14. Clustering:
k-Means Clustering Algorithm

k-Means Clustering: Overview

Facts

It is a simple partition based clustering algorithm

It divides a number of data points into a number of clusters, with each cluster represented by a centroid

It has been proposed by MacQueen in 1967

Algorithm Overview

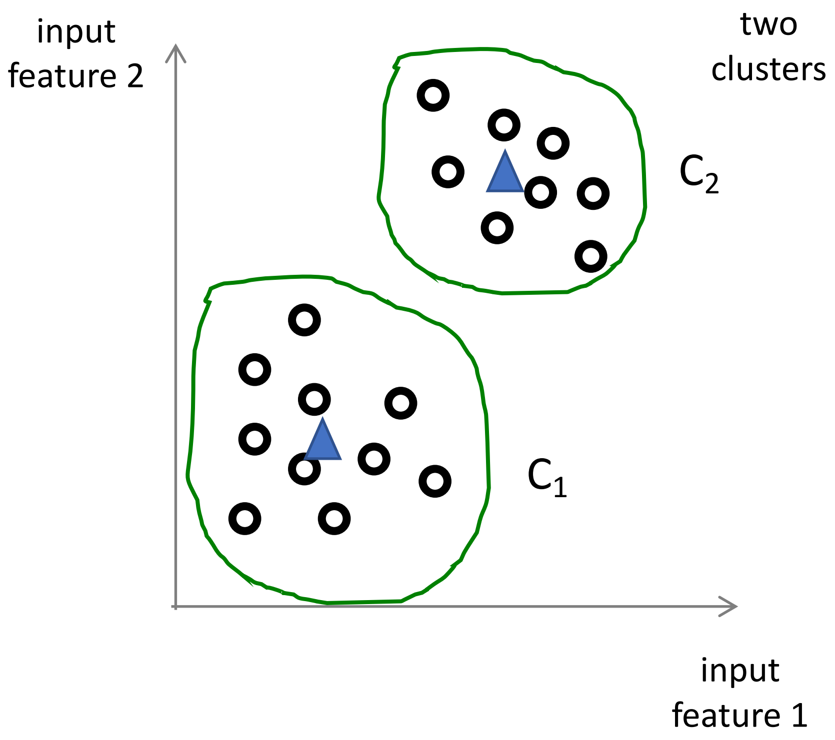
Input:

- The collection of data points
- The number of clusters

Output:

- The attribution of the data points to different clusters

Example



k-Means Clustering: How Does it Work?

Data

A collection of data points $\boldsymbol{x}_{\boldsymbol{i}} = \left(x_{i,k}, x_{i,2}, \ldots, x_{i,m}\right)$, $\mathrm{i} = 1,2, \ldots, n$, that one wishes to organize onto K different clusters, where each data point is characterized by different features.

Distance Function

Given two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we can define their similarity via their squared Euclidean distance:

$$d\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sum_{l=1}^{m} \left(x_{i,l} - x_{j,l}\right)^2$$

Cost Function

Given the K clusters $C_1, C_2, \ldots, C_K$, we can then define the cost function

$$J(C_1, C_2, \ldots, C_K) = \frac{1}{K} \cdot \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{c}_k) \quad \text{with} \quad \boldsymbol{c}_k = \frac{1}{|C_k|} \cdot \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$$

Optimization Problem

The clustering problem can then be posed as follows:

$$C_1^*, C_2^*, \ldots, C_K^* = \underset{C_1, C_2, \ldots, C_K}{\operatorname{argmin}} J(C_1, C_2, \ldots, C_K)$$

Computationally very complex problem

$\Downarrow$

heuristic algorithm

k-Means Clustering: Algorithm

k-Means Clustering Algorithm

This is a very simple iterative algorithm involving various steps:

- The very first step involves choosing some initial cluster representatives (e.g. by choosing K random data samples from the dataset)

- The next steps involve iterating between two of operations

  - Attribution of data samples to clusters

$$\boldsymbol{x} \in C_k \Leftarrow d(\boldsymbol{x}, \boldsymbol{c}_k) < d(\boldsymbol{x}, \boldsymbol{c}_l), \forall l \neq k$$

Minimize $J(C_1, C_2, \dots, C_K)$ keeping constant the centroids

  - Re-computation of cluster centroids

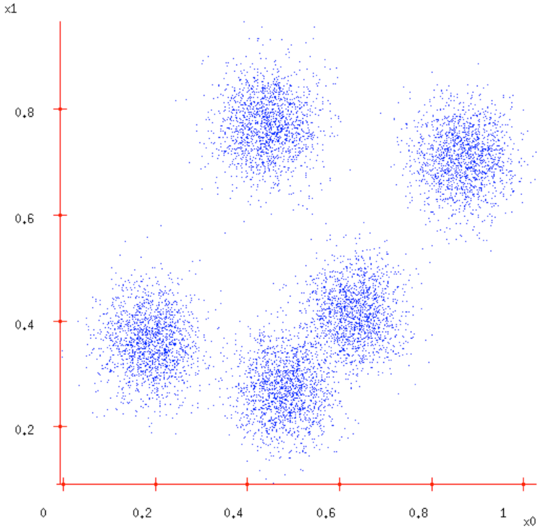
$$\boldsymbol{c}_k = \frac{1}{|C_k|} \cdot \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$$

Chose centroids that minimize $J(C_1, C_2, \dots, C_K)$

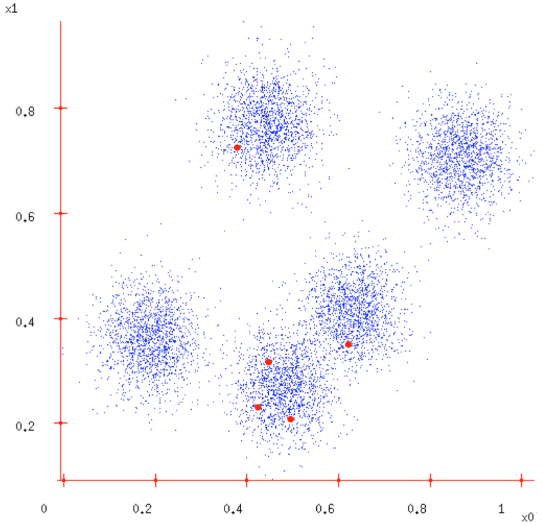
- until convergence (e.g. until nothing changes)

k-Means Clustering: Example

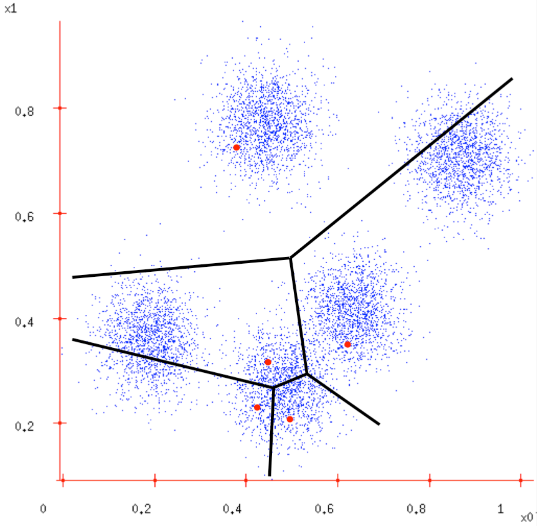
Example



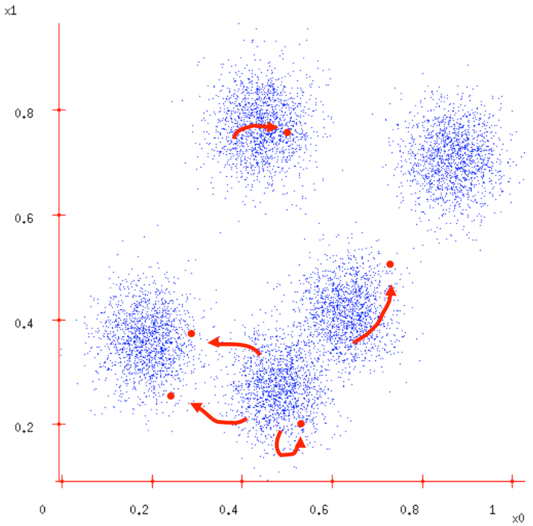
Initialization of Centroids



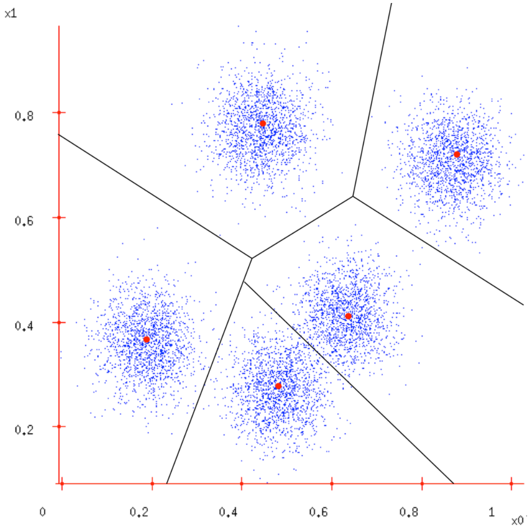
Attribution of Data to Clusters



Re-Computation of Centroids



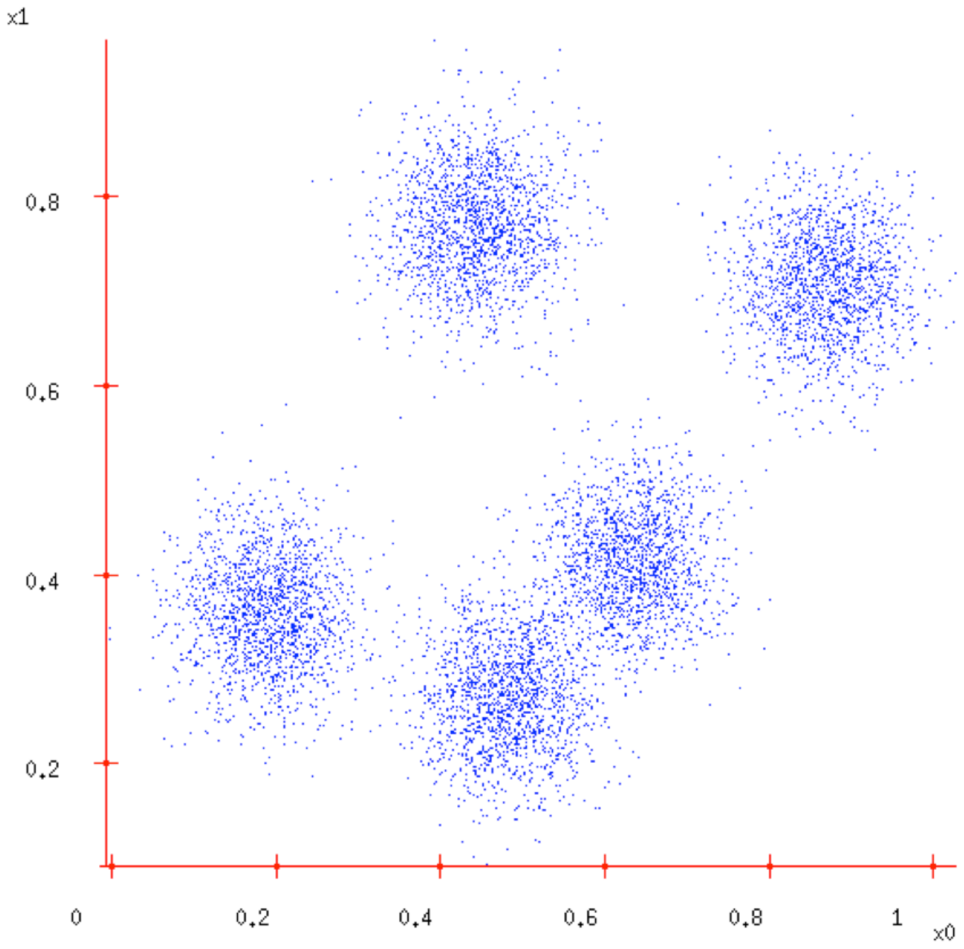
Re-Attribution of Data to Clusters



This general process is iterated until nothing changes

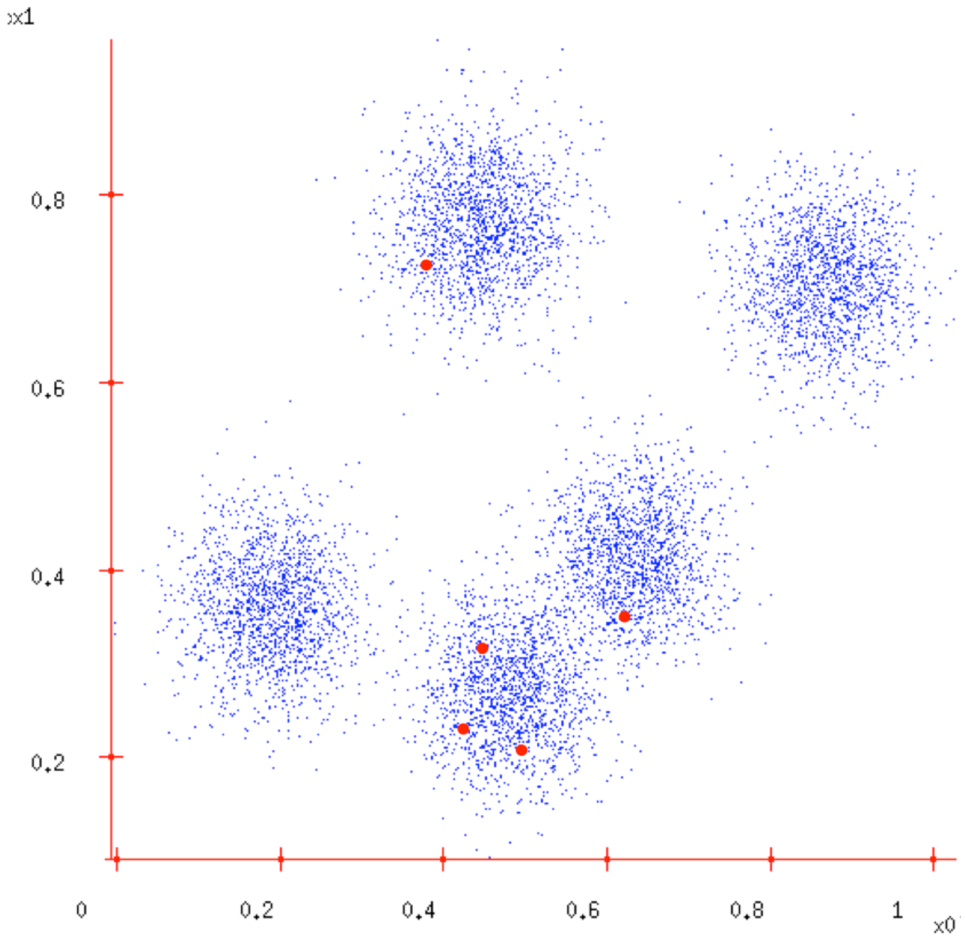
k-Means Clustering: Example

Example



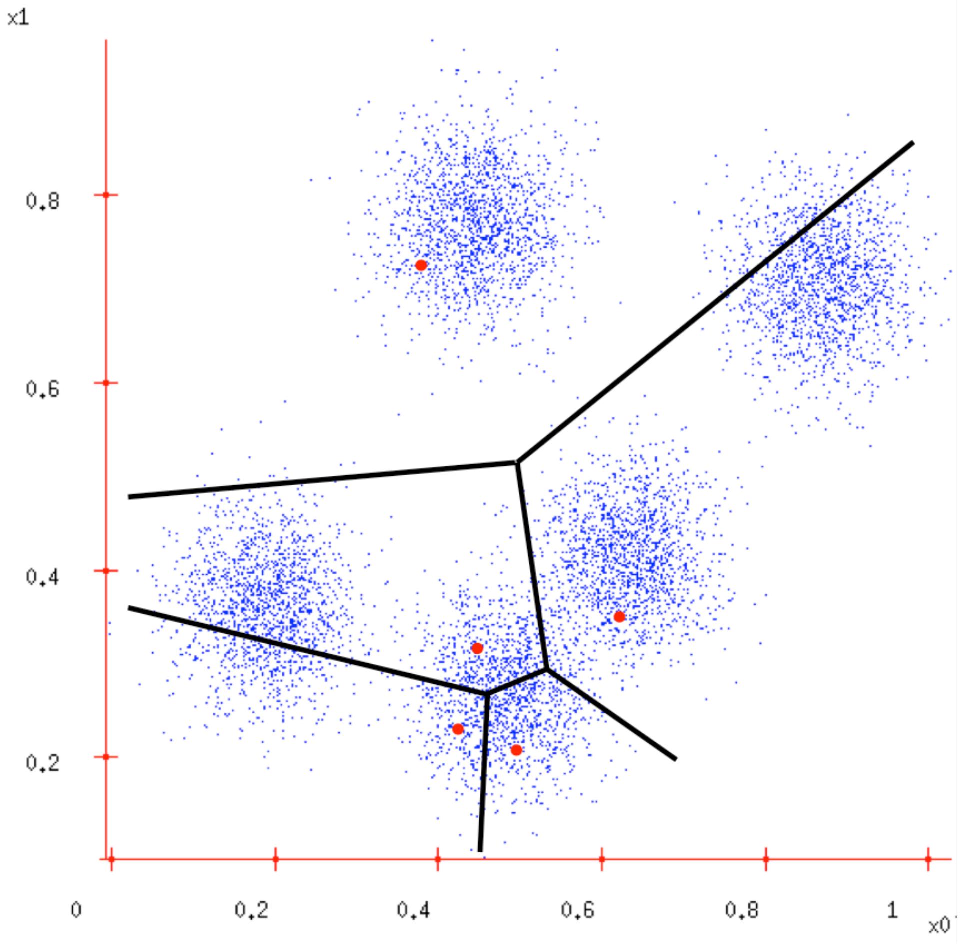
k-Means Clustering: Example

Initialization of Centroids



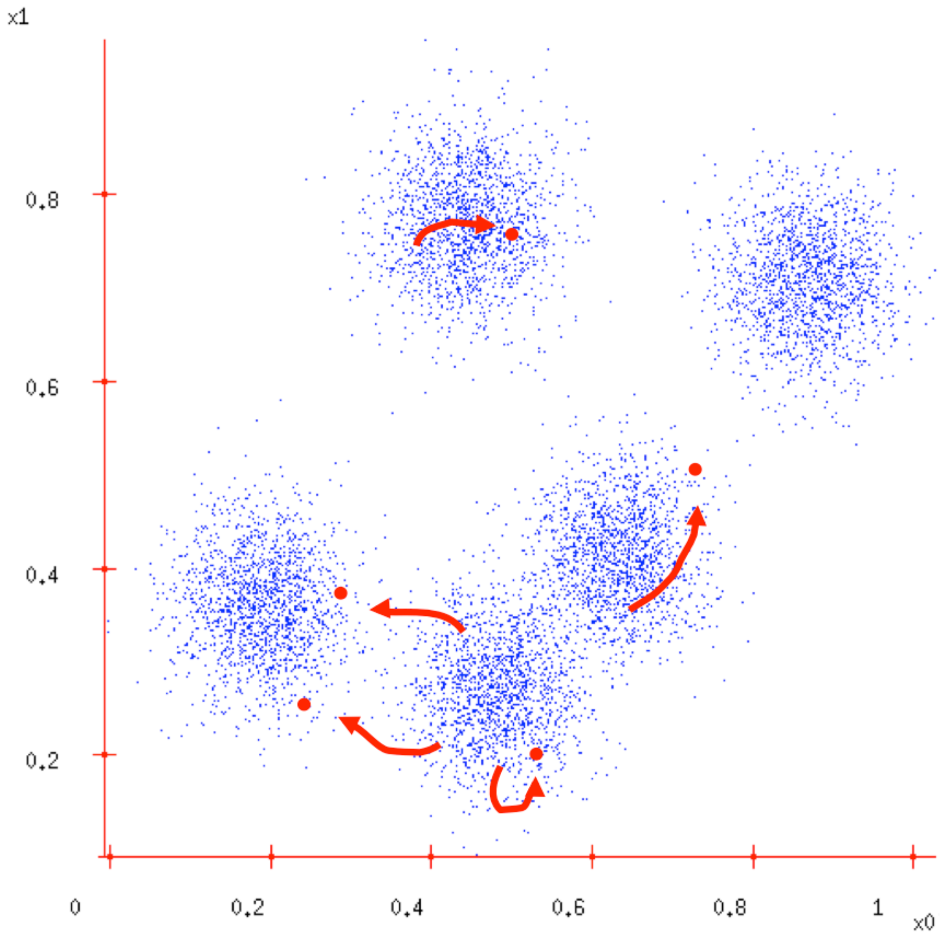
k-Means Clustering: Example

Attribution of Data to Clusters



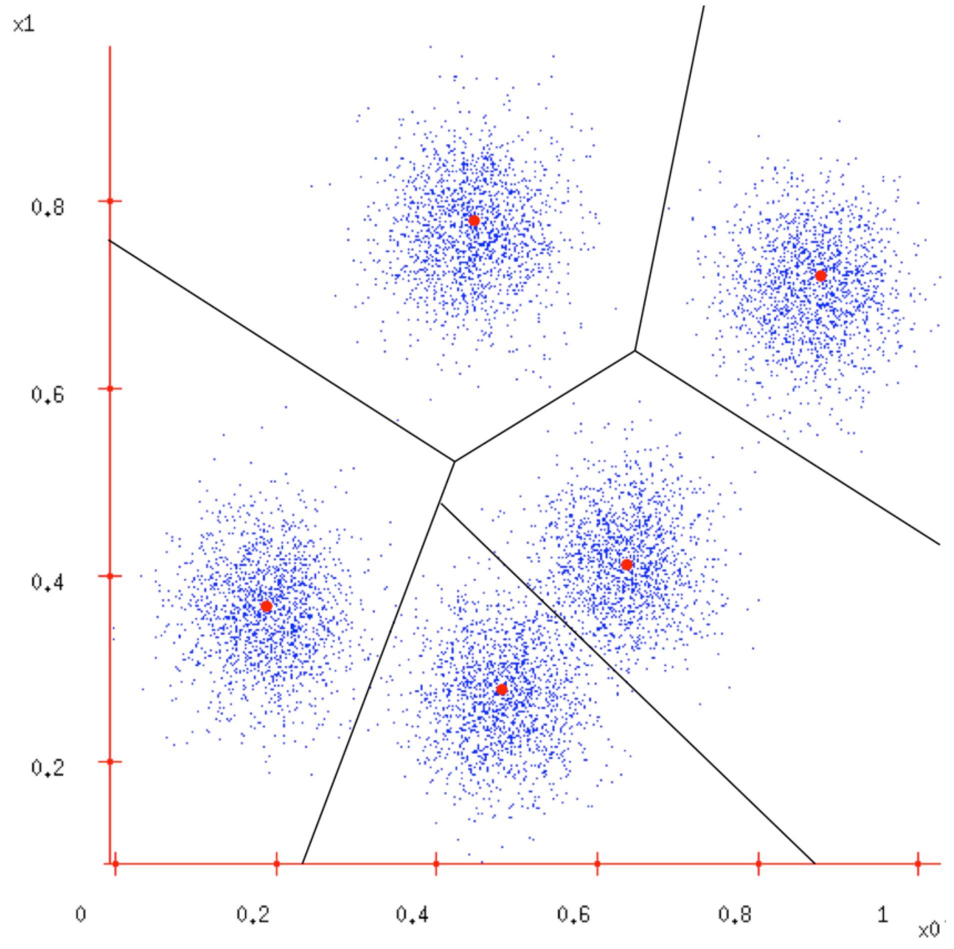
k-Means Clustering: Example

Re-Computation of Centroids



k-Means Clustering: Example

Re-Attribution of Data to Clusters



k-Means Clustering: Considerations

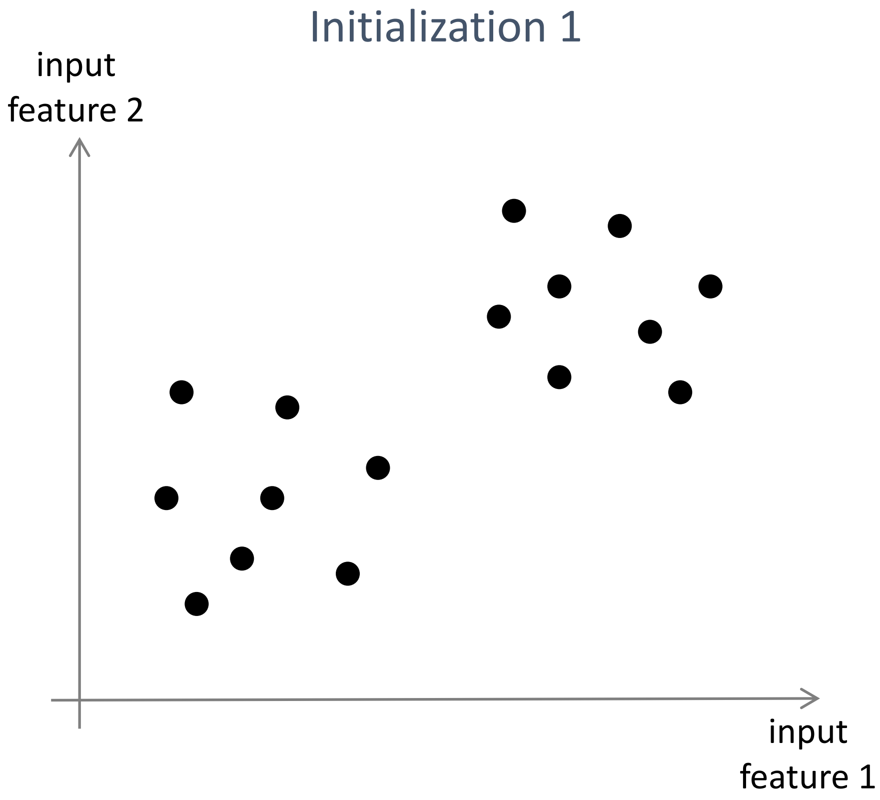
Does the algorithm terminate?

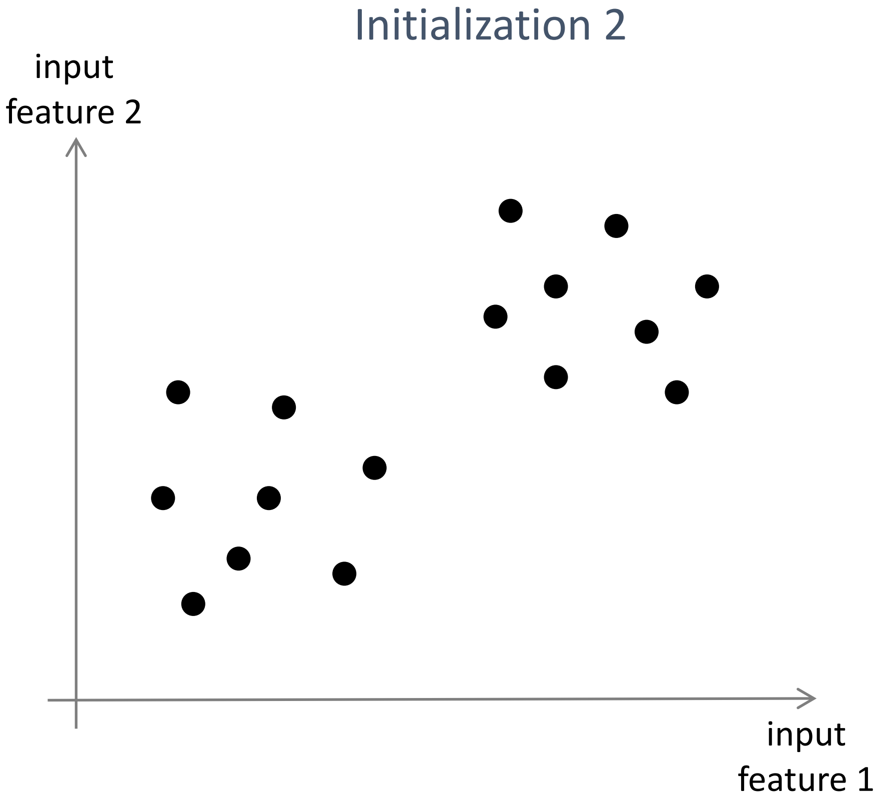
The k-means clustering algorithm always converges to a solution (cluster assignment), with the cost decreasing from iteration to iteration.

However, the resulting cluster assignment may be a local minimum of the cost function *in lieu* of the global one.

k-Means Clustering: Considerations

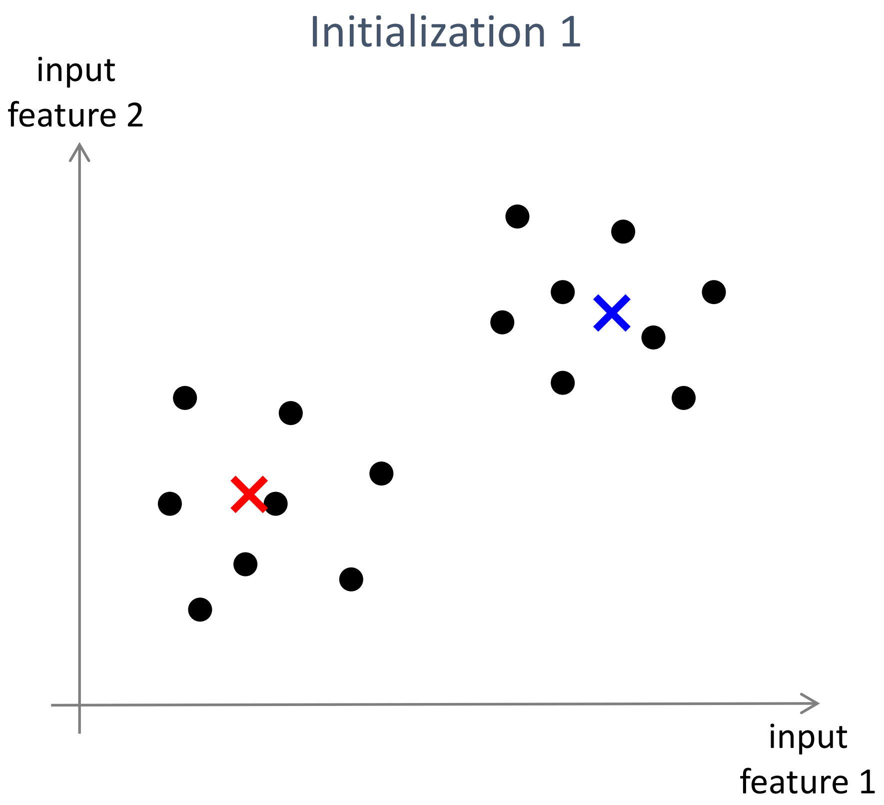
What is the impact of initialization?

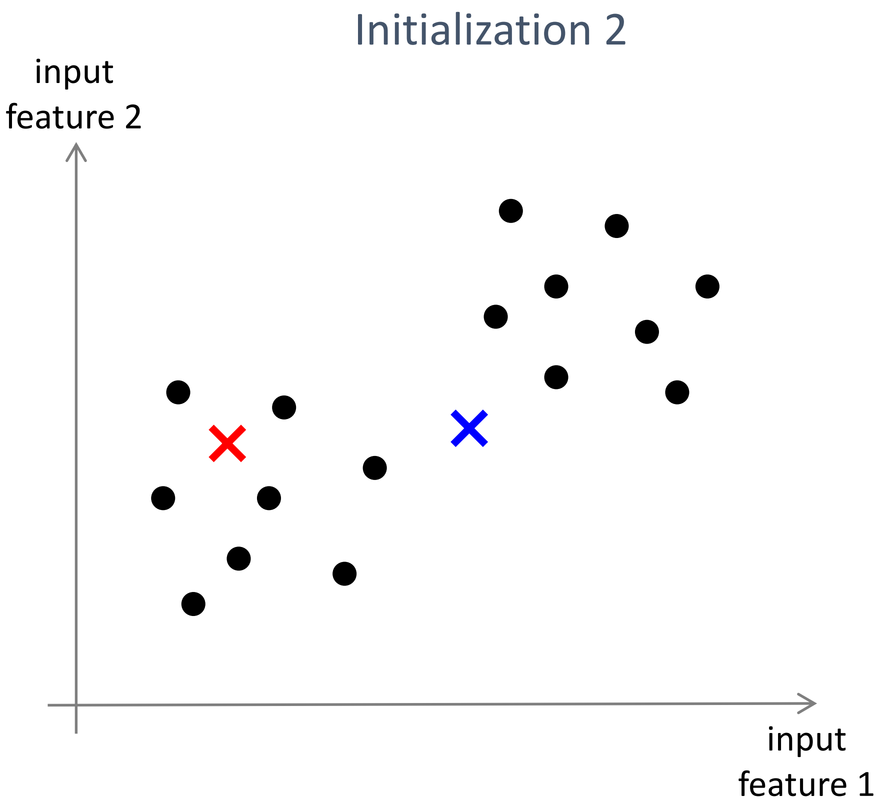




k-Means Clustering: Considerations

What is the impact of initialization?

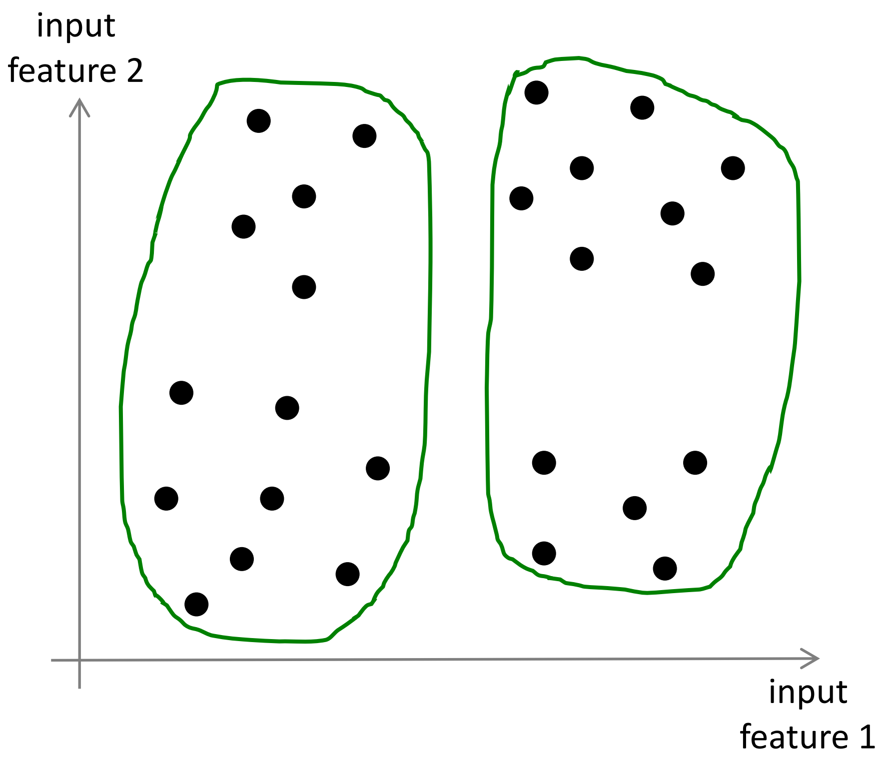


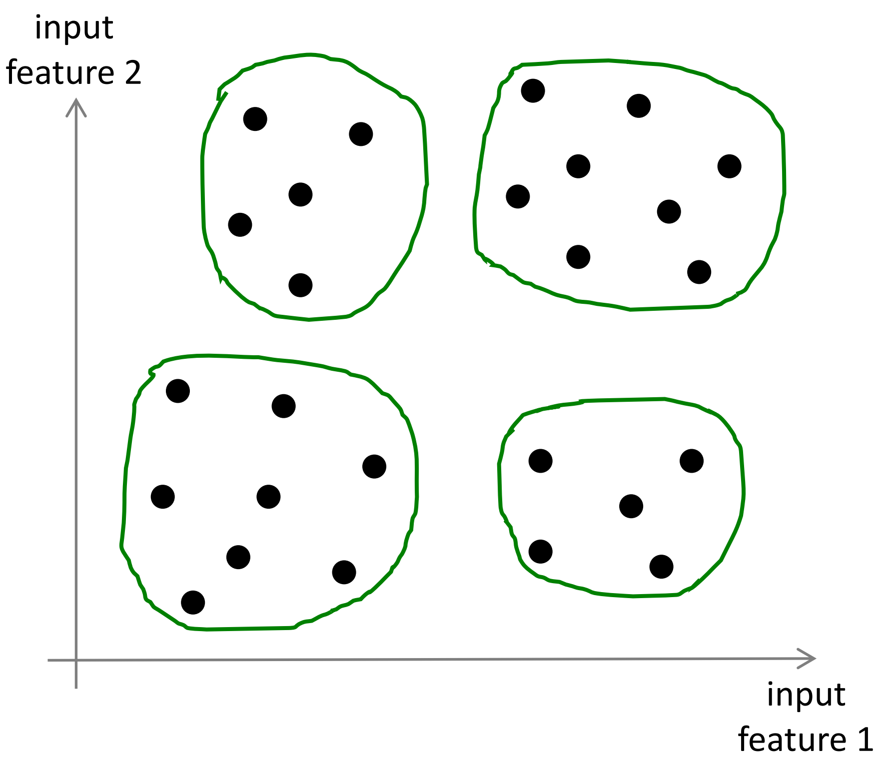


Solution: Run the algorithm with multiple initializations, but select the result with minimum cost

k-Means Clustering: Considerations

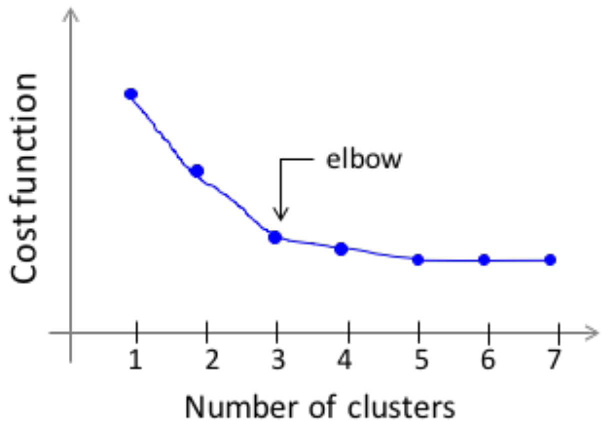
How to select the number of clusters?

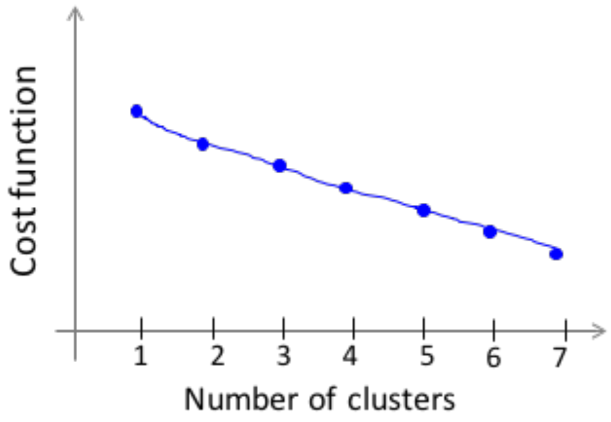




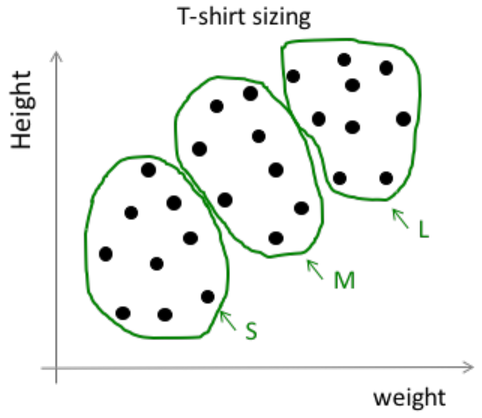
k-Means Clustering: Considerations

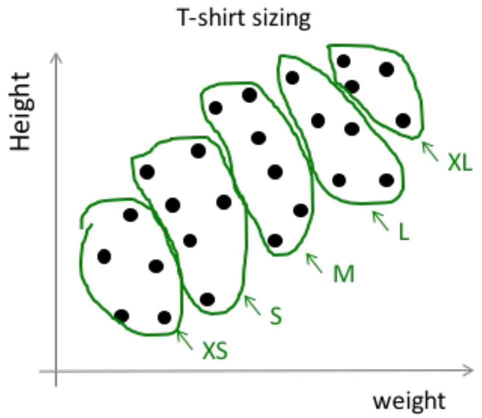
How to select the number of clusters?





Elbow Method





Application Informed