Statistical Learning with High-dimensional Data

Pr. Charles BOUVEYRON

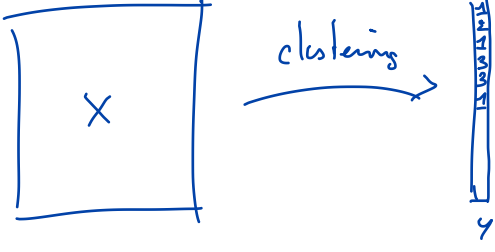
Professor of Statistics and Artificial Intelligence
Chair of the Institut 3IA Côte d'Azur
Deputy Scientific Director of the Institut 3IA Côte d'Azur
Head of the Maasai research team
Université Côte d'Azur & Inria

charles.bouveyron@inria.fr - @cbouveyron

Clustering :

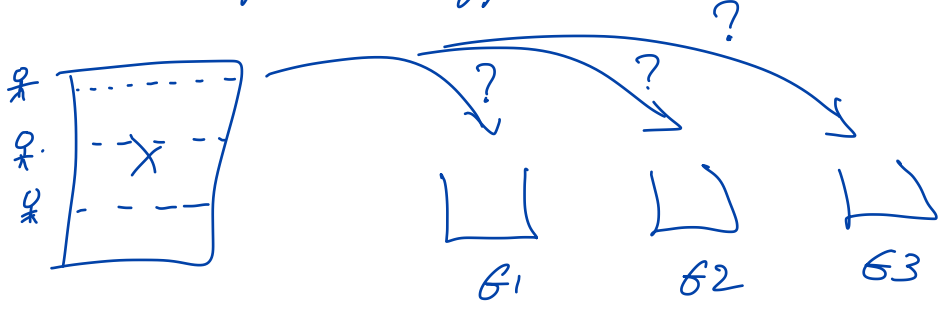
The goal of clustering is to be able to predict from X a categorical variable Y

X —learning→ Y = the group memberships of the observations



⚠ From the conceptual point of view, clustering is far more difficult than supervised classif.

Clustering is based on a simple idea of assigning individuals to groups (such that people who behave similarly are in the same group) but which have a combinatorial aspect which makes the problem difficult!



Clustering methods ?

- k means
- hierarchical clustering
- EM algorithm for GMM
- ...

The k-means algorithm:

The algo: for a fixed value k

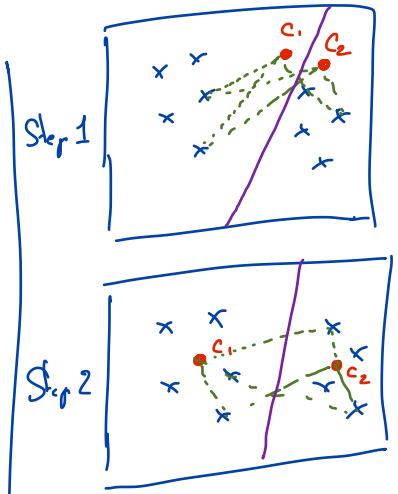
initialization: . pick k centers

loop: (i) compute the distances between all individuals and the k centers

(ii) assign each individual to the closest center and recompute the centers of each group.

stop: when the groups are stable (no more changes!)

$k = 2$



↳ stop: because the group assignements are already stable!

K-means is very simple and works very well most of the time, but requires to select k externally.

One approach to select k is to rely on the elbow method computed on the evolution of a clustering quality criterion:

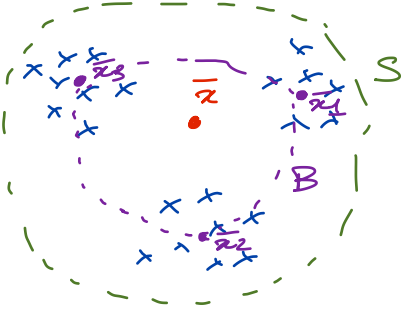
$$J(k) = \frac{B(k)}{S}$$

between variance

which has to be maximized

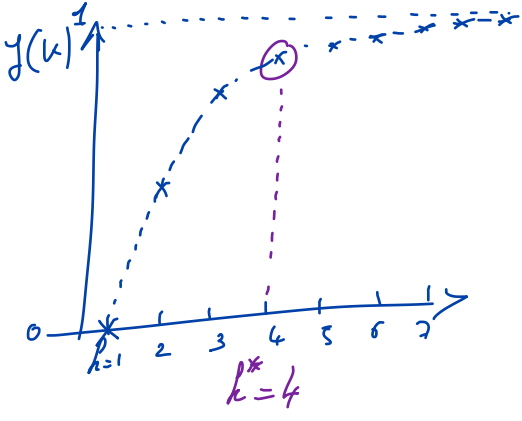
$J(k) \to 1$

variance of the data



$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^t(x_i - \bar{x}), \quad B = \frac{1}{n}\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^t(\bar{x}_k - \bar{x})$$

To select the most appropriate k, we compute $J(k)$ for all possible values of k for the k-means and we look for the elbow of the curve.



Here $k^* = 4$ realises a compromise between an efficient clustering in terms of $J(k)$ with a small number of clusters.

Summary on k-means:

⊕ simple and efficient algorithm (→ reference technique)

⊕ the centers can be easily interpreted as the average individual of each group.

⊕ in combination to the elbow method, you can select k.

⊖ it is sensitive to the initialization and to outliers.

⊖ k-means tends to produce clusters of the same size.

Hierarchical clustering:

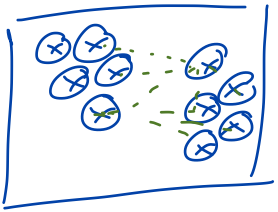
The algo:

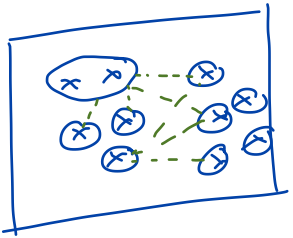
initialization: each individual is assigned to its own group.

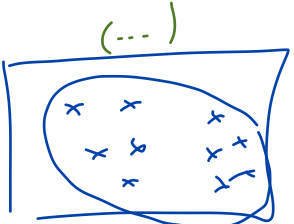
loop: (i) compute the distances between all groups

(ii) gather the two groups that are the closest ones

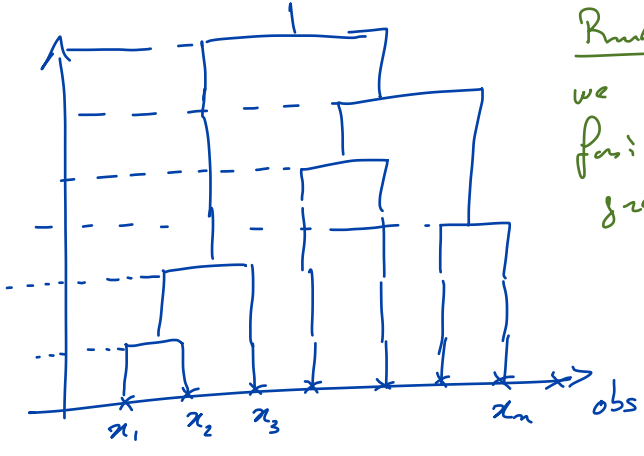
stop: when it remains only one group.







HC is generating a hierarchy of clustering with $k \in [1, n]$ groups. Interestingly, the hierarchy can be easily visualized thanks to a dendrogram:



Rmk: at each level, we have a unique fusion between two groups.

HC is in fact a family of algorithms, because we have several possibilities for the distance between groups.

- the **complete linkage**:

$$d(A,B) = \max \{ d(a,b) \,/\, a \in A, b \in B \}$$

- the **single linkage**:

$$d(A,B) = \min \{ d(a,b) \,/\, a \in A, b \in B \}$$

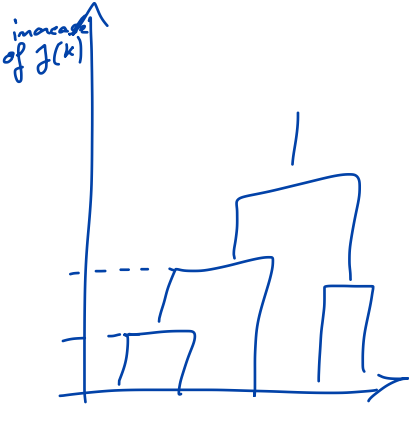
- the **centroid dist**:

$$d(A,B) = d(\bar{a}, \bar{b})$$

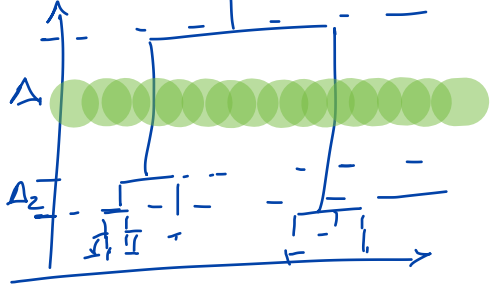
- the **Ward dist**:

$$d(A,B) = \frac{d(\bar{a}, \bar{b})}{\frac{1}{n_A} + \frac{1}{n_B}}$$

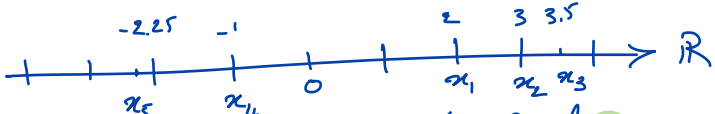
in HC, the elbow technique can also be applied to select k. Interestingly, it has been incorporated in most dendrograms produced by softwares like R



we pick the clustering with the largest gap



Exercise :

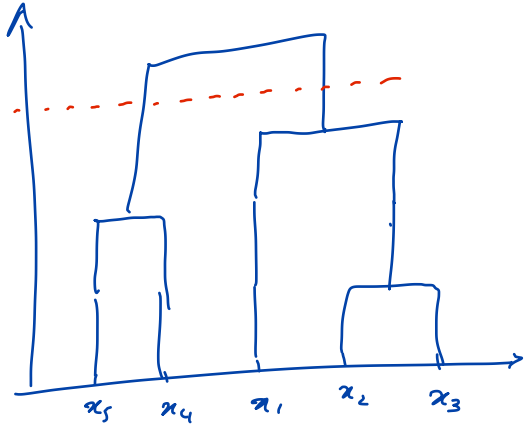


↳ apply HC on these data with the complete linkage

	x_2	x_3	x_4	x_5
x_1	1	1.5	3	4.25
x_2		0.5	4	5.25
x_3			4.5	5.75
x_4				1.25

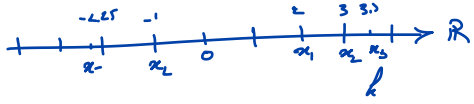
	x_{23}	x_4	x_5
x_1	1.5	3	4.25
x_{23}		4.5	5.75
x_4			1.25

	x_{23}	x_{45}
x_1	1.5	4.25
x_{23}		5.75



$$d(x_1, x_{23}) = \max\{d(x_1, x_2), d(x_1, x_3)\}$$
$$= \max\{1, 1.5\} = 1.5$$

Let's now apply k-means on the same data ($k = 2$) with x_1 and x_2 as initial centers.

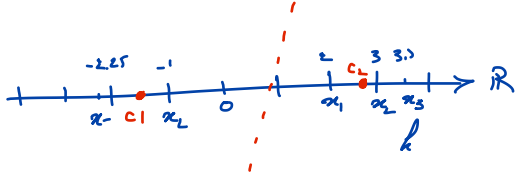


	x_1	x_2	x_3	x_4	x_5
$c_1 = x_1$	0	1	1.5	3	4.25
$c_2 = x_2$	1	0	0.5	4	5.25

↳ $c_1 = \text{mean}(x_1, x_4, x_5) = -0.41$

$c_2 = \text{mean}(x_2, x_3) = 3.25$

	x_1	x_2	x_3	x_4	x_5
c_1	2.41	3.41	3.91	0.39	1.64
c_2	1.25	0.25	0.25	4.25	5.5



↳ $c_1 = \text{mean}(x_4, x_5)$

$c_2 = \text{mean}(x_1, x_2, x_3)$

Summary on hierachical clustering:

⊕ the dendrogram is a very appreciated tool for visualizing the clustering hierarchy

⊕ a flexible tool thanks to the different group distances

⊕ the ability to choose K thanks to the elbow method directly within the dendrogram.

⊖ high algorithmic complexity ($\mathcal{O}(n^2 \log n)$)
↳ not possible to use it for $n > 10\,000$

DANGER some people use kmean as a pretreatment before applying hclust on the centers.