**K-Means Clustering: From A to Z**

Everything you need to know about K-means clustering

[Azika Amelia](https://medium.com/@azika?source=post_page-----f6242a314e9a--------------------------------)

[Sep 27, 2018·7 min read](https://towardsdatascience.com/k-means-clustering-from-a-to-z-f6242a314e9a?source=post_page-----f6242a314e9a--------------------------------)

Data is essential for data science (as if the name isn’t suggestive enough). With tons of data being generated every millisecond, it’s no surprise that most of this data is unlabeled. But that’s okay, because there are different techniques available to make do with unlabeled datasets. In fact, there’s an entire domain of Machine Learning called “Unsupervised Learning” that deals with unlabeled data.

Sometimes we just want to see how the data is organized, and that’s where clustering comes into play. Even though it’s mostly used of unlabeled data, but it works just fine for labeled data as well. The word ‘clustering’ means grouping similar things together. The most commonly used clustering method is K-Means (because of it’s simplicity).

This post explains how K-Means Clustering work (in depth), how to measure the quality of clusters, choose the optimal number of K, and mentions other clustering algorithms.

# The Concept

Imagine you’re opening a small book store. You have a stack of different books, and 3 bookshelves. Your goal is place similar books in one shelf. What you would do, is pick up 3 books, one for each shelf in order to set a theme for every shelf. These books will now dictate which of the remaining books will go in which shelf.

Every time you pick a new book up from the stack, you would compare it with those first 3 books, and place this new book on the shelf that has similar books. You would repeat this process until all the books have been placed.

Once you’re done, you might notice that changing the number of bookshelves, and picking up different initial books for those shelves (changing the theme for each shelf) would increase how well you’ve grouped the books. So, you repeat the process in hopes of a better outcome.

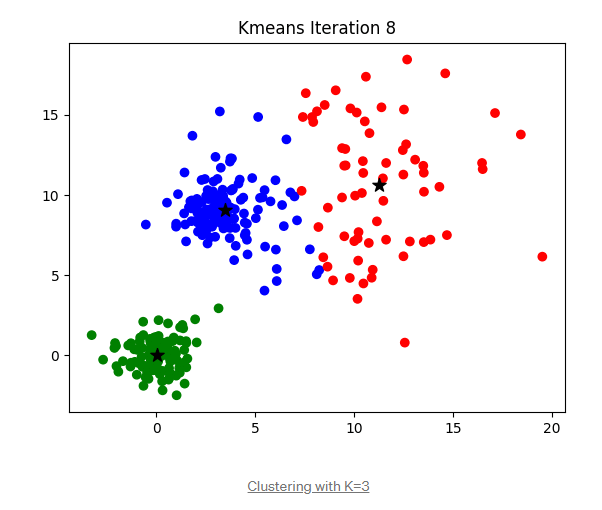
K-Means algorithm works [something just like this.](https://www.youtube.com/watch?v=FM7MFYoylVs)

# The Algorithm

K-means clustering is a good place to start exploring an unlabeled dataset. The K in K-Means denotes the number of clusters. This algorithm is bound to converge to a solution after some iterations. It has 4 basic steps:

1. Initialize Cluster Centroids (Choose those 3 books to start with)
2. Assign datapoints to Clusters (Place remaining the books one by one)
3. Update Cluster centroids (Start over with 3 different books)
4. Repeat step 2–3 until the stopping condition is met.

You don’t have to start with 3 clusters initially, but 2–3 is generally a good place to start, and update later on.



## 1. Initialize K & Centroids

As a starting point, you tell your model how many clusters it should make. First the model picks up K, (let K = 3) datapoints from the dataset. These datapoints are called cluster centroids.

Now there are different ways you to initialize the centroids, you can either choose them at random — or sort the dataset, split it into K portions and pick one datapoint from each portion as a centriod.

## 2. Assigning Clusters to datapoints

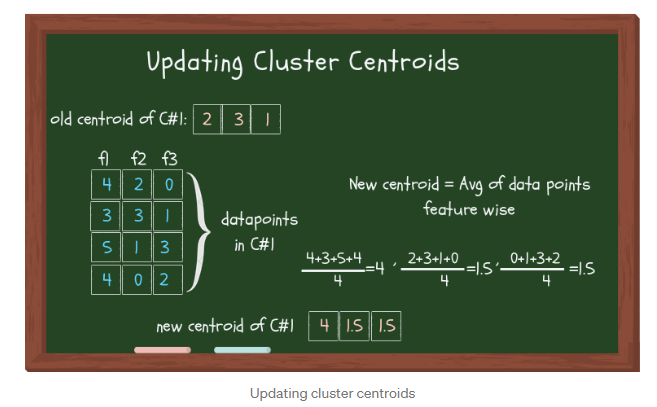
From here on wards, the model performs calculations on it’s own and assigns a cluster to each datapoint. Your model would calculate the distance between the datapoint & all the centroids, and will be assigned to the cluster with the nearest centroid. Again, there are different ways you can calculate this distance; all having their pros and cons. Usually we use the L2 distance.

The picture below shows how to calculate the L2 distance between the centeroid and a datapoint. Every time a datapoint is assigned to a cluster the following steps are followed.

## 

## 3. Updating Centroids

Because the initial centroids were chosen arbitrarily, your model the updates them with new cluster values. The new value might or might not occur in the dataset, in fact, it would be a coincidence if it does. This is because the updated cluster centorid is the average or the mean value of all the datapoints within that cluster.



Now if some other algo, like K-Mode, or K-Median was used, instead of taking the average value, mode and median would be taken respectively.

## 4. Stopping Criterion

Since step 2 and 3 would be performed iteratively, it would go on forever if we don’t set a stopping criterion. The stopping criterion tells our algo when to stop updating the clusters. It is important to note that setting a stopping criterion would not necessarily return THE BEST clusters, but to make sure it returns reasonably good clusters, and more importantly at least **return**some clusters, we **need** to have a stopping criterion.

Like everything else, there are different ways to set the stopping criterion. You can even set multiple conditions that, if met, would stop the iteration and return the results. Some of the stopping conditions are:

1. The datapoints assigned to specific cluster remain the same (takes too much time)
2. Centroids remain the same (time consuming)
3. The distance of datapoints from their centroid is minimum (the thresh you’ve set)
4. Fixed number of iterations have reached (insufficient iterations → poor results, choose max iteration wisely)

# Evaluating the cluster quality

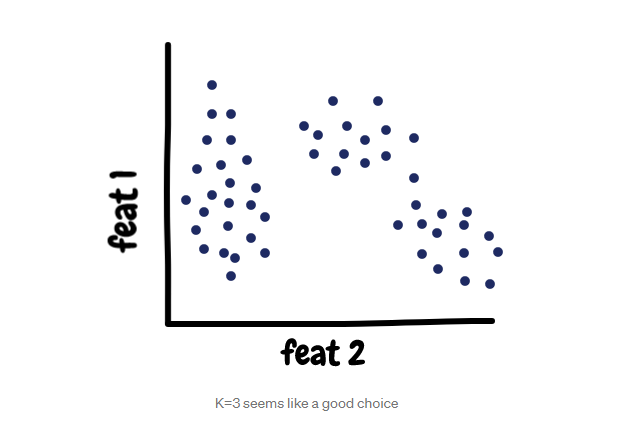
The goal here isn’t just to make clusters, but to make good, meaningful clusters. Quality clustering is when the datapoints within a cluster are close together, and afar from other clusters.

The two methods to measure the cluster quality are described below:

1. **Inertia:** Intuitively, inertia tells how far away the points within a cluster are. Therefore, a small of inertia is aimed for. The range of inertia’s value starts from zero and goes up.
2. **Silhouette score:** Silhouette score tells how far away the datapoints in one cluster are, from the datapoints in another cluster. The range of silhouette score is from -1 to 1. Score should be closer to 1 than -1.

# How many clusters?

You have to specify the number of clusters you want to make. There are a few methods available to choose the optimal number of K. The direct method is to just plot the datapoints and see if it gives you a hint. As you can see in the figure below, making 3 clusters seems like a good choice.

 Other method is to use the value of inertia. The idea behind good clustering is having a small value of inertia, and small number of clusters.

The value of inertia decreases as the number of clusters increase. So, its a trade-off here. Rule of thumb: The elbow point in the inertia graph is a good choice because after that the change in the value of inertia isn’t significant.

## 

## Naming your Clusters

When you’ve formed a cluster, you give a name it, and all the datapoints in that cluster are assigned this name as their label. Now your dataset has labels! You can perform testing using these labels. To find insights about your data, you can see what similarity do the datapoints within a cluster have, and how does it differ from other clusters.

## Assigning a cluster to a new data point

Once you’ve finalized your model, it can now assign a cluster to a new data point. The method of assigning cluster remains same, i.e., assigning it to the cluster with the closet centroid.

# ⚠️ Warning!

It’s important to preprocess your data before performing K-Means. You would have to convert your dataset into numerical values if it is not already, so that calculations can be performed. Also, applying feature reduction techniques would speed up the process, and also improve the results. These steps are important to follow because K-Means is sensitive to outliers, just like every other algo that uses average/mean values. Following these steps alleviate these issues.

I used to be very intimidated by clustering and any unsupervised algorithm in general, because I knew very little about it. I remember the first time I had to use K-Means for a music recommendation engine that I was working on, I kept thinking how would I test the final model without the labels. In this post, I tired to pack every important thing about K-Means, and mentioned the alternatives. I hope this article helps you, if you ever find yourself in the situation that I was in.