Clustering

Patrice Koehl
Department of Biological Sciences
National University of Singapore

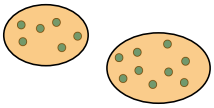
http://www.cs.ucdavis.edu/~koehl/Teaching/BL5229
koehl@cs.ucdavis.edu

Clustering is a hard problem

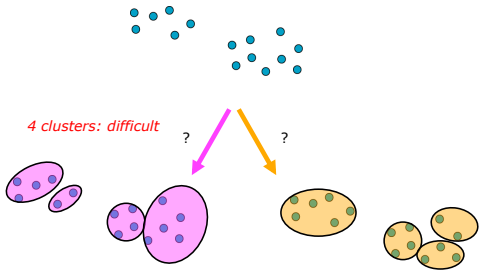
Many possibilities; What is best clustering ?

Clustering is a hard problem

2 clusters: easy

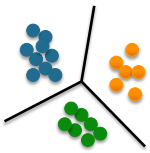


Clustering is a hard problem



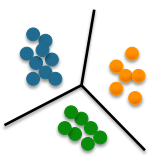
Many possibilities; What is best clustering ?

Clustering



- Hierarchical clustering
- K-means clustering
- How many clusters?

Clustering

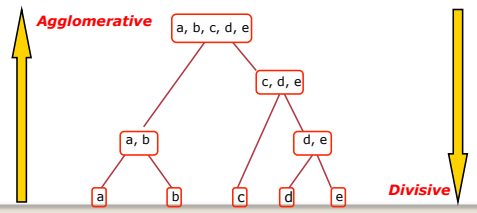


- Hierarchical clustering
- K-means clustering
- How many clusters?

Hierarchical Clustering

To cluster a set of data D={P_1, P_2, …,P_N}, hierarchical clustering proceeds through a series of partitions that runs from a single cluster containing all data points, to N clusters, each containing 1 data points.

Two forms of hierarchical clustering:



Agglomerative hierarchical clustering techniques

- *Starts with N independent clusters: {P_1}, {P_2}, …,{P_N}*
- *Find the two closest (most similar) clusters, and join them*
- *Repeat step 2 until all points belong to the same cluster*

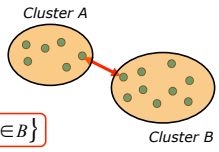
Methods differ in their definition of inter-cluster distance (or similarity)

Agglomerative hierarchical clustering techniques

1) Single linkage clustering

Distance between closest pairs of points:

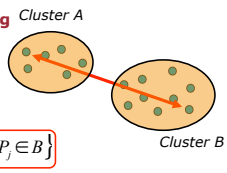
$$d(A,B) = \min\left\{d(P_i,P_j), P_i \in A, P_j \in B\right\}$$



2) Complete linkage clustering

Distance between farthest pairs of points:

$$d(A,B) = \max\left\{d(P_i,P_j), P_i \in A, P_j \in B\right\}$$

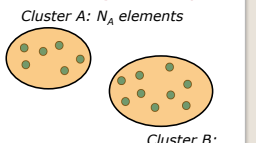


Agglomerative hierarchical clustering techniques

3) Average linkage clustering

Mean distance of all mixed pairs of points:

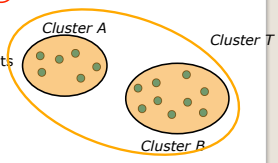
$$d(A,B) = \frac{\sum_{P_i \in A} \sum_{P_j \in B} d(P_i, P_j)}{N_A N_B}$$



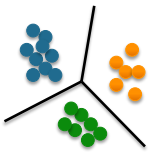
4) Average group linkage clustering

Mean distance of all pairs of points

$$d(A,B) = \frac{\sum_{P_i \in T} \sum_{P_j \in T} d(P_i, P_j)}{N_T^2}$$



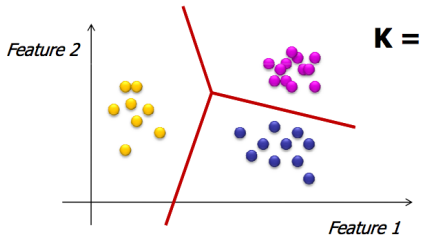
Clustering



- Hierarchical clustering
- K-means clustering
- How many clusters?

K-means clustering

The k-means algorithm partitions the data into k mutually exclusive clusters

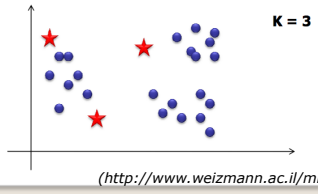


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters, K
- Randomly choose initial positions of K centroids

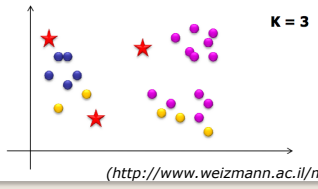


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)

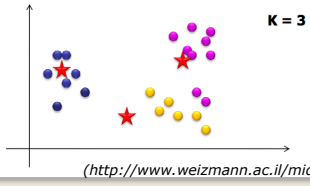


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)
- Re-compute centroid positions
- If solution converges → Stop!

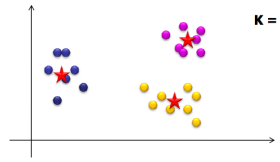


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

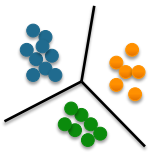
Algorithm description

- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)
- Re-compute centroid positions
- **If solution converges → Stop!**



(http://www.weizmann.ac.il/midrasha/courses/)

Clustering



- Hierarchical clustering
- K-means clustering
- How many clusters?

Cluster validation

Clustering is hard: it is an unsupervised learning technique. Once a Clustering has been obtained, it is important to assess its validity!

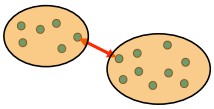
The questions to answer:

- Did we choose the right number of clusters?
- Are the clusters compact?
- Are the clusters well separated?

To answer these questions, we need a quantitative measure of the cluster sizes:

- intra-cluster size
- Inter-cluster distances

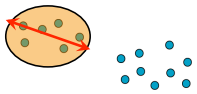
Inter cluster size



Several options:

- Single linkage
- Complete linkage
- Average linkage
- Average group linkage

Intra cluster size



Several options:

For a cluster S, with N members and center C:

-Complete diameter: $\Delta(S) = \max_{(x,y)\in S^2}\left(d(x,y)\right)$

-Average diameter: $\Delta(S) = \frac{1}{N(N-1)}\sum_{\substack{(x,y)\in S \\ x\neq y}} d(x,y)$

-Centroid diameter: $\Delta(S) = \frac{2}{N}\sum_{x\in S} d(x,C)$

Cluster Quality

For a clustering with K clusters:

1) Dunn's index

$$D = \min_{1\le i\le K}\left(\min_{\substack{1\le j\le K \\ j\neq i}}\left\{\frac{\delta(S_i,S_j)}{\max_{1\le k\le K}\left(\Delta(S_k)\right)}\right\}\right)$$

Large values of D correspond to good clusters

2) Davies-Bouldin's index

$$DB = \frac{1}{K}\max_{i\neq j}\left(\frac{\Delta(S_i)+\Delta(S_i)}{\delta(S_i,S_j)}\right)$$

Low values of DB correspond to good clusters

Cluster Quality: Silhouette index

Define a quality index for each point in the original dataset:

- For the ith object, calculate its average distance to all other objects in its cluster. Call this value ai.
- For the ith object and any cluster not containing the object, calculate the object's average distance to all the objects in the given cluster.
  Find the minimum such value with respect to all clusters;
  call this value bi.
- For the ith object, the silhouette coefficient is

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

Cluster Quality: Silhouette index

Note that:

$$-1 \le s(i) \le 1$$

- s(i) = 1, i is likely to be well classified
- s(i) = -1, i is likely to be incorrectly classified
- s(i) = 0, indifferent

Cluster Quality: Silhouette index

Cluster silhouette index:

$$S(X_i) = \frac{1}{N}\sum_{j=1}^{N} s(j)$$

Global silhouette index:

$$GS = \frac{1}{K}\sum_{i=1}^{N} S(X_i)$$

Large values of GS correspond to good clusters