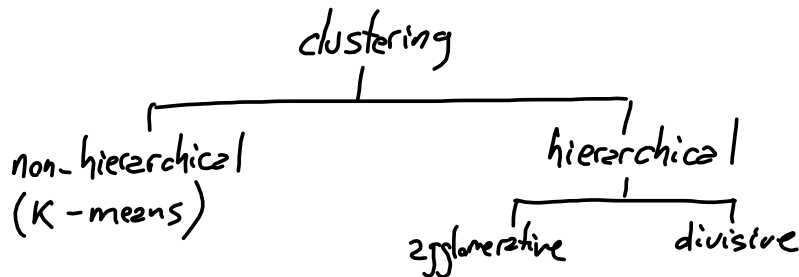


D matrix of dissimilarities

$$D = \begin{pmatrix} 0 & d(1,2) \\ d(2,1) & \ddots \\ & \ddots & 0 \end{pmatrix}$$

$$D = \frac{D + D^T}{2}$$



GOAL: separate observations in K groups based on their (dis)similarities

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^n \sum_{i'=1}^n d(i, i') &= \frac{1}{2} \sum_{k=1}^K \sum_{c(i)=k} \left(\sum_{c(i')=k} d(i, i') + \sum_{c(i') \neq k} d(i, i') \right) \\ &= \underbrace{\frac{1}{2} \sum_{k=1}^K \sum_{c(i)=k} \sum_{c(i')=k} d(i, i')}_{D_{\text{within}}(c)} + \underbrace{\frac{1}{2} \sum_{k=1}^K \sum_{c(i)=k} \sum_{c(i') \neq k} d(i, i')}_{D_{\text{between}}(c)} \end{aligned}$$

$W(c)$ is the dissimilarity WITHIN the groups

$B(c)$ is the dissimilarity BETWEEN the groups

OBJECTIVE: either minimize $W(c)$ or maximize $B(c)$

Ideally, we would try all possible combinations (allocating n observations in K clusters) and select the best. It is infeasible due to the high number of possibilities:

$$s(n, K) = \frac{1}{K!} \sum_{k=1}^K (-1)^{K-k} \binom{K}{k} k^n$$

$$s(10, 4) = 34105$$

$$s(15, 4) \approx 10^{10}$$

We need a more clever algorithm, and the most famous is K -means ^{arguably}

K -means

- it has been developed for quantitative variables, and it is based on the Euclidean distance. Therefore

$$W(c) = \sum_{k=1}^K \sum_{c(i)=k} \sum_{c(i')=k} \|x_i - x_{i'}\|^2 = 2 \sum_{k=1}^K n_k \sum_{c(i)=k} \|x_i - m_k\|^2$$

where n_k is the number of observations in the group k , and m_k is the mean vector of the observations in group k

K -means iteratively assigns observations to the group corresponding to the closest mean, and update the mean

- 1) Choose K and initial values (arbitrarily chosen) for $m_k, k=1, \dots, K$
- 2) Repeat:
 - a) for $i=1, \dots, n$ assign x_i to the group k , s.t. $k = \arg \min_c \|x_i - m_c\|^2$
 - b) for $k=1, \dots, K$ compute m_k as the mean of the observation belonging to the k -th group
 until the results stabilize

Step 2a assures us that $W(c)$ is minimum given m_k

Step 2b assures us that $W(c)$ is minimum given the groups (observation assigned)

in this example
 • often the result does not depend on the initial placing on m_k . But there is not guarantee that we find the global minimum.

- since k -means is based on the Euclidean distance, it tends to find spherical structures

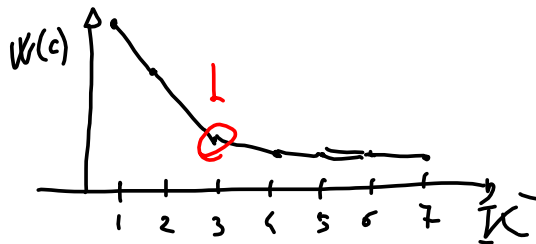
↑
 the result strongly depend on the choice of the Euclidean distance

- strong limitation: it strongly depends on the choice of K (also m_k , but less)
- only applies to quantitative variables

↳ to solve this issue, one can use K -medoids (centroids). The procedure can be based on different procedures, e.g. GAP-statistics

Choice of K

- increasing means decreasing $W(c)$ both in the training and test set → no cross-validation
- idea: at certain point, the decrease in $W(c)$ by adding +1 to K is not 'large enough' to justify a new group

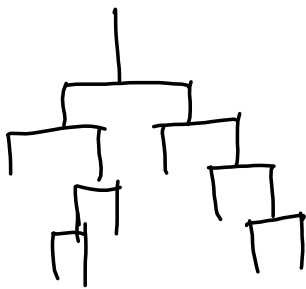


elbow rule

GAP statistic: we compare the $\frac{\log(W(c))}{W(c)}$ with what we would obtain in the case of uniformly distributed observations and choose K with the largest GAP

Hierarchical clustering

- one of the main limitations of k-means is the choice of K
- alternative approach:



← dendrogram (tree-fashion representation of the clustering algorithm)

← in the leaves there are the observations

- agglomerative clustering: start from n clusters (each observation is in its own cluster) and aggregate those that are less dissimilar step-by-step, until we have a single cluster with all observations
- divisive clustering: start with all observations in a single cluster, and separate step-by-step those that are the most dissimilar, until we get n clusters (one for each observation)

To perform hierarchical clustering we need to define the dissimilarity between groups (clusters). When $K=n$, obviously we use $d(i, i')$. For groups C and C'

- single link: $d_s(C, C') = \min_{i \in C, i' \in C'} d(i, i')$

- complete link: $d_c(C, C') = \max_{i \in C, i' \in C'} d(i, i')$

- average link: $d_n(C, C') = \frac{1}{n_C n_{C'}} \sum_{i \in C} \sum_{i' \in C'} d(i, i')$

Single link tends to work better with 'filiform' structures
Complete link tends to work better with spherical structures.

Choice of k

- the idea is the same of k-means: find the point when adding more clusters/groups does not decrease 'sufficiently' the $W(C)$

The dendrograms are constructed in a way that the length of the vertical lines are proportional to the differences in $W(C)$. A strategy can be to cut where the lines are longer