K-Means Clustering

**Unsupervised Learning**

Cluster Algorithm:

A way to group data into segments that can be classified.

Once more, we view data x as vectors in

* Applicable to Market Segmentation

One such algorithm is:

K-Means Algorithm:

At a conceptual level the algorithm goes as such.

Randomly initialize k cluster centroids (as many groups as required/desired)

Iterative step{

* Assign closest datapoints to these cluster centroids
* Calculate the mean position of the points grouped
* Reposition Centroids

Repeat until Convergence.

}

Formally:

1. Randomly initialize K cluster centroids:
2. Repeat{

for i = 1:m

for k = 1:K

}

K-means for non-separate clusters will still automatically segment the dataset.

Optimisation Objective:

Cost Function:

Random Initialisation:

Due to random initialization, it is very much possible for cluster centroids to get stuck in local optima, resulting in less ideal clustering as seen below.

A picture containing sky, text

Description automatically generated

Hence it is often a good idea to iterate through a number of random initializations and select the cluster that computes the lowest value on the cost function.

Optimizing the number of clusters:

Choosing the number of clusters, there are certain scenarios where the the number of clusters is small and discrete, such that iterating the number of clusters would show a sudden drop in the cost function ‘a elbow’

A close up of a map

Description automatically generated

But often it is ambiguous.

Hence a better way:

K means clustering is usually used for some form of downstream purpose.

Evaluate the clustering based on a metric for that later purpose.