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.

Consider Initial Centroids as

(i) $C_1 = 25$; $C_2 = 81$

cluster1 = {1, 4, 9, 16, 25, 36, 49}

cluster2 = {81, 64, 100}

New centroids are $C_1 = 20$, $C_2 = 81.67$

cluster1 = {1, 4, 9, 16, 25, 36, 49}

cluster2 = {64, 81, 100}

Here, '0' points are reclassified after obtaining new centroids.

(ii) $C_1 = 1$; $C_2 = 25$

cluster1 = {1, 4, 9}

cluster2 = {25, 16, 36, 49, 64, 81, 100}

New centroids are $C_1 = 4.67$, $C_2 = 53$

cluster1 = {1, 4, 9, 16, 25}

cluster2 = {36, 49, 64, 81, 100}

New centroids are $C_1 = 11$, $C_2 = 66$

cluster1 = {1, 4, 9, 16, 25, 36}

cluster2 = {49, 64, 81, 100}

New centroids are $C_1 = 15.17$, $C_2 = 73.5$

cluster1 = {1, 4, 9, 16, 25, 36, 49}

cluster2 = {64, 81, 100}

(iii) $C_1 = 4$, $C_2 = 16$

cluster1 = {1, 4, 9}

cluster2 = {16, 25, 36, 49, 64, 81, 100}

(iv) $C_1 = 9$, $C_2 = 36$

cluster 1 = {1, 4, 9, 16}

cluster 2 = {25, 36, 49, 64, 81, 100}

New centroids are $C_1 = 7.5$, $C_2 = 59.17$

cluster 1 = {1, 4, 9, 16, 25}

cluster 2 = {36, 49, 64, 81, 100}

Here, only one point is reclustered after obtaining new centroids.

→ Centroids $C_1 = 9$, $C_2 = 36$ only one point will be reclustered.

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.

② Given centroids : $C_1 = (0,0)$, $C_2 = (100, 40)$

(a) $(50, 18)$

Manhattan distance : with $C_1 \Rightarrow |50-0| + |18-0|$

$= 50 + 18 = 68$

with $C_2 \Rightarrow 50 + 22 = 72$.

It falls under cluster 1 according to MD.

Euclidean distance : $C_1 \Rightarrow \sqrt{(50-0)^2 + (18-0)^2} = 53.14$

$C_2 = \sqrt{(50-100)^2 + (18-40)^2} = 54.62$.

It falls under cluster 1 according to ED.

(b) $(55, 5)$

MD : $C_1 = |55-0| + |5-0| = 60$

$C_2 = |55-100| + |5-40| = 70$.

$\Big\} \rightarrow$ cluster 1

ED : $C_1 = \sqrt{(55-0)^2 + (5-0)^2} = 55.22$

$C_2 = \sqrt{(55-100)^2 + (5-40)^2} = 57.008$

$\Big\} \rightarrow$ cluster 1

(c) (63, 8)

MD: $C_1 = |63-0| + |8-0| = 63+8 = 71$
$C_2 = |63-100| + |8-40| = 69$
} cluster 2

ED: $C_1 = \sqrt{(63-0)^2 + (8-0)^2} = 63.5$
$C_2 = \sqrt{(63-100)^2 + (8-40)^2} = 48.9$
} cluster 2

(d) (53, 15)

MD: $C_1 = |53-0| + |15-0| = 68$
$C_2 = |53-100| + |15-40| = 72$
} cluster 1

ED: $C_1 = \sqrt{(53-0)^2 + (15-0)^2} = 55.08$
$C_2 = \sqrt{(53-100)^2 + (15-40)^2} = 53.23$
} cluster 2.

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.

③ Given points are 1, 4, 9, 16, 25, 36, 49, 64.

(i) In the first step, each point is considered as a cluster by itself.

Here, 1 and 4 have minimum distance. So, we will merge 1 + 4 into a single cluster with centroid

$\frac{1+4}{2} = 2.5$

Now, new clusters will be

$\overset{2.5}{(1, 4)}$, 9, 16, 25, 36, 49, 64

(ii) Now, cluster containing {1, 4} with centroid 2.5 have minimum distance with 9.

we will merge 1, 4 + 9 into a cluster with centroid $\frac{1+4+9}{3} = 4.67$

Now, new clusters are :

(1, 4, 9) [4.67], 16, 25, 36, 49, 64

(iii) Minimum distance is between 16 and 25. So, we will merge them into a cluster with centroid

$$\frac{16+25}{2} = 20.5$$

Now, new clusters are :

(1, 4, 9) [4.67], (16, 25) [20.5], 36, 49, 64

(iv) 36 & 49 - min $\Rightarrow$ centroid $= \frac{36+49}{2} = 42.5$.

$\Rightarrow$ (1, 4, 9) [4.67], (16, 25) [20.5], (36, 49) [42.5] 64

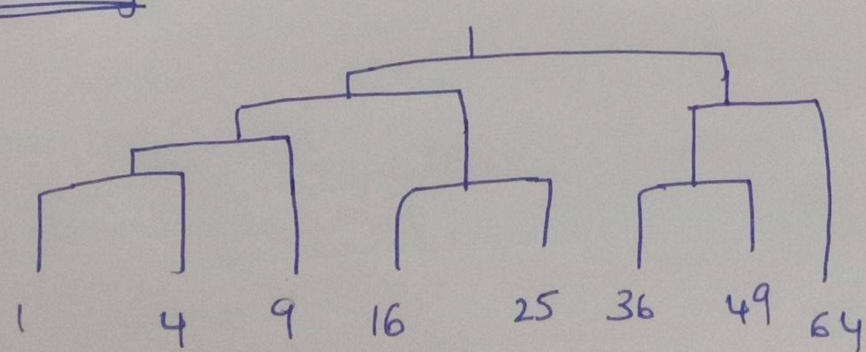
(v) 4.67 & 20.5 - min $\Rightarrow$ centroid $= \frac{4.67+20.5}{2} = 12.585$

((1, 4, 9) [4.67], (16, 25) [20.5]) [12.585], (36, 49) [42.5] 64

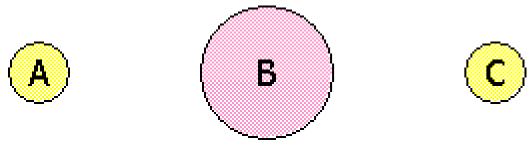
(vi) 42.5 & 64 - min $\Rightarrow$ centroid $= \frac{42.5+64}{2} = 53.25$

((1, 4, 9) [4.67], (16, 25) [20.5]) [12.585], ((36, 49) [42.5], 64) [53.25]

Dendogram :



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?

④ Total no. of points from (-10,0) to (10,0) on the X-axis are 20 points.

* x, y, z are chosen randomly.
  We can select any point from those 20 points.
  So, their probabilities will be

$$P(x) = \frac{1}{20}, \quad P(y) = \frac{1}{20}, \quad P(z) = \frac{1}{20}.$$

* 'A' contains 1000 points.
  Probability that 'A' is correct is

$$P(x) * P(\text{selecting } 1000 \text{ points})$$

$$= \frac{1}{20} * \frac{1000}{10000}$$

$$= 0.005$$

* 'C' contains 1000 points.
  ⇒ Probability that 'C' is correct is 0.005

* A & C together contain 2000 points.
  Probability that both A & C are correct is

$$[P(x) \cdot P(y)] * P[\text{selecting } 2000 \text{ points}]$$

$$= \frac{2}{20} * \frac{2000}{10000}$$

$$= 0.02$$

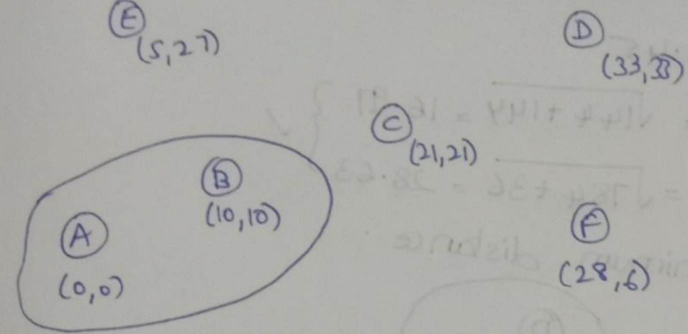
Question 5: Perform a hierarchical clustering of the following six points:

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?

(5) Given points are

A(0,0), B(10,10), C(21,21), D(33,33), E(5,27), F(28,6)

(1) A & B have minimum distance out of all clusters. So, we will ~~xxxx~~ merge them into a single cluster.



(2) Distance between {A,B} & E = $\sqrt{25+729}$ = 27.45

{A,B} & C = $\sqrt{441+441}$ = 29.69.

C & F = 16.55

C & D = $\sqrt{144+144}$ = 16.97

F & D = $\sqrt{25+729}$ = 27.45

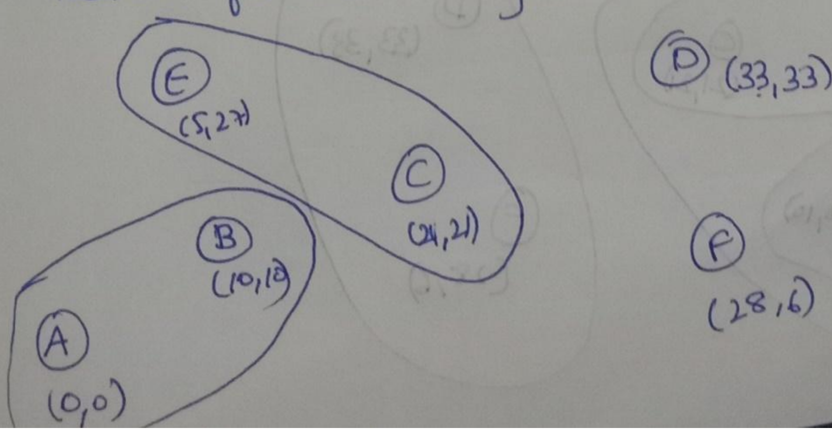
E & C = $\sqrt{196+36}$ = 15.23

E & D = $\sqrt{784+36}$ = 28.63

{A,B} & F = $\sqrt{324+16}$ = 18.43

∴ E & C have minimum distance.

Here, we are following complete-link proximity measure for calculating distance between clusters.



(3) Distance between $\{A, B\} + \{E, C\} = \sqrt{441 + 441} = 29.69$

$\{A, B\} + F = \sqrt{784 + 36} = 28.63$

$\{E, C\} + F = \sqrt{49 + 225} = 16.55$ ✓

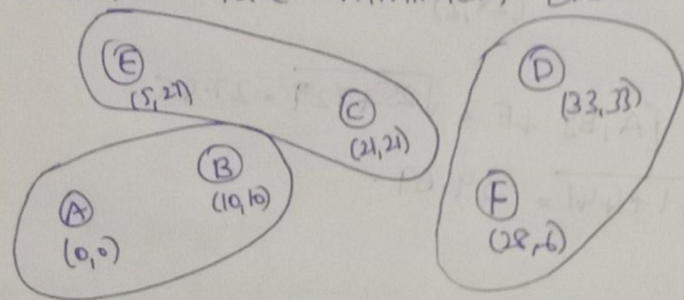
$\sqrt{484 + 441} = 30.41$

$F + D = 27.45$

$\{E, C\} + D = \sqrt{144 + 144} = 16.97$ ✓

$= \sqrt{784 + 36} = 28.63$

$\therefore$ F + D have minimum distance.



(4) Distance between $\{A, B\} + \{E, C\} = 29.69$

$\{A, B\} + \{F, D\} = \sqrt{1089 + 1089} = 46.66$

$\{E, C\} + \{F, D\} = \sqrt{784 + 36} = 28.63$

$= \sqrt{144 + 144} = 16.97$ ✓

$= \sqrt{49 + 225} = 16.55$

$= \sqrt{484 + 441} = 30.41$

$\therefore \{A, B\} + \{E, C\}$ have minimum distance.