Clustering

The K-means clustering algorithm

Flat clustering

- Cluster objects/genes/samples into K clusters
- In the following, we will consider clustering **genes** based on their expression profiles
- K: number of clusters, a user defined argument
- Two example algorithms
  - *K*-means
  - Gaussian mixture model-based clustering

Notation for K-means clustering

- K number of clusters
- N_k Number of elements in cluster k
- x_i p-dimensional expression profile for i^{th} gene
- $X = \{x_1, \cdots, x_N\}$ is the collection of N gene expression profiles to cluster
- f_k Center of the k^{th} cluster
- $C(i)$ Cluster assignment (1 to K) of i^{th} gene

K-means clustering

- Hard-clustering algorithm
- Dissimilarity measure is the Euclidean distance
- Minimizes *within-cluster scatter* defined as

$$\sum_{k=1}^{K} \sum_{i:C(i)=k} \left\| x_i - f_k \right\|^2$$

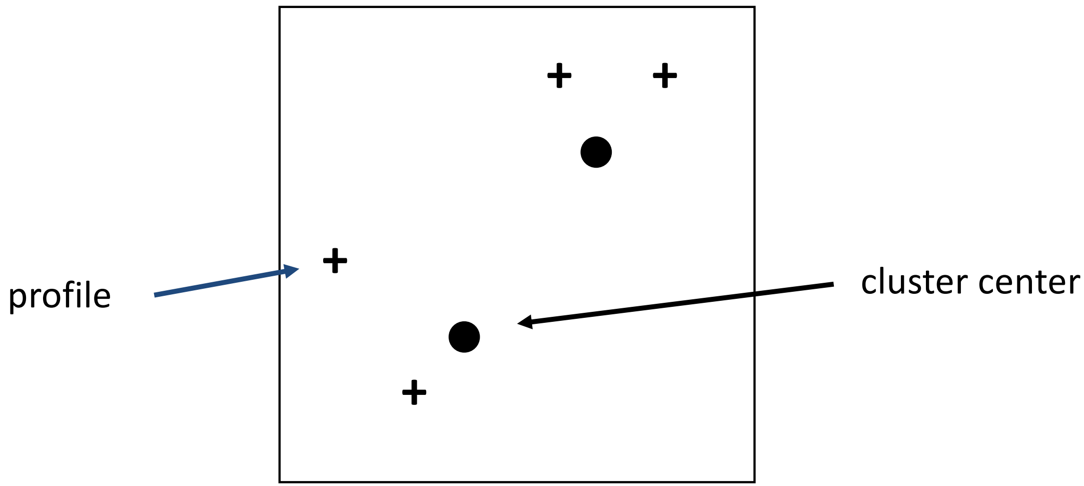
Sum over all genes assigned to cluster k

Center of cluster k

- This minimization is an NP-hard problem in general
- The K-means algorithm is an efficient heuristic

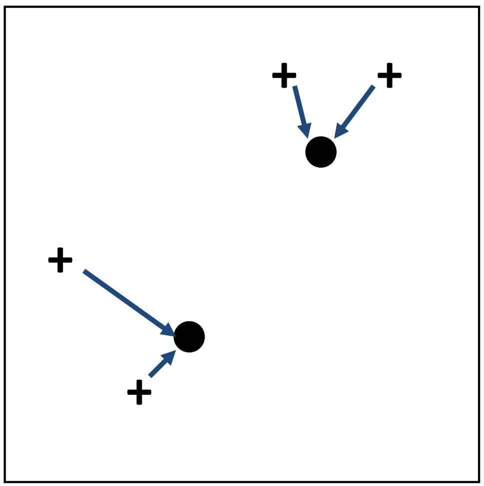
K-means clustering

- consider an example in which our vectors have 2 dimensions

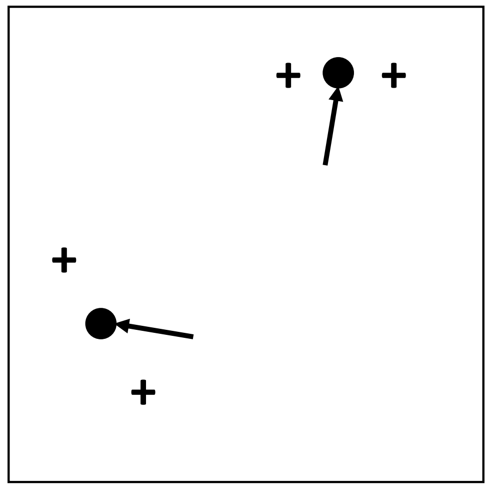


K-means clustering

- each iteration involves two steps
  - assignment of profiles to clusters
  - re-computation of the cluster centers (means)



assignment



re-computation of cluster centers

K-means algorithm

- Input: K, number of clusters, a set $X=\{x_1,.. x_N\}$ of data points, where x_i are p-dimensional vectors
- Initialize
  - Select initial cluster means $f_1, \dots, f_K$
- Repeat until convergence
  - Assign each x_i to cluster $C(i)$ such that

$$C(i) = \text{argmin}_{1 \leq k \leq K} ||x_i - f_k||^2$$

  - Re-estimate the mean of each cluster based on new members

K-means: updating the mean

- To compute the mean of the k^{th} cluster

$$f_k = \frac{1}{N_k} \sum_{i:C(i)=k} x_i$$

Number of genes in cluster k

All genes in cluster k

$$f_{kj} = \frac{1}{N_k} \sum_{i:C(i)=k} x_{ij}$$

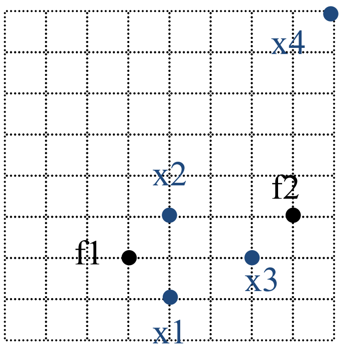
K-means stopping criteria

- Assignment of objects to clusters don’t change
- Fix the max number of iterations
- Optimization criterion changes by a small value

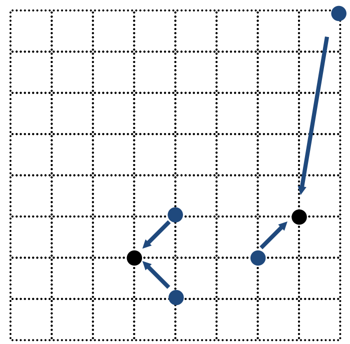
K-means Clustering Example

Given the following 4 instances and 2 clusters initialized as shown.

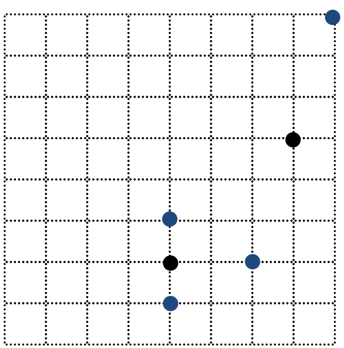
$$\text{dist}(x_i, x_j)^2 = \left\| x_i - x_j \right\|^2$$



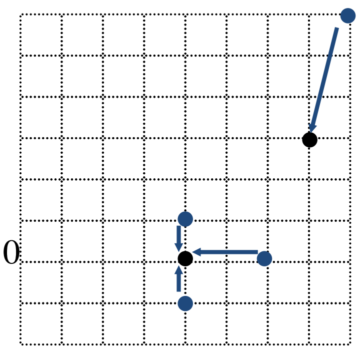
$$dist(x_1, f_1)^2 = 2, \quad dist(x_1, f_2)^2 = 13$$
$$dist(x_2, f_1)^2 = 2, \quad dist(x_2, f_2)^2 = 9$$
$$dist(x_3, f_1)^2 = 9, \quad dist(x_3, f_2)^2 = 2$$
$$dist(x_4, f_1)^2 = 61, \quad dist(x_4, f_2)^2 = 26$$



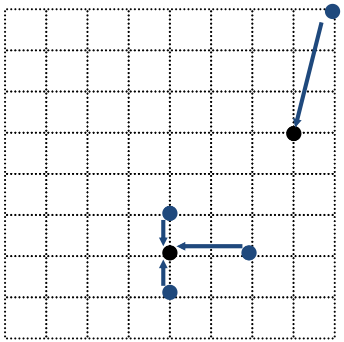
$$f_1 = \left(\frac{4+4}{2}, \frac{1+3}{2}\right) = (4,2)$$
$$f_2 = \left(\frac{6+8}{2}, \frac{2+8}{2}\right) = (7,5)$$



$$dist(x_1, f_1)^2 = 1, \quad dist(x_1, f_2)^2 = 25$$
$$dist(x_2, f_1)^2 = 1, \quad dist(x_2, f_2)^2 = 13$$
$$dist(x_3, f_1)^2 = 4, \quad dist(x_3, f_2)^2 = 10$$
$$dist(x_4, f_1)^2 = 52, \quad dist(x_4, f_2)^2 = 10$$

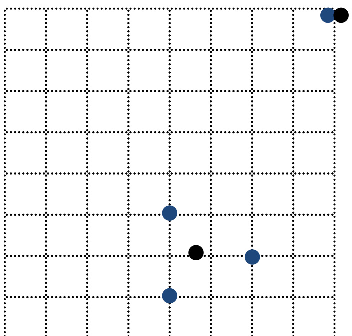


K-means Clustering Example (Continued)



$$f_1 = \left(\frac{4+4+6}{3}, \frac{1+3+2}{3}\right) = (4.67, 2)$$

$$f_2 = \left(\frac{8}{1}, \frac{8}{1}\right) = (8, 8)$$



assignments remain the same,
so the procedure has converged

Summary

- K-means is a simple flat clustering method
- Heuristic – not guaranteed to find optimal clustering
- Iterative method alternating between
  - Assigning profiles to closest cluster centers
  - Updating location of cluster centers
- Sensitive to initial cluster centers