Practice

This document has 10 questions.

Question-1

Statement

Which of the following sequences is correct for K-Means algorithm?

1. Assign each data point to the nearest cluster centres.
2. Re-assign each point to nearest cluster centres.
3. Assign cluster centres randomly.
4. Re-compute cluster centres.
5. Specify the number of clusters.

Options

(a)

3, 5, 1, 4, 2

(b)

5, 3, 1, 2, 4

(c)

5, 3, 1, 4, 2

(d)

3, 5, 2, 4, 1

(e)

None of these

Answer

(c)

Solution

The correct sequence of steps used in k-means algorithm is:

- Specify the number of clusters.
- Assign cluster centres randomly.
- Assign each data point to the nearest cluster centres.
- Re-compute cluster centres.
- Re-assign each point to nearest cluster centres.

Question 2

Statement

If $F(z_1^t, z_2^t, \ldots, z_n^t)$ represents the value of objective function in iteration t of Lloyd's algorithm, then which of the following is true?

Options

(a)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) > F(z_1^t, z_2^t, \ldots, z_n^t)$

(b)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) < F(z_1^t, z_2^t, \ldots, z_n^t)$

(c)

$F(z_1^{t+1}, z_2^{t+1}, \ldots, z_n^{t+1}) = F(z_1^t, z_2^t, \ldots, z_n^t)$

Answer

(b)

Solution

In Lloyd's algorithm, in each iteration, the data points change their cluster only if they find a cluster center which is closer to them as compared to their existing cluster's center. Therefore, every re-assignment results in the reduction of the objective function value, which is represented by $F(z_1^t, z_2^t, \ldots, z_n^t)$for iteration t.

Question-3

Statement

If μ_1 and μ_2 are means of two clusters in k-means, then the boundary between the two clusters will be

Options

(a)

Perpendicular to the line joining μ_1 and μ_2 and at the point $\dfrac{(\mu_1 + \mu_2)^2}{2}$

(b)

Parallel to the line joining μ_1 and μ_2 and at the point $\dfrac{(\mu_1 + \mu_2)^2}{2}$

(c)

Perpendicular to the line joining μ_1 and μ_2 and at the point $\dfrac{\mu_1 + \mu_2}{2}$

(d)

Parallel to the line joining μ_1 and μ_2 and at the point $\dfrac{\mu_1 + \mu_2}{2}$

Answer

(c)

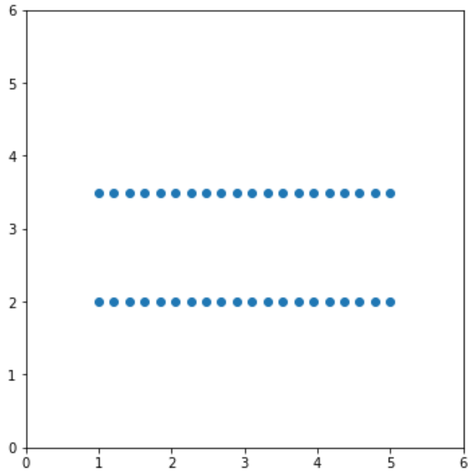
Solution

The clusters in k-means are separated by half-spaces. These half-spaces are formed by the perpendicular bisector of the line joining the clusters' centres. Therefore, if μ_1 and μ_2 are the centres of two clusters, the boundary between these clusters will be perpendicular to the line joining μ_1 and μ_2 and at the midpoint of this line, which is $\dfrac{\mu_1 + \mu_2}{2}$

Question-4

Statement

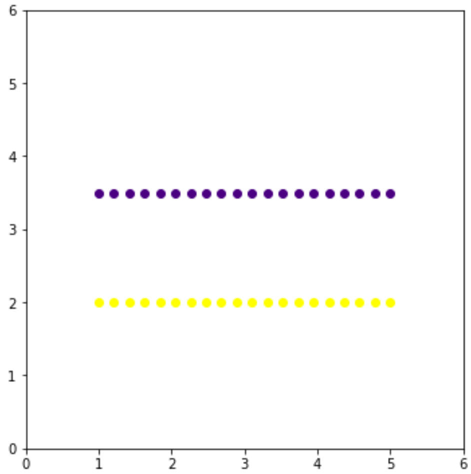
Consider the following data points:



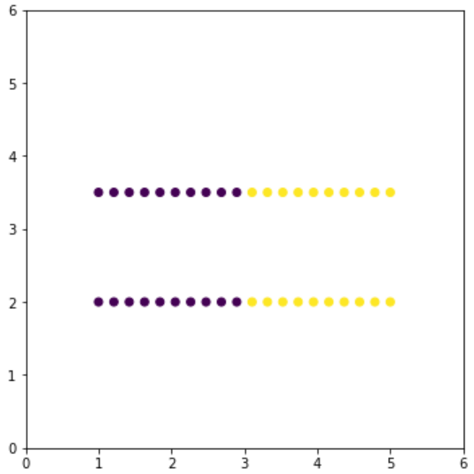
Assume K-means is run on these points with k = 2. Which of the following are expected to be the clusters formed out of K-means?

Options

(a)



(b)



(c)

Depends on cluster center initializations and the distance between the two lines.

Answer

(c)

Solution

The kind of clusters obtained will indeed depend on the cluster initializations. In the specific case of the mid-points of each line being chosen as the initial clusters, the clusters obtained will be of the type of Option (a). In all other cases, the clusters will be of the option (b) type.

Another point to note is that the clusters obtained will depend on the distance between the lines as well. In the case where the distance between the two lines is much larger than the gap between the points on the lines, the cluster center initializations will not matter and we will get clusters of type (a).

Question-5

Statement

In the initalization step of k-means++, the squared distances from the closest mean for 5 points x_1, x_2, x_3, x_4, x_5 are: 25, 67, 89, 24, 56. In this context, which of the following is true?

Options

(a)

Any point out of of x_1, x_2, x_3, x_4, x_5 may be chosen uniformly at random as next mean.

(b)

Certainly x_3 will be chosen as its distance from closest mean is largest.

(c)

x_3 will be chosen with the highest probability, but we are not sure whether this point will definitely be chosen.

Answer

(c)

Solution

K-means++ performs a smart initialization of cluster centers. The first cluster center is chosen uniformly at random out of all data points. To choose the next cluster center, the squared distances of all the remaining data points from the first cluster center are computed. These squared distances become the scores for these data points, which are further normalized and are treated as probabilities for being chosen as the next data point.

In the given question, although the score of x_3 is the maximum, it will result into the highest probability for x_3 to be chosen as the next cluster center. However, since it is probabilistic, we can not guarantee that x_3 will certainly be chosen as the next cluster center.

Question-6

Statement

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

Options

(a)

The partition configurations can not repeat themselves.

(b)

After doing the reassignments, we might get the same partition configuration again.

(c)

Objective function after making the re-assignments strictly reduces.

(d)

Objective function after making the re-assignments may increase.

(e)

Change of value of objective function indicates that the partition configuration has changed.

(f)

For partitioning n data points across k partitions, Lloyd's algorithm takes k^n iterations to converge.

Answer

(a), (c), (e)

Solution

(a) - (e) In Lloyd’s algorithm, in each iteration, re-assignments happen only for those data points, which find a mean that is closer to them than their own mean. Hence, the value of the objective function strictly reduces, resulting in a new partition. Since the value of objective function can not be same for any two partitions, this means that the partitions can not repeat themselves.

(f) k^n is only the upper limit of number of possible partitions for k clusters and n data points.

Question-7

Statement

For 1000 data points, out of k = 1, 10 and 100, which value of k is likely to result in the maximum value of the objective function?

Options

(a)

1

(b)

10

(c)

100

(d)

Insufficient information. Depends on data.

Answer

(a)

Solution

If all data points are in the same cluster (i. e., $k = 1$), the value of the objective function will be high, as there will be only one mean, and every point's distance will be measured from this one mean. As the value of k is increased, due to the presence of more means, the distances of the data points from these means will reduce.

Therefore, $k = 1$ will result in the maximum value of objective function.

Question-8

Statement

For 100 data points, if k = 100, what will be the value of the objective function?

Options

(a)

100

(b)

0

(c)

100*100

Answer

(b)

Solution

For 100 data points, if $k = 100$, this means that every point is in their own cluster. In this case, since the point itself will represent the mean in each cluster, the distance of each point from its mean will be zero, resulting in zero value of the objective function.

Question-9

Statement

Choose the correct statements:

Options

(a)

In k-means algorithm, all cluster initializations lead to the same result.

(b)

One initialization might get stuck in local minima, while another may lead to global minima.

(c)

One initialization may converge while another may not.

(d)

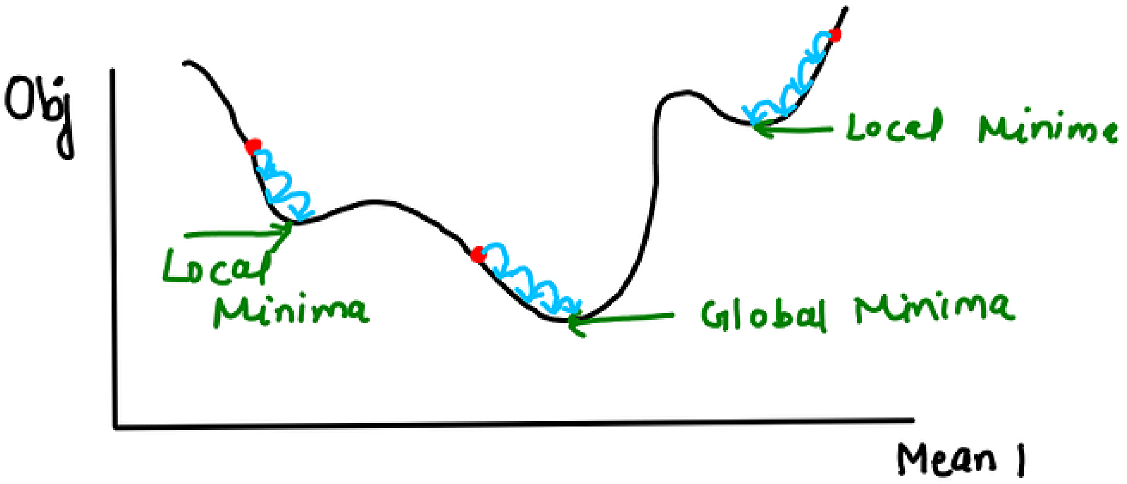
The initialization of cluster centres may affect the number of iterations K-means takes to converge.

Answer

(b), (d)

Solution

The following image shows how different initial cluster means may result into k-means converging in either local minima or global minima. Further, depending on the initial cluster means, the number of iterations required for K-means to converge might vary.



Question-10

Statement

Outliers are data points that deviate significantly from the rest of data points. Knowing the way Lloyd’s algorithm works, do you think it is sensitive to outliers?

Options

(a)

Yes

(b)

No,

Answer

(a)

Solution

Since K-means is based on computing means and euclidean distances, the presence of outliers may affect the inherent clustering present in the ‘non-outliers’ .