Homework 9

① Ex 7) B) More centroids should be allocated to the less dense region. The distance in a more dense region from the centroid is small compared to the less dense region. Hence, in order to minimize SSE, we will have more centroids in the less dense region to minimize the distance from the centroids.

Ex. 12)

a) advantages: * the size of the clusters won't be an issue here; ✓ it is likely that it will find the correct clusters (for the most part) without the help of the user → K-means needs to know how many clusters to do;

disadvantages: it might consider noise data/ points as new clusters and it will be wrong;

b) - Find a way to get rid of the noise. This will make the leader algorithm much better.
- Work with only one leader at a time to make more accurate predictions

---

Ex. 16) single and complete link hierarchical clustering;

① single link

distance = 1 - similarity

=> Distance matrix

	p1	p2	p3	p4	p5
p1	0				
p2	0.90	0			
p3	0.59	0.26	0		
p4	0.45	0.53	0.56	0	
p5	0.65	0.02	0.15	0.24	0

$\Rightarrow D(\{p_2\} \to \{p_5\}) = 0.02$ is the smallest $\Rightarrow$ cluster

$\Rightarrow D(\{p_2, p_5\} \to \{p_1\}) = \min(D(\{p_2\} \to \{p_1\}), D(\{p_5\} \to \{p_1\})) =$

$= \underline{0.65}$

$D(\{p_2, p_5\} \to \{p_3\}) = \underline{0.15}$

$D(\{p_2, p_5\} \to \{p_4\}) = \underline{0.24}$

$\Rightarrow$ p1 p2,p5 p3 p4

p1	0			
p2,p5	0.65	0		
p3	0.59	0.15	0	
p4	0.45	0.24	0.56	0

$\Rightarrow D(\{p_2, p_5\} \to \{p_3\}) = 0.15$ is the smallest $\Rightarrow$ cluster

$D(\{p_2, p_3, p_5\} \to \{p_1\}) = 0.90, 0.59, 0.65 = \underline{0.59}$

$D(\{p_2, p_3, p_5\} \to \{p_4\}) = 0.53, 0.56, 0.24 = \underline{0.24}$

$\Rightarrow$ p1 p2,p3,p5 p4

p1	0		
p2,p3,p5	0.59	0	
p4	0.45	0.24	0

$\Rightarrow D(\{p_2, p_3, p_5\} \to \{p_4\}) = 0.24$ smallest

$D(\{p_2, p_3, p_4, p_5\} \to \{p_1\}) = 0.90, 0.59, 0.45, 0.65 = \underline{0.45}$

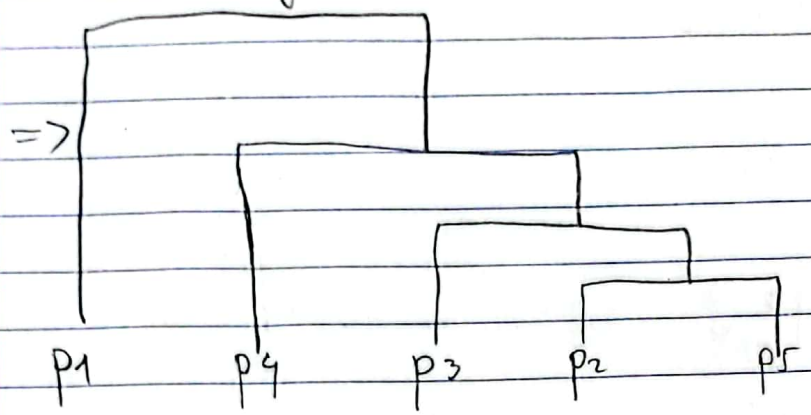
p2, p3, p4, p5 p1

p2, p3, p4, p5 0

D

p1 0.45

Dendrogram



p1 p4 p3 p2 p5

(2) Complete link

	p1	p2	p3	p4	p5
p1	0				
p2	0.90	0			
p3	0.59	0.26	0		
p4	0.45	0.53	0.56	0	
p5	0.65	0.02	0.15	0.24	0

$D(\{p1\} \to \{p2\}) = 0.90$ is max => cluster

$D(\{p1, p2\} \to \{p3\}) = 0.59$

$D(\{p1, p2\} \to \{p4\}) = 0.53$

$D(\{p1, p2\} \to \{p5\}) = 0.65$

	p1,p2	p3	p4	p5
p1,p2	0			
p3	0.59	0		
p4	0.53	0.56	0	
p5	0.65	0.15	0.24	0

$D(\{p1,p2\} \rightarrow \{p5\}) = 0.65$ is max

$\Rightarrow D(\{p1,p2,p5\} \rightarrow \{p3\}) = 0.59, 0.26, 0.15 = \underline{0.59}$

$D(\{p1,p2,p5\} \rightarrow \{p4\}) = 0.45, 0.53, 0.24 = \underline{0.53}$

$\Rightarrow$

	p1,p2,p5	p3	p4
p1,p2,p5	0		
p3	0.59	0	
p4	0.53	0.56	0

$D(\{p1,p2,p5\} \rightarrow \{p3\}) = 0.59$ is max

$\Rightarrow D(\{p1,p2,p3,p5\} \rightarrow \{p4\}) = \underline{0.56}$

$\Rightarrow$

	p1,p2,p3,p5	p4
p1,p2,p3,p5	0	
p4	0.56	0

Dendrogram