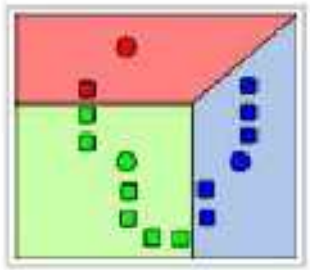
K-Means Clustering

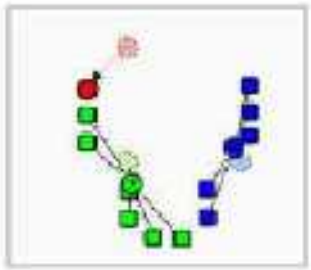
K-means Clustering (DHS 10.4.3)

- K-means Clustering

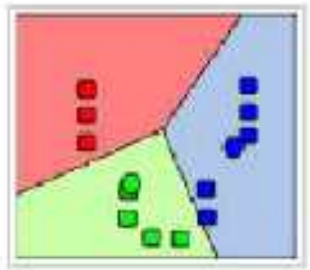
Shows the initial randomized centroids and a number of points.



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.

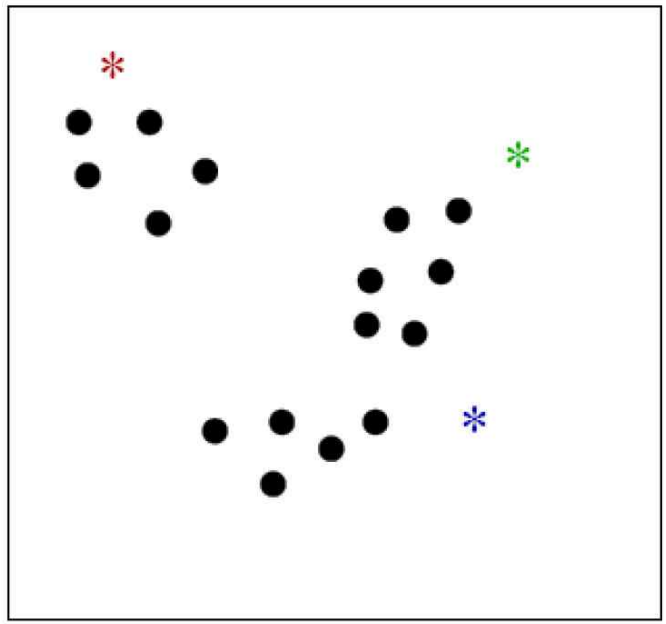
K-Means Algorithm

- 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.

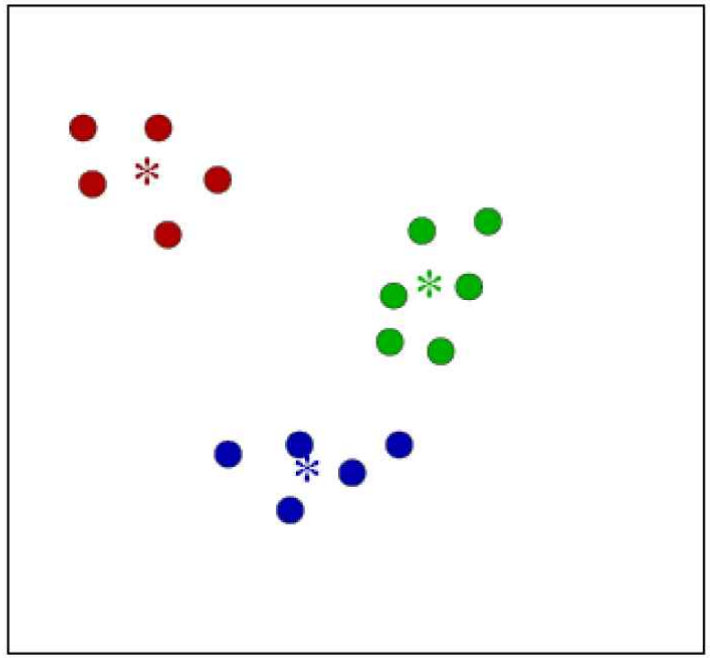


Initialize representatives ("means")

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

- J. MacQueen, Some methods for classification and analysis of multivariate observations," Proc. of the Fifth Berkeley Symp. On Math. Stat. and Prob., vol. 1, pp. 281-296, 1967.
- E. Forgy, Cluster analysis of multivariate data: efficiency vs. interpretability of classification," Biometrics, vol. 21, pp. 768, 1965.
- D. J. Hall and G. B. Ball, ISODATA: A novel method of data analysis and pattern classification," Technical Report, Stanford Research Institute, Menlo Park, CA, 1965.

- The history of k-means type of algorithms (LBG Algorithm, 1980) R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325-2384, October 1998. (Commemorative Issue, 1948-1998)

K-means Clustering – Details

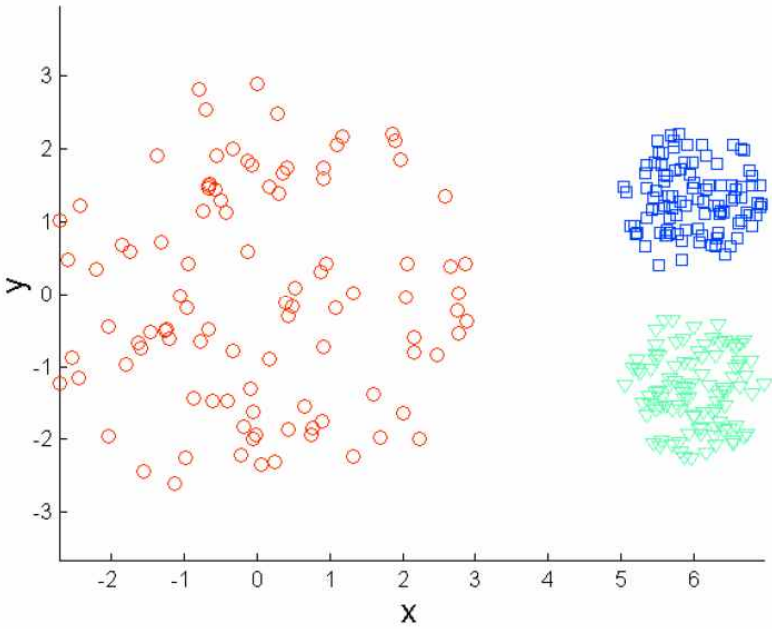
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters, I = number of iterations, d = number of attributes
  - Easily parallelized
  - Use kd-trees or other efficient spatial data structures for some situations
    - Pelleg and Moore (X-means)
- Sensitivity to initial conditions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

Limitations of K-means

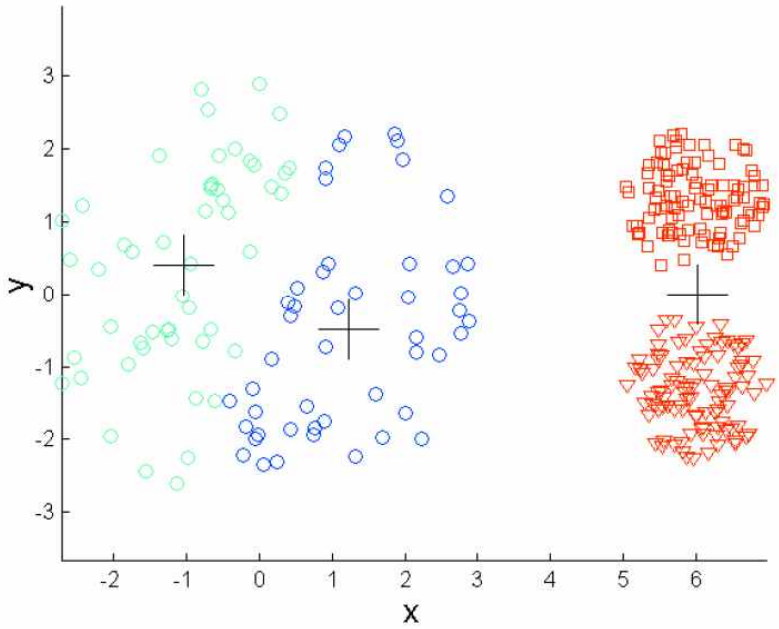
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- Problems with outliers
- Empty clusters

Limitations of K-means: Differing Density

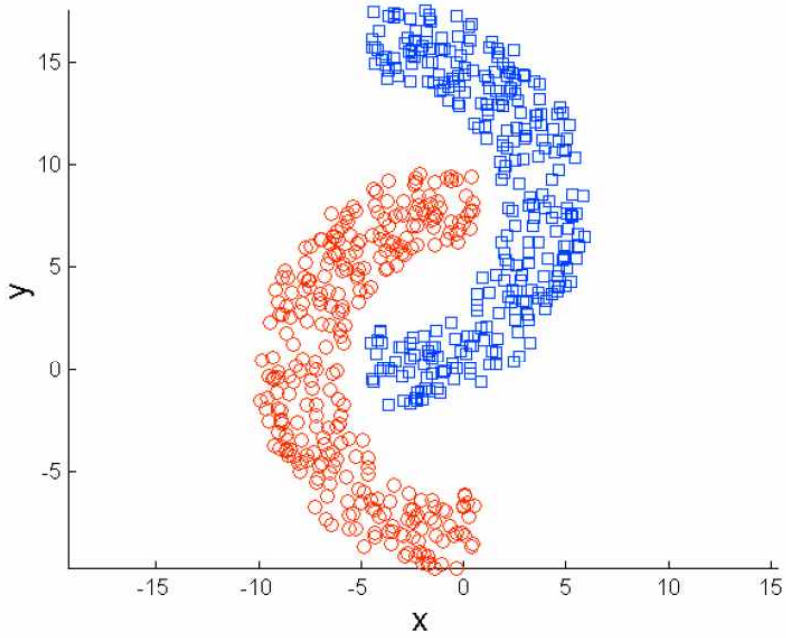


Original Points

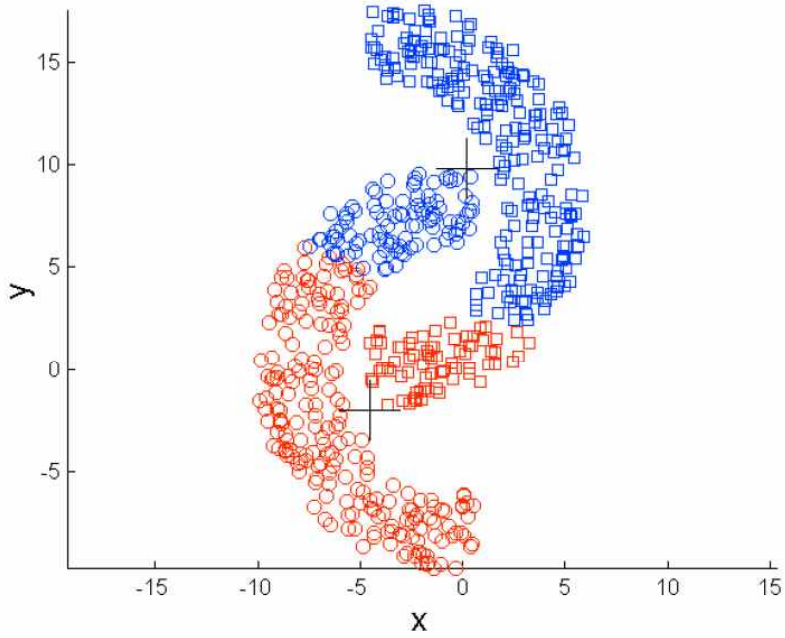


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

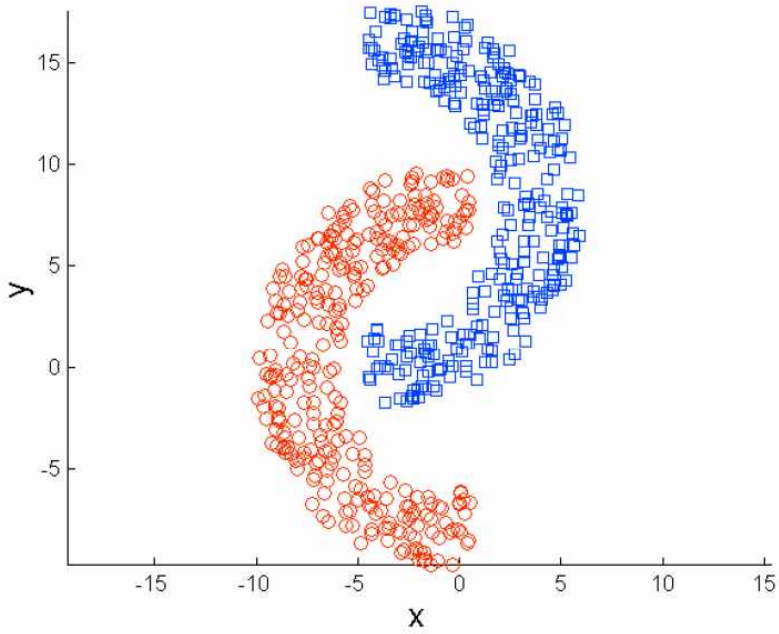


Original Points

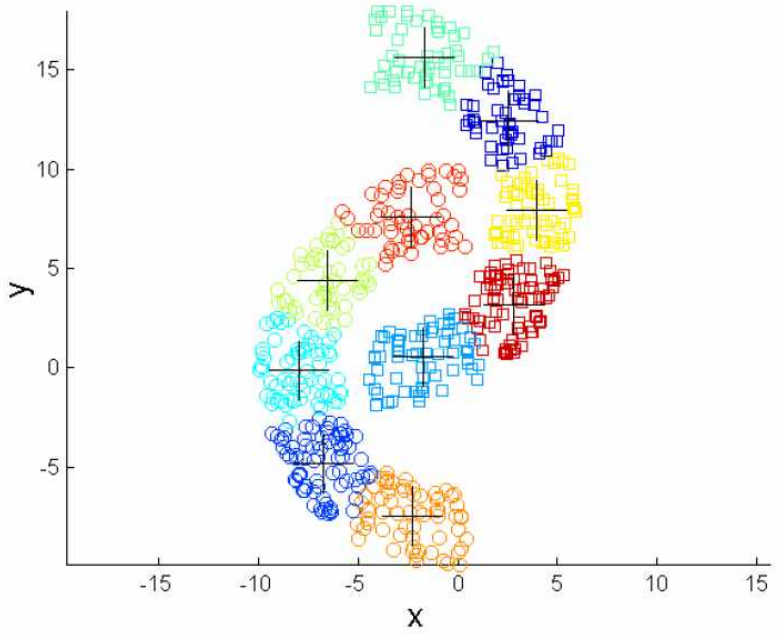


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points



K-means Clusters

Solutions to Initial Centroids Problem

- Multiple runs
- Cluster a sample first
- ….

- Bisecting K-means
  - Not as susceptible to initialization issues

Generalizing K-means

- Model based k-means
  - "means" are probabilistic models"
    - (unified framework, Zhong & Ghosh, JMLR 03)

- Kernel k-means
  - Map data to higher dimensional space
  - Perform k-means clustering
  - Has a relationship to spectral clustering
    - Inderjit S. Dhillon, Yuqiang Guan, Brian Kulis: Kernel k-means: spectral clustering and normalized cuts. KDD 2004: 551-556

K-mean Research

- Almost every aspect of K-means has been modified
  - Distance measures
  - Centroid and objective definitions
  - Overall process
  - Efficiency Enhancements
  - Initialization

K-means Research

- New centroid and objective definitions
  - Fuzzy c-means
    - An object belongs to all clusters with a some weight
    - Sum of the weights is 1
    - J. C. Bezdek (1973). Fuzzy Mathematics in Pattern Classification, PhD Thesis, Cornell University, Ithaca, NY.
  - Harmonic K-means
    - Use harmonic mean instead of standard mean
    - Zhang, Bin; Hsu, Meichun; Dayal, Umeshwar, K-Harmonic Means - A Data Clustering Algorithm, HPL-1999-124

K-mean Research

- New Distance measures
  - Euclidean was the initial measures
  - Use of cosine measure allows k-means to work well for documents
  - Correlation, L1 distance, and Jaccard measures also used
  - Bregman divergence measures allow a k-means type algorithm to apply to many distance measures
    - **Clustering with Bregman Divergences**
      A. Banerjee, S. Merugu, I. Dhillon and J. Ghosh.
      Journal of Machine Learning Research (JMLR) (2005).