2	1	2	4	1
4	8	7	56	64	49
5	6	5	30	36	25
Total	$\sum X = 30$	$\sum Y = 23$	$\sum XY = 182$	$\sum X_i^2 = 220$	$\sum Y_i^2 = 157$

Pearson's correlation coefficient for gene P53 and Mdm2, r = $\frac{5*182-(30*23)}{\sqrt{[5*220-900][5*157-529]}}$

r = $\frac{910-690}{\sqrt{51200}}$

r = 0.97

Similarly, after calculating the correlation coefficient for other genes, the resultant matrix is:

	P53	Mdm2	Bcl2	CyclinE	Caspase 8
P53	1	0.97	-0.43	0.89	-0.7
Mdm2	0.97	1	-0.54	0.79	-0.795
Bcl2	-0.43	-0.54	1	-0.21	0.93
CyclinE	0.89	0.79	-0.21	1	-0.45
Caspase 8	-0.7	-0.795	0.93	-0.45	1

(2) Clustering genes using Hierarchical clustering based on Euclidean distance and Centroid linkage:

Suppose, P53 = G1, Mdm2 = G2, Bcl2 = G3, CyclinE = G4, Caspase 8 = G5. Initially there are five clusters (each gene forms a cluster)

Step1:

Calculating Euclidean distance between each cluster:

Example: Distance between C1 (G1) and C2 (G2):

d(C1,C2) = $\sqrt{(10-9)^2 + (4-1)^2 + (2-1)^2 + (8-7)^2 + (6-5)^2}$

= $\sqrt{1 + 9 + 1 + 1 + 1}$

= $\sqrt{13}$

= 3.6

Similarly, after calculating the Euclidean distance between other clusters, the distance matrix is:

	C1 = G1	C2 = G2	C3 = G3	C4 = G4	C5 = G5
C1 = G1	0	-	-	-	-
C2 = G2	3.61	0	-	-	-
C3 = G3	11.05	12.61	0	-	-
C4 = G4	5.1	6.9	7.2	0	-
C5 = G5	11.67	13.15	**2.45**	7.07	0

Here, the smallest distance is between cluster C3 and C5, so they form a new cluster together.

So new set of clusters, **S = {C1, C2, C3, C4}**

S = {(G3, G5), G1, G2, G4}

Centroid of new cluster C1 = {(2+2)/2, (10+10)/2, (4+6)/2, (5+4)/2, (9+8)/2}

= {2, 10, 5, 4.5, 8.5}

Step2:

	C1 = G3, G5	C2 = G1	C3 = G2	C4 = G4
C1 = G3, G5	0	-	-	-
C2 = G1	11.29	0	-	-
C3 = G2	12.83	**3.61**	0	-
C4 = G4	7.04	4.24	6.86	0

Here, the smallest distance is between cluster C2 and C3, so they form a new cluster together.

So new set of clusters, **S = {C1, C2, C3 }**

S = {(G3, G5), (G1, G2), G4}

Centroid of new cluster C2 = {(10+9)/2, (4+1)/2, (2+1)/2, (8+7)/2, (6+5)/2}

= {9.5, 2.5, 1.5, 7.5, 5.5}

Step3:

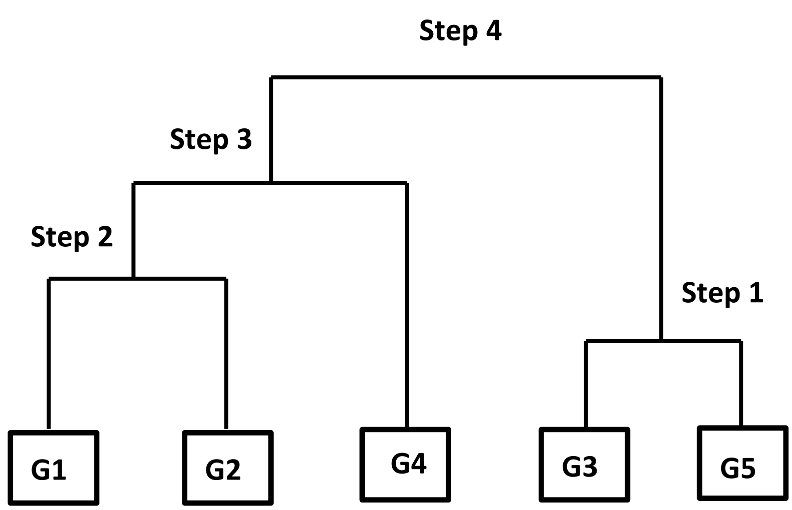
	C1 = G3, G5	C2 = G1, G2	C3 = G4
C1 = G3, G5	0	-	-
C2 = G1, G2	11.95	0	-
C3 = G4	7.04	**5.77**	0

Here, the smallest distance is between cluster C2 and C3, so they form a new cluster together.
So new set of clusters, **S = {C1, C2}**
S = {(G3, G5), (G1, G2, G4)}
Centroid of new cluster, C2 = {(10+9+7)/3, (4+1+6)/3, (2+1+5)/3, (8+7+6)/3, (6+5+6)/3}
= {8.67, 3.67, 2.67, 7, 5.67}

Step4:

	C1 = G3, G5	C2 = G1, G2, G4
C1 = G3, G5	0	-
C2 = G1, G2, G4	**10.21**	0

Here, cluster C1 and C2 form a new cluster together.
So new set of clusters, **S = {C1}**
S = {(G1, G2, G3, G4, G5)}



(3) Suppose, P53 = G1, Mdm2 = G2, Bcl2 = G3, CyclinE = G4, Caspase 8 = G5. Here, k =2; means there will be 2 clusters, C1 and C2. Initial vectors are p53 (G1) and bcl2 (G3).

So, initially, C1 = G1 = {10, 4, 2, 8, 6}

C2 = G3 = {2, 10, 4, 5, 9}

Round-1:

Here, G1 belongs to C1 and G3 belongs to C2.

Calculating Euclidean distance for G2:

- Distance between C1 and G2:

d(C1,G2) = $\sqrt{(10-9)^2+(4-1)^2+(2-1)^2+\ (8-7)^2+\ (6-5)^2}$

= $\sqrt{1+9+1+\ 1+\ 1}$

= $\sqrt{13}$

= 3.6

- Distance between C2 and G2:

d(C2,G2) = $\sqrt{(2-9)^2+(10-1)^2+(4-1)^2+\ (5-7)^2+\ (9-5)^2}$

= $\sqrt{49+81+9+\ 4+\ 16}$

= $\sqrt{159}$

= 12.61

As, d(C1,G2) < d(C2,G2) , so, G2 will go to cluster C1.

Calculating Euclidean distance for G4:

- Distance between C1 and G4:

d(C1,G4) = $\sqrt{(10-7)^2+(4-6)^2+(2-5)^2+\ (8-6)^2+\ (6-6)^2}$

= 5.1

- Distance between C2 and G4:

d(C2,G4) = $\sqrt{(2-7)^2+(10-6)^2+(4-5)^2+\ (5-6)^2+\ (9-6)^2}$

= 7.21

As, d(C1,G4) < d(C2,G4) , so, G4 will go to cluster C1.

Calculating Euclidean distance for G5:

- Distance between C1 and G5:

d(C1,G5) = $\sqrt{(10-2)^2+(4-10)^2+(2-6)^2+\ (8-4)^2+\ (6-8)^2}$

= 11.67

- Distance between C2 and G5:

d(C2,G5) = $\sqrt{(2-2)^2+(10-10)^2+(4-6)^2+\ (5-4)^2+\ (9-8)^2}$

= 2.45

As, d(C2,G5) < d(C1,G5) , so, G5 will go to cluster C2.

Finally, C1 = {G1, G2, G4} , Centroid = {(10+9+7)/3, (4+1+6)/3, (2+1+5)/3, (8+7+6)/3, (6+5+6)/3}

= {8.67, 3.67, 2.67, 7, 5.67}

C2 = {G3, G5}, Centroid = {(2+2)/2, (10+10)/2, (4+6)/2, (5+4)/2, (9+8)/2}

= {2, 10, 5, 4.5, 8.5}

Round-2:

Calculating Euclidean distance for G1:

- Distance between C1 and G1:

d(C1,G1) = $\sqrt{(8.67-10)^2+(3.67-4)^2+(2.67-2)^2+\ (7-8)^2+\ (5.67-6)^2}$

= 1.85

- Distance between C2 and G1:

d(C2,G1) = $\sqrt{(2-10)^2+(10-4)^2+(5-2)^2+\ (4.5-8)^2+\ (8.5-6)^2}$

= 11.3

As, d(C1,G1) < d(C2,G1) , so, G1 will go to cluster C1.

Calculating Euclidean distance for G2:

- Distance between C1 and G2:

d(C1,G2) = $\sqrt{(8.67-9)^2+(3.67-1)^2+(2.67-1)^2+\ (7-7)^2+\ (5.67-5)^2}$

= 3.24

- Distance between C2 and G2:

d(C2,G2) = $\sqrt{(2-9)^2+(10-1)^2+(5-1)^2+\ (4.5-7)^2+\ (8.5-5)^2}$

= 12.83

As, d(C1,G2) < d(C2,G2) , so, G2 will go to cluster C1.

Calculating Euclidean distance for G3:

- Distance between C1 and G3:

d(C1,G3) = $\sqrt{(8.67-2)^2+(3.67-10)^2+(2.67-4)^2+\ (7-5)^2+\ (5.67-9)^2}$

= 10.07

- Distance between C2 and G4:

d(C2,G3) = $\sqrt{(2-2)^2+(10-10)^2+(5-4)^2+\ (4.5-5)^2+\ (8.5-9)^2}$

= 1.22

As, d(C2,G3) < d(C1,G3) , so, G3 will go to cluster C2.

Calculating Euclidean distance for G4:

- Distance between C1 and G4:

d(C1,G4) = $\sqrt{(8.67-7)^2+(3.67-6)^2+(2.67-5)^2+\ (7-6)^2+\ (5.67-6)^2}$

= 3.84

- Distance between C2 and G4:

d(C2,G4) = $\sqrt{(2-7)^2+(10-6)^2+(5-5)^2+\ (4.5-6)^2+\ (8.5-6)^2}$

= 7.04

As, d(C1,G4) < d(C2,G4) , so, G4 will go to cluster C1.

Calculating Euclidean distance for G5:

- Distance between C1 and G5:

d(C1,G5) = $\sqrt{(8.67-2)^2+(3.67-10)^2+(2.67-6)^2+\ (7-4)^2+\ (5.67-8)^2}$

= 10.5

- Distance between C2 and G4:

d(C2,G5) = $\sqrt{(2-2)^2+(10-10)^2+(5-6)^2+\ (4.5-4)^2+\ (8.5-8)^2}$

= 1.22

As, d(C2,G5) < d(C1,G5) , so, G5 will go to cluster C2.

Finally, C1 = {G1, G2, G4} , Centroid = {(10+9+7)/3, (4+1+6)/3, (2+1+5)/3, (8+7+6)/3, (6+5+6)/3}

= {8.67, 3.67, 2.67, 7, 5.67}

C2 = {G3, G5}, Centroid = {(2+2)/2, (10+10)/2, (4+6)/2, (5+4)/2, (9+8)/2}

= {2, 10, 5, 4.5, 8.5}

As C1 and C2 remain unchanged, the loop will break here.

Final Result:

C1 = {G1, G2, G4}, C2 = {G3, G5}

Question 2:

ALDC (automatical local density clustering) algorithm is a recently proposed algorithm which has been proposed by Xuanzuo et al. in 2017 [1]. It is based on the concept of LDC (local density clustering) [2] and DBSCAN [3]. The steps of this algorithm [1] are given below:

Steps:

- *Step I*: Calculating the Euclidean distance between every point in data set
- *Step II*: Calculating the local density and the distance deviation of every points
- *Step III*: Calculating the product of local density and distance deviation
- *Step IV*: Expanding the difference between the potential cluster center and the remaining points to solve the problem of getting wrong number of cluster centers
- *Step V*: Using the measure criterion to capture the cluster center

- *Step VI*: Assigning the remaining points to their nearest neighbor which has higher density

Implementation:

Data set, D = {(1, 1), (2, 1), (7, 8), (8, 11), (9, 95), (8, 10), (0.5,1)}

D = {P1, P2, P3, P4, P5, P6, P7}

Step1:

Euclidean distance matrix:

	P1	P2	P3	P4	P5	P6	P7
P1	0						
P2	1	0					
P3	9.22	8.6	0				
P4	12.21	11.66	3.16	0			
P5	11.67	11.01	2.5	1.8	0		
P6	11.4	10.8	2.24	1	1.12	0	
P7	0.5	1.5	9.6	12.5	12.02	11.7	0

Step2:

Consider cut-off value, d_c = 1.5

Local density calculation:

$$\rho_i = \sum_{j \in I_U \setminus \{i\}} \chi(d_{ij} - d_c)$$

$$\chi(a) = \begin{cases} 1 & \text{a}<0 \\ 0 & \text{a}\geq 0 \end{cases}$$

For P1, ρ_1 = $\chi(1-1.5) + \chi(9.22-1.5) + \chi(12.21-1.5) + \chi(11.67-1.5) + \chi(11.4-1.5) + \chi(0.5-1.5)$ = 1 + 0 + 0 + 0 + 0 + 1

For P2, ρ_2 = 1 + 0 + 0 + 0 + 0 + 0 = 1

For P3, ρ_3 = 0 + 0 + 0 + 0 + 0 + 0 = 0

For P4, ρ_4 = 0 + 0 + 0 + 0 + 1 + 0 = 1

For P5, ρ_5 = 0 + 0 + 0 + 0 + 1 + 0 = 1
For P6, ρ_6 = 0 + 0 + 0 + 1 + 1 + 0 = 2
For P7, ρ_7 = 1 + 0 + 0 + 0 + 0 + 0 = 1

Distance deviation calculation:

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j>\rho_i} (d_{ij}) & i \text{ is not the highest density point} \\ \max\limits_{j} (d_{ij}) & i \text{ is the highest density point} \end{cases}$$

δ_1 = 12.21 ; P1 highest density point, so considering maximum distance
δ_2 = 1 ; P2 not highest density point, so considering minimum distance from highest density points
δ_3 = 2.23 ; P3 not highest density point
δ_4 =1 ; P4 not highest density point
δ_5 = 1.12 ; P5 not highest density point
δ_6 = 11.72 ; P6 highest density point
δ_7 = 0.5 ; P7 not highest density point

Step3:

Product of local density and distance deviation calculation:

$$\gamma_i = \rho_i * \delta_i$$

γ_1 = 24.42, γ_2 = 1, γ_3 = 0 , γ_4 = 1, γ_5 = 1.12 , γ_6 = 23.44 , γ_7 = 0.5

Step4:

Expanding the difference between the potential cluster center:

$$E_i = \Sigma_{j \in I_U \setminus \{i\}} \sqrt{(\gamma_i - \gamma_j)^2}$$

$E_1 = \sqrt{(24.42-1)^2 + (24.42-0)^2 + (24.42-1)^2 + (24.42-1.12)^2 + (24.42-23.44)^2 + (24.42-0.5)^2}$
= 119.46
Similarly,
E_2 = 47.48, E_3 = 119.46, E_4 = 47.48, E_5 = 119.46, E_6 = 47.48, E_7 = 47.48

Step5:

Measure criterion calculation:

$$Z_i = e^{-E_i}$$

Z_1 = 1.32E-52, Z_2 = 2.39717E-21, Z_3 = 4.39056E-23, Z_4 = 2.39717E-21, Z_5 = 2.12609E-21, Z_6 = 1.77E-50, Z_7 = 5.35E-22

Let dividing line = 1E-40; all the points below dividing line are considered as cluster center. So here we have two cluster centers.
So, for 1st cluster C1, center = P1 and for 2nd cluster C2, center = P6

Step6:

After assigning the remaining points to their nearest cluster center (using Euclidean distance matrix), we get:
1st Cluster, C1 = {P1, P2, P7} and 2nd Cluster, C2 = {P3, P4, P5, P6}

References:

[1] https://ieeexplore.ieee.org/document/7978726
[2] https://science.sciencemag.org/content/344/6191/1492
[3] https://dl.acm.org/citation.cfm?id=3001507

(2) Advantages and disadvantages of Automatical local density clustering algorithm is given below:

Advantages:

i) Number of clusters doesn't have to be predefined like K-means algorithm
ii) Like k-means algorithm, here initially cluster centers are not considered arbitrarily, the centers are calculated based on local density.
iii) It doesn't need to iterate again and again and calculate the center of newly formed clusters like K-means and Hierarchical; it assigns the remaining points without any iteration.
iv) Unlike k-means and Hierarchical, order of data has not effect on the result.
v) Time complexity less than k-means and hierarchical.

Disadvantages:

i) Result depends on the consideration of cut-off value and divide-line value
ii) K-means and hierarchical easier to implement
iii) Input parameters difficult to determine and result can be sensitive to input parameters