K-means

K-Means is a type of unsupervised learning algorithm. As with all unsupervised algorithms, it works with unlabelled data.
Before explaining K-Means, it is useful to clarify the difference between supervised and unsupervised learning.

- Supervised learning is based on datasets containing labelled data, where each row includes an input X and its corresponding output Y.
- Unsupervised learning, on the other hand, works with unlabelled data, where each row contains only the input X without any associated output.

The real difference lies in the type of data being used. Labelled data is harder to obtain, while unlabelled data is much more abundant and easier to collect. The challenge, however, is that the way the model learns from these data changes significantly.

In our case, we will focus only on unlabelled data, which is what K-Means uses. An example of a dataset containing unlabelled data is the following:

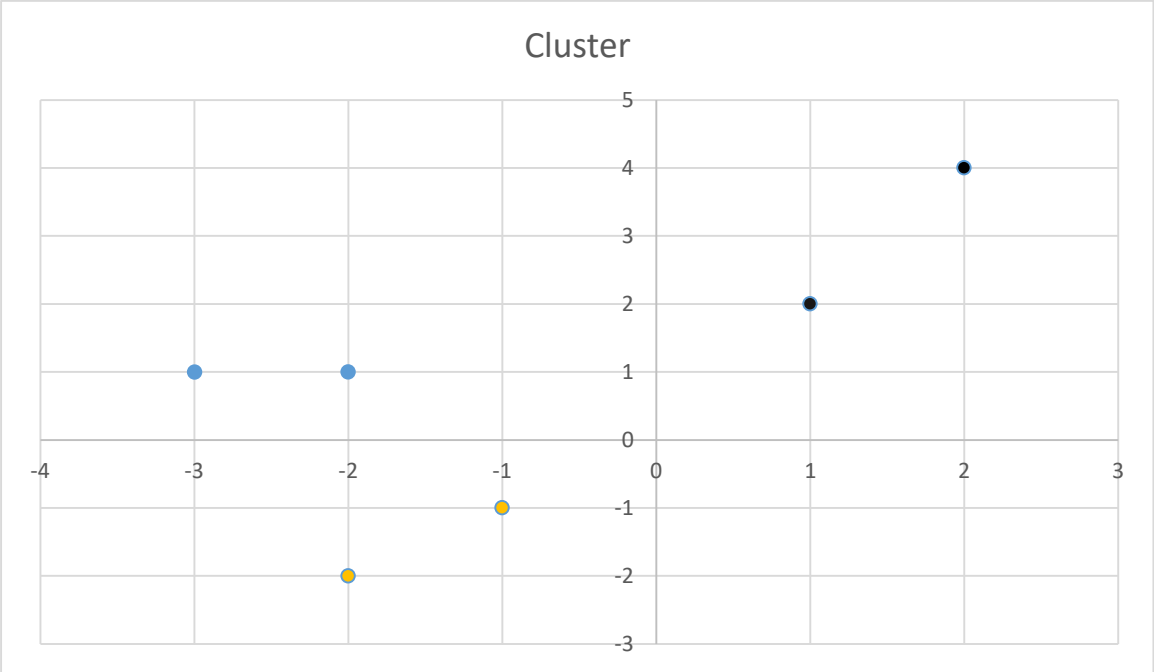
1- (1,2)
2- (2,4)
3- (-1,-1)
4- (-2,-2)
5- (-3,1)
6- (-2,1)

In this case, each X is a point with coordinates (x, y). However, we do not know any additional information about them, since there is no associated output.

The goal of unsupervised learning is to allow the model to autonomously learn the characteristics of the inputs. What the model must do is create groups of inputs that share similar characteristics. These groups of similar inputs are called clusters.

Why is this useful? Because once the groups are created (starting from the training set), the model will be able to take a new, unseen input, recognize its characteristics, and assign it to the correct cluster.

A graphical example of clusters, based on the dataset introduced earlier, is as follows:



In this case, three clusters are formed:

- The black group (1st quadrant)
- The blue group (2nd quadrant)
- The yellow group (3rd quadrant)

As you can see, each group contains inputs with similar characteristics. For instance, all points in the black cluster are in the first quadrant, those in the blue cluster are in the second, and so on.

This illustrates why unsupervised learning is so powerful: by using simple unlabelled data, it is able to learn their characteristics and create meaningful clusters.

K-Means is simply an algorithm that allows us to create these clusters starting from a dataset of unlabelled data. The procedure is relatively straightforward.

First, k indicates the number of clusters we want to create. Once this is defined, exactly k centroids are generated (with random coordinates, for example). Centroids are simply points (which can also be represented graphically). These centroids will serve as the “centers” for creating the clusters.

Once the centroids are initialized, the algorithm enters a loop that continues until the correct configuration is found. Within this loop, two main steps are repeated: cluster creation and centroid repositioning.

Cluster Creation:

To create the clusters, each point is assigned to the nearest centroid (using the Euclidean distance). Thus, for every point P in the training set, we assign the closest centroid.

Centroid Repositioning:

After the clusters are formed, centroids are repositioned. This step is crucial for finding the correct clusters and centroid configuration.

The new coordinates of each centroid are computed as the average of the coordinates of the points assigned to it:

$$\text{Xnew} = \frac{\sum x}{n} \qquad \text{Ynew} = \frac{\sum y}{n}$$

For example, if a centroid has the points (1,1) and (2,1) assigned to it:

$$x = \frac{1+2}{2} = 1.5 \qquad y = \frac{1+1}{2} = 2$$

So the new centroid position will be (1.5, 1). This procedure is repeated for every centroid.

In summary:

1. Randomly generate k centroids
2. Enter a loop
3. At each iteration: first assign points to the nearest centroid (form clusters), then reposition centroids
4. Stop when the configuration stabilizes (no points change clusters or centroids stop moving)

Error in K-Means:

To measure error in K-Means, we calculate the average Euclidean distance between all points and their assigned centroids:

$$\frac{\sum \sqrt{(xc - x)^2 + (yc - y)^2}}{N}$$

Where:

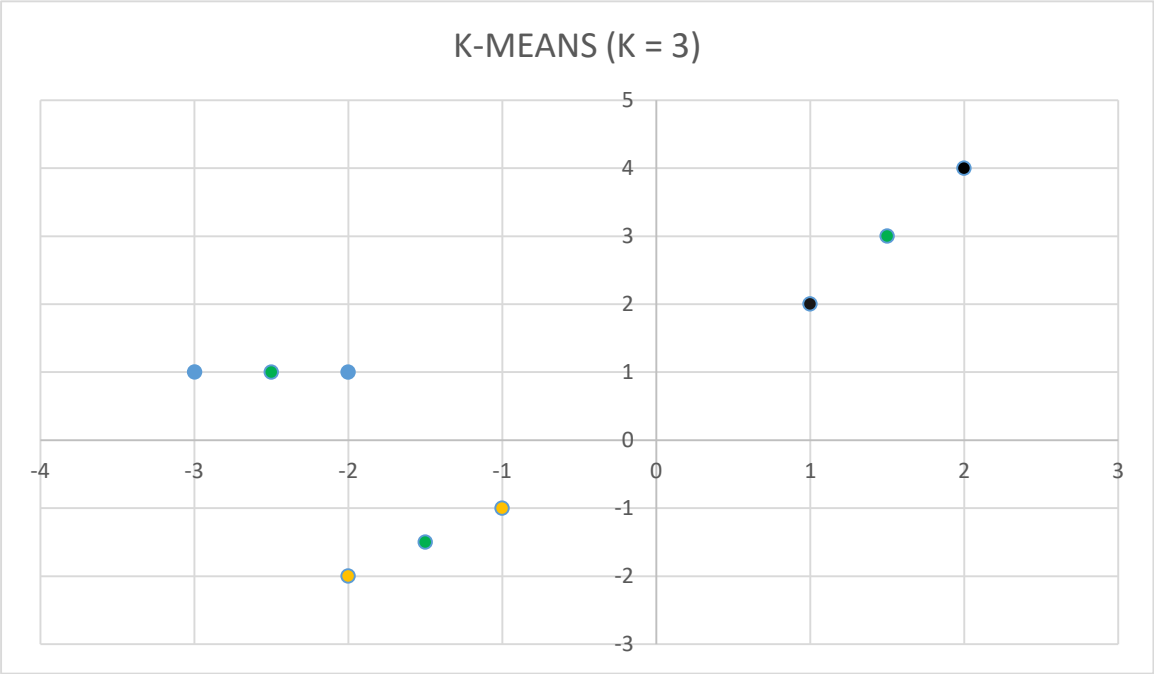
- (xc,yc)= coordinates of the centroid assigned to point P
- (x,y) = coordinates of point P
- N = total number of unlabelled data points

The smaller this value, the better the centroids are positioned.

Limitation of K-Means:

The main limitation of K-Means is that centroids are initialized randomly. This means the algorithm may not always find the best possible configuration. With different initial positions, the algorithm could converge to a solution with lower error.

Example K-means with the previous dataset:



The centroids are coloured in green.