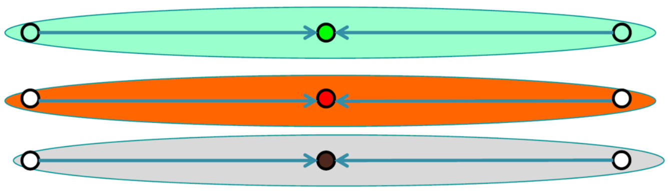
Clustering

K-means

- partition $(x_1, x_2, \ldots, x_n), x_i \in \mathbb{R}^d$ into $(S_1, S_2, \ldots, S_k)$ k classes

$$\arg\min_S \sum_{i=0}^{k} \sum_{k \in S_i} ||x - \mu_i||^2 \tag{1}$$

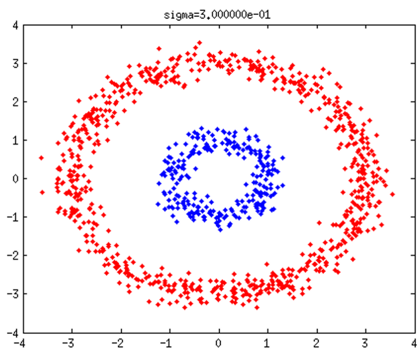
- In general , NP-hard
- Heuristic: Lloyd's algorithm
  - Randomized initialization
  - Recursively, assign points to the closest center and recompute the center
  - A local optima is shown as follows:



  - Terminates since decreasing average distance each iteration

Spectral Graph Clustering

- ℓ_p not the best in some scenarios
- In general, we shall define similarities between points



- **Intra-group** edges have **large** weights
- **Inter-group** edges have **small** weights

Different types of graphs

- ϵ-neighborhood $w_{ij} = 1$ iff ϵ-close
- KNN graph
- fully connected with w_{ij} self defined

Graph Laplacian

- D is the diagonal matrix $D = \text{diag}(d_1, d_2, \ldots, d_n)$.
- A is the adjacent matrix
- **Graph Laplacian:** $L = D - A$.

Theorem Let G be an undirected graph with nonnegative weights.

- $\#$ zero eigenvalues of L = $\#$ connected components in G
- L is symmetric and also positive semidefinite
- Free to assume: $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$
-