

★ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



# The Math Behind K-Means Clustering

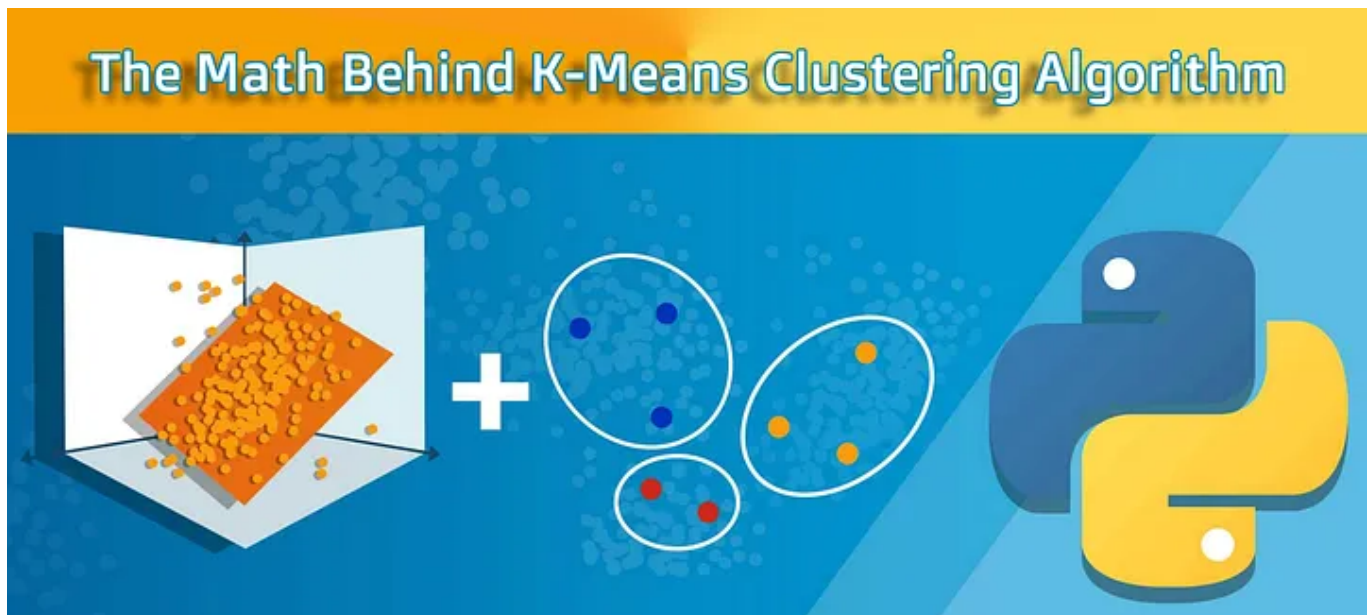


Dharmaraj · [Follow](#)

5 min read · Jan 26, 2022



19



## Introduction

**K**-Means Clustering is an Unsupervised Learning Algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters or groups that need to be created in the

process, as if  $K=5$ , there will be five clusters, and for  $K=10$ , there will be ten clusters, and so on. The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for  $K$  center points.
- Assigns each data point to its closest k-center. Groups assign based on  $k$  center points by measuring the distance between  $k$  points and data points.

In this blog, we are going to learn about the math behind the K-Means Clustering so if you want to learn how to implement K-Means Clustering please check my other blog [here with source code](#).

### **K-Means Clustering Algorithm-**

K-Means Clustering Algorithm involves the following steps:

**Step 1:** Calculate the number of  $K$  (Clusters).

**Step 2:** Randomly select  $K$  data points as cluster center.

**Step 3:** Using the Euclidean distance formula measure the distance between each data point and each cluster center.

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

**Step 4:** Assign each data point to that cluster whose center is nearest to that data point.

**Step 5:** Re-compute the center of newly formed clusters. The center of a cluster is computed by taking the mean of all the data points contained in that cluster.

**Step 6:** Keep repeating the procedure from Step 3 to Step 5 until any of the following stopping criteria is met-

- If data points fall in the same cluster
- Reached maximum of iteration
- The newly formed cluster does not change in center points

## Example

Lets consider we have cluster points P1(1,3) , P2(2,2) , P3(5,8) , P4(8,5) , P5(3,9) , P6(10,7) , P7(3,3) , P8(9,4) , P9(3,7).

First, we take our K value as 3 and we assume that our Initial cluster centers are P7(3,3), P9(3,7), P8(9,4) as C1, C2, C3. We will find out the new centroids after 2 iterations for the above data points.

## Step 1

Find the distance between data points and Centroids. which data points have a minimum distance that points moved to the nearest cluster centroid.

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

## Iteration 1

Calcualte the distance between data points and K (C1,C2,C3)

$$C1P1 \Rightarrow (3,3)(1,3) \Rightarrow \sqrt{[(1-3)^2 + (3-3)^2]} \Rightarrow \sqrt{4} \Rightarrow 2$$

$$C2P1 \Rightarrow (3,7)(1,3) \Rightarrow \sqrt{[(1-3)^2 + (3-7)^2]} \Rightarrow \sqrt{20} \Rightarrow 4.5$$

$$C3P1 \Rightarrow (9,4)(1,3) \Rightarrow \sqrt{[(1-9)^2 + (3-4)^2]} \Rightarrow \sqrt{65} \Rightarrow 8.1$$

For P2,

$$C1P2 \Rightarrow (3,3)(2,2) \Rightarrow \sqrt{[(2-3)^2 + (2-3)^2]} \Rightarrow \sqrt{2} \Rightarrow 1.4$$

$$C2P2 \Rightarrow (3,7)(2,2) \Rightarrow \sqrt{[(2-3)^2 + (2-7)^2]} \Rightarrow \sqrt{26} \Rightarrow 5.1$$

$$C3P2 \Rightarrow (9,4)(2,2) \Rightarrow \sqrt{[(2-9)^2 + (2-4)^2]} \Rightarrow \sqrt{53} \Rightarrow 7.3$$

For P3,

$$C1P2 \Rightarrow (3,3)(5,8) \Rightarrow \sqrt{[(5-3)^2 + (8-3)^2]} \Rightarrow \sqrt{29} \Rightarrow 5.3$$

$$C2P2 \Rightarrow (3,7)(5,8) \Rightarrow \sqrt{[(5-3)^2 + (8-7)^2]} \Rightarrow \sqrt{5} \Rightarrow 2.2$$

$$C3P2 \Rightarrow (9,4)(5,8) \Rightarrow \sqrt{[(5-9)^2 + (8-4)^2]} \Rightarrow \sqrt{32} \Rightarrow 5.7$$

Similarly for other distances..

Data Points	Centroid (3,3)	Centroid (3,7)	Centroid (9,4)	Cluster
P1(1,3)	2	4.5	8.1	C1
P2(2,2)	1.4	5.1	7.3	C1
P3(5,8)	5.3	2.2	5.7	C2
P4(8,5)	5.4	5.4	5.1	C3
P5(3,9)	6	2	7.9	C2
P6(10,7)	8.1	7	3.2	C3
P7(3,3)	0	4	6.1	C1
P8(9,4)	6.1	6.7	0	C3
P9(3,7)	4	0	6.7	C2

Cluster 1 => P1(1,3) , P2(2,2) , P7(3,3)

Cluster 2 => P3(5,8) , P5(3,9) , P9(3,7)

Cluster 3 => P4(8,5) , P6(10,7) , P8(9,4)

Now, We re-compute the new clusters and the new cluster center is computed by taking the mean of all the points contained in that particular cluster.

New center of Cluster 1 =>  $(1+2+3)/3$  ,  $(3+2+3)/3$  => 2,2.7

New center of Cluster 2 =>  $(5+3+3)/3$  ,  $(8+9+7)/3$  => 3.7,8

New center of Cluster 3 =>  $(8+10+9)/3$  ,  $(5+7+4)/3$  => 9,5.3

Iteration 1 is over. Now, let us take our new center points and repeat the same steps which are to calculate the distance between data points and new center points with the Euclidean formula and find cluster groups.

## Iteration 2

Calculate the distance between data points and K (C1,C2,C3)

C1(2,2.7) , C2(3.7,8) , C3(9,5.3)

$C1P1 \Rightarrow (2,2.7)(1,3) \Rightarrow \sqrt{[(1-2)^2 + (3-2.7)^2]} \Rightarrow \sqrt{1.1} \Rightarrow 1.0$

$C2P1 \Rightarrow (3.7,8)(1,3) \Rightarrow \sqrt{[(1-3.7)^2 + (3-8)^2]} \Rightarrow \sqrt{32.29} \Rightarrow 4.5$

$C3P1 \Rightarrow (9,5.3)(1,3) \Rightarrow \sqrt{[(1-9)^2 + (3-5.3)^2]} \Rightarrow \sqrt{69.29} \Rightarrow 8.3$

Similarly for other distances..

Data Points	Centroid (2,2.7)	Centroid (3.7,8)	Centroid (9,5.3)	Cluster
P1(1,3)	1.0	4.5	8.3	C1

Open in app ↗



Search

Write



P5(3,9)	6.4	1.2	7.0	C2
P6(10,7)	9.1	6.4	1.9	C3
P7(3,3)	1.0	5.0	6.4	C1
P8(9,4)	7.1	6.6	1.3	C3
P9(3,7)	4.4	1.2	6.2	C2

Cluster 1  $\Rightarrow$  P1(1,3) , P2(2,2) , P7(3,3)

Cluster 2  $\Rightarrow$  P3(5,8) , P5(3,9) , P9(3,7)

Cluster 3  $\Rightarrow$  P4(8,5) , P6(10,7) , P8(9,4)

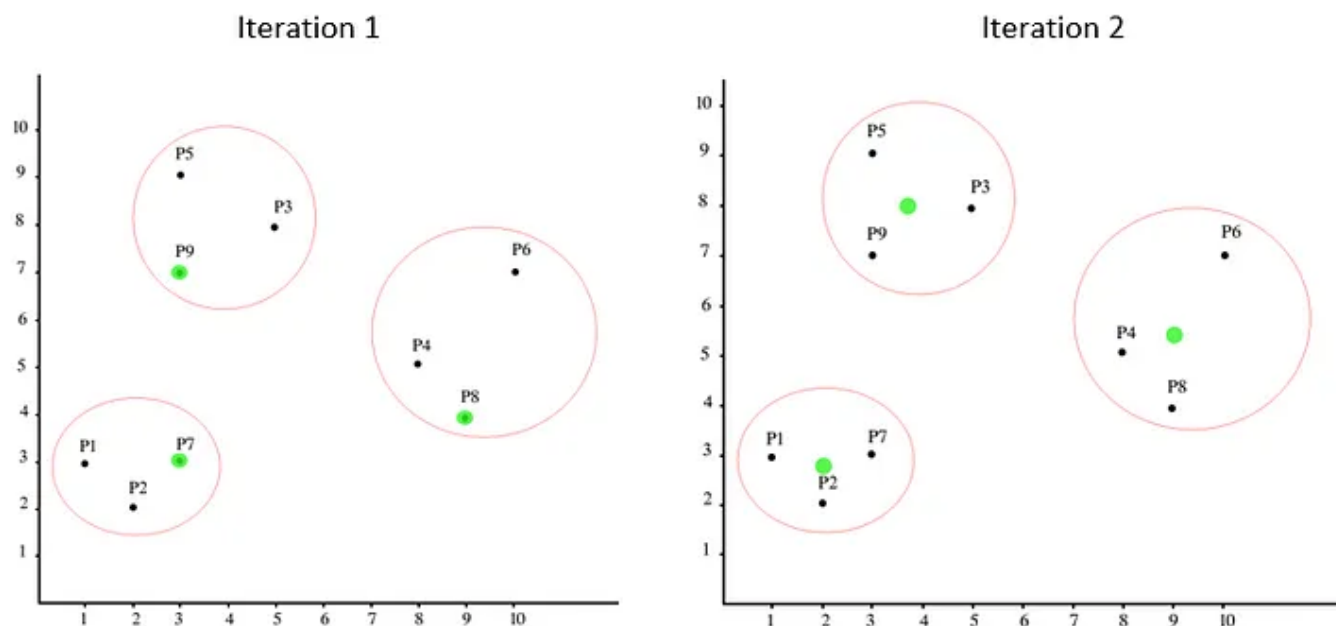
Center of Cluster 1  $\Rightarrow (1+2+3)/3$  ,  $(3+2+3)/3 \Rightarrow 2, 2.7$

Center of Cluster 2  $\Rightarrow (5+3+3)/3$  ,  $(8+9+7)/3 \Rightarrow 3.7, 8$

Center of Cluster 3  $\Rightarrow (8+10+9)/3$  ,  $(5+7+4)/3 \Rightarrow 9, 5.3$

We got the same centroid and cluster groups which indicates that this dataset has only 2 groups. K-Means clustering stops iteration because of the same cluster repeating so no need to continue iteration and display the last iteration as the best cluster groups for this dataset.

The Below graph explained the difference between iterations 1 and 2. We can see centroids (green dot) changed in the 2nd Iteration.



Green Dot — Center of the Cluster which we have found in above table C1, C2, C3

## Conclusion

I hope you are clear about the math steps behind the K-Means Clustering. In this blog, we took a small number for the dataset so we are given a k value is 3 and took 2 iterations. In real-time, the dataset feature will be maximum in that case we should use the Elbow method (WCSS) to get the perfect K(Cluster groups) value. Check out some of my previous blogs:

### Implementation of K-Means Clustering

Introduction

[medium.com](https://medium.com)

### Linear Regression - LSM

Linear Regression is used to find the relationship between a dependent variable and the independent variable. There are...

[medium.com](https://medium.com)



## Implementation of Linear Regression using Python

Introduction

medium.com

*Have doubts? Need help? Contact me!*

**LinkedIn:** <https://www.linkedin.com/in/dharmaraj-d-1b707898>

**Github:** <https://github.com/DharmarajPi>

Data Science

Machine Learning

K Means Clustering

Clustering

Artificial Intelligence



**Written by Dharmaraj**

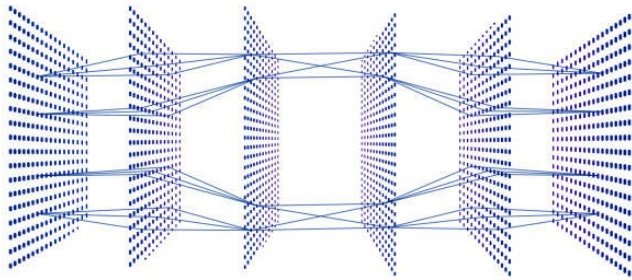
Follow



381 Followers

I have worked on projects that involved Machine Learning, Deep Learning, Computer Vision, and AWS. <https://www.linkedin.com/in/dharmaraj-d-1b707898/>

## More from Dharmaraj

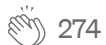


Dharmaraj

## Convolutional Neural Networks (CNN)—Architectures Explained

Introduction

5 min read · Jun 1, 2022



274



2



...



Dharmaraj

## Image classification and prediction using transfer learning

In this blog, we will implement the image classification using the VGG-16 Deep...

4 min read · Apr 3, 2022



16

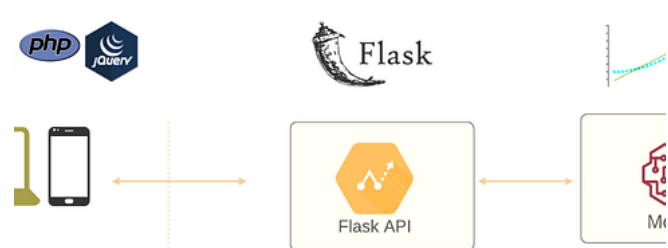


...



Dharmaraj

## Text Recognition and Extraction In Images



Dharmaraj

## Deploying Deep Learning Model using Flask API

In this post, I will show you how to extract text from an image using OpenCV and OCR. This...

4 min read · Oct 17, 2021



17



2



308




1



See all from Dharmaraj

## Recommended from Medium



 Nirmal Sankalana

## K-means Clustering: Choosing Optimal K, Process, and Evaluation...


In today's data-driven world, businesses and researchers encounter a huge amount of...

16 min read · Sep 19, 2023



2



 Megha Natarajan

## Deciphering Optimal Clusters: Elbow Method vs. Silhouette...

When stepping into the realm of unsupervised learning, k-means clustering...

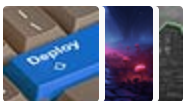
6 min read · Oct 16, 2023



3



### Lists



#### Predictive Modeling w/ Python

20 stories · 783 saves



#### Practical Guides to Machine Learning

10 stories · 907 saves



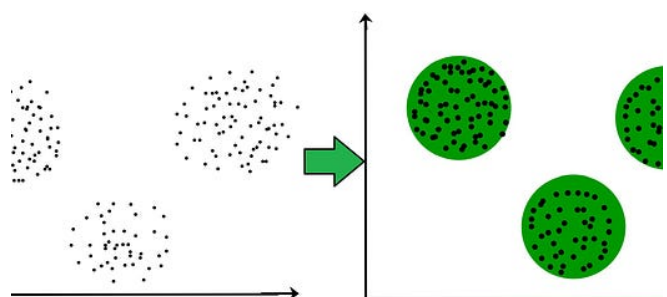
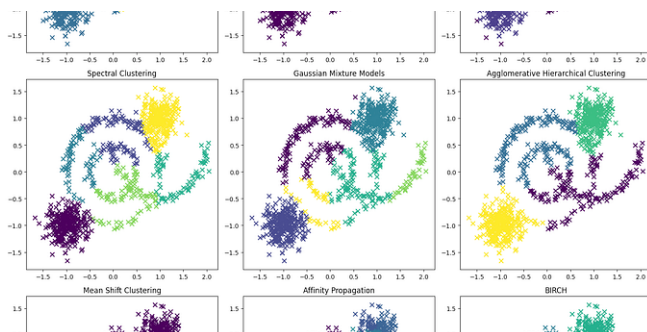
#### Natural Language Processing

1094 stories · 560 saves



#### data science and AI

39 stories · 41 saves





Sina Nazeri

## Comparing The-State-of-The-Art Clustering Algorithms

Let's generate complex data and try different clustering algorithms

9 min read · Jul 19, 2023



106



Kasun Dissanayake in Towards Dev

## Machine Learning Algorithms(14) —K-Means Clustering and...

In this article, we are learning about K-means Clustering and Hierarchical Clustering. In...

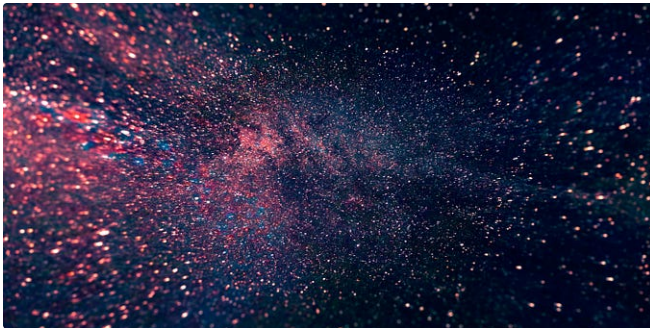
19 min read · Jan 6



462



1



Manish Shivanandhan in TuringTalks

## K-Means Clustering: The Key to Unveiling Hidden Patterns in Your...

K-means clustering is a powerful technique that helps discover hidden patterns and...

★ · 6 min read · Jan 3



1



Rukshan Pramoditha in Data Science 365

## 3 Easy Steps to Perform Dimensionality Reduction Using...

Running the PCA algorithm twice is the most effective way of performing PCA

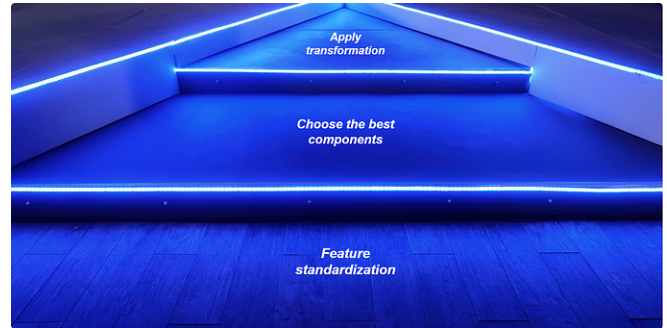
★ · 10 min read · Jan 3, 2023



152



2

[See more recommendations](#)