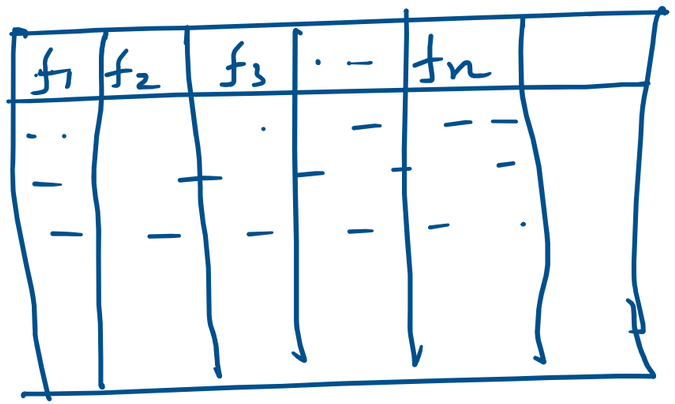
Finding Optimal Number of clusters in kMeans

Even before we start the algorithm of k-mean we need to furnish 'k'. (number of clusters)

Now, how many clusters are there is not known in advance.

So k value is not known in advance

clusters



How can we select optimum number of cluster?

Elbow method

Elbow method:-

1. First choose a range of values of k. (say, $k = 2, 3, 4, \ldots, 10$)
2. For each value of k obtain the clustering.
3. Calculate sum of squared distance (SSD) from the cluster.

$$d_1^{(i)} = dist\left(x_1^{(i)}, c_1\right)$$

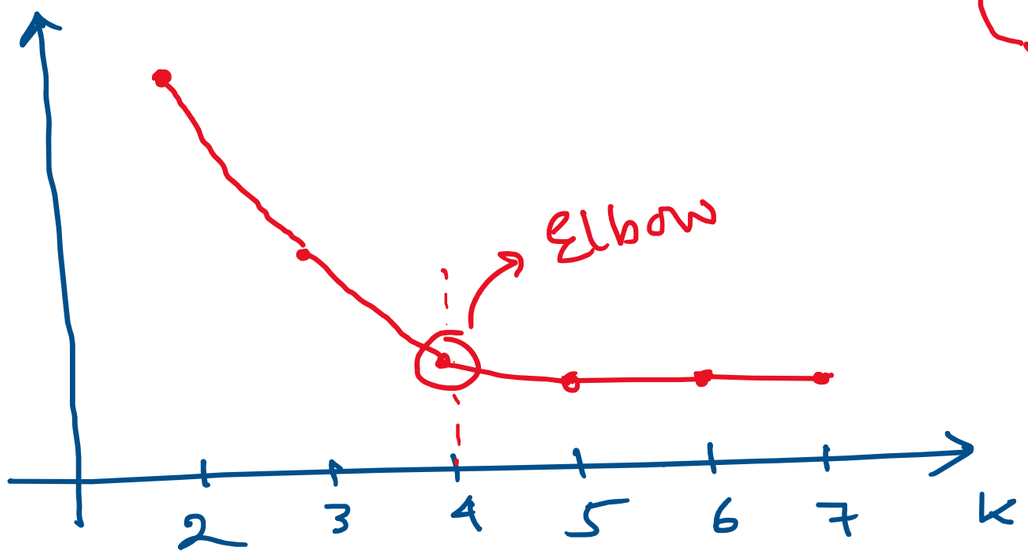
$$SSD_1 = \sum_i \left(d_1^{(i)}\right)^2 = \sum_{x_1 \in C_1} \left[dist\left(x_1, c_1\right)\right]^2$$

$$d_2^{(j)} = dist\left(x_2^{(j)}, c_2\right)$$

$$SSD_2 = \sum_j \left(d_2^{(j)}\right)^2$$

$$SSD_1 + SSD_2 = SSD$$

k	SSD
2	
3	
4	
5	
⋮	
10	



4. Plot SSD vs k

5. We will choose the value of k where we get the elbow.

How to measure goodness of clustering :-

Silhouette Coefficient

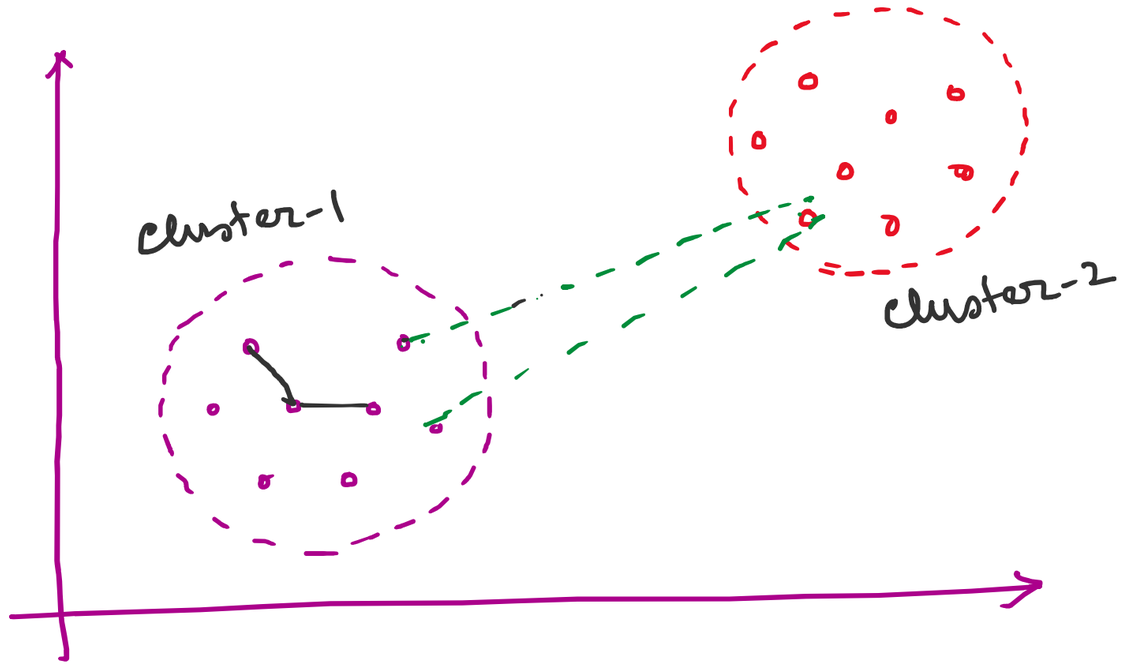
It is a metric used to calculate the goodness of a clustering technique.

Its value ranges from -1 to $+1$

→ 1 : Clusters are well apart from each other & are clearly distinguished.

→ 0 : clusters are overlapping.

→ -1 : clusters are assigned in the wrong way.



$$\text{Silhoutte Score} = \frac{b-a}{\max(a,b)}$$

a = average intra-cluster distance, i.e. average of the distances between each point within a cluster.

b = average inter-cluster distance, i.e. average of the distances between the points of clusters.