Clustering

Question 1: We can cluster in one dimension as well as in many dimensions. In this problem, we are going to cluster numbers on the real line. The particular numbers (data points) are 1, 4, 9, 16, 25, 36, 49, 64, 81, and 100, i.e., the squares of 1 through 10. We shall use a k-means algorithm, with two clusters. You can verify easily that no matter which two points we choose as the initial centroids, some prefix of the sequence of squares will go into the cluster of the smaller and the remaining suffix goes into the other cluster. As a result, there are only nine different clusterings that can be achieved, ranging from {1}{4,9,...,100} through {1,4,...,81}{100}.

We then go through a reclustering phase, where the centroids of the two clusters are recalculated and all points are reassigned to the nearer of the two new centroids. For each of the nine possible clusterings, calculate how many points are reclassified during the reclustering phase. List out pair of initial centroids that results in *exactly one* point being reclassified.

Sol:

Assignment - 9

(81) data points: {1, 4, 9, 16, 25, 36, 49, 64, 81, 100}

iteration 1

M_1, M_2 are 2 randomly selected centroids.

$M_1 = 9$, $M_2 = 64$

initial clusters are

$C_1 = \{4\}$ $C_2 = \{81\}$ $\{64\}$

Calculate Euclidian distance as

$b = \sqrt{(x-a)^2}$ D_1 is distance from M_1

D_2 is " " " M_2

Datapoint	D_1	D_2	cluster
1	8	63	C_1
4	5	60	C_1
9	0	55	C_1
16	7	48	C_1
25	16	39	C_1
36	27	28	C_2
49	40	15	C_2
64	55	0	C_2
81	72	17	C_2
100	91	36	C_2

$C_1 = \{1, 4, 9, 16, 25\}$

$C_2 \{36, 49, 64, 81, 100\}$

iteration 2

$$M_1 = \frac{1+4+9+16+15}{5} = 11$$

Datapoint	D_1	D_2	cluster	
1	10	65	C_1	
4	7	62	C_1	
9	2	57	C_1	
16	5	50	C_1	$C_1 = \{1, 4, 9, 16, 25, 36\}$
25	14	41	C_1	
36	25	30	C_1	$C_2 = \{49, 64, 81, 100\}$
49	48	17	C_2	
64	53	2	C_2	
81	70	15	C_2	
100	89	34	C_2	

iteration 3 : $M_1 = \frac{1+4+9+16+25+36}{6} = 15.16$

$M_2 = \frac{49+64+81+100}{4} = 73.5$

Datapoint	D_1	D_2	cluster	
1	14.6	72.5	C_1	
4	11.16	69.5	C_1	$C_1 = \{1, 4, 9, 16, 25, 36\}$
9	6.16	64.5	C_1	
16	0.84	57.5	C_1	$C_2 \{49, 64, 81, 100\}$
25	9.84	48.5	C_1	
36	20.84	37.5	C_1	
49	48.84	24.5	C_2	
64	65.84	9.5	C_2	
81	84.84	7.5	C_2	
100		28.5	C_2	

As we can see datapoints in cluster C_1 & C_2 are same and

Question 2: Suppose we want to assign points to one of two cluster centroids, either (0,0) or (100,40). Depending on whether we use the L_1 or L_2 norm, a point (x,y) could be clustered with a different one of these two centroids. For this problem, you should work out the conditions under which a point will be clustered with the centoid (0,0) when the L_1 norm is used, but clustered with the centroid (100,40) when the L_2 norm is used. List out those points.

Sol:

Given centroids are (0,0), (100, 40).
Given a point (x, y) which could be clustered with a different one of these two centroids.

L1 norm is the Manhattan Distance and L2 norm is the Euclidean Distance.
After L1 norm and L2 norm are calculated the values of x and y are 55, 5.

When the L1 norm is applied on point (55, 5), the point is clustered with centroid (0, 0). When L2 norm is applied on point (55, 5), the point is clustered with centroid (100, 40).

Question 3: Suppose our data set consists of the perfect squares 1, 4, 9, 16, 25, 36, 49, and 64, which are points in one dimension. Perform a hierarchical chustering on these points, as follows. Initially, each point is in a cluster by itself. At each step, merge the two clusters with the closest centroids, and continue until only two clusters remain. Which centroid of a cluster that exists at some time during this process? Positions are represented to the nearest 0.1.

Sol:

(Q3) Centroid distance is calculated using euclidean distance.

cluster/centroid	1	4	9	16	25	36	49	64
1	0							
4	(3)	0						
9	8	5	0					
16	15	12	7	0				
25	24	21	16	9	0			
36	35	32	27	20	11	0		
49	48	45	40	33	24	13	0	
64	63	60	55	48	39	28	15	0

→ Smallest dist

(2)

cluster/centroid	(1,4)	9	16	25	36	49	64
(1,4) c=2.5	0						
9	6.5	0					
16	13.5	7	0				
25	22.5	16	9	0			
36	33.5	27	20	11	0		
49	46.5	40	33	24	13	0	
64	61.5	55	48	39	28	15	0

(3)

cluster/centroid	(1,4,9)	16	25	36	49	64
(1,4,9) c=4.6	0					
16	11.4	0				
25	20.4	9	0			
36	31.4	20	11	0		
	33	24	13	0		

④

Cluster/centroid	(1,4,9)	(16,25)	36	49	64
(1,4,9) c=4.6	0				
(16,25) c=20.5	15.9	0			
36	31.4	15.5	0		
49	44.4	28.5	13	0	
64	59.4	43.5	28	15	0

⑤

Cluster/centroid	(1,4,9)	(16,25)	(36,49)	64
(1,4,9) c=4.6	0			
(16,25) c=20.5	15.9	0		
(36,49) c=42.5	37.9	22	0	
64	59.4	43.5	21.5	0

⑥

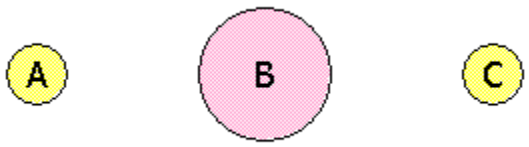
Cluster/centroid	(1,4,9,16,25)	(36,49)	64
(1,4,9,16,25) c=11	0		
(36,49) c=42.5	31.5	0	
64	53	21.5	0

⑦

Cluster/centroid	(1,4,9,16,25)	(36,49,64)
(1,4,9,16,25) c=11	0	
(16,49,64) c=49.6	38.6	0

During the previous centroids are 2.5, 4.6, 20.5,

Question 4: Suppose that the true data consists of three clusters, as suggested by the diagram below:



There is a large cluster B centered around the origin (0,0), with 8000 points uniformly distributed in a circle of radius 2. There are two small clusters, A and C, each with 1000 points uniformly distributed in a circle of radius 1. The center of A is at (-10,0) and the center of C is at (10,0).

Suppose we choose three initial centroids x, y, and z, and cluster the points according to which of x, y, or z they are closest. The result will be three *apparent* clusters, which may or may not coincide with the *true* clusters A, B, and C. Say that one of the true clusters is *correct* if there is an apparent cluster that consists of all and only the points in that true cluster. Assuming initial centroids x, y, and z are chosen independently and at random, what is the probability that A is correct? What is the probability that C is correct? What is the probability that both are correct?

Sol:

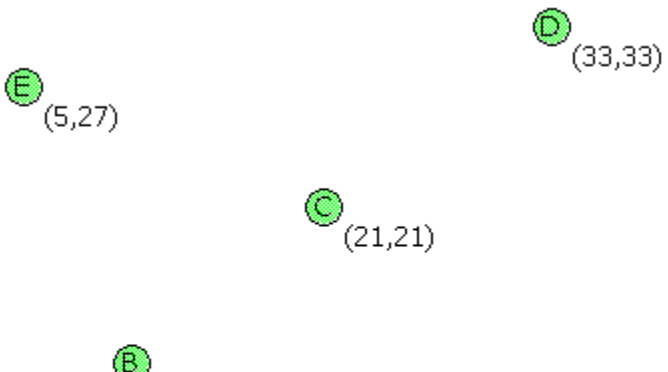
Given centroids are x, y, z
We can assign each of x, y, z to A, B, C in 27 possible ways.

Chance of being in A is 1000/10000 = 0.1
Chance of being in B is 8000/10000 = 0.8
Chance of being in C is 1000/10000 = 0.1

There are 6 different cases to interchange x, y, z in A, B, C which will be total 27.

The probability that A is correct is 24%
The probability that C is correct is 24%
The probability that A & C are correct is 4.8%

Question 5: Perform a hierarchical clustering of the following six points:



using the *complete-link* proximity measure (the distance between two clusters is the largest distance between any two points, one from each cluster). Find out a cluster at some stage of the agglomeration?

Sol:

(Q5) Points – A(0,0), B(10,10), C(21,21), D(33,33), E(5,27), F(22,16)

Dist Matrix, $d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

distance Matrix

	A	B	C	D	E	F
A	0					
B	14	0				
C	30	15	0			
D	47	32	17	0		
E	27	18	17	29	0	
F	29	18	16	28	31	0

Iteration 1

B → A is the smallest distance using complete linkage clustering.

max(d(C,A), d(C,B))

max(30,15) = 30

max(d(D,A), d(D,B))

max(47, 32) = 47

max(27, 18) = 27

max(29, 18) = 29

	(A,B)	C	D	E	F
(A,B)	0				
C	30	0			
D	47	17	0		
E	27	17	29	0	
F	29	(16)	28	31	0

Iteration 2

F → C is smallest dist

	AB	CF	D	E
AB	0			
CF	30	0		
D	47	17	0	
E	27	17	29	0

max(CA, CB, FA, FB)

= 30

iteration 3

E → AB is smallest dist

	ABE	CF	D
ABE	0		
CF	31	0	
D	47	(23)	0

iteration 4

D → CF is smallest dist

	ABE	CDF
ABE	0	
CDF	47	0

* Dendogram representation *