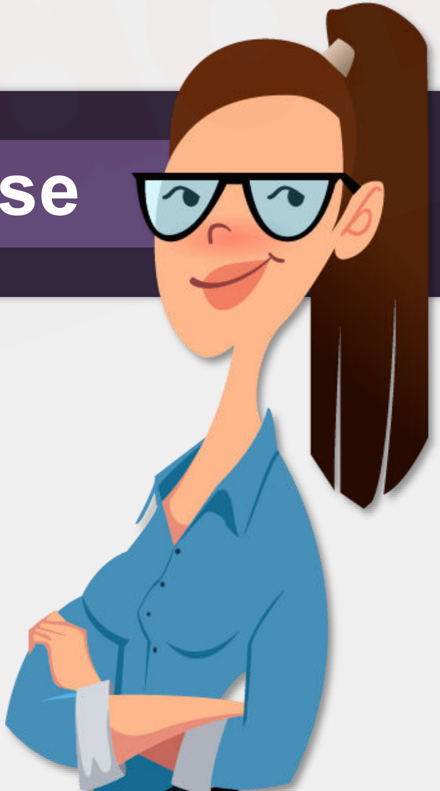
Python Certification Course



Unsupervised Learning: Clustering

Things you will learn after this Session

Unsupervised Learning: Clustering

- What is Clustering?
- Use Case of Clustering
- Types of Clustering

What is Clustering?

“Clustering is the process of dividing the datasets into groups, consisting of similar data-points”

- Points in the same group are as similar as possible
- Points in different group are as dissimilar as possible

Example of Clustering

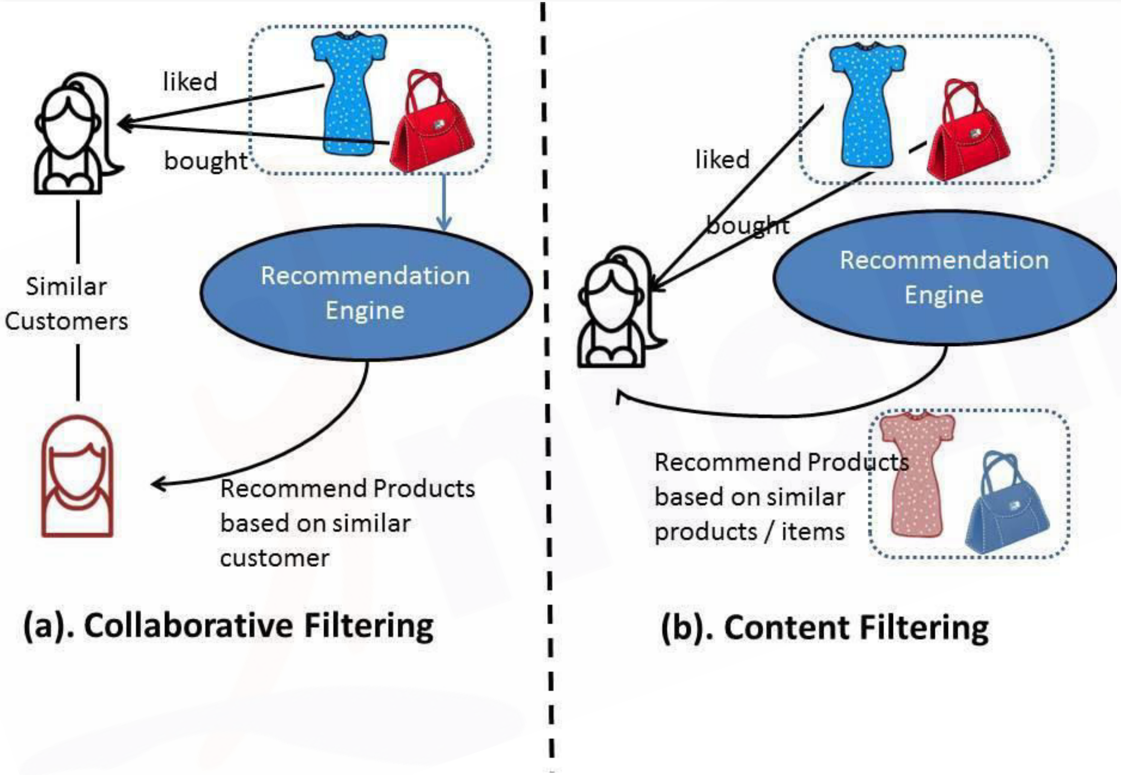
Example 1: Cluster of different colors of FROOT LOOPS

Example 2: Cluster of different colors of Fruits

PAPER
METAL
PLASTIC
E-WASTE
GLASS
ORGANIC

Example 3: Cluster of different types of Garbage

Use-Case of Clustering


liked
bought
Similar Customers
Recommendation Engine
Recommend Products based on similar customer
liked
bought
Recommendation Engine
Recommend Products based on similar products / items

(a). Collaborative Filtering

(b). Content Filtering

amazon

NETFLIX

Recommendation Engine

Where is it used?

amazon

NETFLIX

flickr

Types of Clustering

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering

Types of Clustering

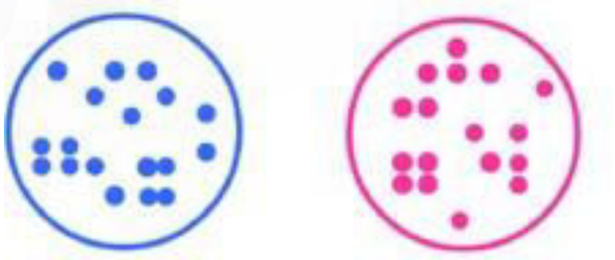
Exclusive Clustering

Overlapping Clustering

Hierarchical Clustering

Exclusive Clustering

- Each data object can only exist in one cluster
- For Example: K-Means Clustering



Types of Clustering

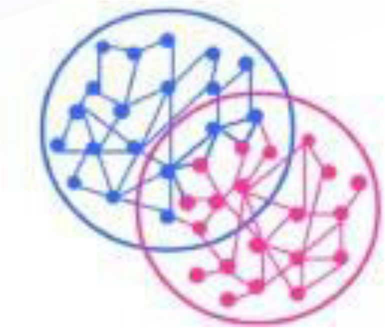
Exclusive Clustering

Overlapping Clustering

Hierarchical Clustering

Overlapping Clustering

- Allows data objects to be grouped in 2 or more clusters
- For Example: Fuzzy/ C-Means Clustering
- In Fuzzy clustering every data object belongs to every cluster



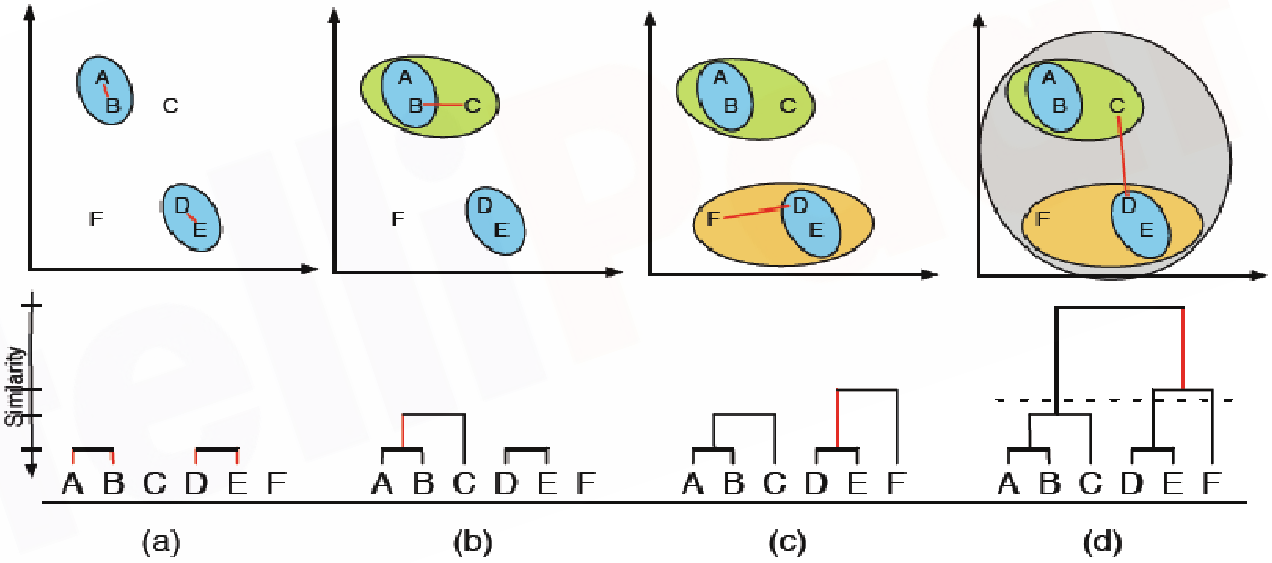
Types of Clustering

Exclusive Clustering

Overlapping Clustering

Hierarchical Clustering

Hierarchical Clustering



Understanding K – Means Clustering

Things you will learn after this Session

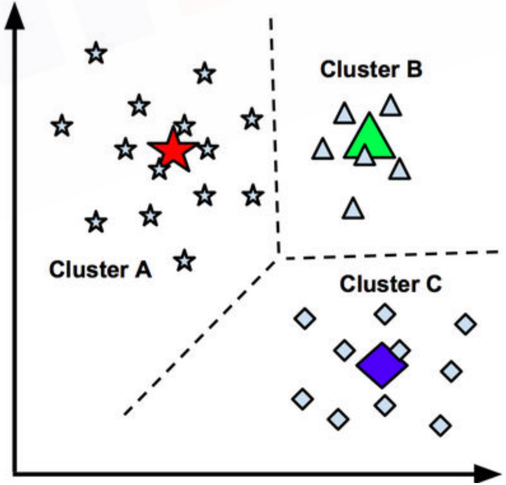
Understanding K – Means Clustering

- What is K – Means Clustering?
- Where can you use it?
- Step by step calculation of K-Means Clustering Algorithm

What is K-Means Clustering?

"**K-Means** is a clustering algorithm which focuses on grouping similar elements or data points into a cluster."

NOTE: 'K' in K-Means represent the number of clusters



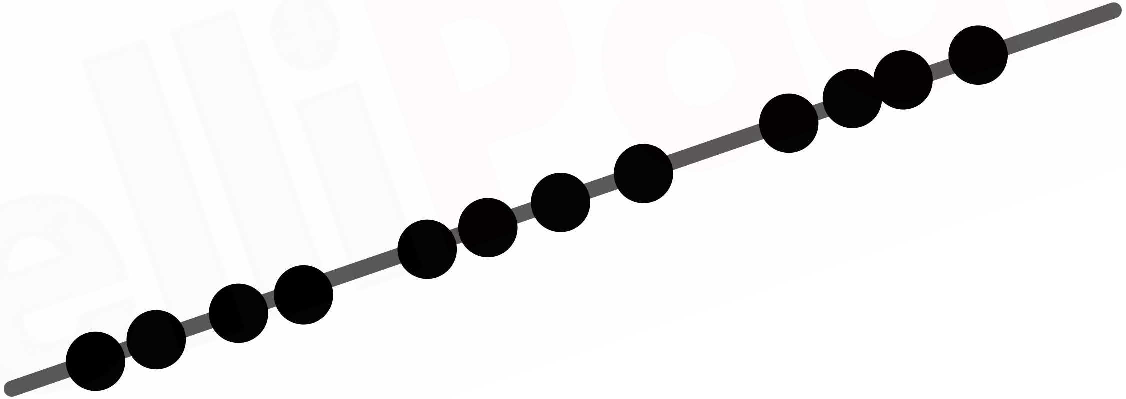
What is K-Means Clustering?

Pile of Laundry

Business Application of K-Means

- Behavioural Segmentation
- Inventory Categorization
- Sorting sensor measurements
- Detecting bots or anomalies

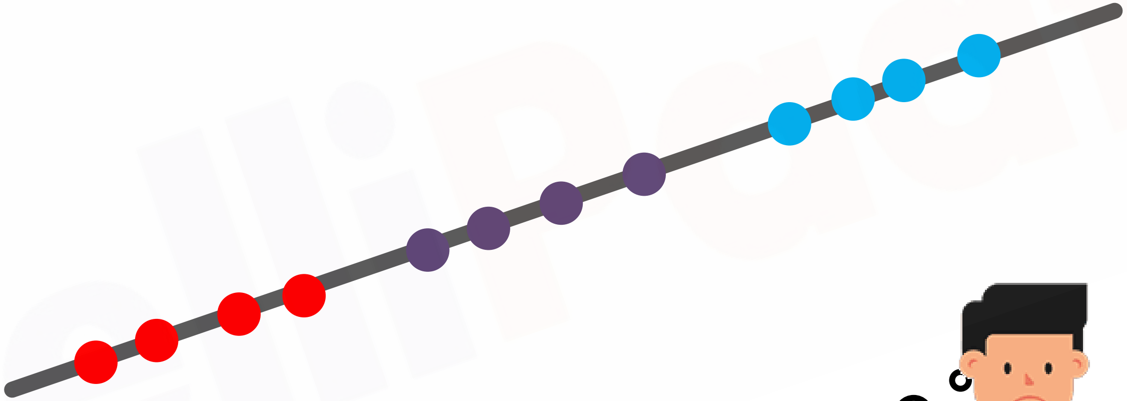
Understand K-Means Algorithm



Number of Clusters = 3

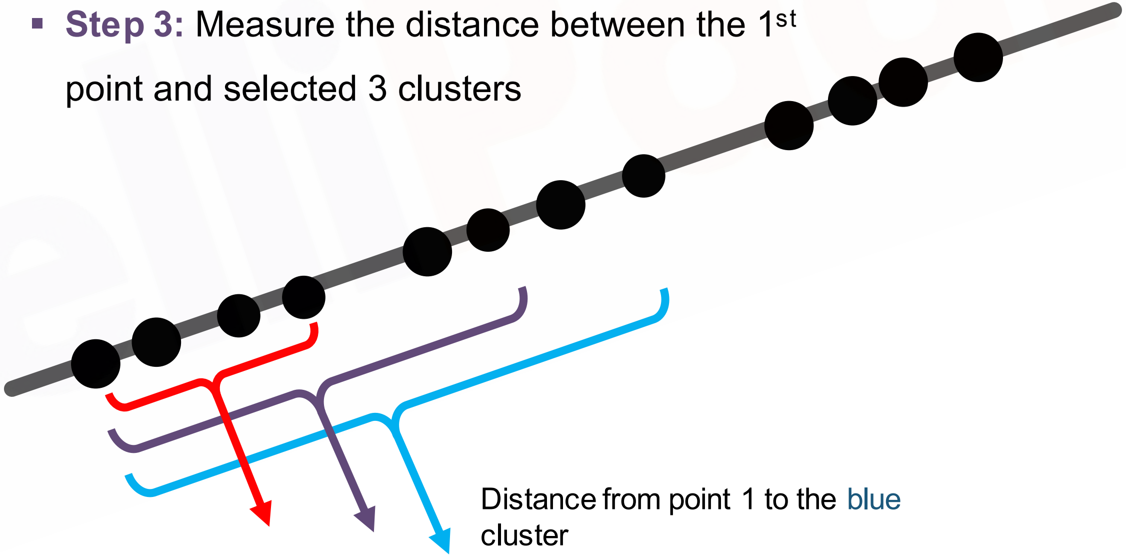
Understand K-Means Algorithm

Number of Clusters, K = 3



Understand K-Means Algorithm

- **Step 1:** Select the number of clusters to be identified, i.e select a value for K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters

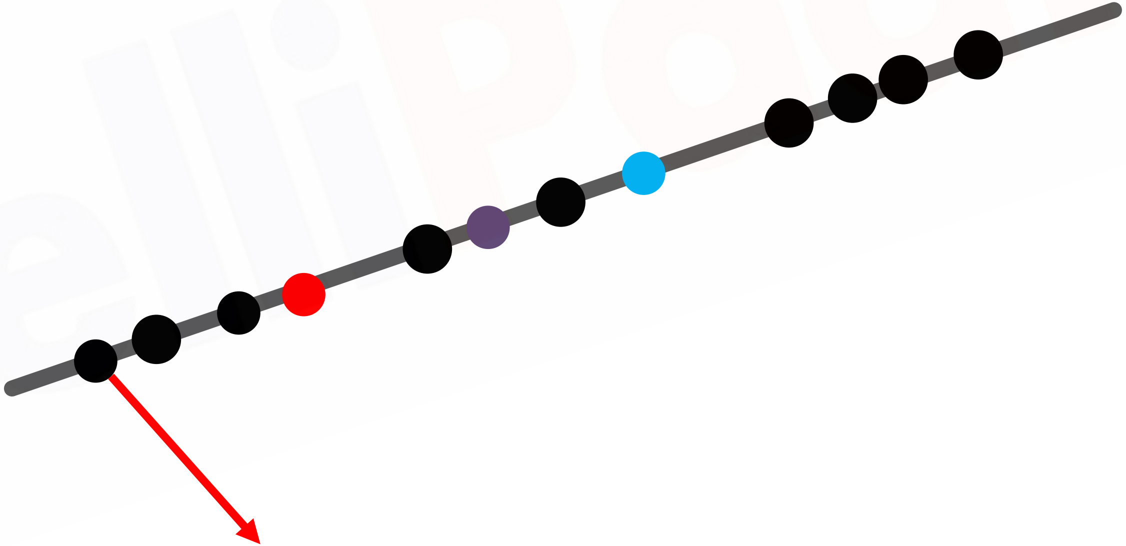


Distance from point 1 to the blue cluster

Distance from point 1 to the red cluster

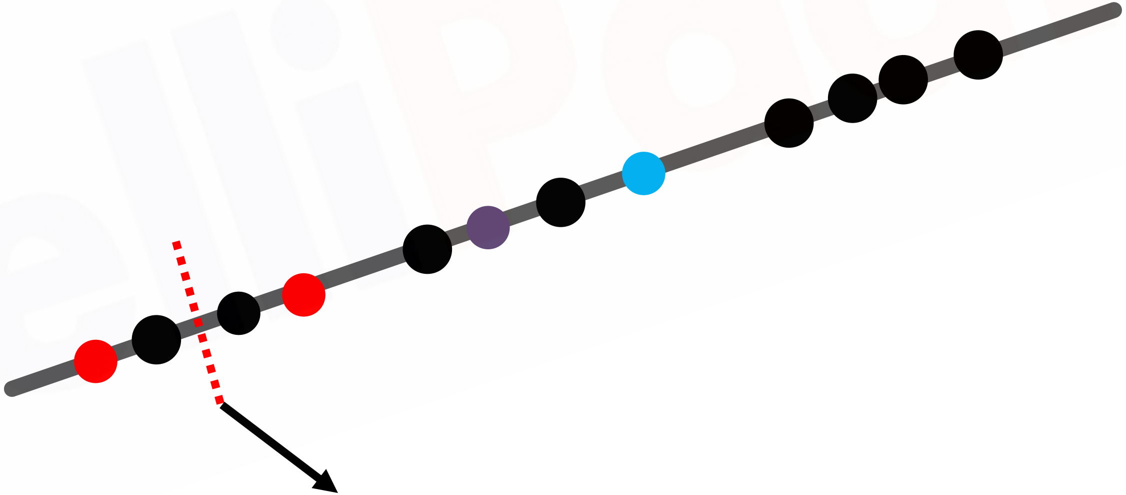
Distance from point 1 to the purple cluster

Understand K-Means Algorithm



Step 4: Assign the 1st point to nearest cluster (red in this case).

Understand K-Means Algorithm

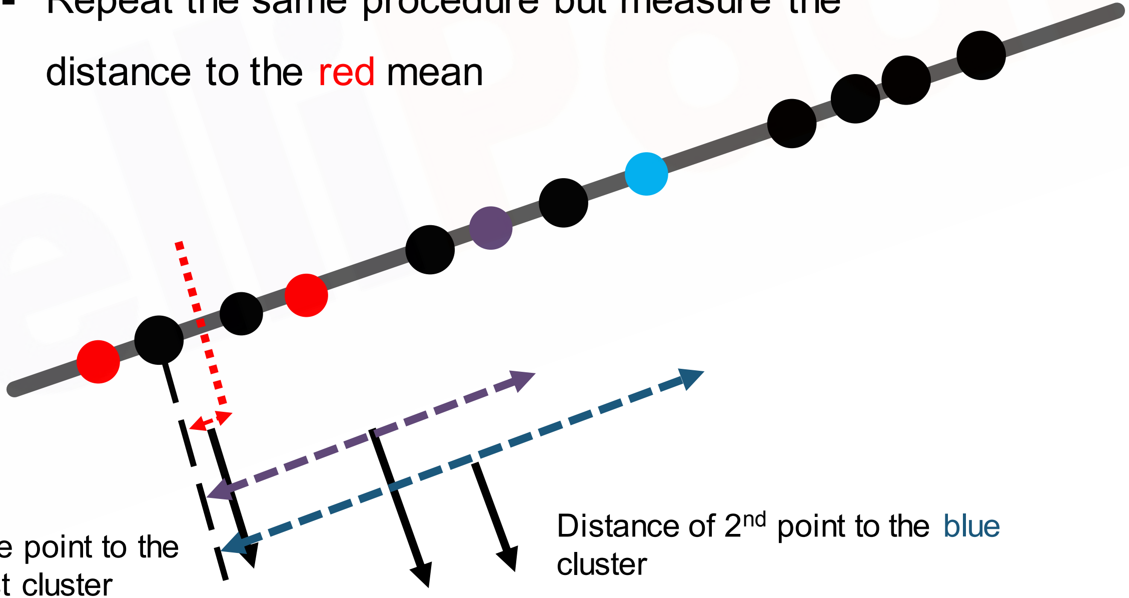


Step 5: Calculate the mean value including the new point for the red cluster

Understand K-Means Algorithm

Find to which cluster does point 2 belongs to, how?

- Repeat the same procedure but measure the distance to the red mean



Add the point to the nearest cluster

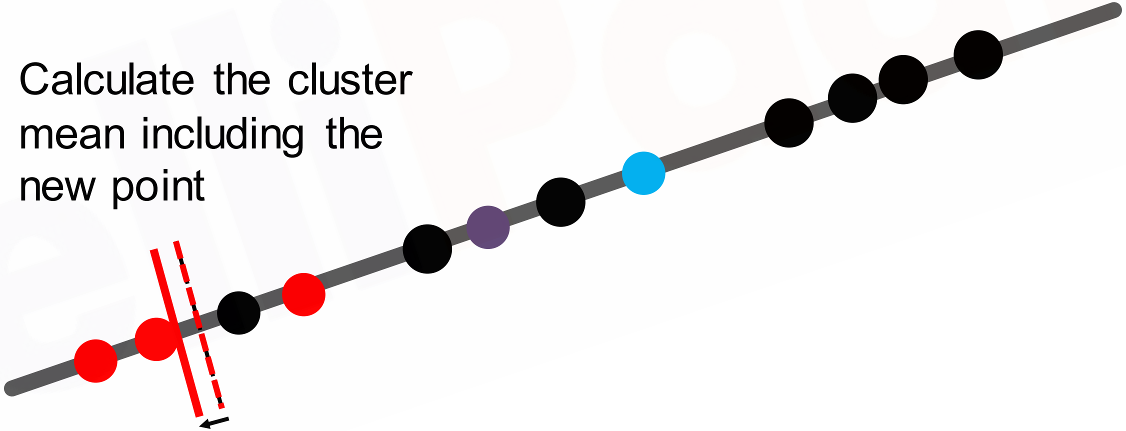
Distance of 2nd point to the red cluster

Distance of 2nd point to the purple cluster

Distance of 2nd point to the blue cluster

Understand K-Means Algorithm

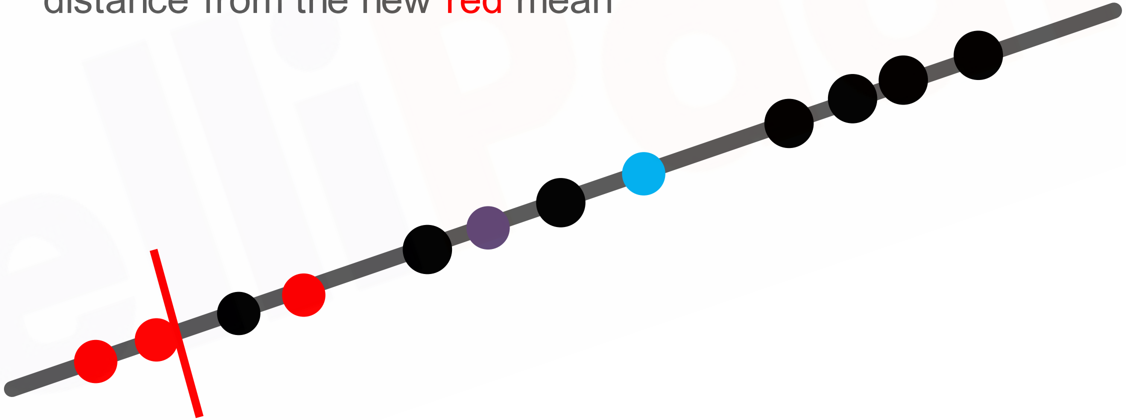
Calculate the cluster mean including the new point



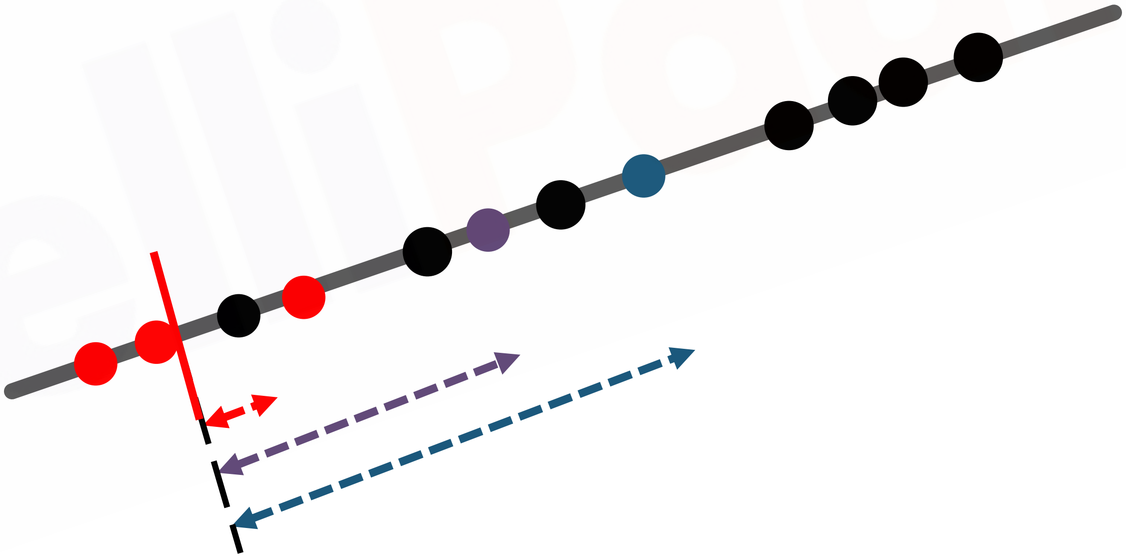
Understand K-Means Algorithm

Find to which cluster does point 3 belongs to, how?

- Repeat the same procedure but measure the distance from the new red mean

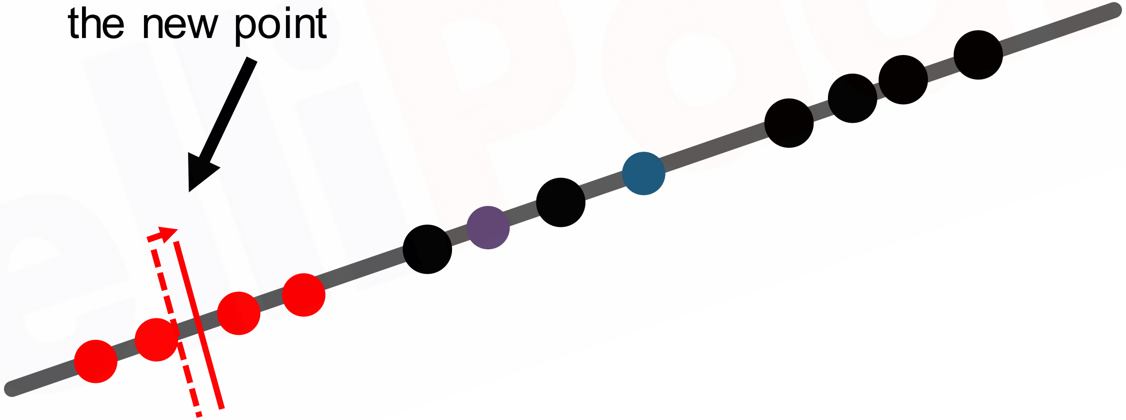


Understand K-Means Algorithm



Measure the distance and add the 3rd point to the cluster(red) having the minimum distance

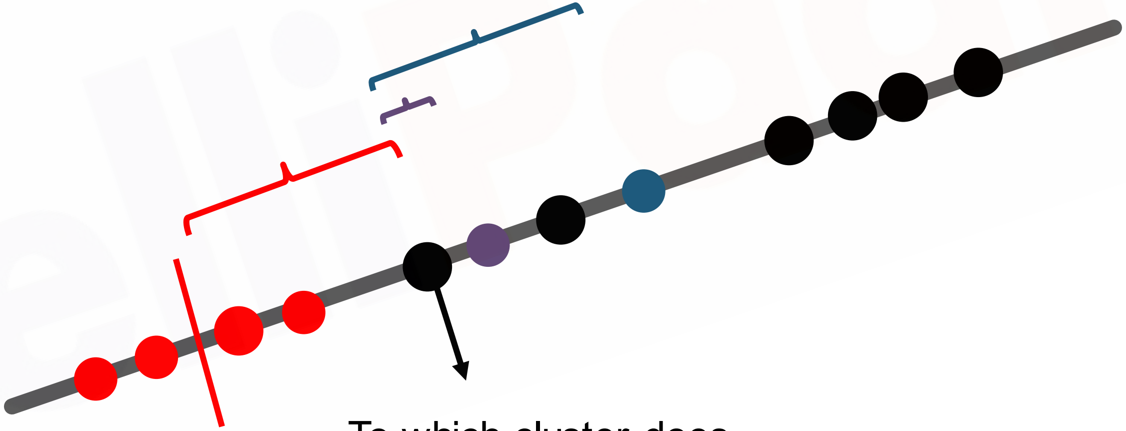
Understand K-Means Algorithm



Calculate the new cluster mean using the new point

Understand K-Means Algorithm

- **Measure the distance**
- Assign the point to the nearest cluster
- Calculate the cluster mean using the new point

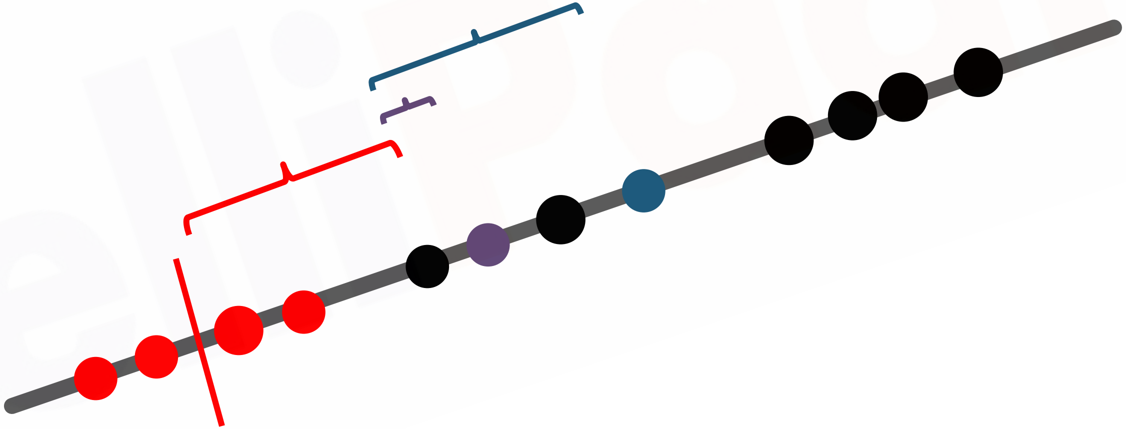


To which cluster does this point belongs to?

REPEAT THE STEPS AGAIN…

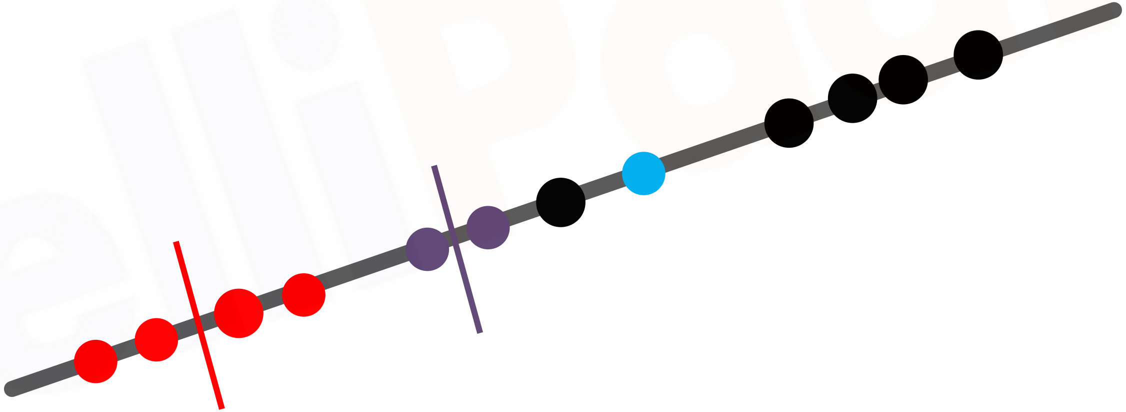
Understand K-Means Algorithm

- Measure the distance
- **Assign the point to the nearest cluster**
- Calculate the cluster mean using the new point



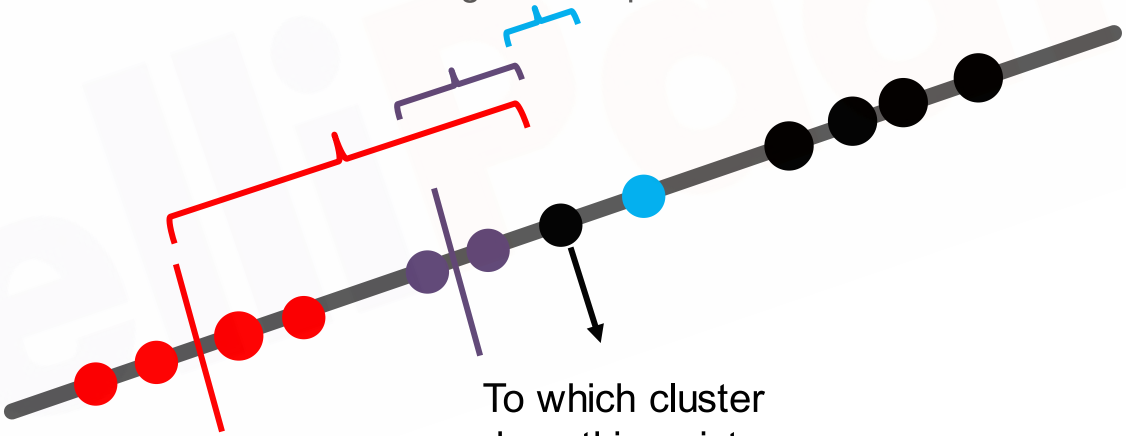
Understand K-Means Algorithm

- Measure the distance
- Assign the point to the nearest cluster
- **Calculate the cluster mean using the new point**



Understand K-Means Algorithm

- **Measure the distance from the cluster mean (centroids)**
- Assign the point to the nearest cluster
- Calculate the cluster mean using the new point

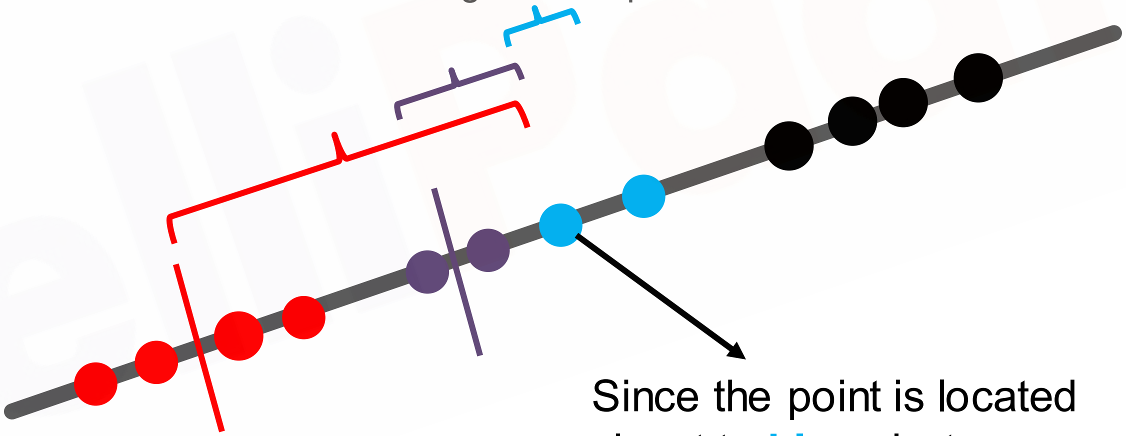


To which cluster does this point belongs to?

REPEAT THE SAME STEPS UNTILL ALL THE CLUSTERS ARE ASSIGNED…

Understand K-Means Algorithm

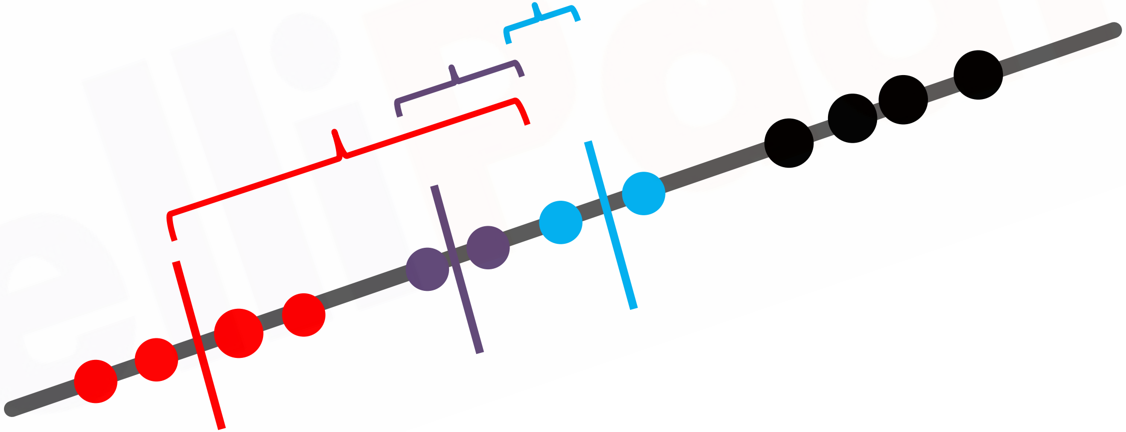
- Measure the distance from the cluster mean (centroids)
- **Assign the point to the nearest cluster**
- Calculate the cluster mean using the new point



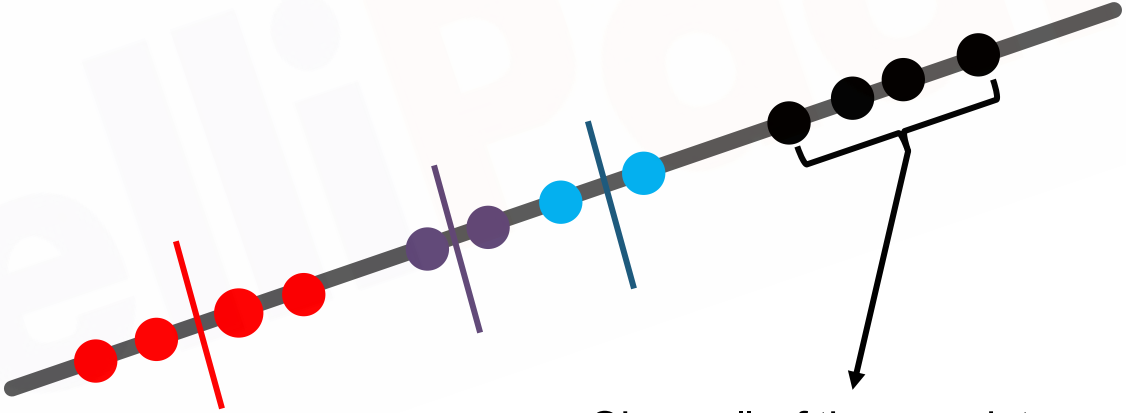
Since the point is located closet to **blue** cluster

Understand K-Means Algorithm

- Measure the distance from the cluster mean (centroids)
- Assign the point to the nearest cluster
- **Calculate the cluster mean using the new point**

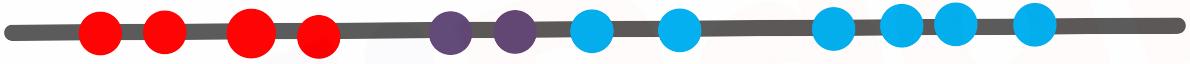


Understand K-Means Algorithm

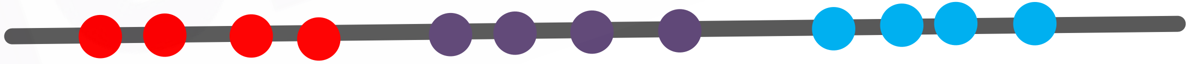


Since all of these points are located closet to **blue** cluster so all of them will be assigned to blue cluster

Understand K-Means Algorithm

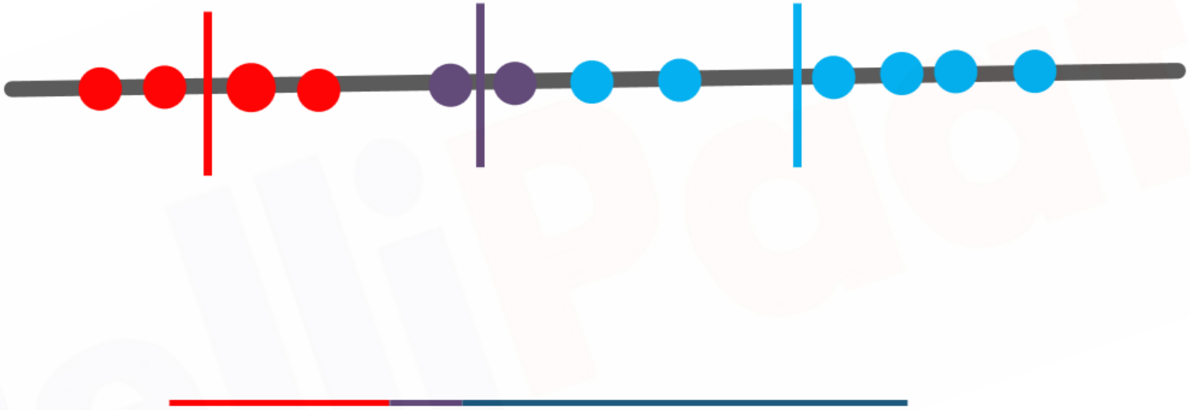


Result from 1st iteration



Original/Expected Result

Understand K-Means Algorithm

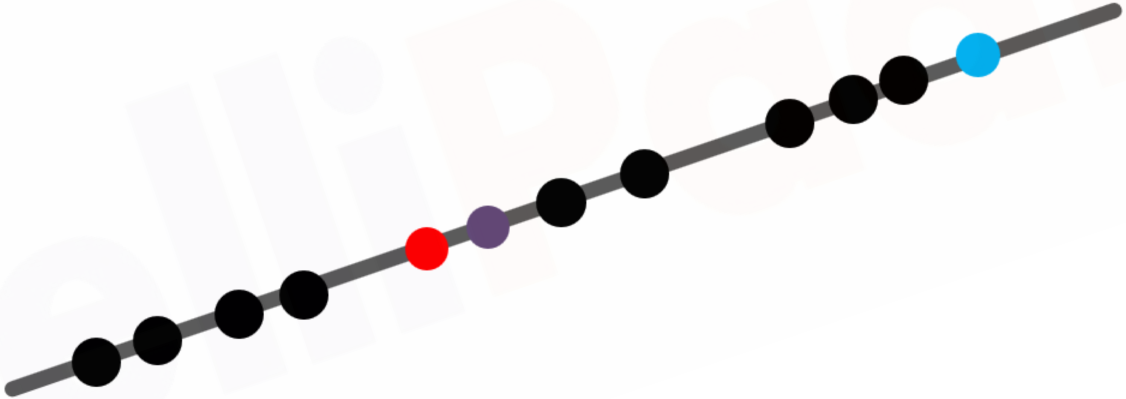


Total variation within the cluster

According to the K-Means Algorithm it iterates over again and again unless and until the data points within each cluster stops changing

Understand K-Means Algorithm

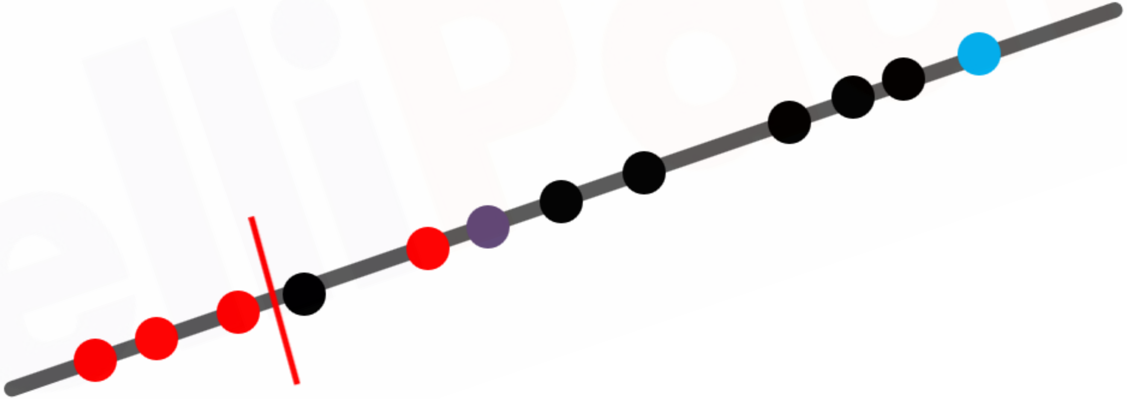
Iteration 2: Again we will start from the scratch. But this time we will select different initial random point (as compared to the 1st iteration)



- **Step 1:** Select the number of clusters to be identified, i.e. K =3 in this case
- **Step 2:** Randomly select 3 distinct data point
- **Step 3:** Measure the distance between the 1st point and selected 3 clusters

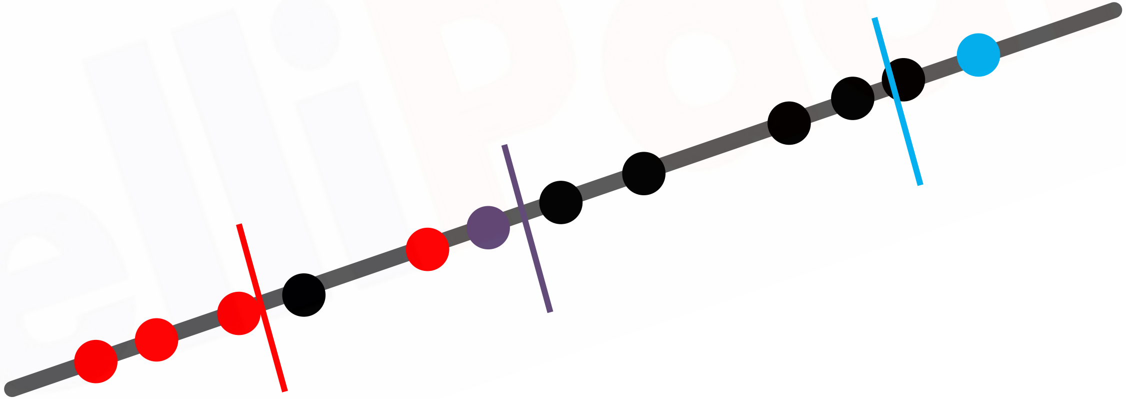
Understand K-Means Algorithm

Algorithm picks 3 initial clusters and adds the remaining points to the cluster with the nearest mean, and again recalculating the mean each time a new point is added to the cluster



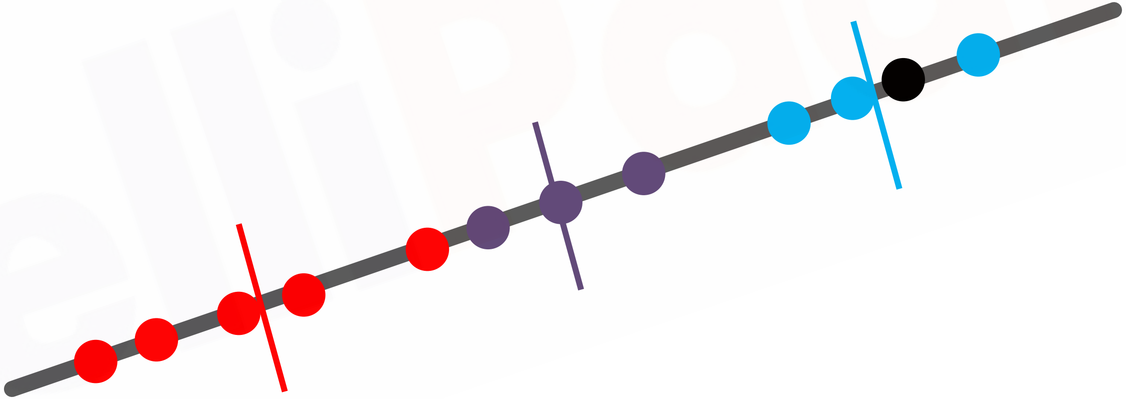
Understand K-Means Algorithm

Algorithm picks 3 initial clusters and adds the remaining points to the cluster with the nearest mean, and again recalculating the mean each time a new point is added to the cluster



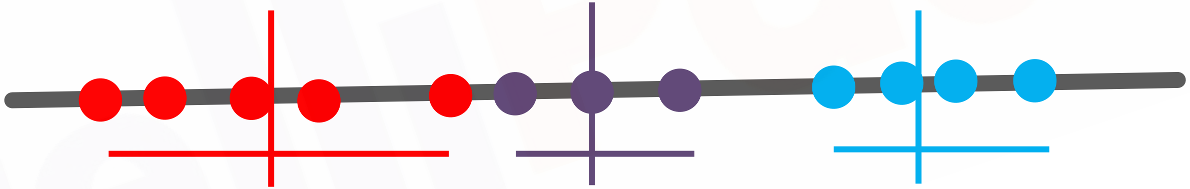
Understand K-Means Algorithm

Algorithm picks 3 initial clusters and adds the remaining points to the cluster with the nearest mean, and again recalculating the mean each time a new point is added to the cluster



Understand K-Means Algorithm

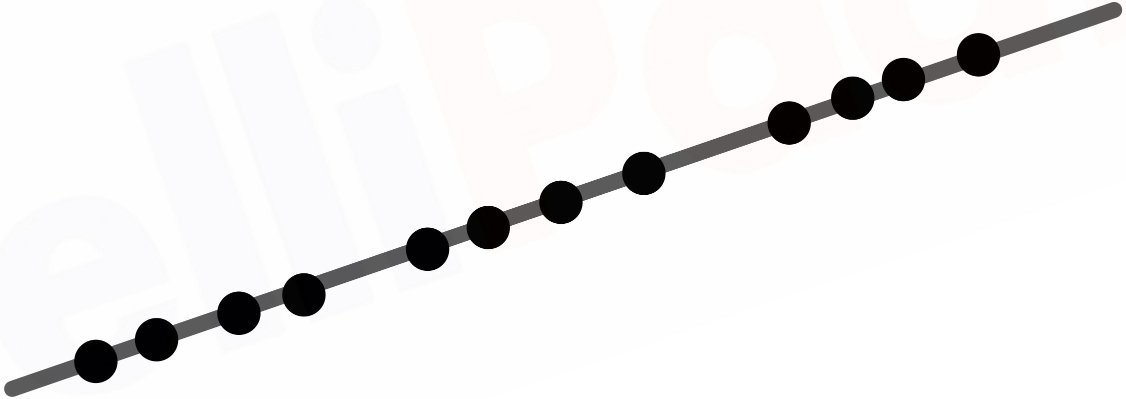
Algorithm picks 3 initial clusters and adds the remaining points to the cluster with the nearest mean, and again recalculating the mean each time a new point is added to the cluster



Total variation within the cluster

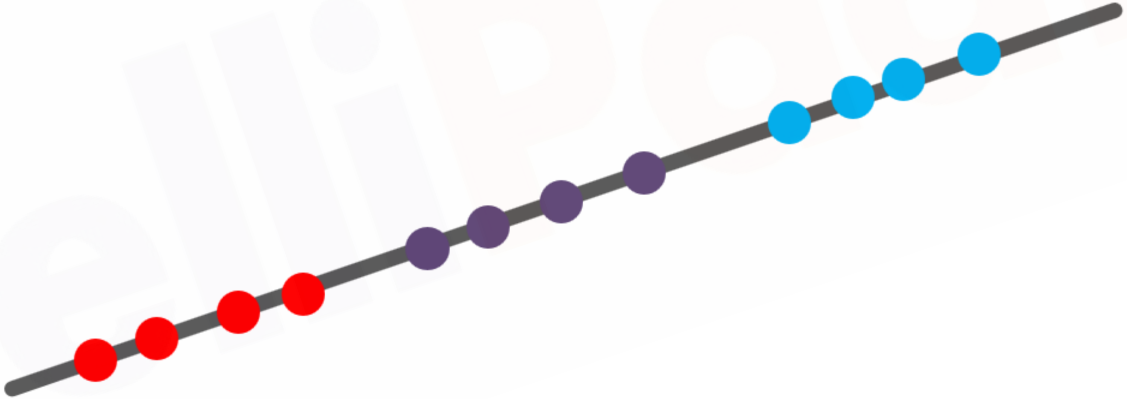
Understand K-Means Algorithm

Iteration 3: Again we will start from the scratch and select different initial random point (as compared to 1st and 2nd iteration)



Pick 3 initial clusters

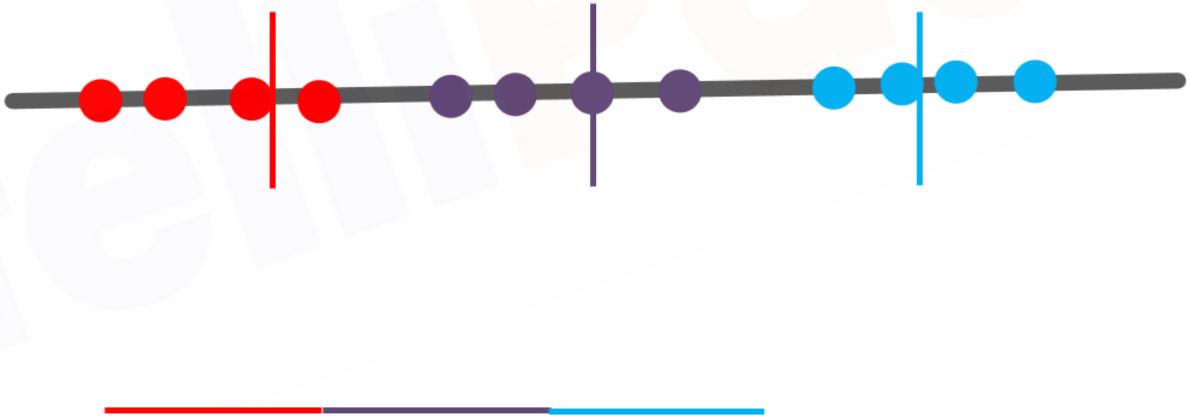
Understand K-Means Algorithm



Cluster the remaining points

Understand K-Means Algorithm

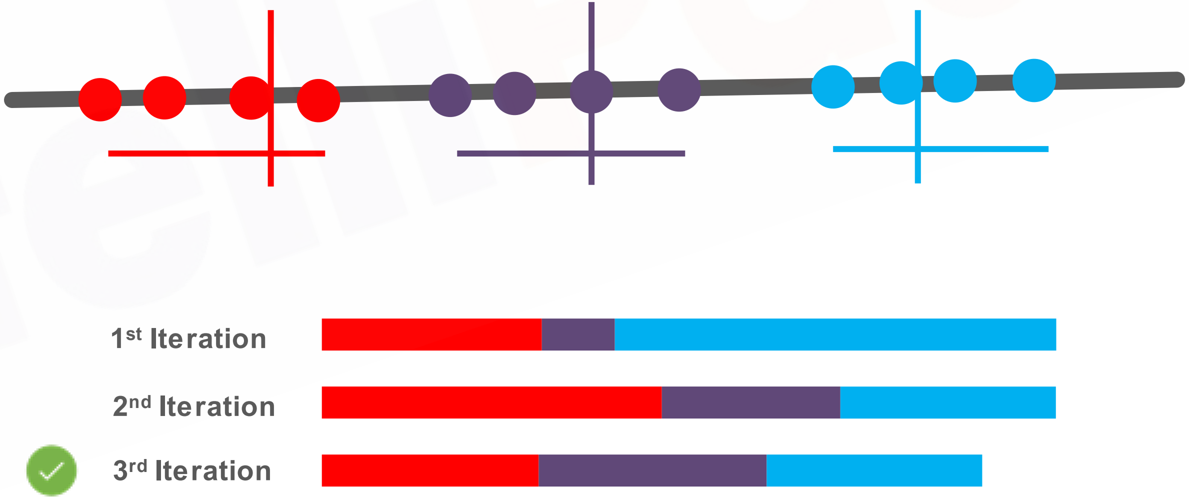
Finally sum the variation within each cluster



Total variation within the cluster

Understand K-Means Algorithm

Finally sum the variation within each cluster



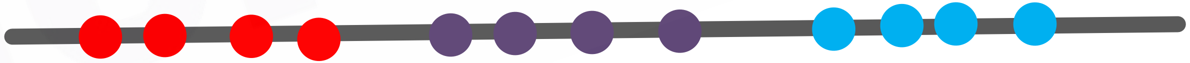
Understand K-Means Algorithm

But how to find the value of ‘K’?

Understand K-Means Algorithm

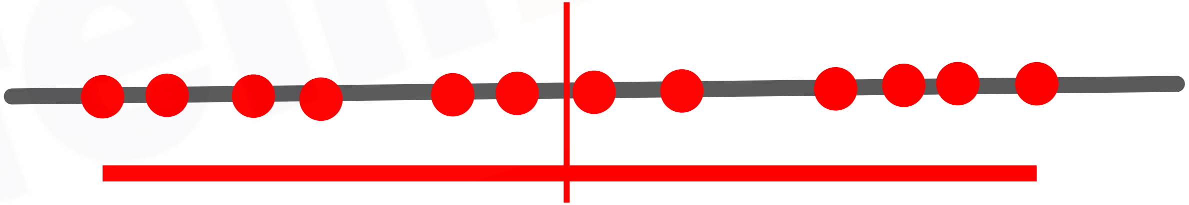
How to find the value of ‘K’?

In the previous scenario k = 3 was known, but what-if we don’t know the exact value of k?



Understand K-Means Algorithm

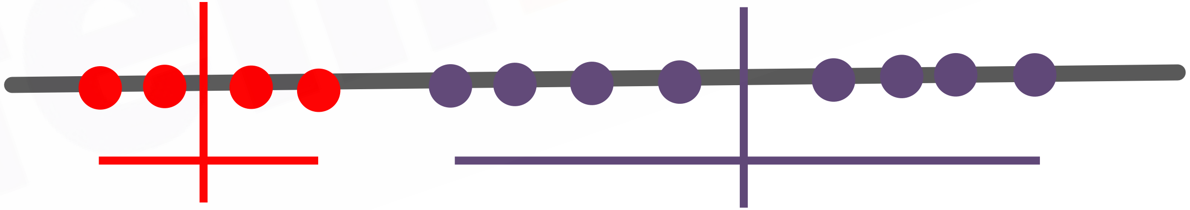
For finding the value of k, you will use hit and trail method, starting from K = 1



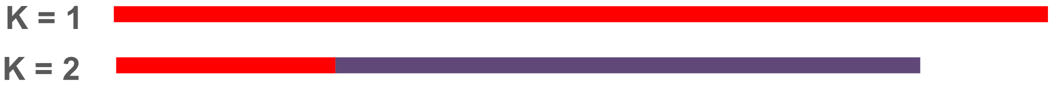
K=1 is the worst case scenario, even you cross-verify it with its total variation(all red)

Understand K-Means Algorithm

Now try with K = 2

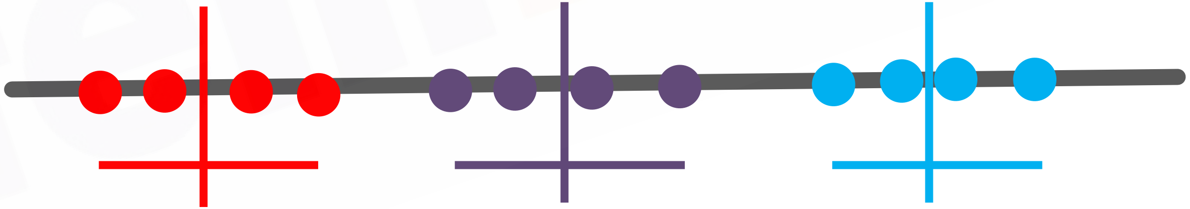


K=2 is still better then K = 1 (Total Variation)

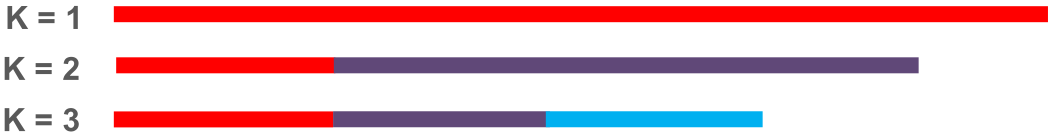


Understand K-Means Algorithm

Now try with K = 3



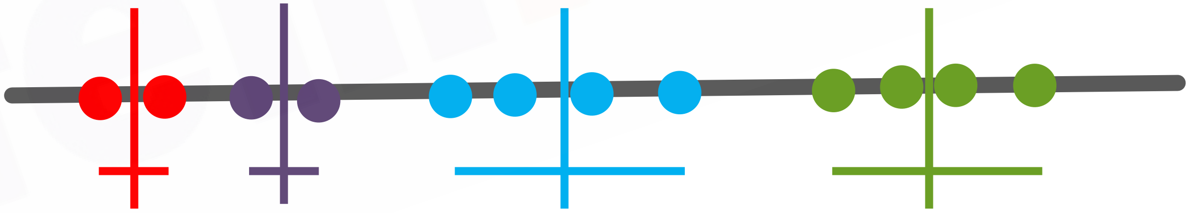
K=3 is still better then K = 2 (Total Variation)



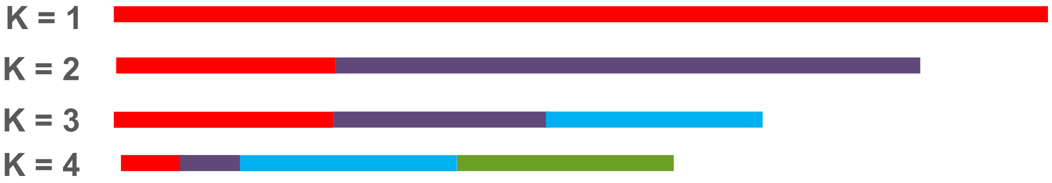
Understand K-Means Algorithm

Now try with K = 4

- Every time you increase the cluster the variation decreases
- If no. of clusters = no. of data points then in that case the variation = 0

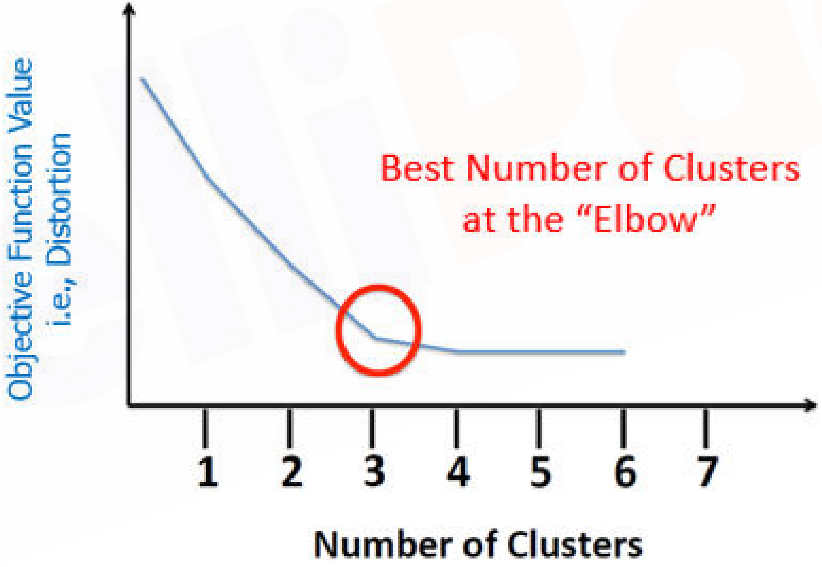


K=4 is still better then K = 3 (Total Variation)



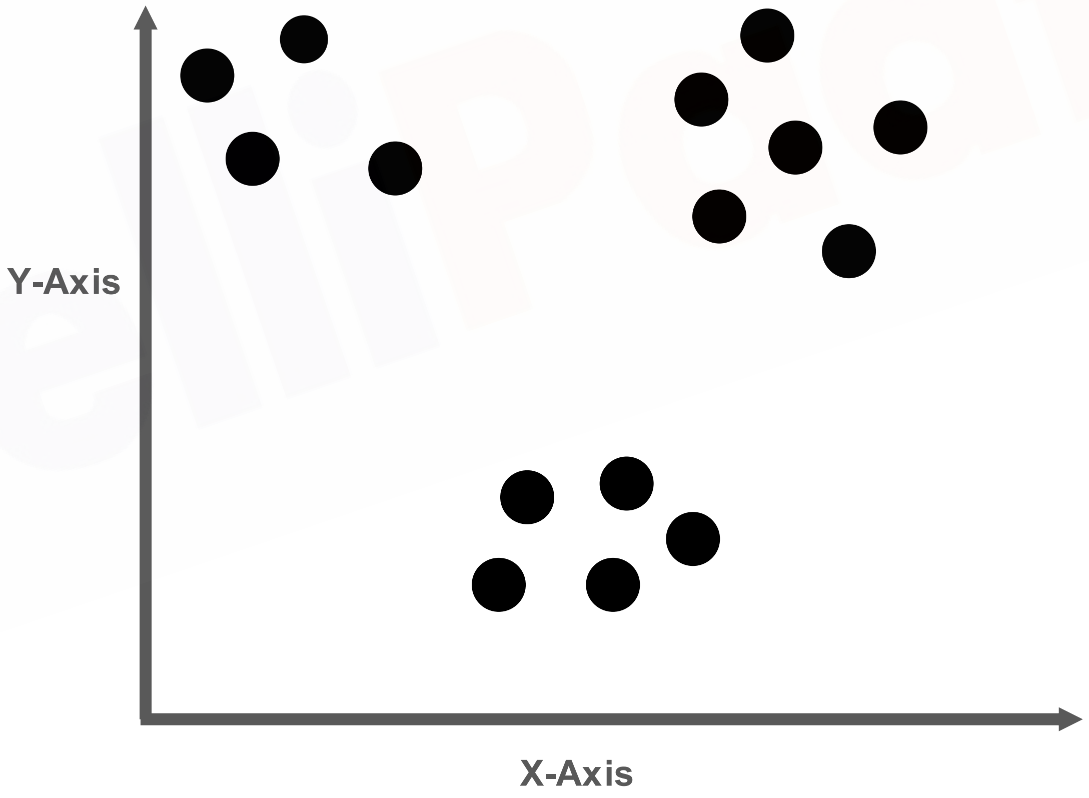
Understand K-Means Algorithm

Finally



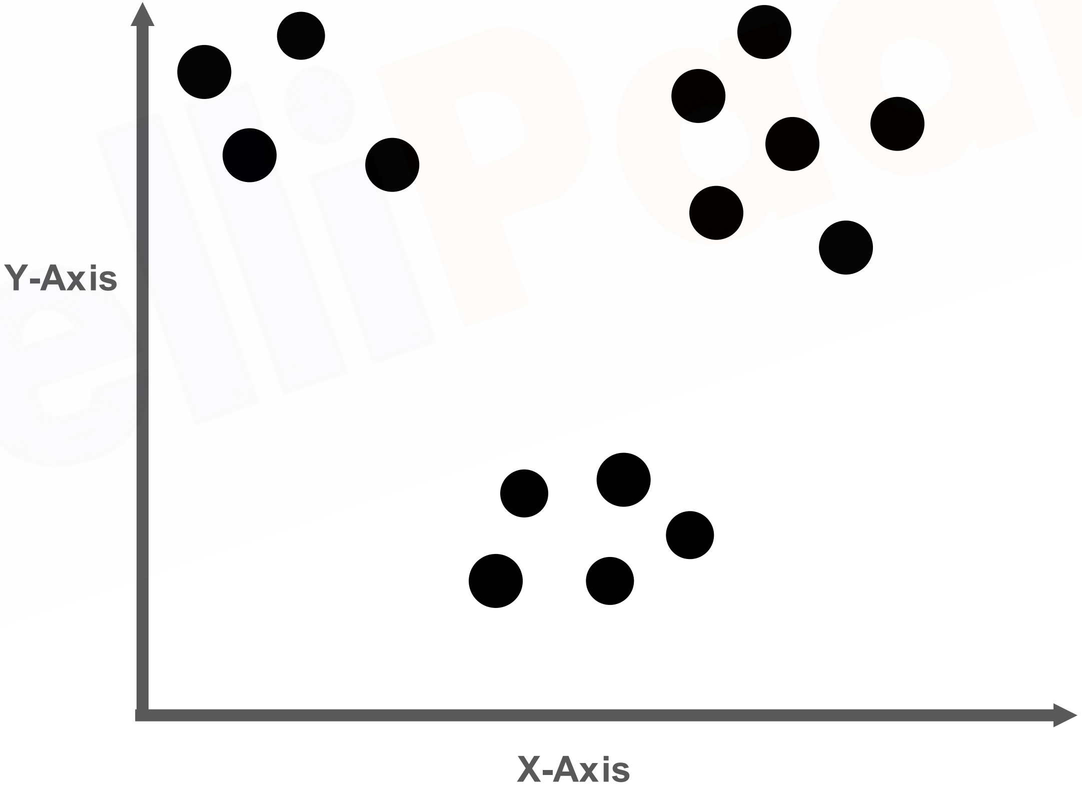
Understand K-Means Algorithm

Now what if we have our data plotted on the X and Y axis



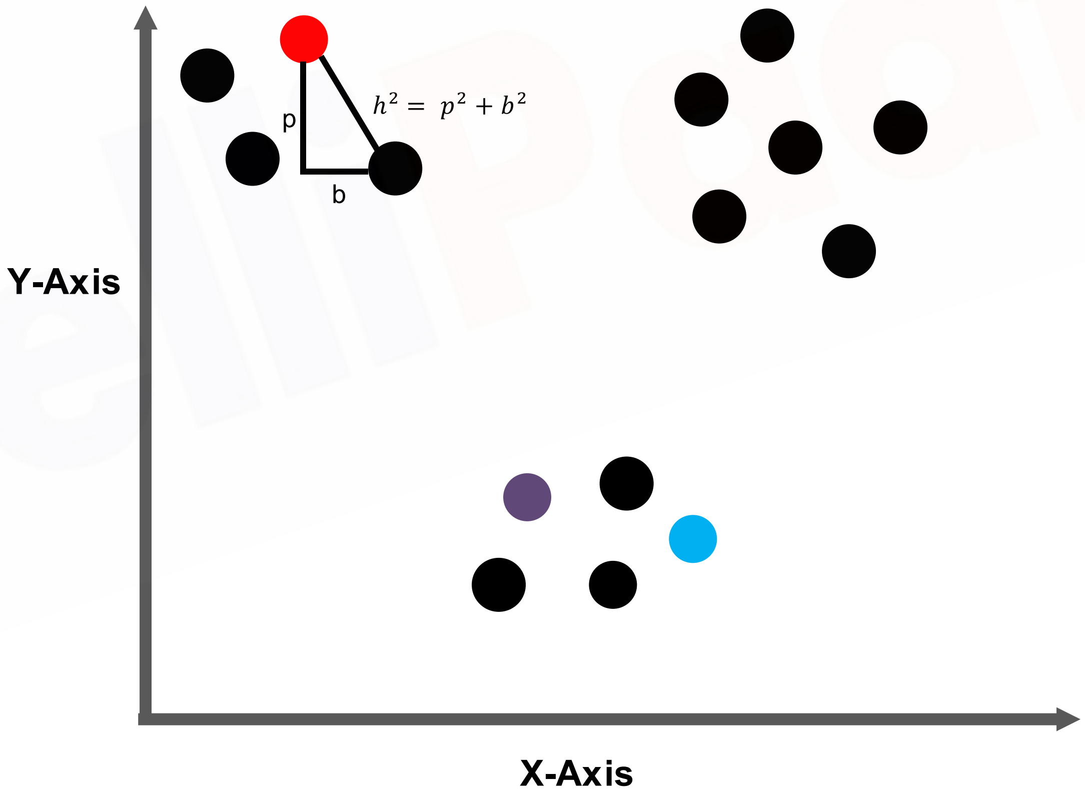
Understand K-Means Algorithm

Similarly, pick initial 3 random points..



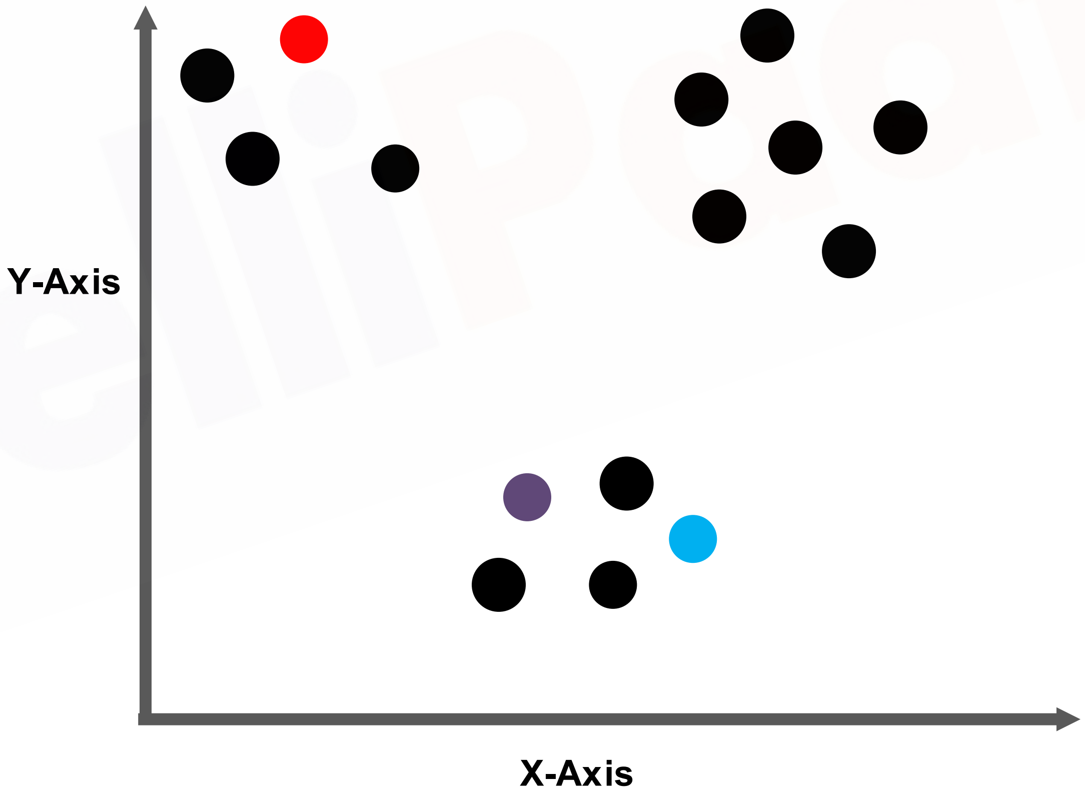
Understand K-Means Algorithm

We will be using the Euclidean distance (in 2D its same as that of a Pythagorean Theorem)



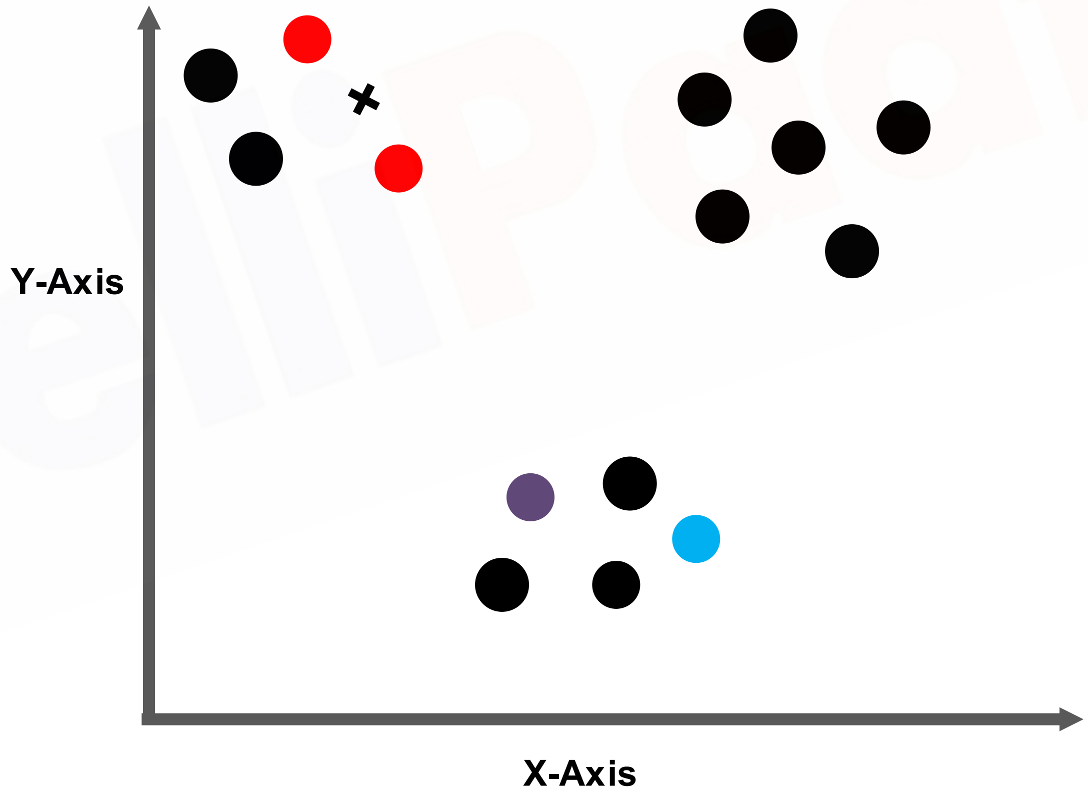
Understand K-Means Algorithm

Again assign the point to the nearest cluster



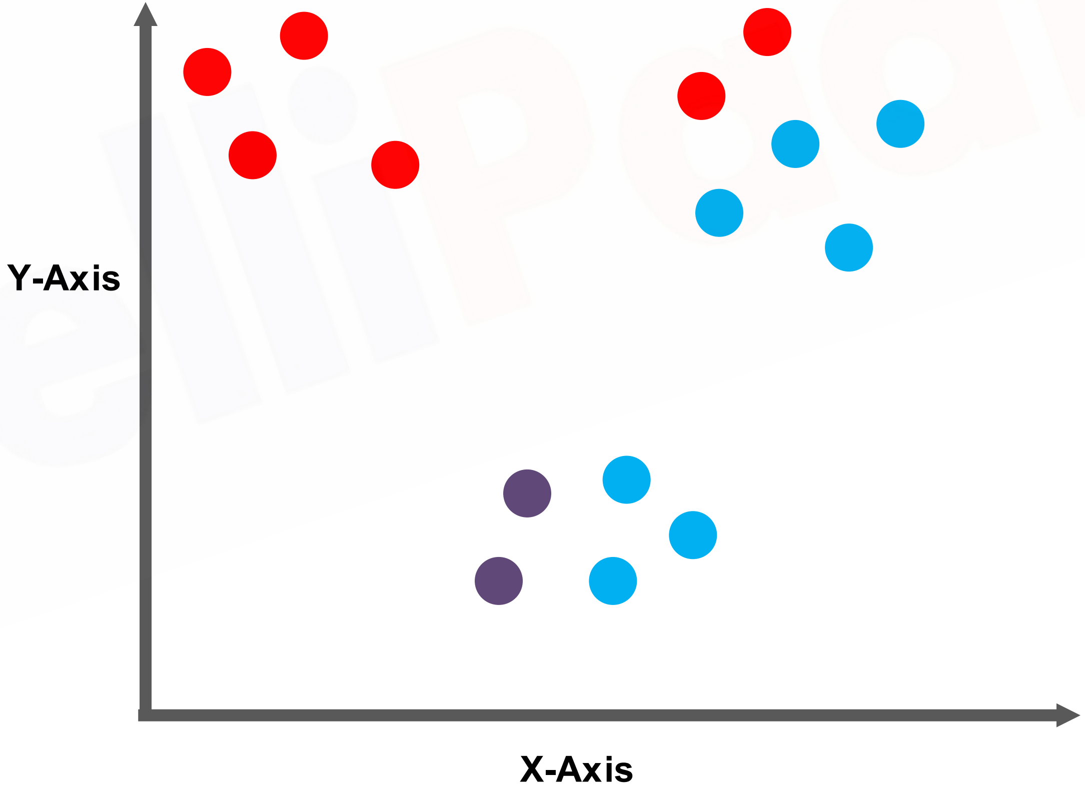
Understand K-Means Algorithm

Finally calculate the centroid (mean of cluster) including the new point



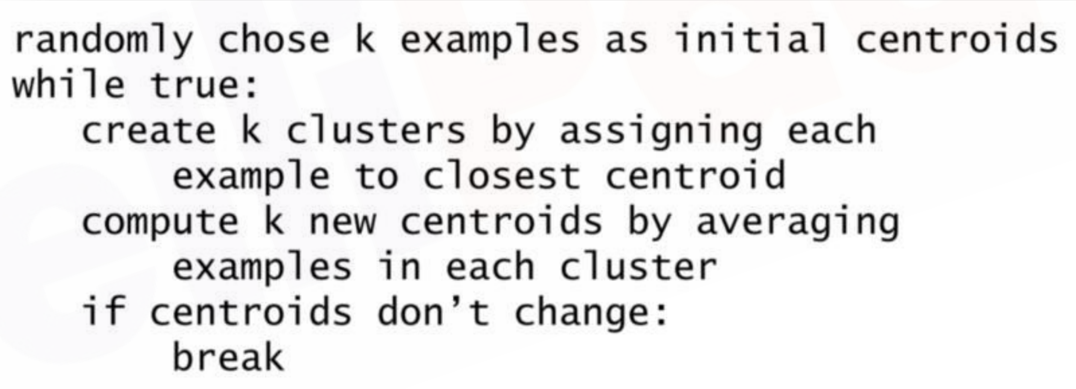
Understand K-Means Algorithm

Finally in first iteration you get something like this…again you have to iterate this process to get the final cluster



Understand K-Means Algorithm

Summarizing the K-Means Algorithm



```
randomly chose k examples as initial centroids
while true:
   create k clusters by assigning each
       example to closest centroid
   compute k new centroids by averaging
       examples in each cluster
   if centroids don't change:
       break
```

K-Means Clustering: Demo

Things you will learn after this Session

K-Means Clustering: Demo

- K – Means Clustering using Python
- K – Means Clustering using sklearn

QUIZ

Quiz 1

What is the minimum no. of variables/ features required to perform clustering?

A 0

B 1

C 2

D 3

Answer 1

What is the minimum no. of variables/ features required to perform clustering?

A	0
[✓ B]	[✓ 1]
C	2
D	3

Quiz 2

In which of the following cases will K-Means clustering fail to give good results?

A Data points with outliers

B Data points with different densities

C Data points with non-convex shapes

D All of the above

Answer 2

In which of the following cases will K-Means clustering fail to give good results?

A Data points with outliers

B Data points with different densities

C Data points with non-convex shapes

[✓ D All of the above]

Quiz 3

Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed for second iteration?

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

B C1: (6,6), C2: (4,4), C3: (9,9)

C C1: (2,2), C2: (0,0), C3: (5,5)

D None of these

Answer 3

Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:

C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed for second iteration?

[✓ A] C1: (4,4), C2: (2,2), C3: (7,7)

B C1: (6,6), C2: (4,4), C3: (9,9)

C C1: (2,2), C2: (0,0), C3: (5,5)

D None of these

Quiz 4

Customer segmentation is an example of

- A Classification
- B Clustering
- C Association
- D None of the above

Answer 4

Customer segmentation is an example of

A Classification

[✓ B Clustering]

C Association

D None of the above

Quiz 5

In K-Means, K stands for ___________

A Data sets

B Number of clusters

C Error function

Answer 5

In K-Means, K stands for ___________

A	Data sets
B	Number of clusters
C	Error function

www.intellipaat.com

CALL US NOW

India : +91-7847955955

US : 1-800-216-8930 (TOLL FREE)

sales@intellipaat.com