The k-means algorithm

(Notes from: Tan, Steinbach, Kumar + Ghosh)

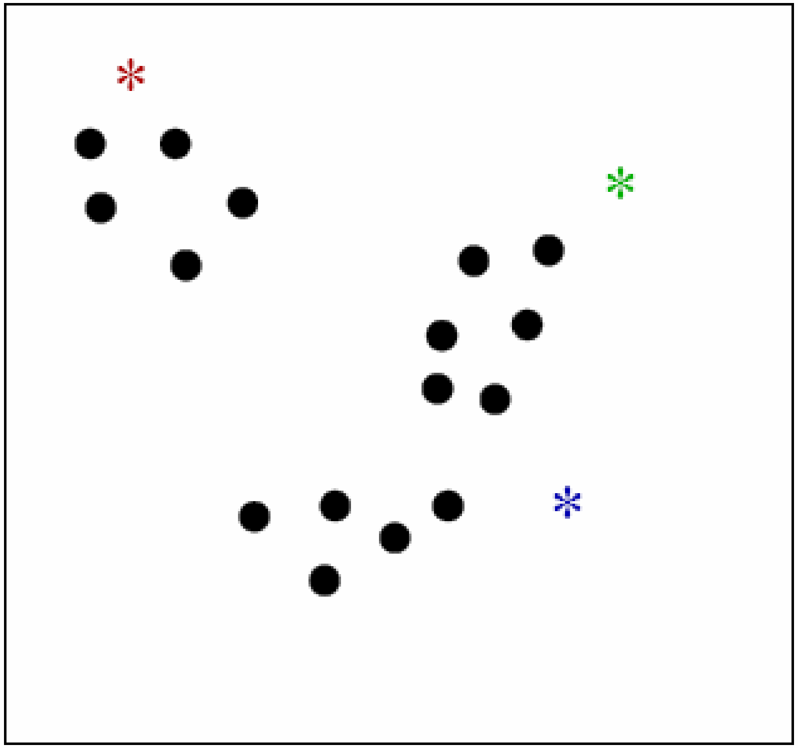
K-Means Algorithm

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

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

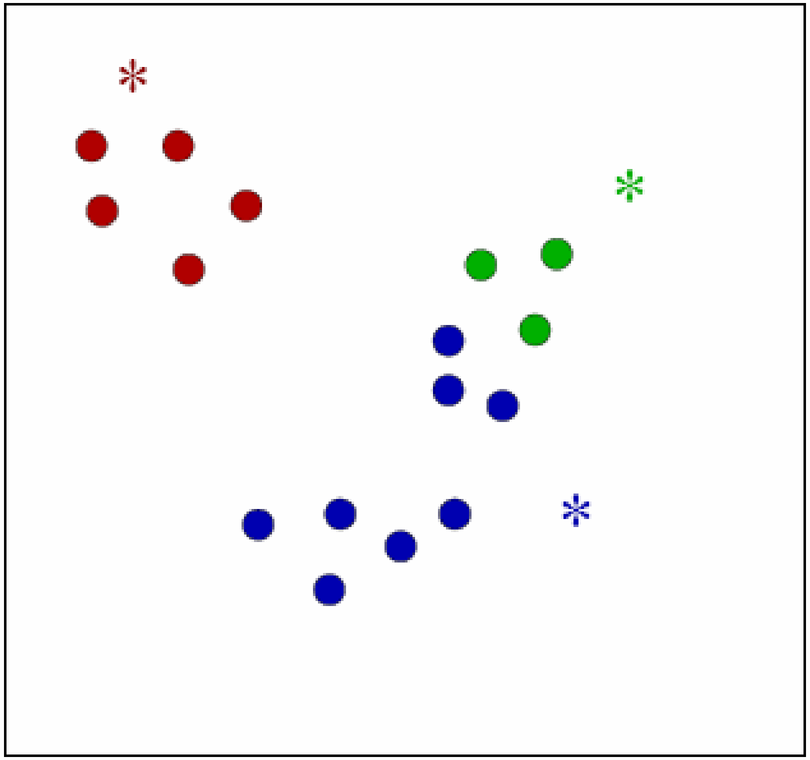
(1) assign each point to nearest mean

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

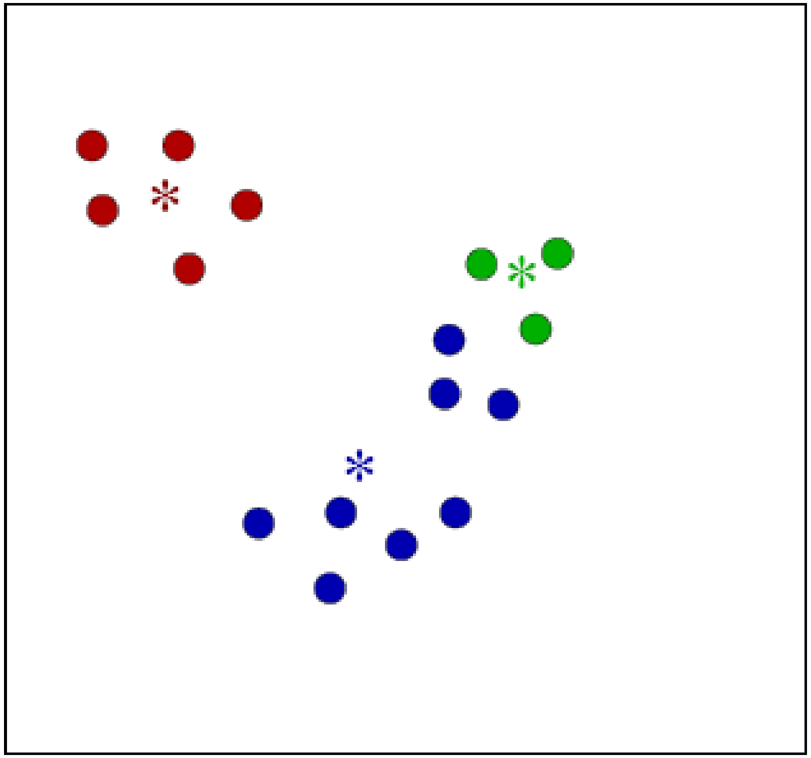


Initialize representatives ("means")

Assignment Step; Means Update



Assign to nearest representative

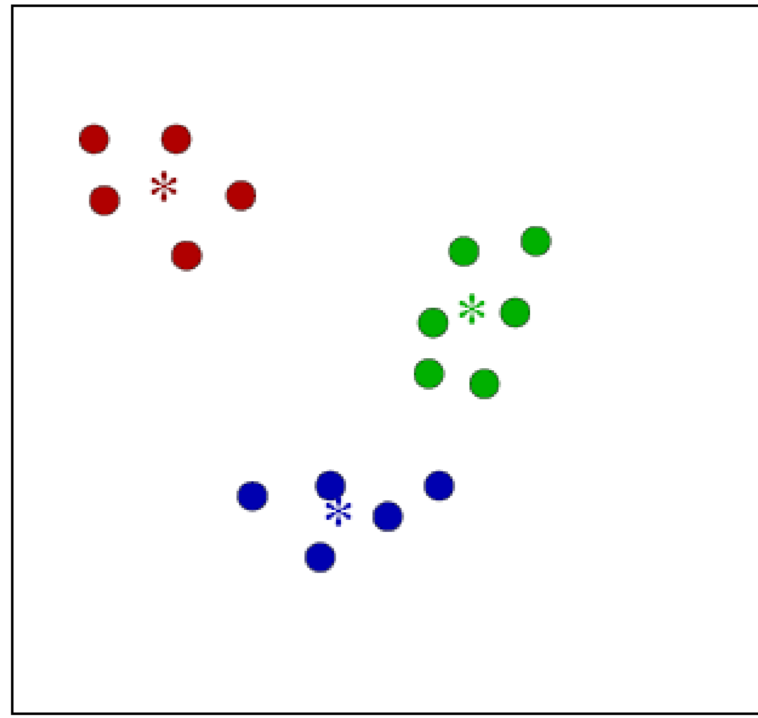


Re-estimate means

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

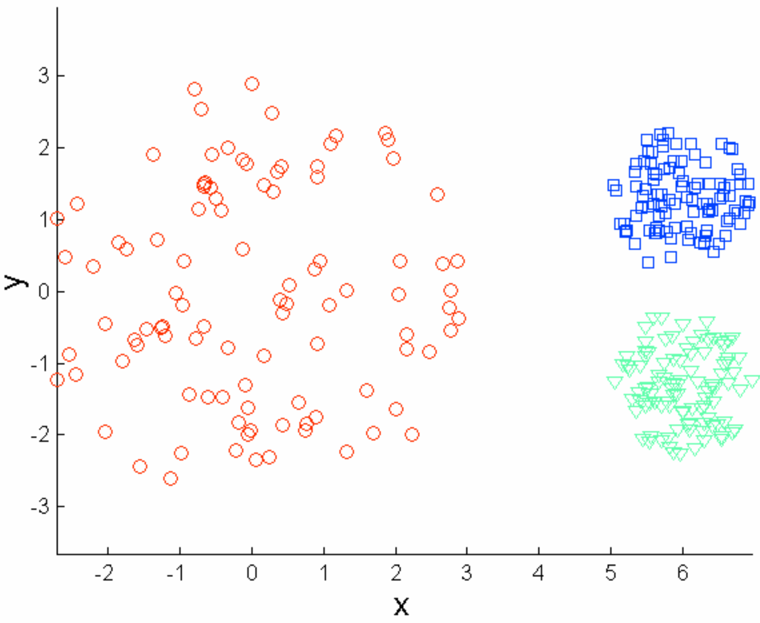
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters, I = number of iterations, d = number of attributes
  - Easily parallelized
  - Use kd-trees or other efficient spatial data structures for some situations
    - Pelleg and Moore (X-means)
- Sensitivity to initial conditions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

Limitations of K-means

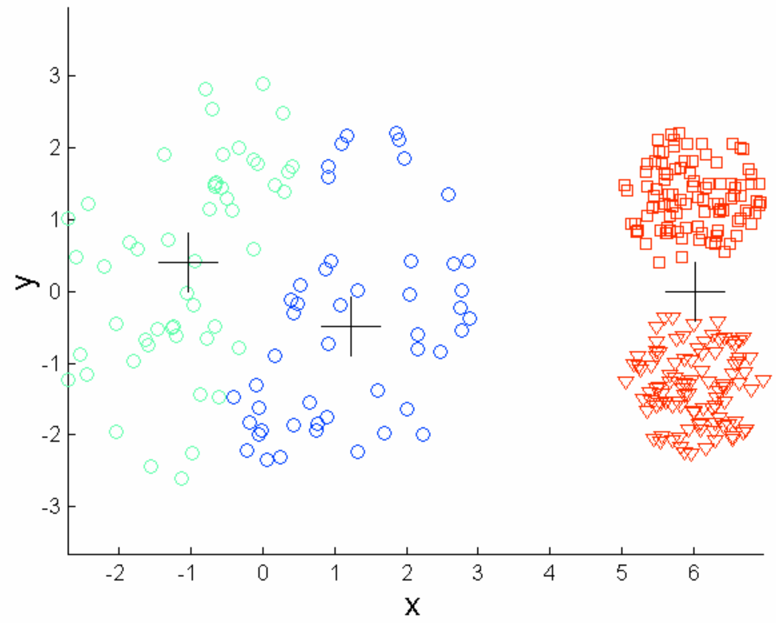
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- Problems with outliers
- Empty clusters

Limitations of K-means: Differing Density

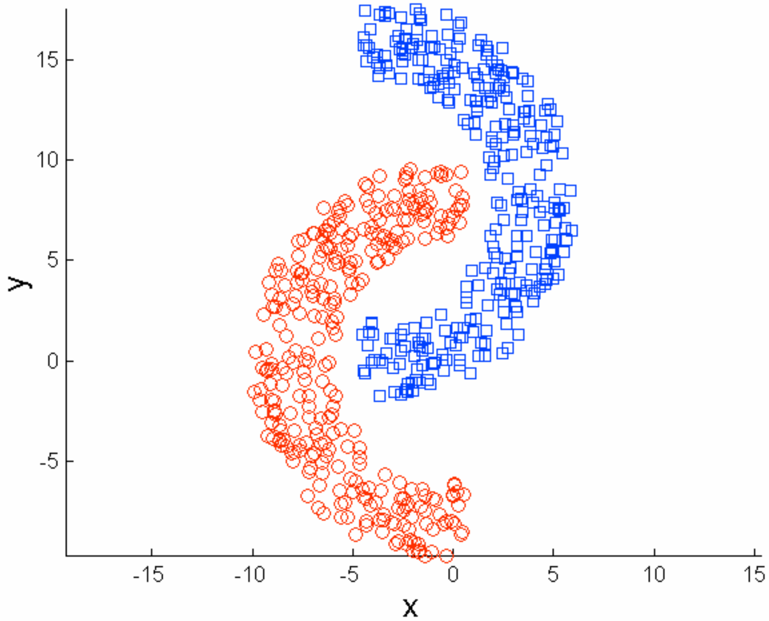


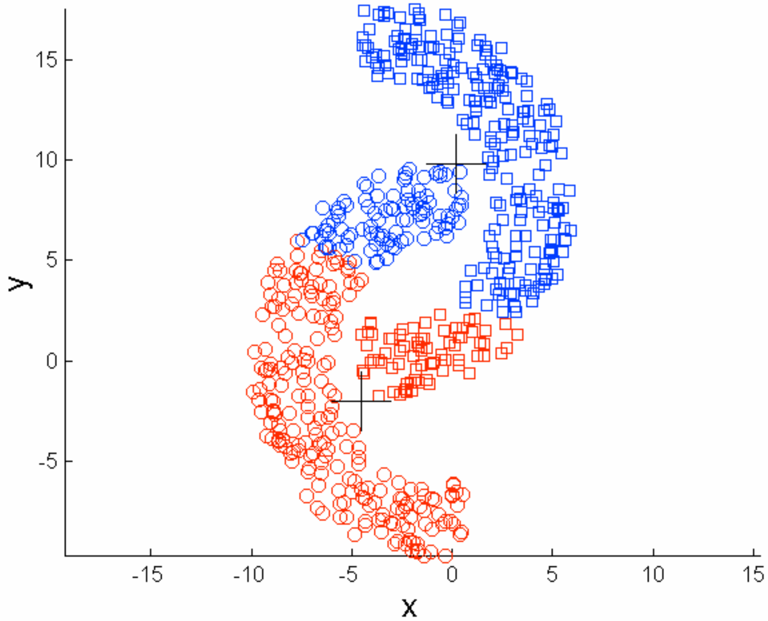
Original Points



K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

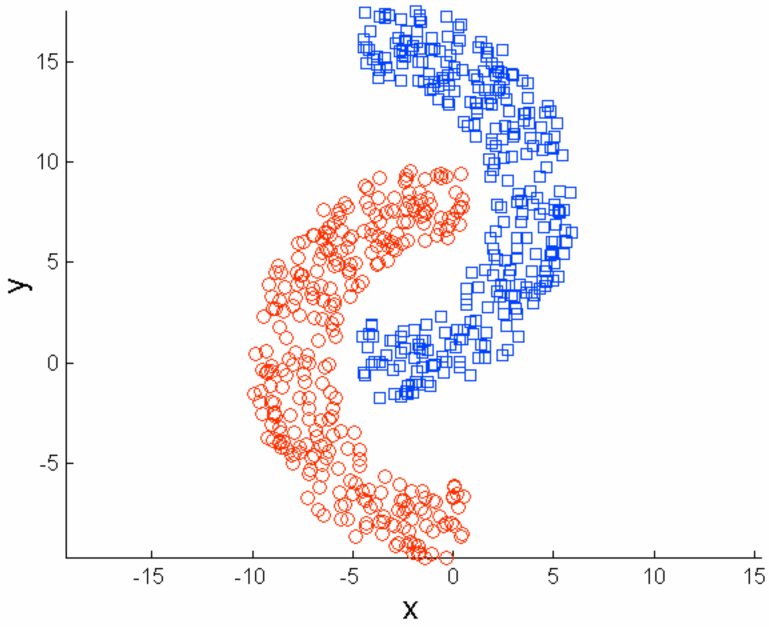




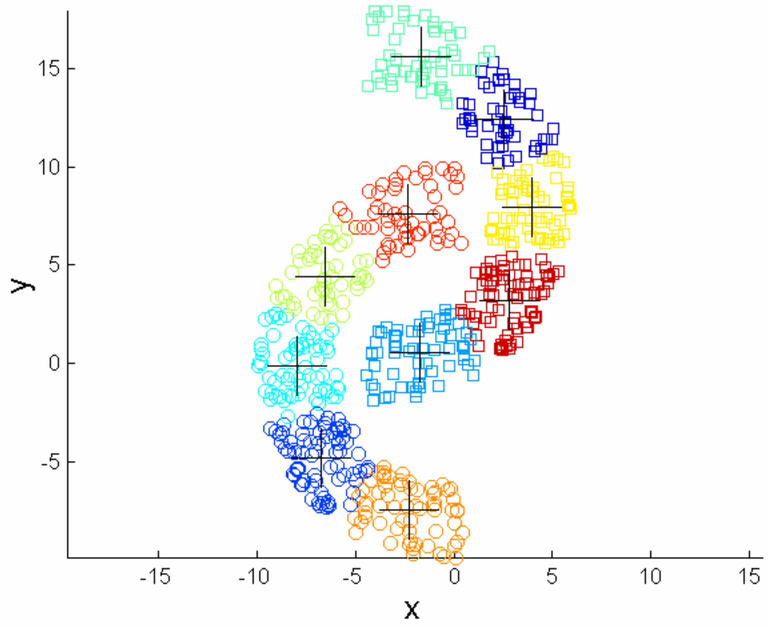
Original Points

K-means (2 Clusters)

Overcoming K-means Limitations



Original Points



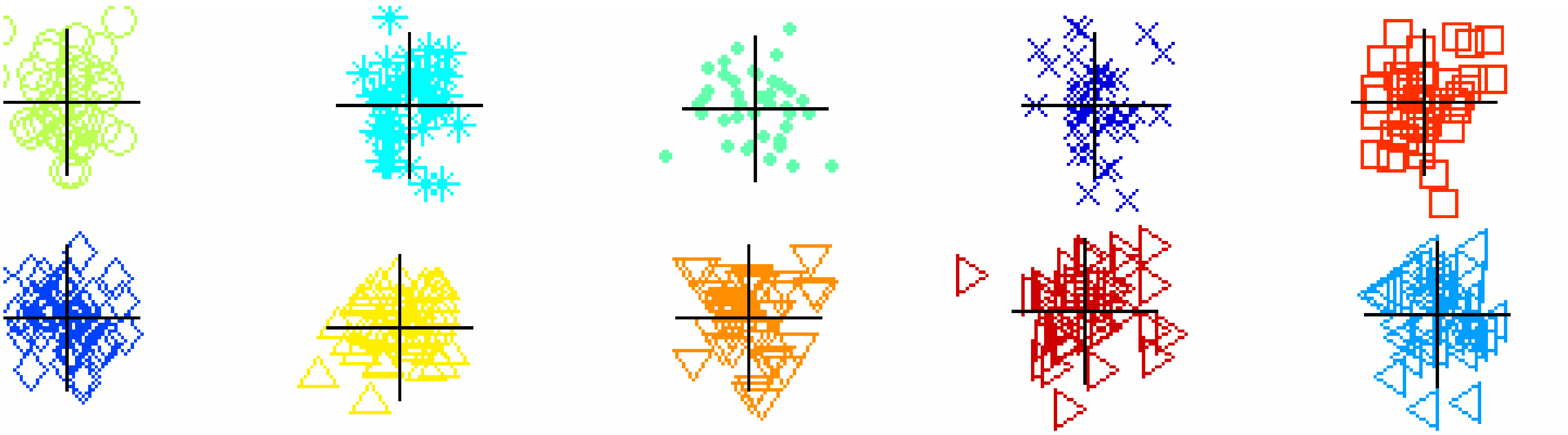
K-means Clusters

Solutions to Initial Centroids Problem

- Multiple runs
- Cluster a sample first
- ….

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

Bisecting K-means Example



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

Clustering with Bregman Divergences

- Banerjee, Merugu, Dhillon, Ghosh, SDM 2004; JMLR 2005
  - Hard Clustering: KMeans-type algo possible for any Bregman Divergence
  - Bijection: convex function <--> Bregman divergence <--> exp. Family
    - Soft Clustering: efficient algo for learning mixtures of any exponential family

Bregman Hard Clustering

- Initialize $\{\mu_h\}_{h=1}^{k}$
- Repeat until *convergence*
  - { Assignment Step }
    Assign x to $\mathcal{X}_h$ if $h = \underset{h'}{\operatorname{argmin}}\ d_\phi(\mathrm{x}, \mu_{h'})$
  - { Re-estimation step }
    For all h

$$\mu_h = \frac{\sum_{\mathrm{x} \in \mathcal{X}_h} p(\mathrm{x})\,\mathrm{x}}{\sum_{\mathrm{x} \in \mathcal{X}_h} p(\mathrm{x})}$$

Algorithm Properties

- Guarantee: Monotonically decreases a global objective function $E[d_\phi(X_h, \mu_h)]$ till convergence
- Scalability: Every iteration is linear in the size of the input
- Exhaustiveness: If such an algorithm exists for a loss function $L(\mathrm{x}, \mu)$, then L has to be a Bregman divergence
- Linear Separators: Hyperplane separators for all Bregman Divergences
- Mixed Data types:
  - Allows appropriate Bregman divergence for subsets of features

Related Areas

- EM clustering
  - K-means is a special case of EM clustering
  - EM approaches provide more generality, but at a cost
  - C. Fraley , and A. E. Raftery, How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis, The Computer Journal 41: 578-588.
- Vector quantization / Compression
  - R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325-2384, October 1998. (Commemorative Issue, 1948-1998)

Related Areas …

- Operations research
  - Facility location problems
- K-medoid clustering
  - L. Kaufman and PJ Rousseeuw. Finding Groups In Data: An Introduction to Cluster Analysis. Wiley-Interscience, 1990.
  - Raymond T. Ng, Jiawei Han: CLARANS: A Method for Clustering Objects for Spatial Data Mining. IEEE Trans. Knowl. Data Eng. 14(5): 1003-1016 (2002)
- Neural Networks
  - Self Organizing Maps (Kohonen)
  - Bishop, C. M., Svens'en, M., and Williams, C. K. I. (1998). GTM: the generative topographic mapping. Neural Computation, 10(1):215--234

General References

- An Introduction to Data Mining, Tan, Steinbach, Kumar, Addision-Wesley, 2005.
  http://www-users.cs.umn.edu/~kumar/dmbook/index.php

- Data Mining: Concepts and Techniques, 2^{nd} Edition, Jiawei Han and Micheline Kamber, Morgan Kauffman, 2006
  http://www-sal.cs.uiuc.edu/~hanj/bk2

- K-means tutorial slides (Andrew Moore)
  http://www.autonlab.org/tutorials/kmeans11.pdf

- CLUTO clustering software
  http://glaros.dtc.umn.edu/gkhome/views/cluto

K-means Research ...

- Efficiency
  - Parallel Implementations
  - Reduction of distance computations
    - Charles Elkan, Clustering with k-means: faster, smarter, cheaper, Keynote talk at the Workshop on Clustering High-Dimensional Data, SIAM International Conference on Data Mining (SDM 2004)
  - Scaling strategies
    - P. S. Bradley, U. Fayyad, and C. Reina, "Scaling Clustering Algorithms to Large Databases", Proc. 4 th International Conf. on Knowledge Discovery and Data Mining (KDD-98). AAAI Press, Aug. 1998
- Initialization
  - P. S. Bradley and U. M. Fayyad. Refining initial points for k-means clustering. In J. Shavlik, editor, Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98), pages 91--99, San Francisco, CA, 1998.
  - Old technique: sample, apply Wards hierarchical clustering to generate k clusters

`

K-mean Research

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

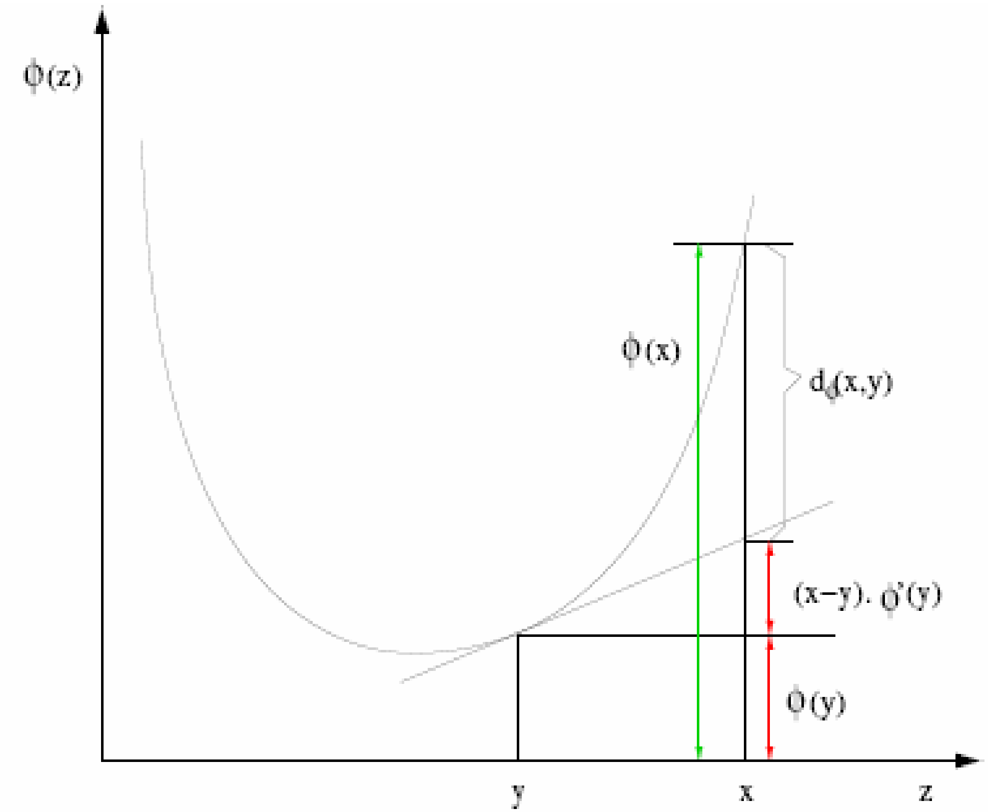
K-mean Research

- 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).

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

Bregman Divergences



ϕ is strictly convex, differentiable

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle$$

Bregman Loss Functions

- $\phi(x) = x^2$ is strictly convex and differentiable on $\mathbb{R}$
  - $D_\phi(x, y) = (x - y)^2$ (squared Euclidean distance)

- $\phi(\mathbf{p}) = \sum_{j=1}^{d} p_j \log p_j$ (negative entropy) is strictly convex and differentiable on the d-simplex
  - $D_\phi(\mathbf{p}, \mathbf{q}) = \sum_{j=1}^{d} p_j \log\left(\frac{p_j}{q_j}\right)$ (KL-divergence)

- $\phi(x) = -\log x$ is strictly convex and differentiable on $\mathbb{R}_{++}$
  - $D_\phi(x, y) = \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1$ (Itakura-Saito distance)