# **Clustering**

## **Unsupervised Learning: Introduction**

Chart, scatter chart

Description automatically generated

Chart, diagram

Description automatically generated

Graphical user interface, website

Description automatically generated

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

## **K-Means Algorithm**

1. **Cluster assignment step**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated Initially assign a random cluster centroid According to the distance, group

the data-points  
It's going through each of the examples, each of these green dots shown here and depending on whether it's closer to the red cluster centroid or the blue cluster centroid, it is going to assign each of the data points to one of the two cluster centroids.

1. **Move the centroid step**

* The other part of K means, in the loop of K means, is the move centroid step, and what we are going to do is, we are going to take the two cluster centroids, that is, the red cross and the blue cross, and we are going to move them to the average of the points coloured the same colour.
* Look at all the red points and compute the average, really the mean of the location of all the red points, and we are going to move the red cluster centroid there.
* Look at all the blue points and compute the average, really the mean of the location of all the blue points, and we are going to move the blue cluster centroid there.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Initial Iteration-1

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Iteration-2 Iteration-3

After running several iterations, the cluster centroid doesn’t move a lot, it remains fixed. 🡪 So we can stop the iteration.

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email, Teams

Description automatically generated

Graphical user interface, text, application

Description automatically generated  
Text

Description automatically generated with medium confidenceText

Description automatically generated

**But what if there is a cluster centroid no points with zero points assigned to it.**

In that case the more common thing to do is to just eliminate that cluster centroid. And if you do that, you end up with K minus one clusters  
instead of k clusters.

Sometimes if you really need k clusters, then the other thing you can do if you have a cluster centroid with no points assigned to it is you can just randomly reinitialize that cluster centroid, but it's more common to just eliminate a cluster if somewhere during K means it with no points assigned to that cluster centroid.

Chart, scatter chart

Description automatically generated

Timeline

Description automatically generated with low confidence

Diagram

Description automatically generated

## **Optimization Objective**

A picture containing text

Description automatically generatedA picture containing diagram

Description automatically generated

Text, letter

Description automatically generated

This cost function is also known as Distortion Cost function /   
Distortion of the K-Means Algorithm.

Text

Description automatically generated

So, we now understand the k-means algorithm as trying to optimize this cost function J, which is also called the distortion function.

Graphical user interface, text, application, email

Description automatically generated  
Text

Description automatically generated with medium confidence

## **Random Initialization**

## **Choosing the number of clusters**