K-means clustering

Definition and Purpose

K-means is a method to **partition observations into 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 (), where each observation is d-dimensional real vector, it aims to partition the observations into cluster

is the mean of points in ,

## Algorithms

Given an initial set of k-means , 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.

Chart, scatter chart

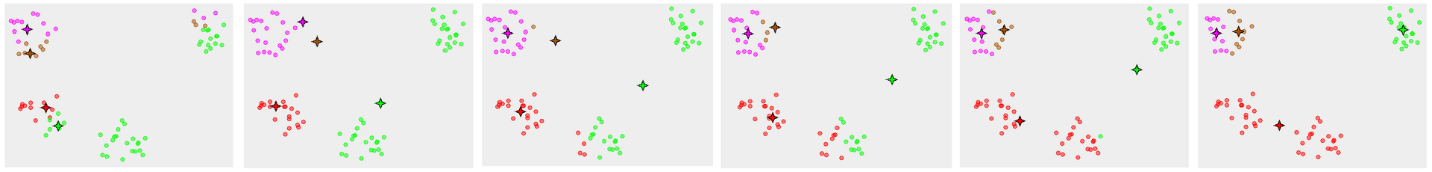
Description automatically generated

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**)



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

A picture containing diagram

Description automatically generated