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 $x \in \mathbb{R}^{n+1} space$

- 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:$\mu_1, \mu_2 \dots, \mu_K$
2. Repeat{

for i = 1:m

$$c^{(i)} = min_k \left\| x^{(i)} - \mu_k \right\|$$

$$assign\ the\ closest\ cluster\ centroid\ to\ the\ datapoint$$

for k = 1:K

$$\mu_k := mean\left(c^{(i)}, \dots c^{(j)}\right)$$

$$the\ average\ of\ points\ assigned\ to\ the\ cluster\ k$$

}

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

Optimisation Objective:

$$c^{(i)} = index\ of\ cluster(1,2, \dots, K) to\ which\ example\ x^{(i)}\ is\ assigned$$

$$\mu_k = cluster\ centroid\ k, \mu_k \in \mathbb{R}^n$$

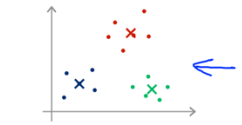
$$\mu_{c^{(i)}} = cluster\ centroid\ to\ which\ eg: x^{(i)}\ is\ assigned$$

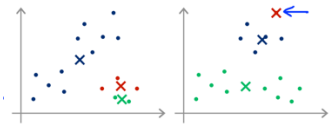
Cost Function:

$$J\left(c^{(1)}, \dots c^{(m)}, \mu_1, \dots, \mu_K\right) = \frac{1}{m} \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

Random Initialisation:

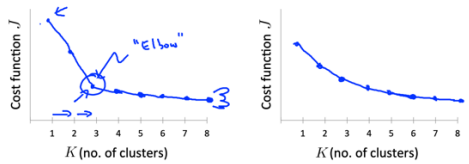
Due to random initialization, it is very much possible for cluster centroids to get stuck in local optima, resulting in less ideal clustering as seen below.





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’

Elbow method:



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.