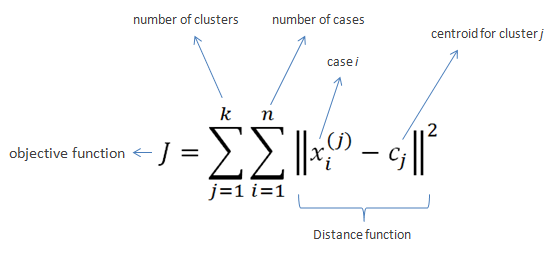
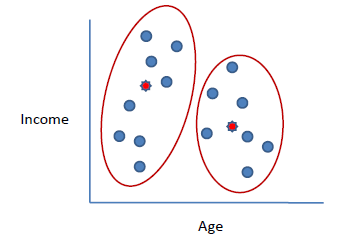
**K-Means Clustering**

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:



Algorithm

* Clusters the data into k groups where k is predefined.
* Select k points at random as cluster centers.
* Assign objects to their closest cluster center according to the Euclidean distance function.
* Calculate the centroid or mean of all objects in each cluster.
* Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.



K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the results are sensitive to initialization and often terminates at a local optimum. Unfortunately, there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of overfitting.