k-Means

k-Means clustering is a partitioning method. The function partitions the data set of N items into k disjoint subsets S_j that contain N_j items so that they are as close to each other as possible according a given distance measure. Each cluster in the partition is defined by its N_j members and by its centroid λ_j. The centroid for each cluster is the point to which the sum of distances from all items in that cluster is minimized. Thus, we can define the k-means algorithm as an iterative process to minimize $E = \sum_1^k \sum_{n \in S_j} d\left(x_n, \lambda_j\right)$, where x_n is a vector representing the n^{th} item, λ_j is the centroid of the item in S_j and d is the distance measure. The k-means algorithm moves items between clusters until E cannot be decreased further.

Algorithm

1. The algorithm works by randomly selecting k centroids.
2. All items are assigned to the cluster whose centroid is the closest to them.
3. The new cluster centroid needs to be updated to account for the items who have been added or removed from the cluster and the membership of the items to the cluster updated.
4. This operation continues until there are no further items that change their cluster membership.

Shortcomings

- it assumes prior knowledge of the data in order to choose the appropriate k
- the final clusters are very sensitive to the selection of the initial centroids;
- it can produce empty cluster
- it has problems when clusters are of differing sizes, densities, and non-globular shapes
- it also has problems when the data contains outliers