UNSUPERVISED LEARNING 2011

LECTURE :K-MEANS

Rita Osadchy

Some slides are due to Eric Xing, Olga Veksler

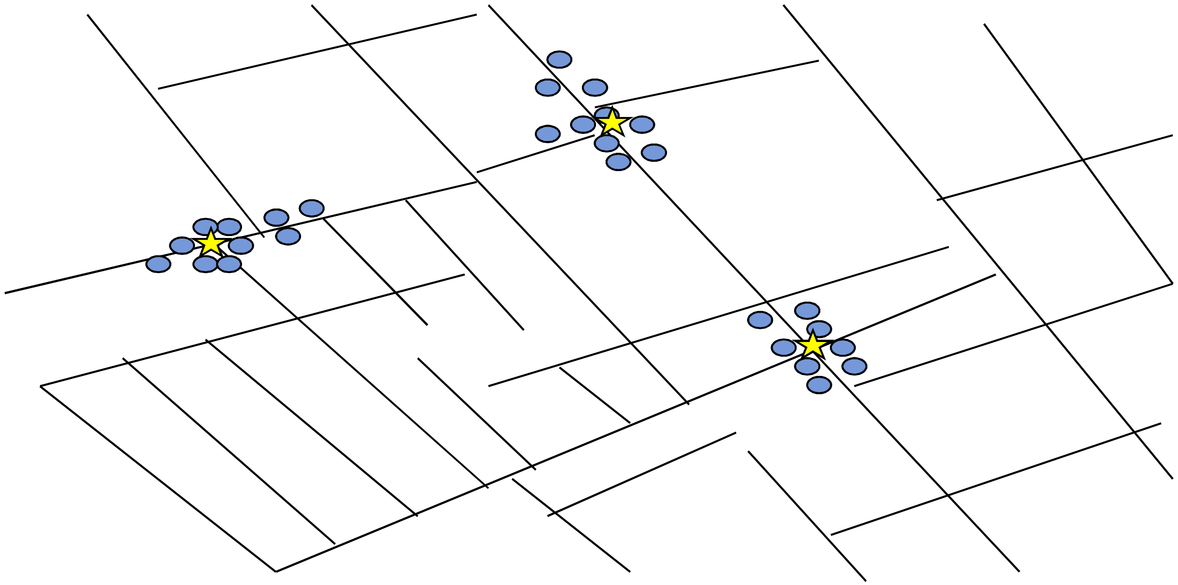
What is clustering?

- Input:
  - Training samples $\{x_1, \ldots, x_m\} \in \Re^n$
  - No labels y_i are given

- Goal: group input samples into classes of similar objects – cohesive “clusters.”
  - high intra-class similarity
  - low inter-class similarity
  - It is the commonest form of unsupervised learning

First (?) Application of Clustering

- John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution.



From: Nina Mishra HP Labs

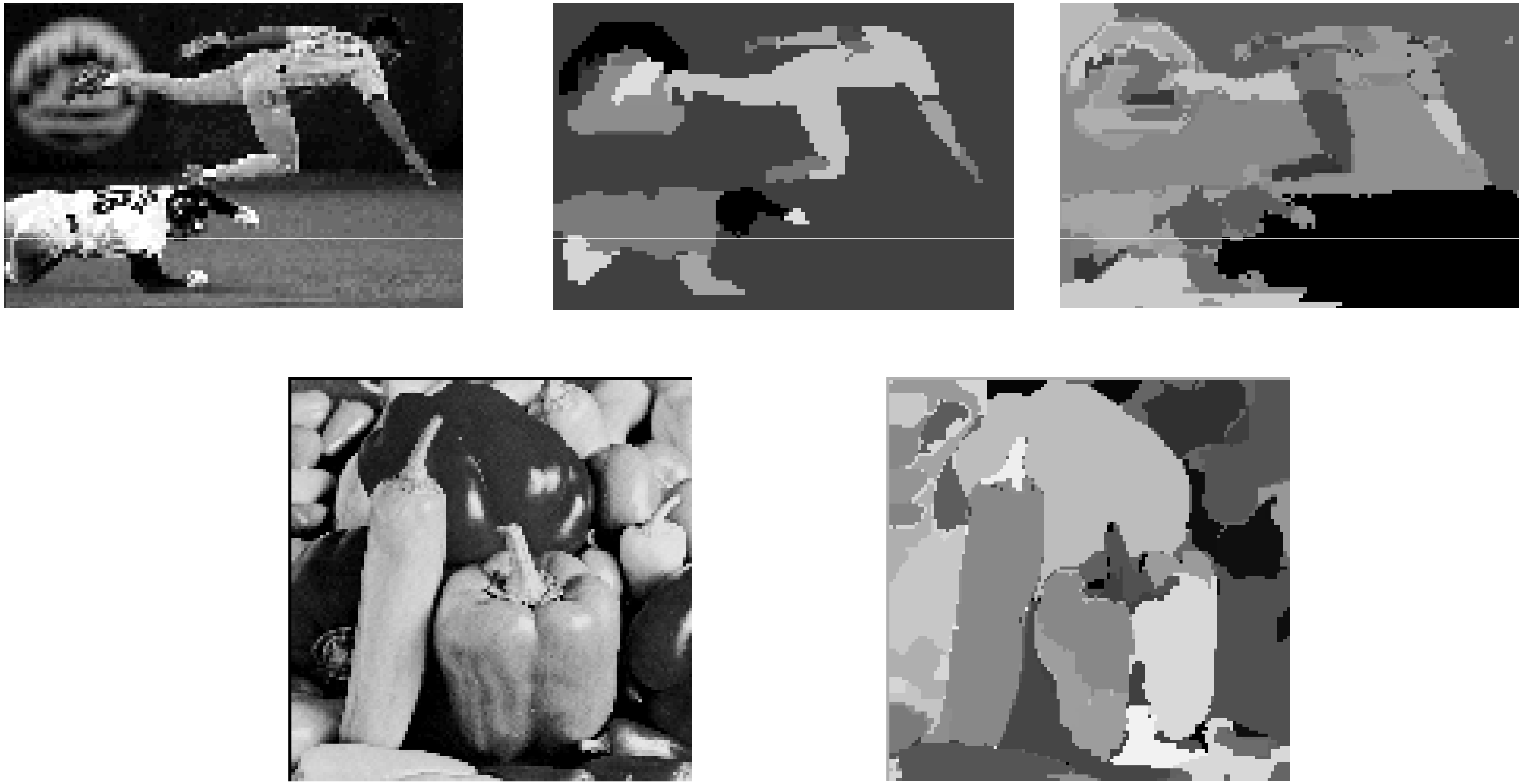
Application of Clustering

- Astronomy
  - SkyCat: Clustered $2x10^9$ sky objects into stars, galaxies, quasars, etc based on radiation emitted in different spectrum bands.

From: Nina Mishra HP Labs

Applications of Clustering

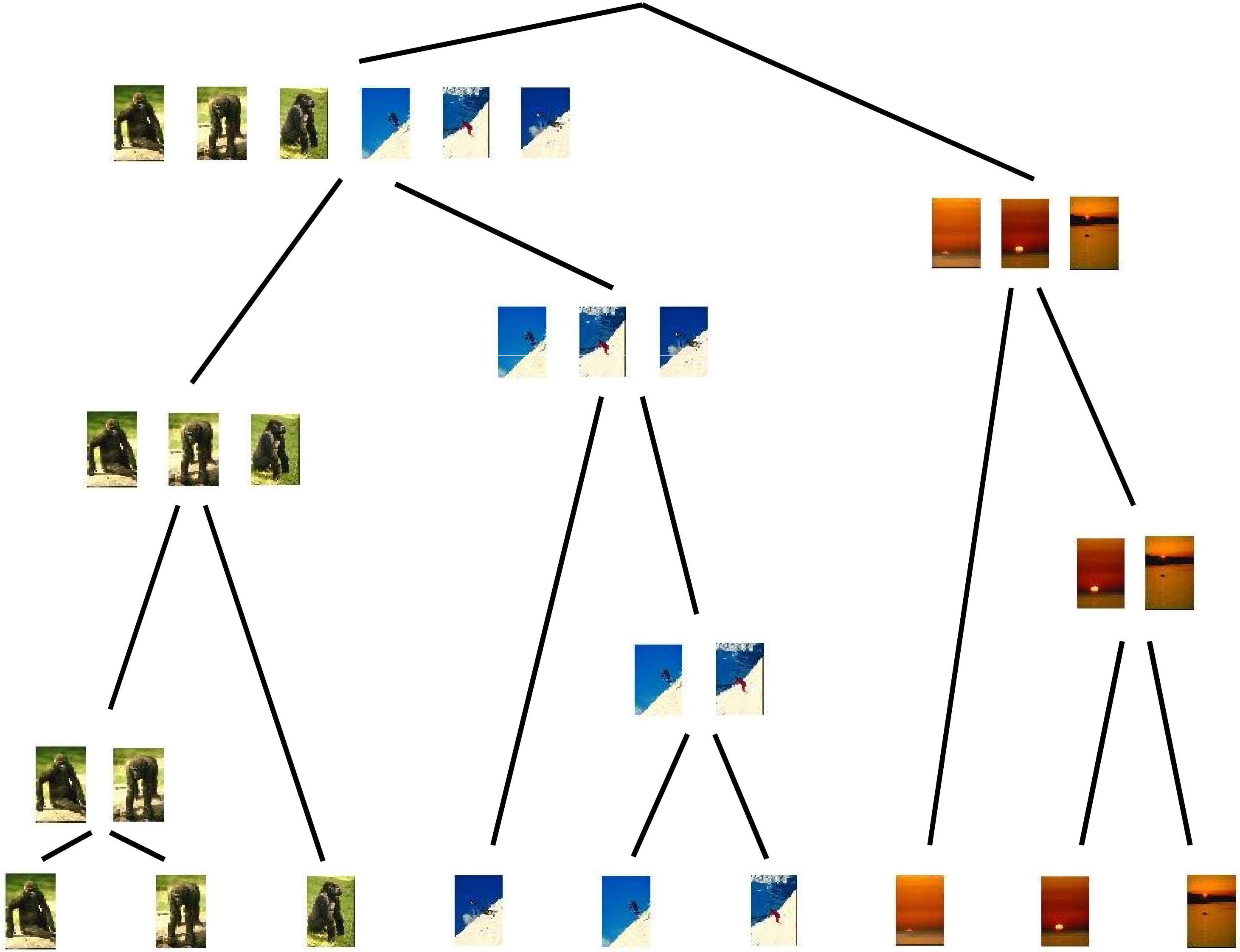
- Image segmentation
  - Find interesting “objects” in images to focus attention at



From: Image Segmentation by Nested Cuts, O. Veksler, CVPR2000

Applications of Clustering

- Image Database Organization
  - for efficient search



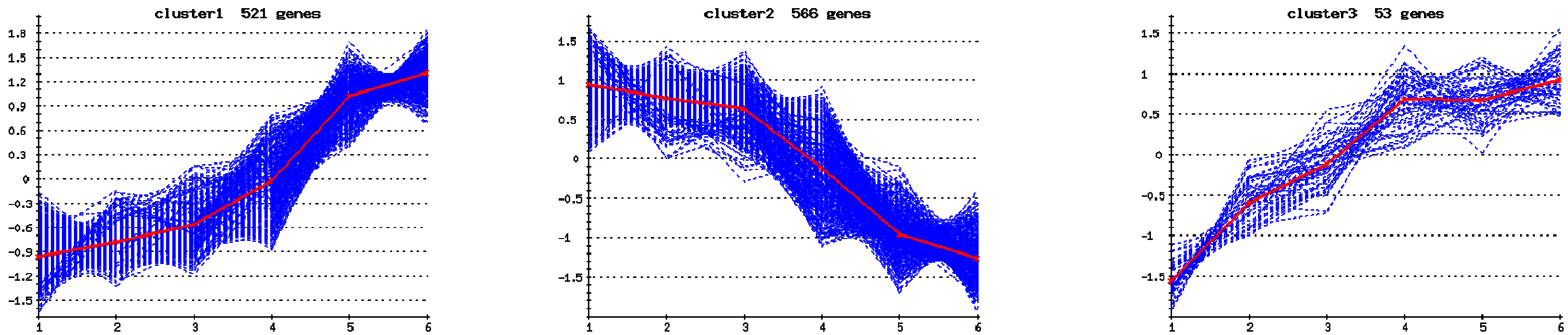
Applications of Clustering

- Data Mining
  - Technology watch
    - Derwent Database, contains all patents filed in the last 10 years worldwide
    - Searching by keywords leads to thousands of documents
    - Find clusters in the database and find if there are any emerging technologies and what competition is up to
  - Marketing
    - Customer database
    - Find clusters of customers and tailor marketing schemes to them

Applications of Clustering

- gene expression profile clustering
  - similar expressions , expect similar function

U18675 4CL -0.151 -0.207 0.126 0.359 0.208 0.091 -0.083 -0.209
M84697 a-TUB 0.188 0.030 0.111 0.094 -0.009 -0.173 -0.119 -0.136
M95595 ACC2 0.000 0.041 0.000 0.000 0.000 0.000 0.000 0.000
X66719 ACO1 0.058 0.155 0.082 0.284 0.240 0.065 -0.159 -0.010
U41998 ACT 0.096 -0.019 0.070 0.137 0.089 0.038 0.096 -0.070
AF057044 ACX1 0.268 0.403 0.679 0.785 0.565 0.260 0.203 0.252
AF057043 ACX2 0.415 0.000 -0.053 0.114 0.296 0.242 0.090 0.230
U40856 AIG1 0.096 -0.106 -0.027 -0.026 -0.005 -0.052 0.054 0.006
U40857 AIG2 0.311 0.140 0.257 0.261 0.158 0.056 -0.049 0.058
AF123253 AIM1 -0.040 0.002 -0.202 -0.040 0.077 0.081 0.088 0.224
X92510 AOS 0.473 0.560 0.914 0.625 0.375 0.387 0.019 0.141



From:De Smet F., Mathys J., Marchal K., Thijs G., De Moor B. & Moreau Y. 2002.
Adaptive Quality-based clustering of gene expression profiles, Bioinformatics, **18**(6), 735-746.

Applications of Clustering

- Profiling Web Users
  - Use web access logs to generate a feature vector for each user
  - Cluster users based on their feature vectors
  - Identify common goals for users
    - Shopping
    - Job Seekers
    - Product Seekers
    - Tutorials Seekers
  - Can use clustering results to improving web content and design

The k-means clustering algorithm

1. Initialize cluster centroids $\mu_1, \ldots, \mu_k \in \Re^n$ randomly.
2. Repeat until convergence: {

   For every i, set

$$c_i = \arg\min_j \left\| x_i - \mu_j \right\|^2$$

   For each j, set

$$\mu_i = \frac{\sum_{i=1}^{m} 1\{c_i = j\} x_i}{\sum_{i=1}^{m} 1\{c_i = j\}}$$

}

K-means, comments.

- k – the number of clusters

  a parameter of the algorithm.

- μ_i cluster centroids

  represent our current guesses for the positions of the centers of the clusters

- Initialization: pick k random training samples.

  Other initialization methods are also possible.

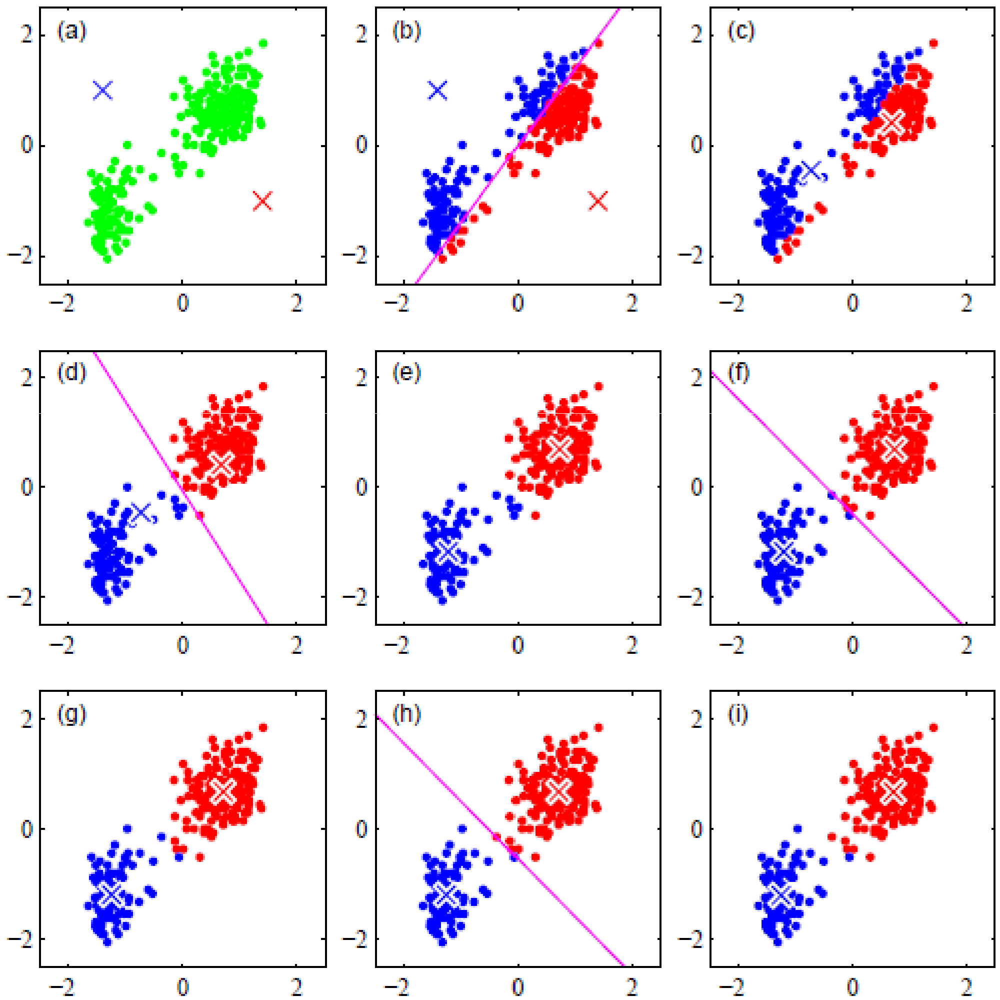
K-means, intuition

- The inner-loop of the algorithm repeatedly carries out two steps:

  (i) “Assigning” each training example x_i to the closest cluster centroid μ_j.

  (ii) Moving each cluster centroid μ_j to the mean of the points assigned to it.

K-means, example



Coordinate Descent

- Minimize a multivariate function F(x) by minimizing it along one direction at a time.
  - Choose search directions from the coordinate directions.
  - Minimizes the F(x) along one coordinate direction at a time, iterating through the list of search directions cyclically.
- Given x^k, the *i*th coordinate of x^{k+1} is given by

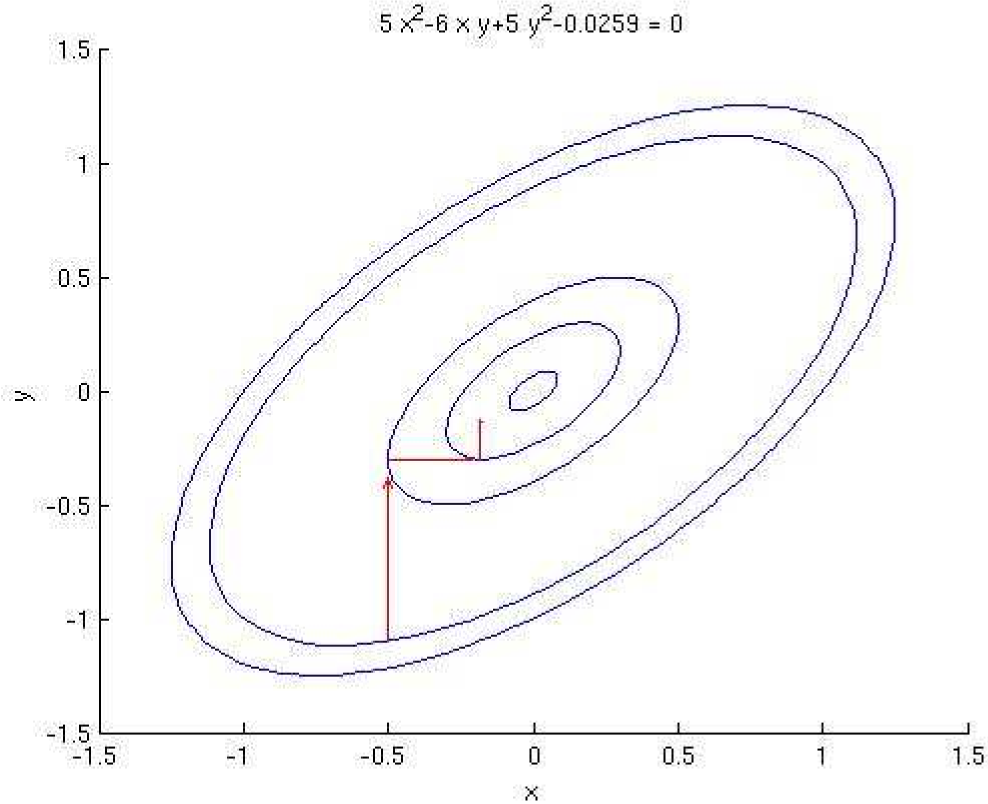
$$x_i^{k+1} = \arg\min_{y \in \mathbf{R}} f\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, y, x_{i+1}^k, \ldots, x_n^k\right);$$

  - Thus one begins with an initial guess x^0 for a local minimum of *F*, and get a sequence $x^0, x^1, x^2, \ldots$ iteratively.
  - By doing line search in each iteration, we automatically have

$$F(x^0) \geq F(x^1) \geq F(x^2), \ldots$$

  - It can be shown that this sequence has similar convergence properties as steepest descent.

Coordinate Descent Example



K-means, convergence

- Define objective function:

$$J(c,\mu) = \sum_{i=1}^{m} \left\| x_i - \mu_{c_i} \right\|^2$$

- k-means is exactly coordinate descent on J.

Inner-loop of k-means repeatedly

- minimizes J with respect to c while holding μ fixed
- minimizes J with respect to μ while holding c fixed.

Thus J must monotonically decrease => value of J must converge.