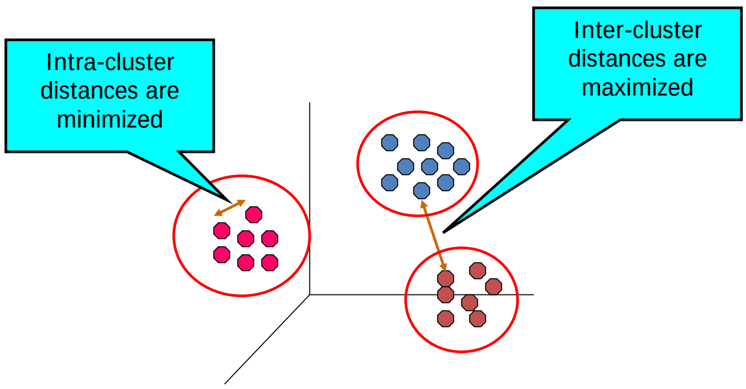
UNIVERSITAT DE BARCELONA

No hierarchical clustering

Prof. Miquel Salicrú
Prof. Sergi Civit

What is clustering?

- A **grouping** of data objects such that the objects **within a group are similar** (or related) to one another **and different from (or unrelated to) the objects in other groups**



Partitioning algorithms

Partitioning algorithms: basic concept

- Construct a partition of a set of ***n*** objects into a set of ***k*** clusters
  - Each object belongs to **exactly one** cluster
  - The number of clusters **k is given in advance**

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters (C_1, C_2, …., C_k) s.t., min sum of squared distance

$$TESS = \Sigma_{i=1}^{k} \Sigma_{j=1}^{|C_i|} d^2(\omega_{ij} - c(i))$$

where c(i) is a centroid or a medoid (representative object) of the group i

Heuristic methods

k-means (MacQueen 1967): Each cluster is represented by the center of the cluster

k-medoids or PAM (Partition around medoids) (Kaufman and Rousseeuw 1987): Each cluster is represented by one of the objects in the cluster

Possible combinations:

$$C(n,k) = \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j} j^n \approx \frac{k^n}{k!}$$

n=25, k=8: C(25, 8) ~ $6.9 \cdot 10^{17}$

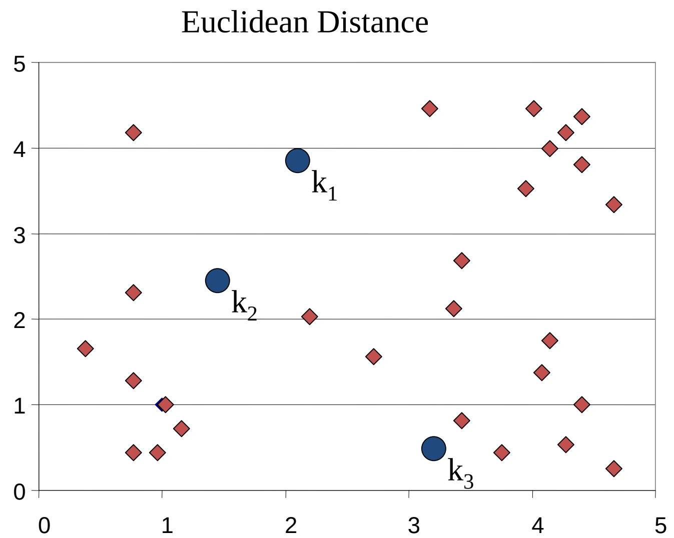
“Time efficiency” problem: Computational resources used by the algorithm

The *K-Means* Clustering Algorithm

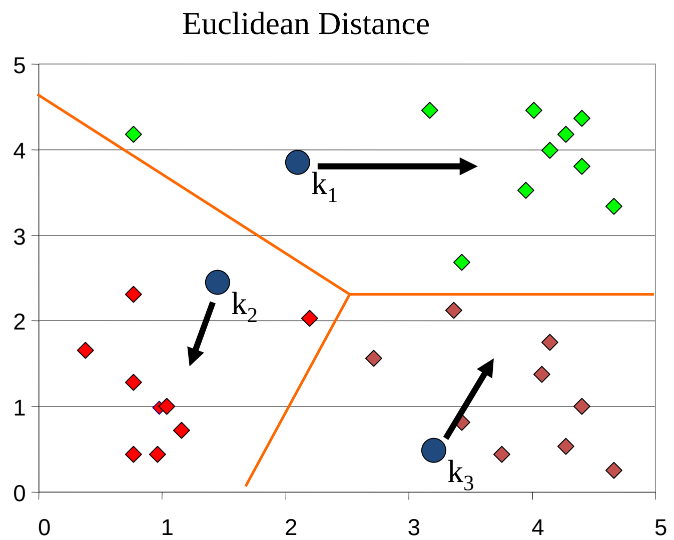
Basic algorithm:

- Decide on a value for k.
- Initialize the k cluster centers (randomly, if necessary).
- Decide the class memberships of the N objects by assigning them to the nearest cluster center (arithmetic average).
- Re-estimate the k cluster centers, by assuming the memberships found above are correct.
- If none of the N objects changed membership in the last iteration, exit. Otherwise go to 3.

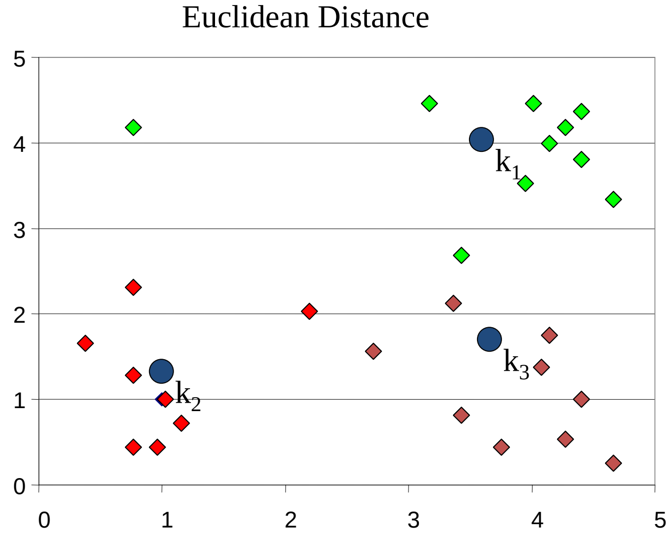
K-means Clustering: Step 1



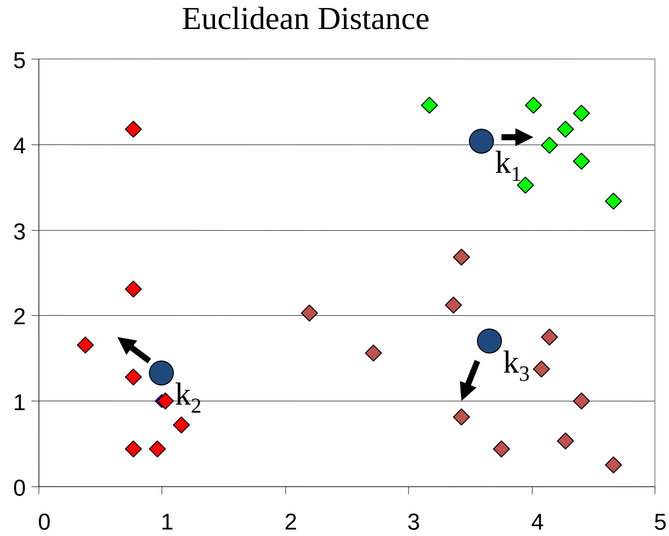
K-means Clustering: Step 2



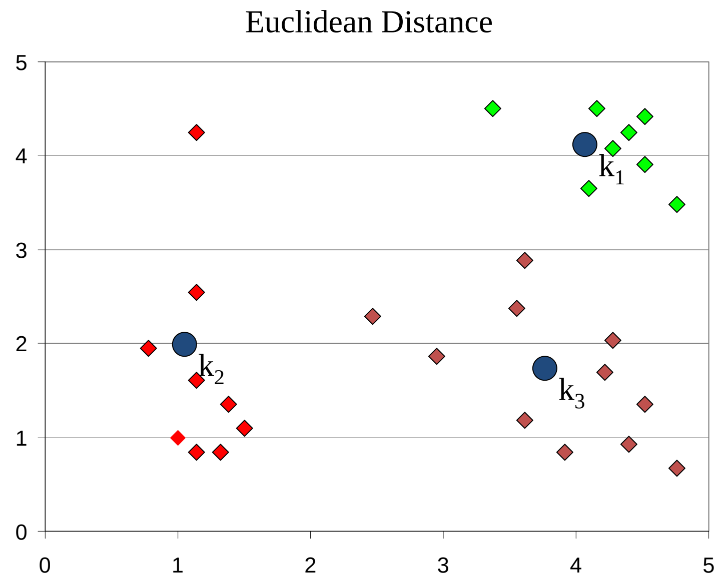
K-means Clustering: Step 3



K-means Clustering: Step 4



K-means Clustering: Step 5



Cluster Validity Indices 1/3

TESS (total error sum of squares).

- Evaluate the within-group variability
- Decrease when increase the number of clusters (TESS=0 with n clusters)
- Criterion: maximize the gradient (variation or drift)

$$\Delta(k-1,k) = \frac{TESS(k-1) - TESS(k)}{TESS(k-1)}$$

Cluster Validity Indices 2/3

Pseudo-F statistics (Calinski-Harabasz, 1974)

- Ratio of the mean sum of squares between groups to the mean sum of squares within group
- Increase according to homogeneity of the groups
- Criterion: maximize F(k)

$$F(k) = \frac{\Sigma_{i=1}^{k} d^2(c(i) - \bar{c})/(k-1)}{\Sigma_{i=1}^{k}\Sigma_{j=1}^{|C_j|} d^2(\omega_{ij} - c(i))/(n-c)}$$

Cluster Validity Indices 3/3

Silhouettes (Rousseuw, 1987)

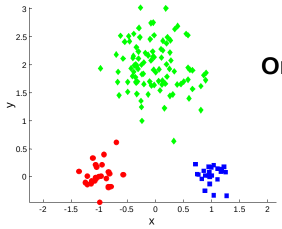
- Based on interdistances differences from the ω_i to cluster/partition (C) and to "nearest neighbor" cluster/partition.
- -1<s(ω_i)<+1: s(ω_i)<0 would be more appropriate if it was clustered in its neighbouring cluster, s(ω_i)=0 between partition, and s(ω_i)>0 means it is well matched.
- Criterion: maximize silhouette statistic

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(\omega_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{b(\omega_i) - a(\omega_i)}{\max\{a(\omega_i), b(\omega_i)\}}$$

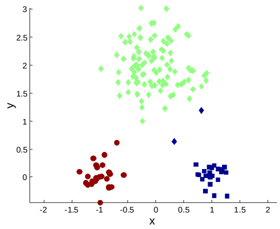
$$a(\omega_i / \omega_i \in C_i) = \frac{1}{|C_i|}\sum_{\omega_j \in C_i} d^2(\omega_i, \omega_j)$$

$$b(\omega_i / \omega_i \in C_i) = \min_{s \neq i}\{\frac{1}{|C_s|}\sum_{\omega_s \in C_s} d^2(\omega_i, \omega_s)\}$$

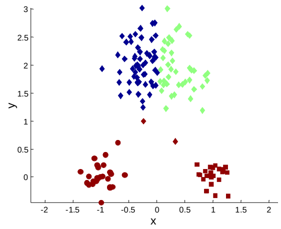
Two different K-means Clusterings



Original Points



Optimal Clustering



Sub-optimal Clustering

Euclidean distance Metric: Relationship

$(P_1, P_2,\ldots, P_n)$ be the set (objects, populations) to classify and $D=(d_{ij})$, a distance
$X= (X_1,\ldots, X_r, X_{r+1},\ldots, X_s)$, displays the *coordinates of the objects in PCoA space*

Populations	Coordinates					
	X_1	…	X_r	X_{r+1}	…	X_s
P_1	x_{11}	…	x_{1r}	$i{\cdot}x_{1r+1}$	…	$i{\cdot}x_{1s}$
P_2	x_{21}	…	x_{2r}	$i{\cdot}x_{2r+1}$	…	$i{\cdot}x_{2s}$
⋮	⋮		⋮	⋮		⋮
P_n	x_{n1}	…	x_{nr}	$i{\cdot}x_{nr+1}$	…	$i{\cdot}x_{ns}$

satisfy the condition (Torgerson, 1952):

$$\mathrm{dist}^2(P_i,P_j)=\sum_{h=1}^{r}(x_{ih}-x_{jh})^2-\sum_{h=r+1}^{s}(x_{ih}-x_{jh})^2=d_{ij}^2$$

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means (random points, the most distant points (from each other),....)
  - Optimization strategy algorithm
  - Objective function (functional, distance, medoid)
  - Robust statistics
  - Handling categorical data: *k-modes* (Huang'98)
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling
- Focusing + spatial data structure (Ester et al., 1995)

PAM (Basic algorithm)

- Decide on a value for k.
- Select ***k*** representative objects arbitrarily (medoids c(1),…, c(k))
- Assign each non-selected object to the most similar representative object
- In each group, compute (Manhattan distance)

$$AD(c(i)) = \frac{1}{|G_i|}\sum_{\omega_{ij}\in G_i} d^{Manh}(\omega_{ij}, c(i))$$

- In each group, swap the medoid (medoids: c'(1),…, c'(k))
- For each pair of c(i) and c'(i),
  - If AD(c'(i))-AD(c(i)) < 0, c(i) is replaced by c'(i)
  - Then assign each non-selected object to the most similar representative object
- repeat steps 5-6 until there is no change

What Is the Problem with PAM?

- Pam is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean
- Pam works efficiently for small data sets but does not **scale well** for large data sets.
  - $O(k(n-k)^2)$ for each iteration

    where n is # of data, k is # of clusters

➔ Sampling based method,

CLARA(Clustering LARge Applications)

CLARA (Clustering Large Applications)

- *CLARA* (Kaufmann and Rousseeuw in 1990)
  - Built in statistical analysis packages, such as S+
- It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- Strength: deals with larger data sets than *PAM*
- Weakness:
  - Efficiency depends on the sample size
  - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased

Fuzzy algorithms

Basic concept

- Construct a partition of a set of ***n*** objects into a set of ***k*** clusters
  - Each object belongs to **one or more** cluster
  - The number of clusters **k is given in advance**

- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters (C_1, C_2, …., C_k) s.t., min sum of squared distance

$$\text{FTESS} = \Sigma_{i=1}^{k} \Sigma_{j=1}^{|C_j|} d^2(\omega_{ij} - c(i)) \cdot u_{ij}^{m}$$

where c(i) is a centroid/medoid of the group i, u_{ij} is the fuzzy membership of object ω_{ij} to the fuzzy set C_i, and $1<m<+\infty$, is a fuzziness exponent which determines the incidence of fuzzy values on the computations

Statistical problem

Minimize the function

$$\text{FTESS} = \Sigma_{i=1}^{k} \Sigma_{j=1}^{|C_j|} d^2(\omega_{ij}, c(i)) \cdot u_{ij}^{m}$$

conditions: $0 \le u_{ij} \le 1 , \quad \sum_{j=1}^{s} u_{ij} = 1 , \quad \sum_{i=1}^{k} u_{ij} > 0$

Solution $u_{ij} = \dfrac{1}{\sum_{s=1}^{k} \left[\dfrac{e_{ij}}{e_{is}} \right]^{2/(m-1)}}$ where $e_{ij} = d(\omega_{ij}, c(j))$

and $c(i) = \dfrac{\sum_j u_{ij}^{m} \cdot x_{ij}}{\sum_j u_{ij}^{m}}$

GEOGRAPHIC REPRESENTATION