Cluster Analysis

Alone we can do so little, together we can do so much

Cluster Analysis

- Introduction
- Pre-requisite concepts
- K-Means Clustering
- Pros and Cons of K-means

Introduction

Cluster Analysis

"Cluster Analysis is a multivariate statistical technique that groups observations on the basis of some of their features or variables they are described by."

In common terms,
Observations in a dataset can be divided into different groups and sometimes this is very useful.

It is a very useful technique for exploring an identifying patterns in the data.

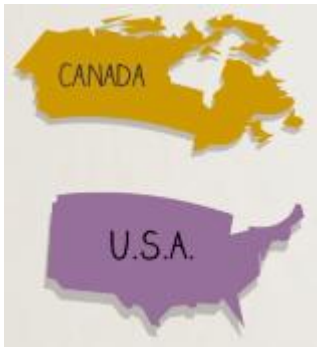
Let's see an example

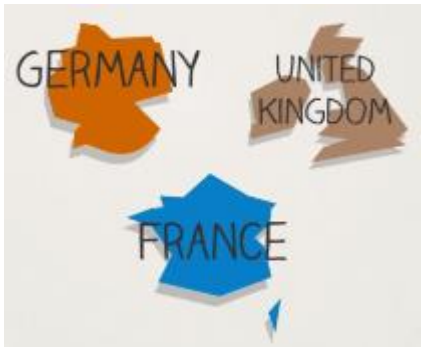
Cluster Analysis

Example:

Lets say we have 6 countries: USA, Canada, UK, Germany, France, Australia

Imagine we have now performed this technique called cluster analysis:
We got 3 groups or clusters.





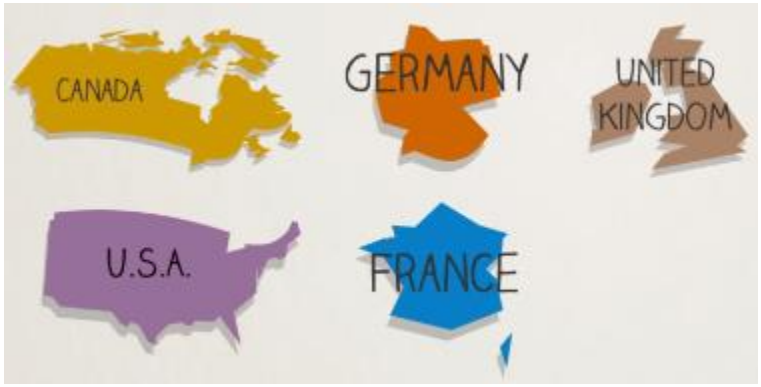
Here, in these clusters, the countries are grouped by geographic proximity.

- US & Canada – North American Countries
- Germany & UK & France – European Countries
- Australia – Oceania Country

Cluster Analysis

Example:

Lets take the same data, now what if we cluster them in 2 clusters.
We get:



Here we can say the countries are clusters according to the hemisphere

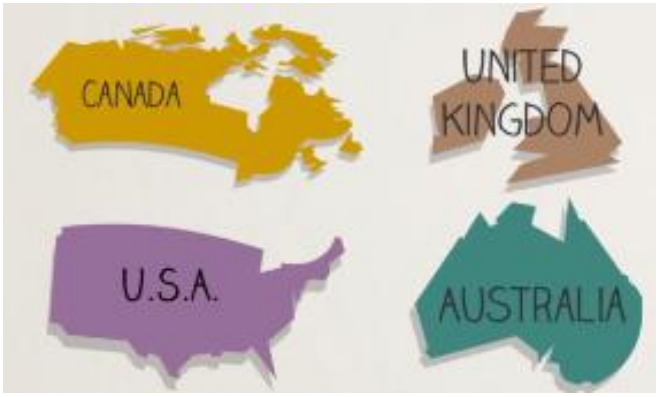
- Northern Hemisphere – Canada, USA, Germany, France, UK
- Southern Hemisphere – Australia

Clustering is easy, isn't it ?

Cluster Analysis

Wait a minute…

Lets take the same data, can also be clustered into two groups as:



Here we can say the countries are clusters according to the language:

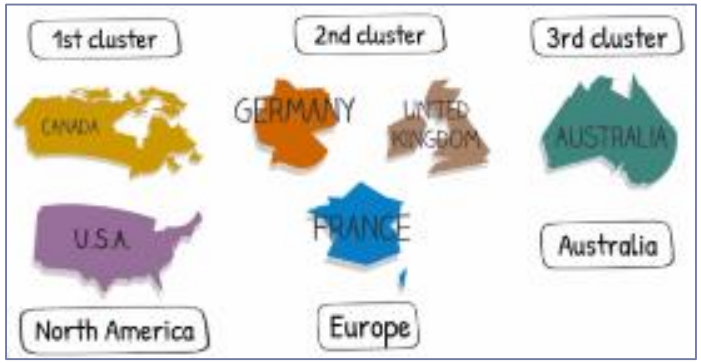
- English as Official Language – Canada, UK, USA, Australia
- English not Official Language – France, Germany

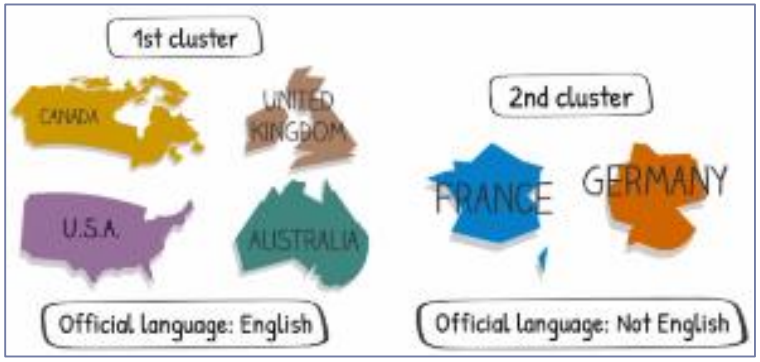
Well in Canada both French and English are official language.

Cluster Analysis

Wait a minute…

We saw we can divide the data into cluster of two in two ways or probably more.





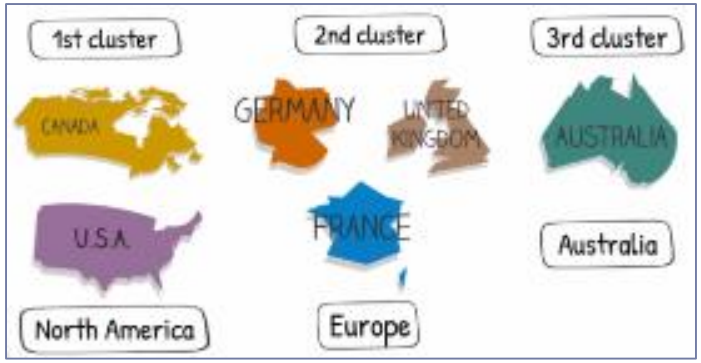
In our case, both the results are perfectly logical.

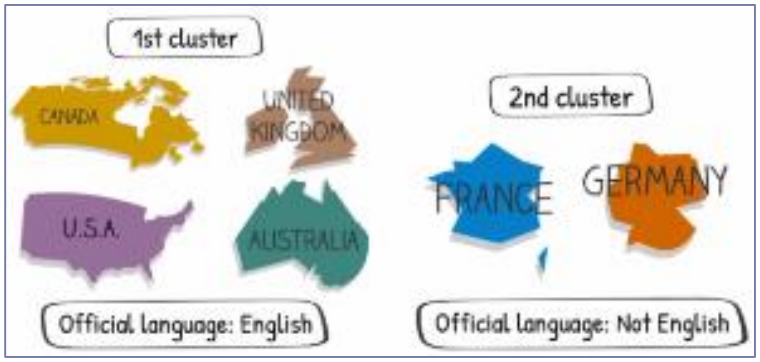
Cluster Analysis

Wait a minute…

Cluster Analysis is extremely intuitive but definitely tricky.
We can group the datasets in required number of clusters in lot many different ways.

Hence we should be very clear about our goal in cluster analysis





Cluster Analysis - Goal

Goal of Cluster Analysis

The goal of cluster analysis / clustering is:

- To maximize the similarity of observations within a cluster
- To maximize the dissimilarity between clusters

Cluster Analysis – Some Applications

Some Examples of cluster analysis are:

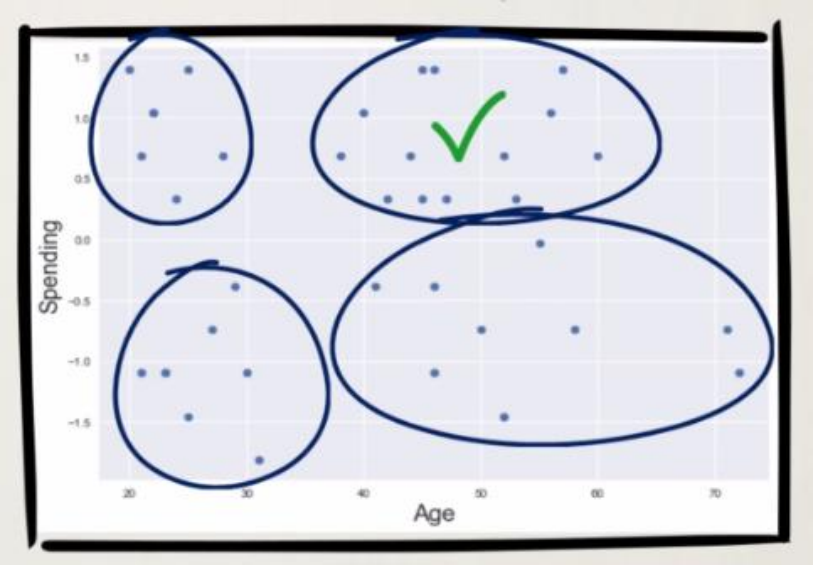
- Market Segmentation
- Image Segmentation

Cluster Analysis – Some Applications

Market Segmentation

Suppose you are appointed as a Data Scientist to a company whose marketing campaigns have been disastrous for quite a while now.
The firm gave you all the data they have gathered and ask you to create the next marketing campaign.

You have no idea who buys the product, so you decided to put some visualizations.



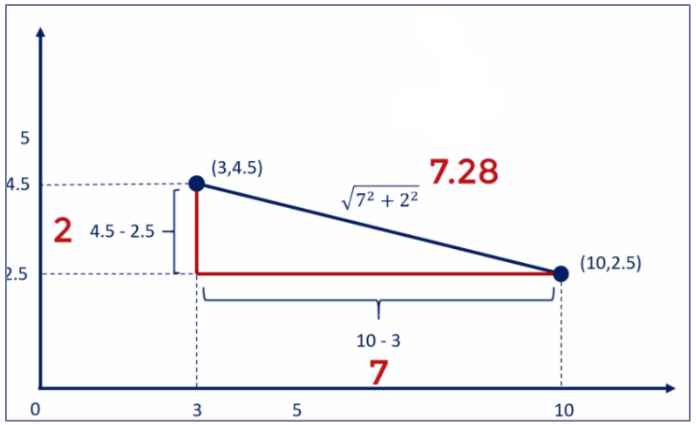
Pre-Requisite Concepts

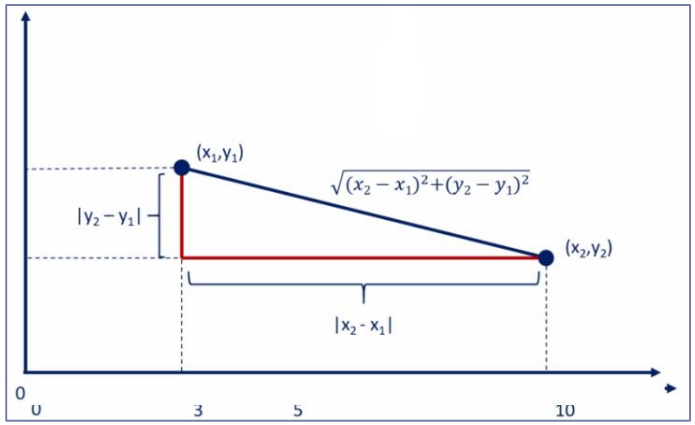
Pre-Requisite Concepts

Some Pre-requisite concepts:

1. Distance between points
2. Centroid

Lets see How to find Distance (Euclidean Distance)





Pre-Requisite Concepts

Some Pre-requisite concepts:

1. Distance between points
2. Centroid

Lets see How to find Distance (Euclidean Distance)s

2D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

3D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

If the coordinates of A are $(a_1, a_2, \ldots, a_n)$ and of B are $(b_1, b_2, \ldots b_n)$

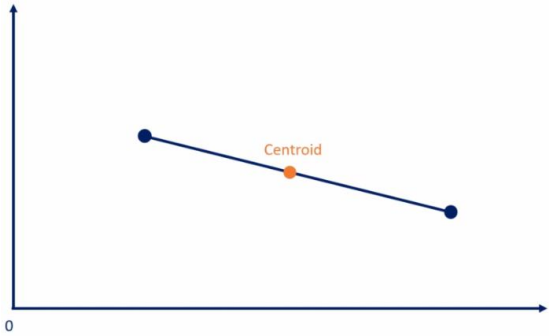
N-dim space: $d(A,B) = d(B,A) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$

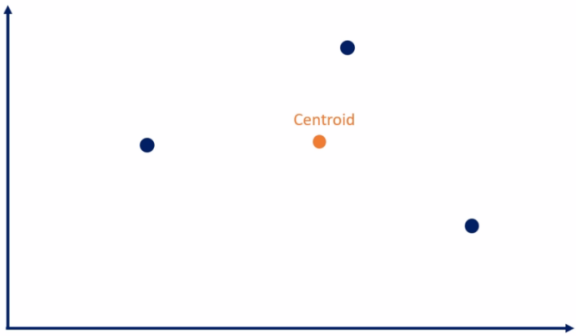
Pre-Requisite Concepts

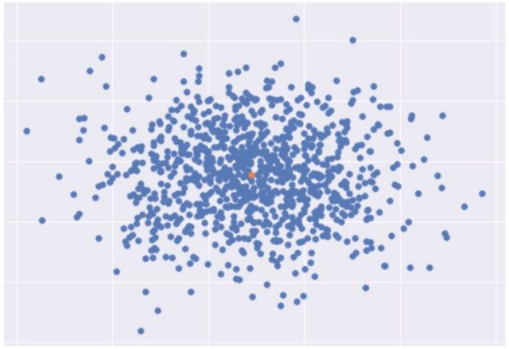
Some Pre-requisite concepts:

1. Distance between points
2. Centroid

Lets see what **Centroid** is ?
Centroid is the mean position of a group of points (aka center of mass)





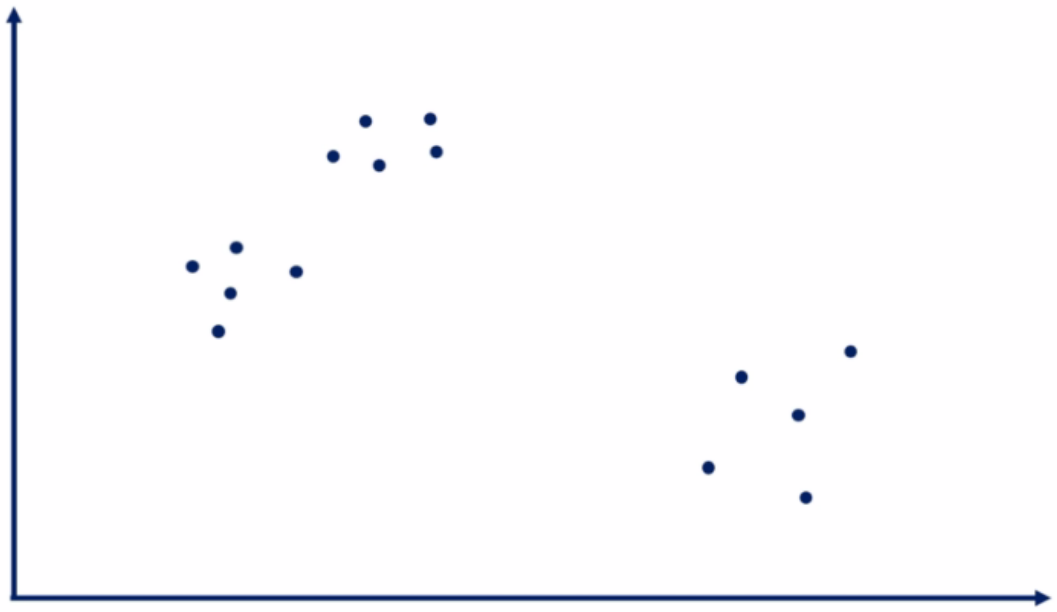


K-Means Clustering

K-Means – How it works

Lets start with an example:

Lets say we have 15 observations across two features.
This is how we plot them:



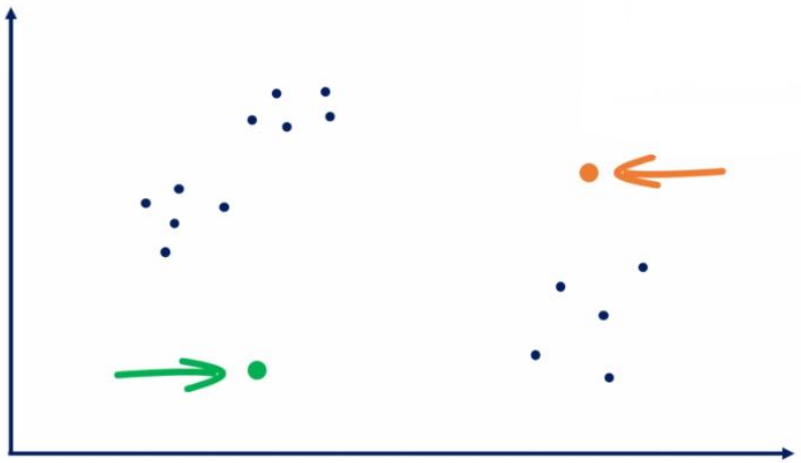
K-Means – How it works

Step 1: Choose the number of clusters

We must choose how many number of cluster we would like to have.
That's what the 'K' stands for in 'K-Means'
Say we choose k=2

Step 2: Specify the cluster seed (initial centroid/s)

A seed is basically a starting centroid.
A seed is chosen at random, or specified by someone who has a prior knowledge about the data



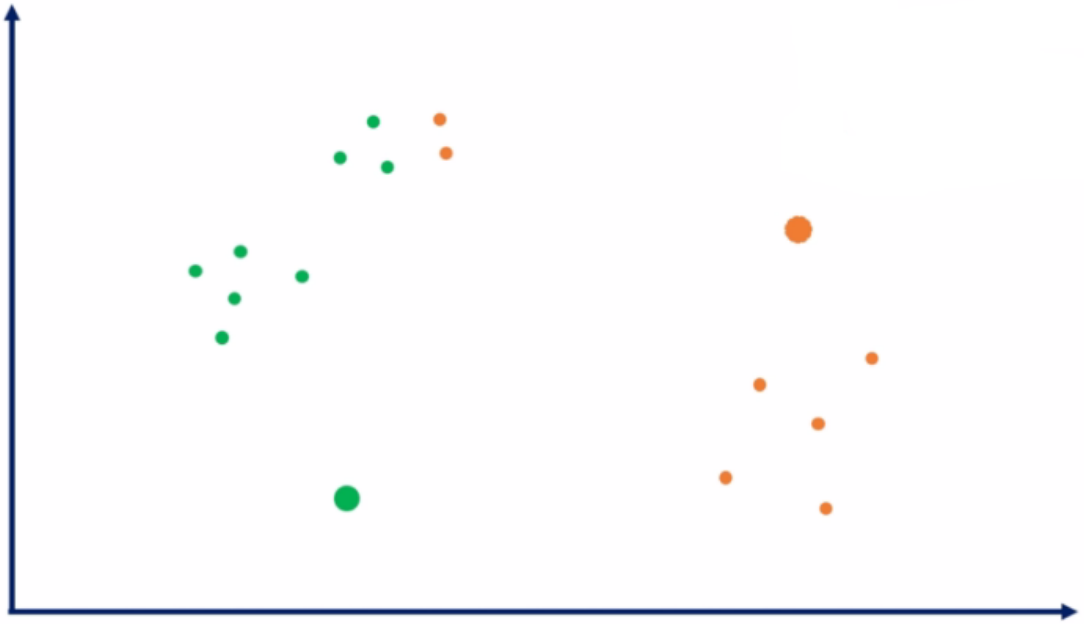
K-Means – How it works

Step 3: Assign each point to a centroid (seed)

This is done based on proximity.

The points closer to green seed will be assigned as green, while the points closer to orange seed will be assigned as orange.

Here we use Euclidean formula to find the distance.



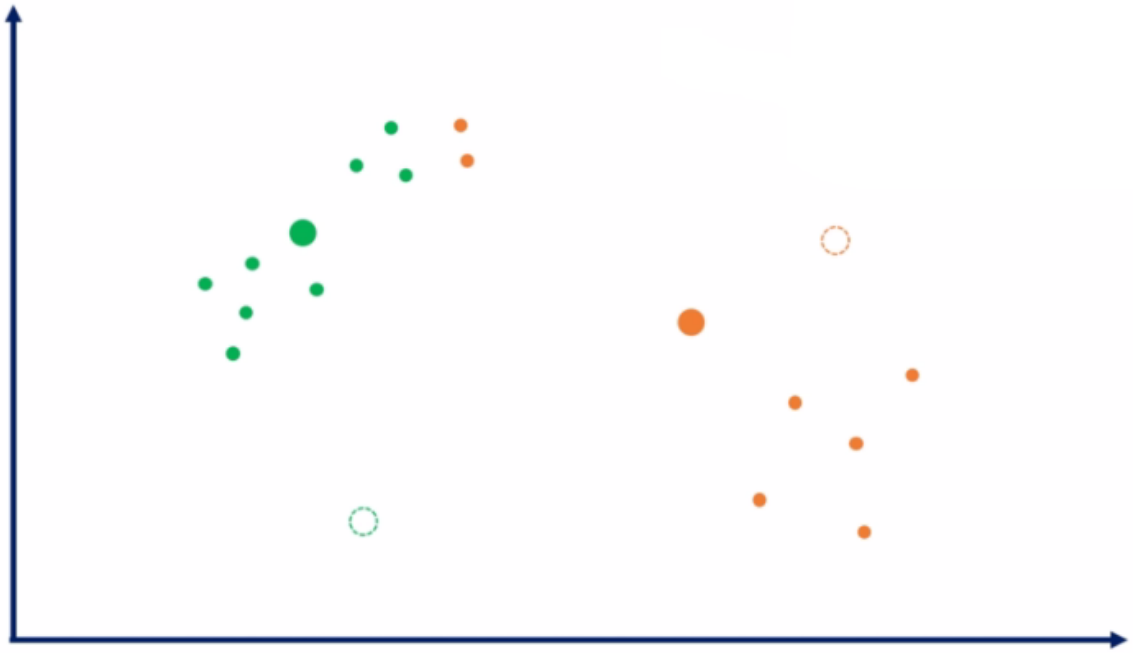
K-Means – How it works

Step 4: Adjust the centroids

The final step to calculate the centroid of the green and orange points.

Here,

Basically we re-adjust the centroid. i.e. we calculate the mean distance from all the similar (green and orange) points and move the centroid to that mean location.

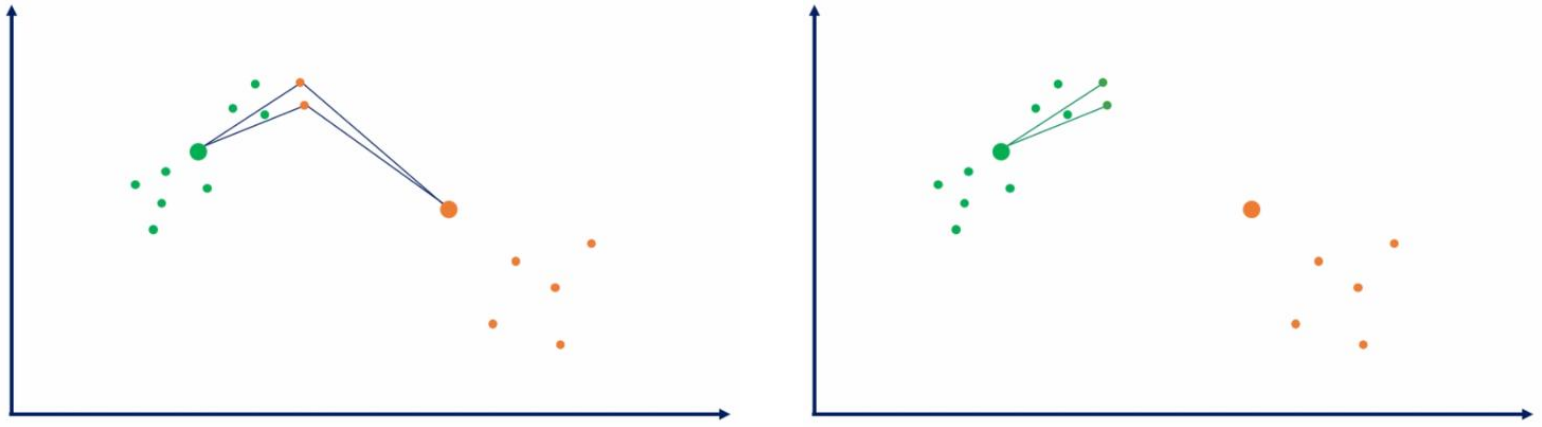


K-Means – How it works

Step 5: Repeat Step 3 and Step 4 Until no re-assignment of points are possible

Repeating Step 3: Now again, calculate the distance of each point from the new centroids and assign accordingly.

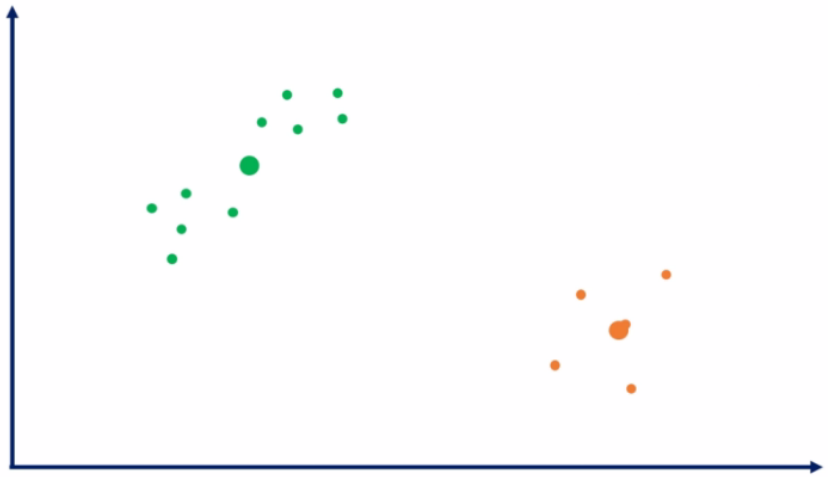
Here we recalculate the distance as shown below and assign the two orange points to green cluster as a part of this step.



K-Means – How it works

Step 5: Repeat Step 3 and Step 4

Repeating Step 4: Calculate the centroids again.
New Centroid points are



We repeat these steps until all the green points are closest to green centroid and all the orange points are closes to the orange centroid and we can no longer re-assign points, and that completes the clustering process.

K-Means – Pros & Cons

Pros:

- Simple to understand
- Fast to cluster
- Widely available (lots of packages available)
- Easy to implement
- Always yields result (also a Con, It may be deceiving as well)

K-Means – Pros & Cons

Cons:

1. We need to pick ‘K’
2. Sensitive to initialization. If seeds are initialized not properly, then the datapoints will be clustered in a non expected way.
3. Sensitive to outliers (Because of Euclidean the outliers are always classified in their own cluster. i.e. If a point is far away, K means will classify it as another cluster rather than outlier.
4. Standardization

Remedies:

Remedy for 1st Con: The elbow method

Remedy for 2nd Con: k-means ++ (sklearn implements this by default). The idea is to run an iterative algorithm in advance before clustering to find good values for initial seeds.

Remedy for 3rd Con: Remove Outliers

K-Means – Standardization

Should we standardize the variables ?

Well it depends,

It is recommended to standardize the variables.

However, If we know that our one variable is inherently important than the other, then we do not standardize.

There is no fixed rule of standardization in K-Means as we had in other algorithms.

It all comes with experience and the knowledge of the data.