Graded

This document has 10 questions.

Question-1

Statement

What would be the correct relationship among the following three quantities?:

- (1) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^t}^t||^2$,
- (2) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^{t+1}}^t||^2$ and
- (3) $\sum_{i=1}^{n} ||x_i - \mu_{z_i^{t+1}}^{t+1}||^2$

  where $\mu_{z_i^t}^t$ and $\mu_{z_i^{t+1}}^{t+1}$ refer to means of cluster z_i in iterations t and $t+1$ respectively. And $\mu_{z_i^{t+1}}^t$ is the mean of the cluster z_i where x_i is going to move in the next (i. e., $(t+1)^{th}$) iteration.

Options

(a)

(1) > (2) < (3)

(b)

(1) < (2) < (3)

(c)

(1) > (2) > (3)

(d)

(1) < (2) > (3)

Answer

(c)

Solution

The first quantity represents the value of objective function in iteration t. the third quantity represents the value of objective function in iteration $t+1$. The second quality represents an intermediate quantity which captures the distance of each data point from the mean that they will be moving towards, in the t+1 iteration. Since in every iteration, the reassignment happens only if a data point has found a closer mean, (3) will be lesser than (1). Further, since every point will want to move towards a closer cluster center in the subsequent iteration, the value of (2) will be between (1) and (3).

Question-2

Statement

Consider that in an iteration t of Lloyd's algorithm, the partition configuration (P^t) is $z_1^t, z_2^t, \ldots, z_n^t$ where each $z_i^t \in \{1, 2, \ldots, k\}$. Assume that the algorithm does not converge in iteration t, and hence some re-assignment happens, thus updating the partition configuration in the next iteration (P^{t+1}) to $z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}$. How can we say that partition configuration P^{t+1} is better than P^t?

Options

(a)

The value of the objective function for P^{t+1} should be more than that for P^t

(b)

The value of the objective function for P^{t+1} should be lesser than that for P^t

(c)

The value of the objective function for P^{t+1} and P^t should be same.

Answer

(b)

Solution

Since in every iteration, the reassignment happens only if a data point has found a closer mean, P^{t+1} will be lesser than P^t.

Question-3

Statement

With respect to Lloyd's algorithm, choose the correct statements:

Options

(a)

At the end of k-means, the objective function settles in a local minima and reaching global minima may not be guaranteed.

(b)

At the end of k-means, the objective function always settles in the global minima.

(c)

The clusters produced by K-means are optimal.

(d)

If the resources are limited and the data set is huge, it will be good to prefer K-means over K-means ++.

(e)

In practice, k should be as large as possible.

Answer

(a), (d)

Solution

(a), (b) K-means may not always settle in a global minima.

(c) Finding optimal clusters is an NP-hard problem. K-means provides approximate clusters.

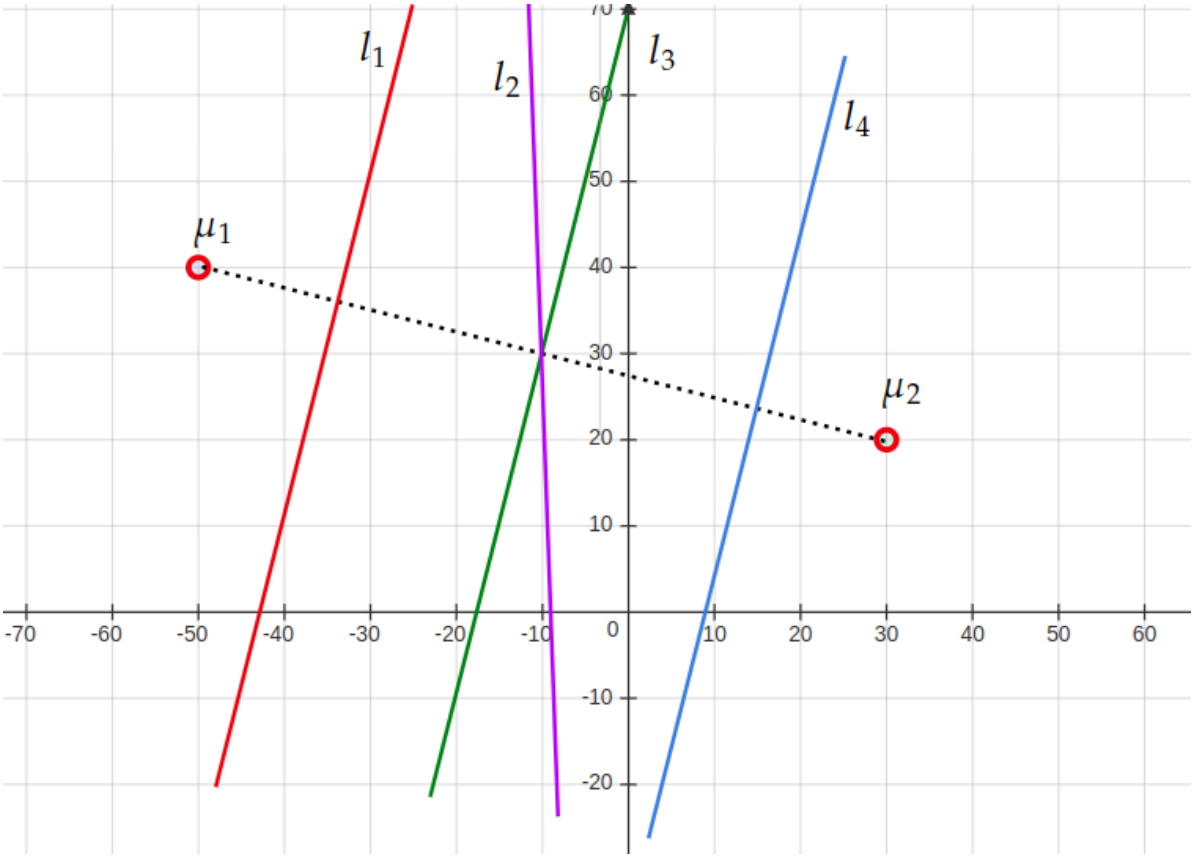
(d) If the dataset is huge. the elaborate initialization step in K-means++ will take a lot of time.

(e) In practice, k should neither be very small nor very large, because in both these cases, we may not be able to uncover groupings present in the data.

Question-4

Statement

Consider two cluster centres μ_1 and μ_2 corresponding to two clusters C_1 and C_2 as shown in the below image. Consider four half spaces represented by lines l_1, l_2, l_3 and l_4. Where would the data points falling in cluster C_1 lie?



Options

(a)

To the left of l_1

(b)

Between l_1 and l_2

(c)

Between l_3 and l_4

(d)

To the left of l_3

(e)

To the left of l_2

Answer

(d)

Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

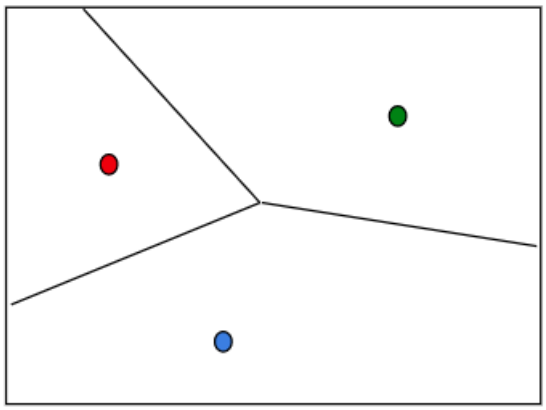
Question-5

Statement

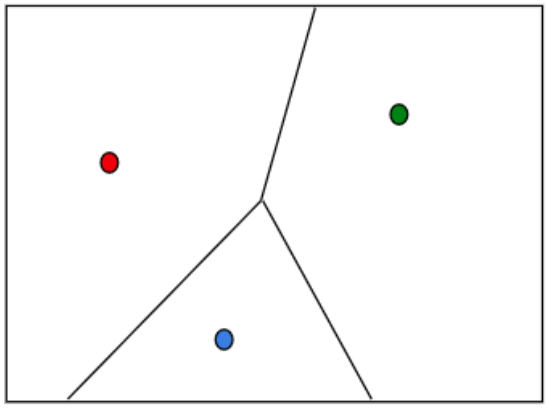
Which of the following best represents a valid voronoi diagram for K-means algorithm with K = 3? (The dots represent the cluster centres of respective clusters.)

Options

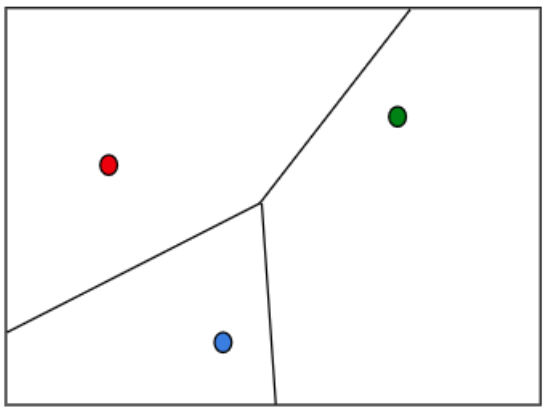
(a)



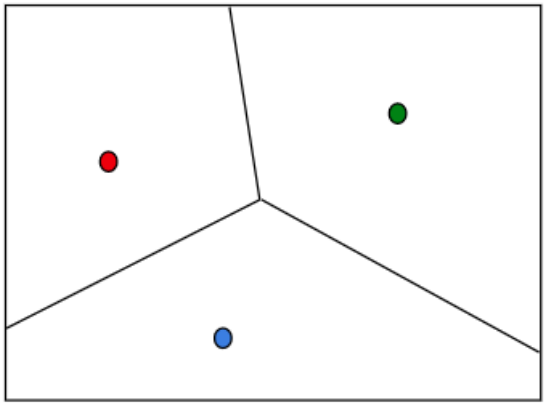
(b)



(c)



(d)



Answer

(d)

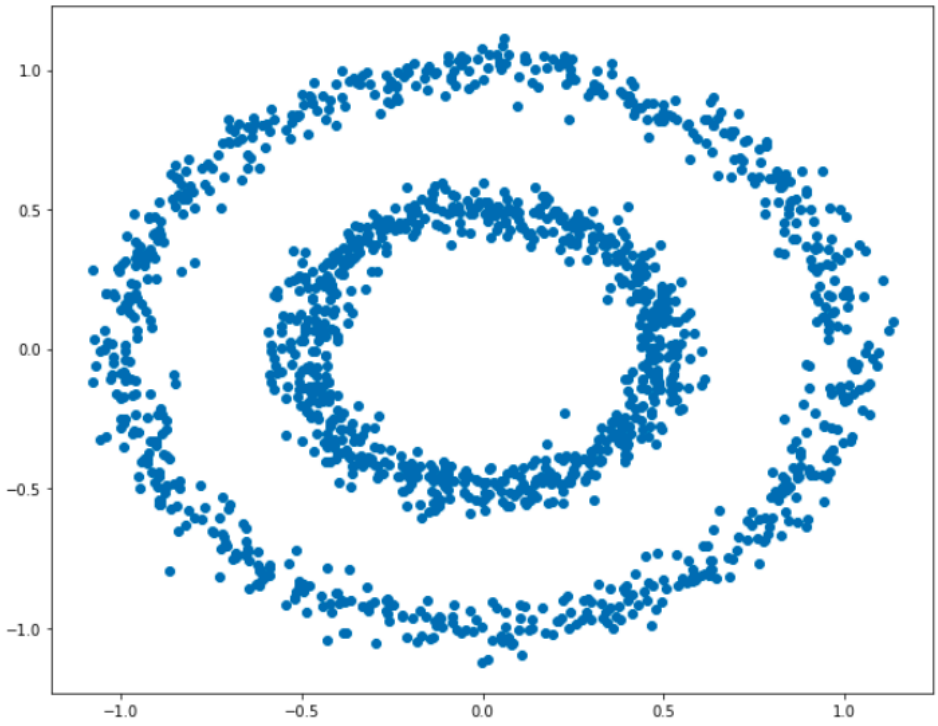
Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

Question-6

Statement

Consider the following data points:

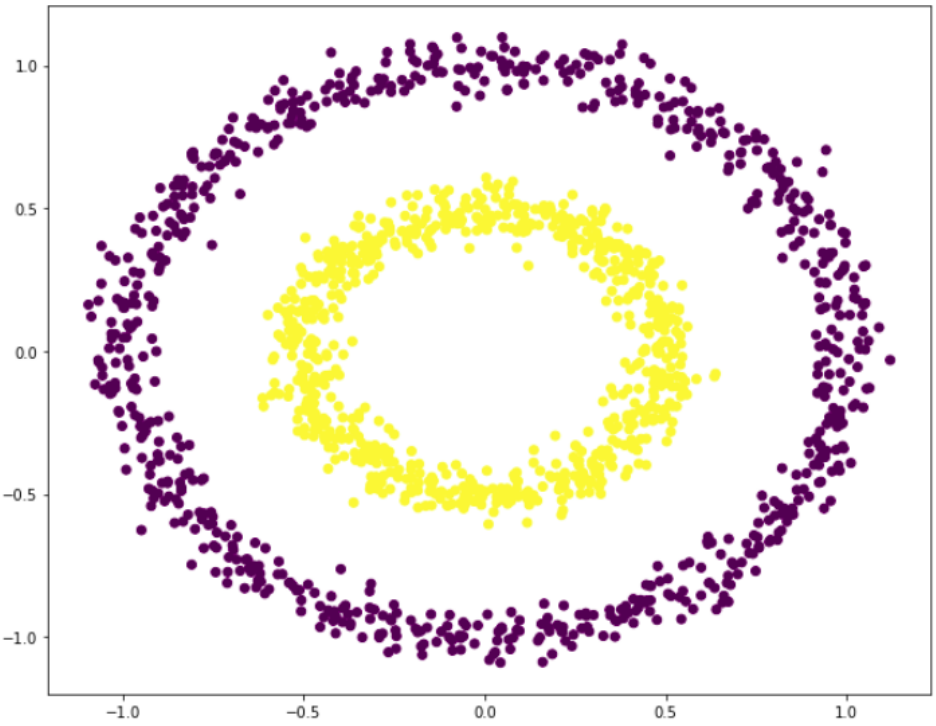


Assume that K-means is applied on this data with k = 2. Which of the following are expected to be the clusters produced?

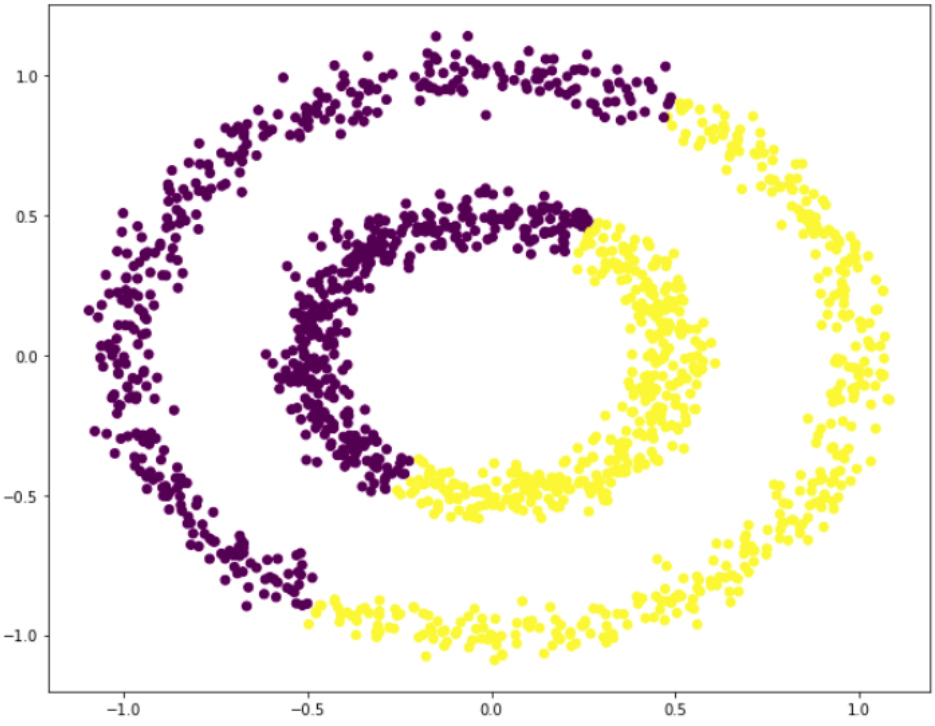
Note: Different colors represent different clusters.

Options

(a)



(b)



Answer

(b)

Solution

Half-spaces are perpendicular bisectors of the line joining the cluster centers.

In the given data, in case of option (a), the cluster centers will coincide, which is something does not happen as a result of applying k-means.

Question-7

Statement

Assume that in the initialization step of k-means++, the squared distances from the closest mean for 10 points $x_1, x_2, \ldots, x_{10}$ are: 25, 67, 89, 24, 56, 78, 90, 85, 35, 95. Which point has the highest probability of getting chosen as the next mean and how much will that probability be?

Options

(a)

x_4, 0.24

(b)

x_4, 0.037

(c)

x_{10}, 0.95

(d)

x_{10}, 0.1475

Answer

(d)

Solution

25+ 67+ 89+ 24+ 56+ 78+ 90+ 85+ 35+ 95 = 644

Probability for $x_{10} = 95/644 = 0.1475$

Probability for $x_4 = 24/644 = 0.037$

Question-8

Statement

Consider 7 data points $x_1, x_2, \ldots, x_7$: {(0, 4), (4, 0), (2, 2), (4, 4), (6, 6), (5, 5), (9, 9)}. Assume that we want to form 3 clusters from these points using K-Means algorithm. Assume that after first iteration, clusters $C1, C2, C3$ have the following data points:

$C1$: {(2,2), (4,4), (6,6)}

$C2$: {(0,4), (4,0)}

$C3$: {(5,5), (9,9)}

After second iteration, which of the clusters is the data point (2, 2) expected to move to?

Options

(a)

C_1

(b)

$C2$

(c)

C_3

(d)

Can't say, it is not deterministic.

Answer

(b)

Solution

C1: (4,4), C2: (2,2), C3: (7,7)

C2 mean is the closest to (2,2) with distance 0.

Question-9

Statement

Which of the following statements are True?

1. K-means is extremely sensitive to cluster center initializations.
2. Bad initialization can lead to poor convergence speed.
3. Bad initialization can lead to bad overall clustering.

Options

(a)

1 and 3

(b)

1 and 2

(c)

2 and 3

(d)

1, 2, and 3

Answer

(d)

Solution

1. Different cluster center initializations may result in different clusters produced by k-means.
2. Some initializations may take more time to converge.
3. Some initializations may converge either in a local minima rather than global minima.

Question-10

Statement

If the data set has two features x_1 and x_2, which of the following are true for K means clustering with k = 3?

1. If x_1 and x_2 have a correlation of 1, the cluster centres will be in a straight line.
2. If x_1 and x_2 have a correlation of 0, the cluster centres will be in straight line.

Options

(a)

1

(b)

2

(c)

None of these. Correlation does not affect cluster centres' position.

Answer

(a)

Solution

If x_1 and x_2 have a correlation of 1, all data points will lie along a line.

Hence the cluster centers will also lie along the same line.