# THE STEP-BY-STEP APPROACH USING K-MEANS CLUSTERING

**Segmentation Overview:**

Before we understand what K-Means clustering is, let's see why to do segmentation and how it is helpful to any business.

Because customers are different.

**A well developed and implemented segmentation can help businesses in many ways:**

- Understanding who your customers are:
- Which are more profitable:
- Opportunities/threats to particular customer groups:
- Identifying business and customer changes over time:
- understand customer needs:

Focused resources rather than mass marketing to the 'Average customer'

That's all!!! We are not here to master segmentation. Let's now see how we can segment the group of customers statistically. There are many statistical methods available. Here, we only consider the K-Means clustering method.

We will understand this method in three steps as follow:

**Step 1: Defining the number of clusters:**

K-means clustering is a type of non-hierarchical clustering where K stands for K number of clusters. Different algorithms are available to get the optimum number of clusters. We will discuss this as we progress.

### Step 2: Define the Centroid of each cluster:

K-means clustering is an iterative procedure to define the clusters. This step basically the starting point as a center of each cluster. Say, we have

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

As a starting point, we define the number of clusters = 2 so there will be 2 center points we need to select. We pick two center points (obs = 3 and obs = 6) as below:

| | | | | |
|---|---|---|---|---|
| center 1 | 7 | 3.2 | 4.7 | 1.4 |
| center 2 | 5.8 | 2.7 | 5.1 | 1.9 |

### Step 3: Calculate the distance of each data points to the center of each cluster:

Once, we are done with the center of each cluster, the next step is to calculate the distance between each data points to the center of the clusters. As you can see, there are 6 observations and 2 centers, there will be a 6*2 matrix as follow:

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |
| **Step 2** | | | | |
| center 1 | 7 | 3.2 | 4.7 | 1.4 |
| center 2 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 1**
K = 2

**Step 3**

| obs | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | 6.7 | 6.9 |
| 2 | 6.8 | 6.6 |
| 3 | 0 | 2.6 |
| 4 | 0.9 | 2.1 |
| 5 | 3.2 | 2.6 |
| 6 | 2.6 | 0 |

Let me unroll the step 3 first obs formula for both the clusters. The same is applicable for all the observations.

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 1**
K = 2

**Step 2**

| | | | | |
|---|---|---|---|---|
| center 1 | 7 | 3.2 | 4.7 | 1.4 |
| center 2 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 3**

| obs | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | =ABS(B2-$B$13) + ABS(C2-$C$13) + ABS(D2-$D$13) + ABS(E2-$E$13) | |
| 2 | 6.8 | 6.6 |
| 3 | 0 | 2.6 |
| 4 | 0.9 | 2.1 |
| 5 | 3.2 | 2.6 |
| 6 | 2.6 | 0 |

First OBS for Cluster 1

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 1**
K = 2

**Step 2**

| | | | | |
|---|---|---|---|---|
| center 1 | 7 | 3.2 | 4.7 | 1.4 |
| center 2 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 3**

| obs | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | 6.7 | =ABS(B2-$B$14) + ABS(C2-$C$14) + ABS(D2-$D$14) + ABS(E2-$E$14) |
| 2 | 6.8 | 6.6 |
| 3 | 0 | 2.6 |
| 4 | 0.9 | 2.1 |
| 5 | 3.2 | 2.6 |
| 6 | 2.6 | 0 |

First OBS for Cluster 2

## Step 4: Allocating final clusters to all the observations:

The last step is to allocate clusters to each observation. We will do that by using the min(cluster 1 value, cluster 2 value) from step 3. So for the first obs, the minimum value is for cluster 1 (6.7 vs. 6.9) so the observation 1 belongs to cluster 1. The minimum value for second obs is for cluster 2 (6.8 vs. 6.6) so the second obs belongs to cluster 2 and so on.

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 1**
K = 2

**Step 2**

| | | | | |
|---|---|---|---|---|
| center 1 | 7 | 3.2 | 4.7 | 1.4 |
| center 2 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 3**

| obs | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | 6.7 | 6.9 |
| 2 | 6.8 | 6.6 |
| 3 | 0 | 2.6 |
| 4 | 0.9 | 2.1 |
| 5 | 3.2 | 2.6 |
| 6 | 2.6 | 0 |

**Step 4**

| Clusters |
|---|
| 1 |
| 2 |
| 1 |
| 1 |
| 2 |
| 2 |

Iteration 1

Cluster allocation after Iteration 1

That's all!!! It seems like we are done with the clustering. Unfortunately NO. We have to follow the same steps to optimize cluster allocation by using different center points. This can be done through different iterations.

Till now, we did only iteration 1 where we picked center points randomly. Now, we will pick the center points based on iteration 1 output as follow:

Observations 1, 3, and 4 belong to cluster 1, and the rest belongs to cluster 2. So, we will take the average of all the points group by cluster to set the new center points in iteration 2.

| **Step 2** | | | | |
|---|---|---|---|---|
| center 1 | 6.16666667 | 3.3 | 3.53333333 | 1.0333333 |
| center 2 | 5.66666667 | 3 | 4.16666667 | 1.5333333 |

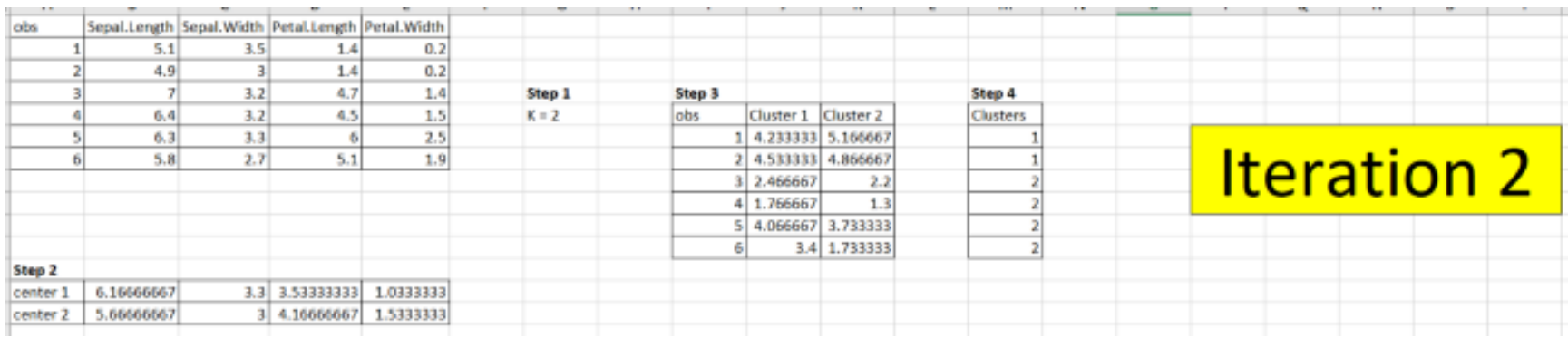
Updated Center points for both the clusters

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

| **Step 2** | | | | |
|---|---|---|---|---|
| center 1 | =AVERAGE(B2,B4,B5) | | 3.53333333 | 1.0333333 |
| center 2 | 5.66666667 | 3 | 4.16666667 | 1.5333333 |

Center points using an average of each point group by cluster

If we update the center points in iteration 2, the distance will also change, and hence, the final cluster will also change as follow:




Cluster allocation after Iteration 2

We again take the average and update the center points in iteration 3. Again the distance and final cluster will get change as follow:

| obs | Sepal.Length | Sepal.Width | Petal.Length | Petal.Width |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | 4.9 | 3 | 1.4 | 0.2 |
| 3 | 7 | 3.2 | 4.7 | 1.4 |
| 4 | 6.4 | 3.2 | 4.5 | 1.5 |
| 5 | 6.3 | 3.3 | 6 | 2.5 |
| 6 | 5.8 | 2.7 | 5.1 | 1.9 |

**Step 1**
K = 2

**Step 2**

| | | | | |
|---|---|---|---|---|
| center 1 | 5 | 3.25 | 1.4 | 0.2 |
| center 2 | 6.375 | 3.1 | 5.075 | 1.825 |

**Step 3**

| obs | Cluster 1 | Cluster 2 |
|---|---|---|
| 1 | 0.35 | 6.975 |
| 2 | 0.35 | 6.875 |
| 3 | 6.55 | 1.525 |
| 4 | 5.85 | 1.025 |
| 5 | 8.25 | 1.875 |
| 6 | 6.75 | 1.075 |

**Step 4**

| Clusters |
|---|
| 1 |
| 1 |
| 2 |
| 2 |
| 2 |
| 2 |

Iteration 3

Cluster allocation after Iteration 3

Wait!!! Did you notice anything? If you see the iteration 2 final cluster allocation and iteration 3 cluster allocation, both are same as follow:

| Clusters |
|---|
| 1 |
| 2 |
| 1 |
| 1 |
| 2 |
| 2 |

Iteration 1

| Clusters |
|---|
| 1 |
| 1 |
| 2 |
| 2 |
| 2 |
| 2 |

Iteration 2

| Clusters |
|---|
| 1 |
| 1 |
| 2 |
| 2 |
| 2 |
| 2 |

Iteration 3

As cluster allocations are not changing even if we perform iteration 3, we can now say that the algorithm is converged and we now have an optimum cluster allocation(that is cluster from iteration 3).