Clustering is the grouping of similar things. There are various clustering methods, for example-

<br>

![image](<https://raw.githubusercontent.com/niteshjindal170988/unsupervised-learning/main/clustering/.scrap/clus0.JPG>)

<br>

Let us say that we have n different data points and K different clusters, such as . For every combination of data point and a cluster, we can compute a distance between them, for example the distance between and or distance between and and so on.

Minimizing the distance between nth data point and kth cluster can be written as:

<br>

However, every data point is not associated with every cluster. Thus, we need to associate the nth data point with the cluster it belongs to, not all the clusters. We will define a latent variable that tells us whether the nth datapoint is associated with cluster K or not. either 0 or 1. If it is 1, then the distance between the nth datapoint and kth cluster matters else not. <br>

Therefore, we can write it as:

We can say that the nth datapoint is associated with cluster K if the distance between the given datapoint is having minimum distance from the Kth cluster vector.

Minimum number of clusters can be one (i.e. all datapoints in a single cluster) and maximum number of clusters can be equal to the number of datapoints in the data.

<br>

Objective Function of K-means Clustering :

<br>

where hyperparameter = k which determines the number of clusters<br>parameter = which determines the mean vector of the cluster<br> Latent Parameter = which determines whether the nth datapoint is associated with cluster K or not. either 0 or 1.<br>Total number of parameters learned = k $\*$ D because each of the kth cluster is a D-dimensional vector.

Hyperparameter tell us how complex the modelling system is and it is provided by the user, in the given case, it is the number of clusters

Model Complexity:

Clustering can be used in following ways:

* Understand the structure in data
* Generate class labels when unknown
* Summarize data points by their cluster center, or mean cluster vector.