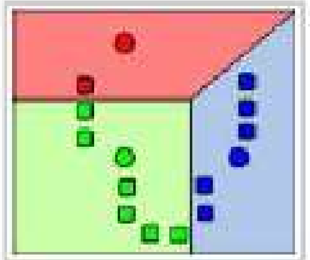
K-Means Clustering

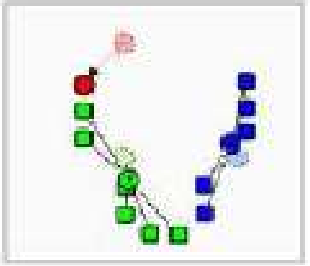
K-means Clustering (DHS 10.4.3)

- K-means Clustering

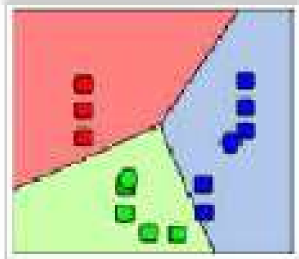
Shows the initial randomized centroids and a number of points.



Points are associated with the nearest centroid.



Now the centroids are moved to the center of their respective clusters.



Steps 2 & 3 are repeated until a suitable level of convergence has been reached.

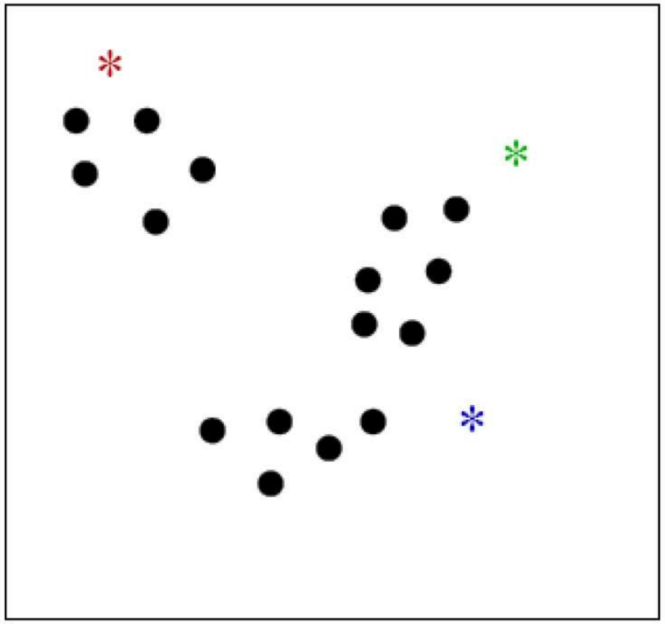
K-Means Algorithm

- K = # of clusters (given); one "mean" per cluster
- Interval data

- Initialize means (e.g. by picking k samples at random)
- Iterate:

(1) assign each point to nearest mean

(2) move "mean" to center of its cluster.

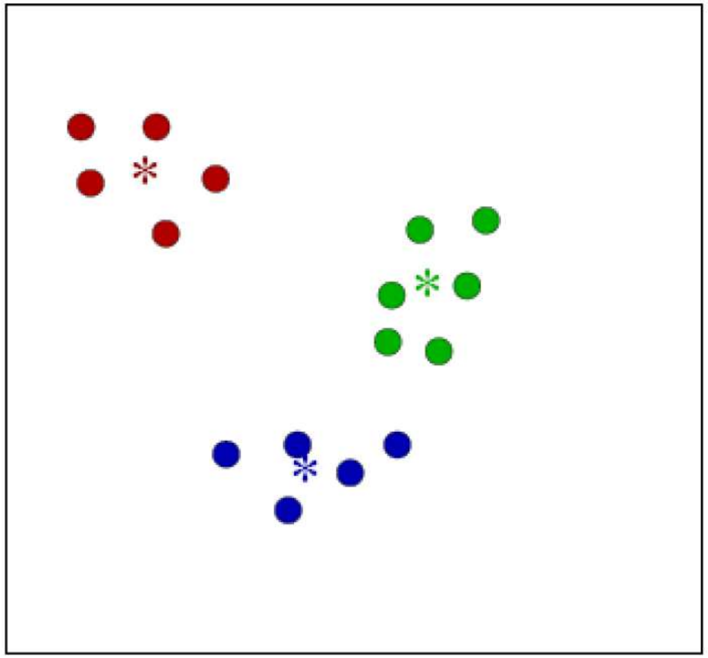


Initialize representatives ("means")

Convergence after another iteration

Complexity:

O(k . n . # of iterations



The objective function is

$$\min_{\{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k\}} \sum_{h=1} \sum_{\mathbf{x}\in\mathcal{X}_h} \|\mathbf{x}-\boldsymbol{\mu}_h\|^2$$

K-means Clustering – Details

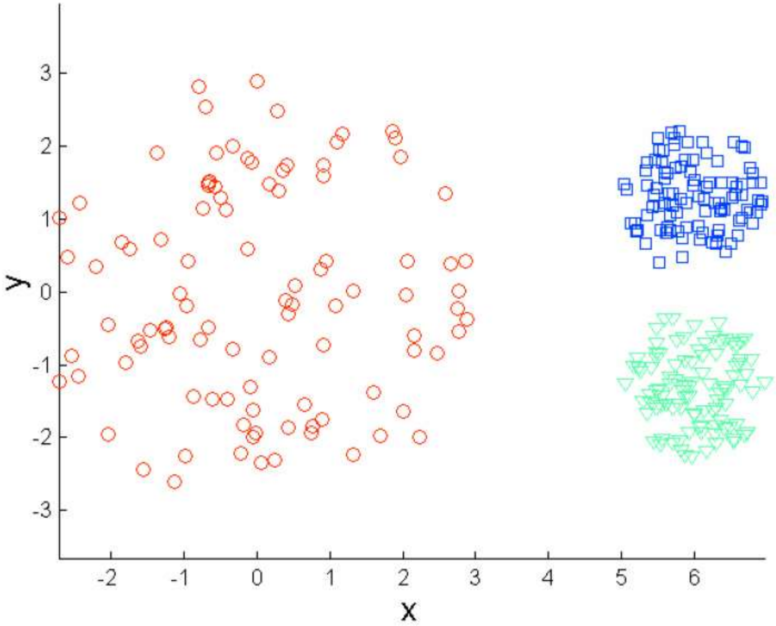
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes
  - Easily parallelized
  - Use kd-trees or other efficient spatial data structures for some situations
    - Pelleg and Moore (X-means)
- Sensitivity to initial conditions
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Limitations of K-means

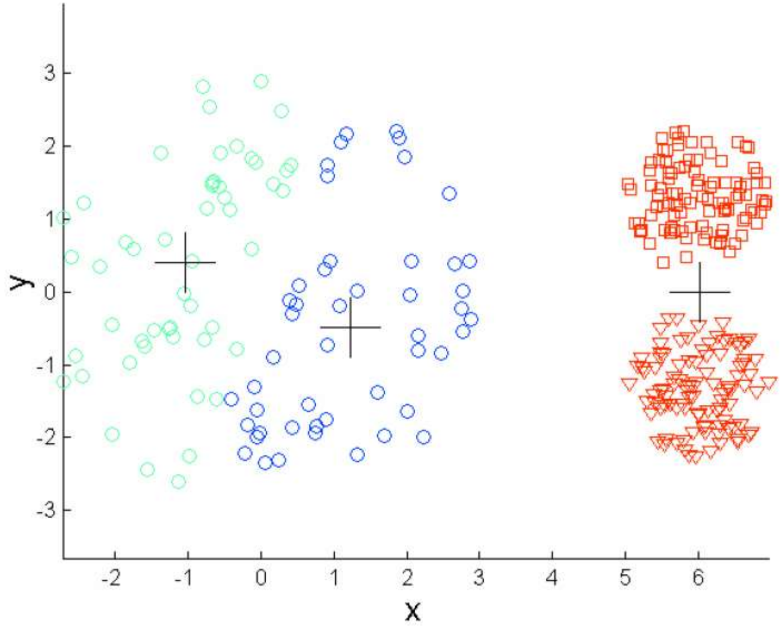
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- Problems with outliers
- Empty clusters

Limitations of K-means: Differing Density

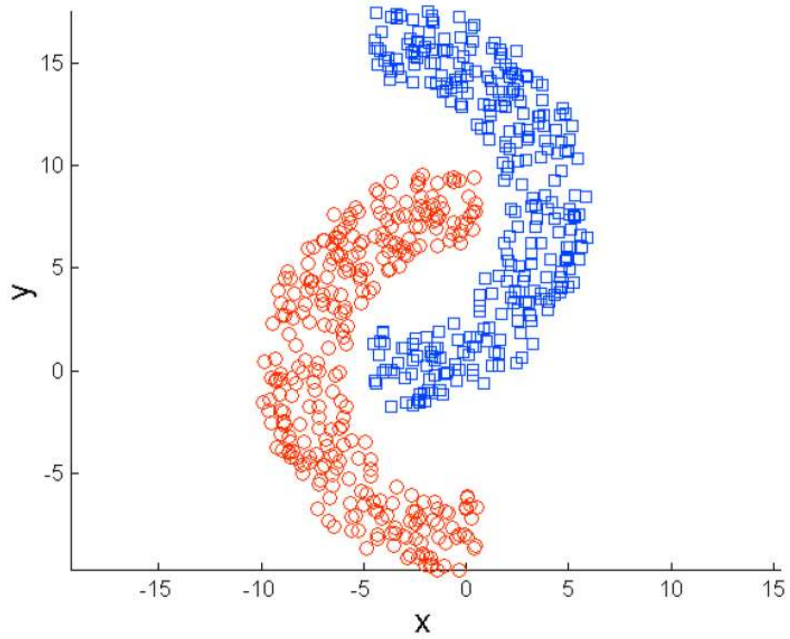


Original Points

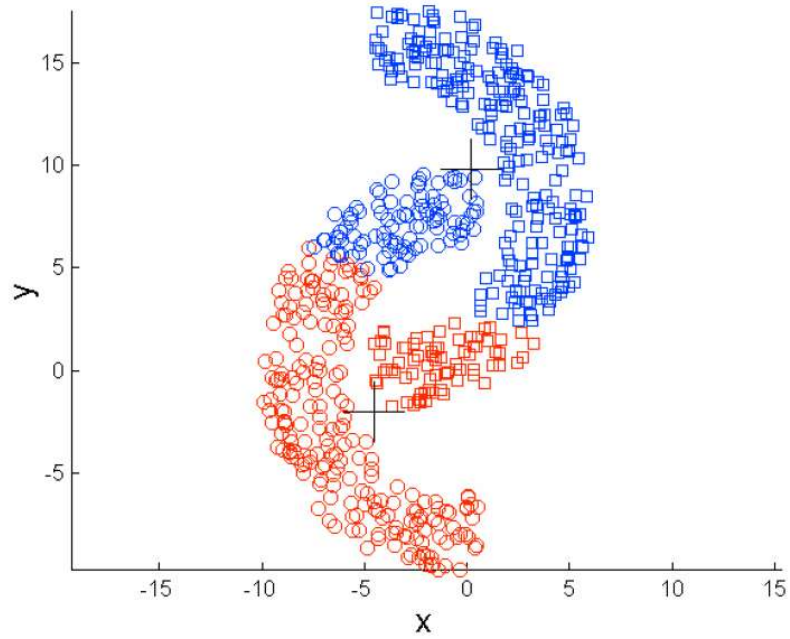


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

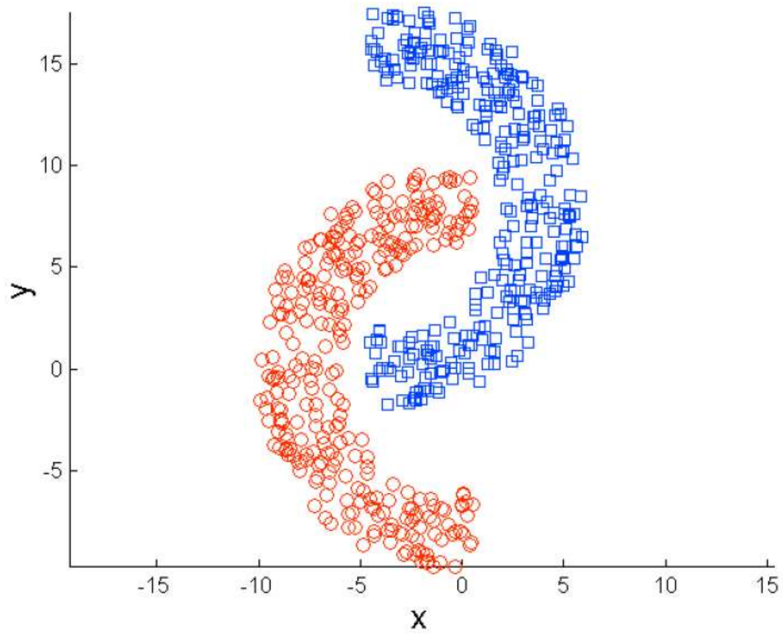


Original Points

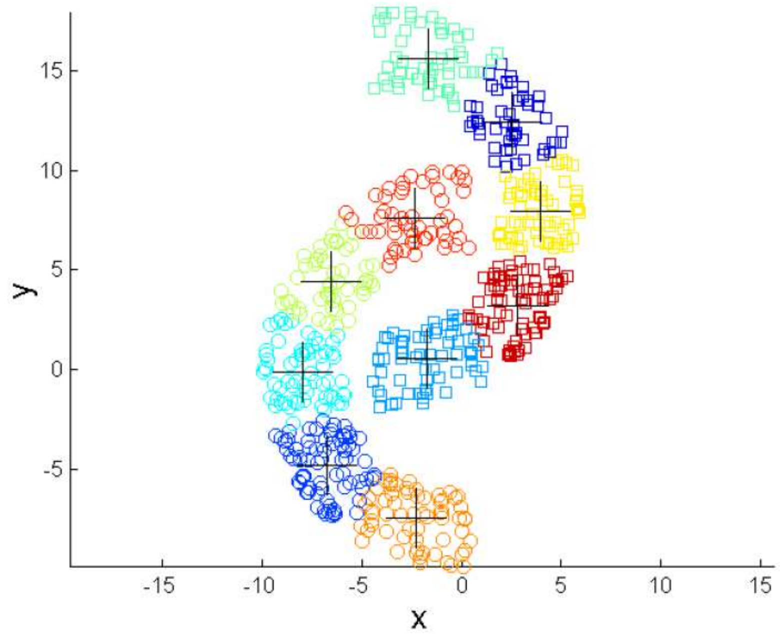


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points



K-means Clusters

Solutions to Initial Centroids Problem

- Multiple runs
- Cluster a sample first
- ….