K-means clustering

Definition and Purpose

K-means is a method to **partition n observations into k clusters** in which each observation belongs to the cluster with the nearest **mean (cluster centers or cluster centroid), serving as a prototype of the cluster**. It minimizes with-cluster variances (squared Euclidean distance), but not regular Euclidean distance (can be solved using k-medians or k-medoids).

In the mathematic expression, given a set of observations $(x_1, x_2, \dots, x_n)$, where each observation is d-dimensional real vector, it aims to partition the n observations into k cluster $S = \{S_1, S_2, \dots, S_k\}$

$$\arg_S \min \sum_{i=1}^{k} \sum_{X \in S_i} \left|\left|X - \mu_i\right|\right|^2 = \arg_S \min \sum_{i=1}^{k} |S_i| \, Var S_i$$

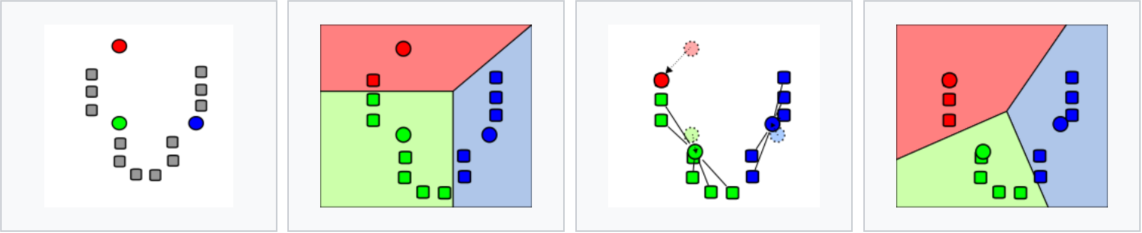
μ_i is the mean of points in S_i,

Algorithms

Given an initial set of k-means $m_1^{(1)}, \dots, m_k^{(1)}$, the algorithm proceeds by alternating between two steps:

1. **Assignment step:** Assign each observation to the cluster with the nearest mean: with the squared Euclidean distance (Mathematically, this means **partitioning the observations according to the Voronoi diagram** centered by the means).
2. Update Step: Recalculate means (centroids) for observations assigned to each cluster.

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

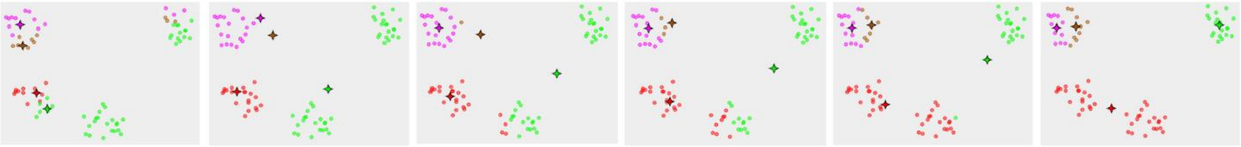
4. Steps 2 and 3 are repeated until convergence has been reached.

The algorithm has converged when the assignments no longer change. **The algorithm is not guaranteed to find the optimum**. Various modifications of k-means such as spherical k-means and k-medoids have been proposed to allow using other distance measures.

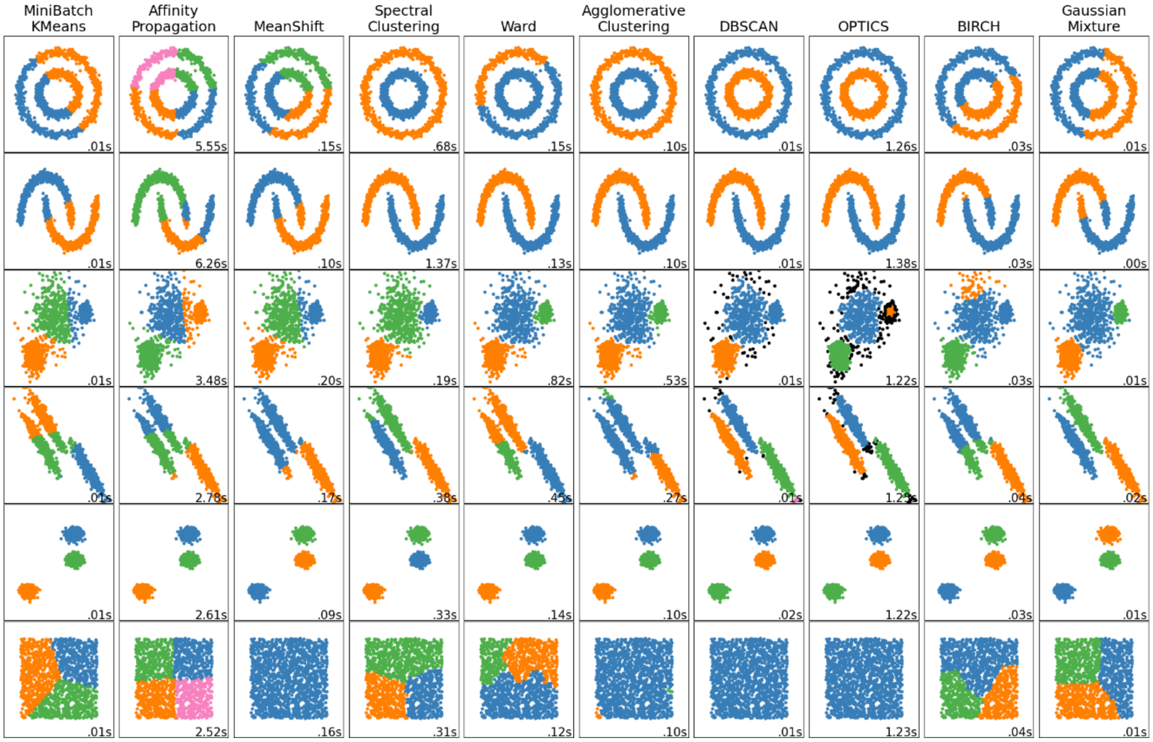
Properties

Three key features of *k*-means that make it efficient are **often regarded as its biggest drawbacks**:

1. The number of clusters k is an input parameter: **an inappropriate choice of k may yield poor results**. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set.
2. Convergence to a **local minimum may produce counterintuitive ("wrong") results** (**wrong initial position**)



3. Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
OPTICS
BIRCH
Gaussian Mixture
.01s
5.55s
.15s
.68s
.15s
.10s
.01s
1.26s
.03s
.01s
.01s
6.26s
.10s
1.37s
.13s
.10s
.01s
1.38s
.03s
.00s
.01s
3.48s
.20s
.19s
.82s
.53s
.01s
1.22s
.03s
.01s
.01s
2.78s
.17s
.36s
.45s
.27s
.01s
1.25s
.04s
.02s
.01s
2.61s
.09s
.33s
.14s
.10s
.02s
1.22s
.03s
.01s
.01s
2.52s
.16s
.31s
.12s
.10s
.01s
1.23s
.04s
.01s