Introduction to statistical learning

2.2 Unsupervised learning: Clustering

V. Lefieux

June 2018

Table of contents

Table of contents

Introduction

- Grouping a set of n data objects (with p variables) into clusters.
- A cluster is a set of elements (individuals):
  - near (similar) to one another in the same cluster,
  - far (dissimilar) to the elements in other clusters.
- Clustering is an **unsupervised** classification

Exemples

- Customers.
- Climatic zones.
- Genes.
- Pictures.
- Textual informations.
- Web pages.

Many partitions of a set I

The total number of partitions of an *n*-element set is rapidly increasing with *n*.
For example, for 4 elements A,B,C et D, there are 15 possible partitions:

ABCD	ABC D	ABD C	ACD B	BCD A
AB CD	AC BD	AD BC	AB C D	AC B D
AD B C	BC A D	BD A C	CD A B	A B C D

Many partitions of a set II

The total number of partitions of an n-element set is given by the Bell number:

$$B_n = \frac{1}{e} \sum_{k=0}^{+\infty} \frac{k^n}{k!} \ .$$

So:

n	4	6	10
p_n	15	203	115 975

The number of partitions of an n-element set into exactly K nonempty parts is asymptotically equivalent to $\frac{K^n}{K!}$.

Aim

In general one consider thousands of individuals, it's impossible to compare all the partitions (with a criterium).

This is why have been developed algorithmic methods to solve this problem.

Methods

- Partitioning-based algorithms: build and evaluate partitions with a criterion.
- Hierarchical-based algorithms: build a hierarchical decomposition.
- Model-based algorithms: a probability distribution is assumed for each of the clusters (e.g Expectation Maximization EM using gaussian mixture models).
- Grid-based algorithms: based on a multiple-level granularity structure
- Density-based algorithms: based on density functions (e.g DBSCAN: Density-Based Spatial Clustering of Applications with Noise).

Partitioning-based and hierarchical-based algorithms

- For both methods a dissimilarity metric (or similarity) between individuals is needed.
- In partitioning-based algorithms, a partitioning criterion is needed.
- In hierarchical-based algorithms, an agglomerative criterion (or divisive criterion) is needed.

Table of contents

Data

p variables measured on n individuals.

The data set that is represented in terms of an $n \times p$ table:

$$\mathbb{X} = \left(x_i^j\right)_{i \in \{1,\dots,n\}, j \in \{1,\dots,p\}} ,$$

where the n rows are the individuals and the p columns are the variables.

x_i^j: value of X^j measured on individual i.

Dataset matrix

		Variables				
		1	...	j	...	p
Individuals	1	x_1^1	...	x_1^j	...	x_1^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	x_i^1	...	x_i^j	...	x_i^p
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	n	x_n^1	...	x_n^j	...	x_n^p

Individuals and variables

Commonly individual i refers to:

$$X_i = \left(x_i^1, \ldots, x_i^p\right)^\top$$

and variable j to:

$$X^j = \left(x_1^j, \ldots, x_n^j\right)^\top .$$

Weights

The sample should be representative: a miniature of the population it comes from. If not, one assign to each individual i a weight ω_i (e.g from a survey design):

- $\forall i \in \{1, \dots, n\} : \omega_i > 0$,
- $\sum_{i=1}^{n} \omega_i = 1$.

One consider the matrix:

$$W = \operatorname{diag}(\omega_1, \dots, \omega_n) \ .$$

Usually weights are uniform:

$$\forall i \in \{1, \dots, n\} : \omega_i = \frac{1}{n} \ ,$$

that is:

$$W = \frac{1}{n} \mathrm{I}_n \ .$$

Barycenter

The barycenter of the data set is:

$$G = \mathbb{X}^\top W \mathbf{1}_n = \sum_{i=1}^{n} \omega_i X_i$$

where $\mathbf{1}_n$ is a n dimensional all ones vector.

Metrics

One consider a metric d defined on $\Omega \times \Omega$ where Ω is the set of individuals.

One usually consider properties for 3 individuals $(i_1, i_2, i_3) \in \Omega^3$:

1. $d(i_1, i_2) \geq 0$.
2. $d(i_1, i_2) = d(i_2, i_1)$.
3. $d(i_1, i_2) = 0 \quad \Rightarrow \quad i_1 = i_2$.
4. $d(i_1, i_2) \leq d(i_1, i_3) + d(i_3, i_2)$.
5. $d(i_1, i_2) \leq \max\left(d(i_1, i_3), d(i_3, i_2)\right)$.

Vocabulary

Dissimilarity	(1)	(2)	(3)		
Distance	(1)	(2)	(3)	(4)	
Ultrametric distance	(1)	(2)	(3)		(5)

Similarity

A similarity is a function s defined on $\Omega \times \Omega$ such for 2 individuals (i_1, i_2):

1. $s(i_1, i_2) \geq 0$.
2. $s(i_1, i_2) = s(i_2, i_1)$.
3. $s(i_1, i_2) \leq s(i_1, i_1)$.
4. $s(i_1, i_1) = s(i_2, i_2) := s_{max}$.

Based on a similarity s, it's possible to build a dissimilarity d. For example:

$$d(i_1, i_2) = s_{max} - s(i_1, i_2) .$$

Metric choice

- Many metrics can be found in literature.
- The choice of a metric depends essentially on the nature of the variables.

Classic metrics: quantitative case I

- Minkoswski distance:

$$d\left(i_1, i_2\right) = \left(\sum_{j=1}^{p} \left|x_{i_1}^{j} - x_{i_2}^{j}\right|^{q}\right)^{\frac{1}{q}} .$$

- Manhattan distance (Minkoswski, $q = 1$):

$$d\left(i_1, i_2\right) = \sum_{j=1}^{p} \left|x_{i_1}^{j} - x_{i_2}^{j}\right| .$$

- Chebychev distance:

$$d\left(i_1, i_2\right) = \max_{j \in \{1,\dots,p\}} \left|x_{i_1}^{j} - x_{i_2}^{j}\right| .$$

Classic metrics: quantitative case II

Euclidean distance (Minkoswski, $q = 2$) is one of the most used in practice:

$$d\left(i_1, i_2\right) = \left[\sum_{j=1}^{p} \left(x_{i_1}^{j} - x_{i_2}^{j}\right)^2\right]^{\frac{1}{2}} \ .$$

One can also write:

$$d^2\left(i_1, i_2\right) = \left(X_{i_1} - X_{i_2}\right)^\top \left(X_{i_1} - X_{i_2}\right) \ .$$

Classic metrics: quantitative case III

We consider M euclidean distance:

$$d_M^2(i_1, i_2) = (X_{i_1} - X_{i_2})^\top M (X_{i_1} - X_{i_2})$$

where M is a non negative matrix.

Distance d_M is induced by the norm:

$$\forall x \in \mathbb{R}^p : \|x\|_M^2 = x^\top M x \ .$$

The inner product is:

$$\forall (x, y) \in \mathbb{R}^p \times \mathbb{R}^p : \langle x, y \rangle_M = x^\top M y \ .$$

Classic metrics: quantitative case IV

Variables are usually normalized:

- Sebestyen distance:

$$d^2(i_1, i_2) = (X_{i_1} - X_{i_2})^\top D (X_{i_1} - X_{i_2})$$

  where D is a positive diagonal matrix, for example:

$$D = D_{\frac{1}{s^2}} := \operatorname{diag}\left(\frac{1}{s_1^2}, \ldots, \frac{1}{s_p^2}\right) .$$

- Mahalanobis distance:

$$d^2(i_1, i_2) = (X_{i_1} - X_{i_2})^\top S^{-1} (X_{i_1} - X_{i_2})$$

  where S is the variance-covariance matrix.

Classic metrics: binary case I

In the case of binary variables, with values in $\{0, 1\}$, consider for 2 individuals i_1 et i_2:

- a: number of attributes that both i_1 et i_2 don't own,
- b: number of attributes that i_2 owns but not i_1,
- c: number of attributes that i_1 owns but not i_2,
- d: number of attributes that both i_1 et i_2 own.

		i_2	
		0	1
i_1	0	a	b
	1	c	d

Classic metrics: binary case II

Classic **similarities** are:

- Jaccard:

$$d\left(i_1, i_2\right) = \frac{d}{b + c + d} \; .$$

- Sokal, Sneath and Anderberg:

$$d\left(i_1, i_2\right) = \frac{d}{2(b + c) + d} \; .$$

- Czekanowski, Sørensen and Dice:

$$d\left(i_1, i_2\right) = \frac{2d}{b + c + 2d} \; .$$

- Russel et Rao:

$$d\left(i_1, i_2\right) = \frac{d}{a + b + c + d} \; .$$

- Ochiai:

$$d\left(i_1, i_2\right) = \frac{d}{\sqrt{d + b}\sqrt{d + c}} \; .$$

Table of contents

Assumptions

Consider a partition that individuals are in a partition (in a partition, individuals are all in a unique cluster) with K clusters $\{C_1, \ldots, C_K\}$.

The cluster C_k has weight:

$$\mu_k = \sum_{i \in C_k} \omega_i$$

and barycenter:

$$G_k = \frac{1}{\mu_k} \sum_{i \in C_k} \omega_i X_i \ .$$

Inertia relative to a point A

Inertia relative to a point A ($A \in \mathbb{R}^p$) is:

$$\mathcal{I}_A = \sum_{i=1}^{n} \omega_i \, d_M^2\left(i, A\right) .$$

Huygens theorem gives:

$$\mathcal{I}_A = \mathcal{I}_G + d_M^2\left(A, G\right) .$$

Total inertia

Total inertia is:

$$\mathcal{I}_{tot} = \sum_{i=1}^{n} \omega_i \, d_M^2\left(i, G\right) \, .$$

Intraclass inertia

Intraclass inertia (between inertia) is:

$$\mathcal{I}_{intra} = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} \omega_i \, d_M^2 \left(i, G_k \right) \, .$$

Interclass inertia

Interclass inertia (within inertia) is:

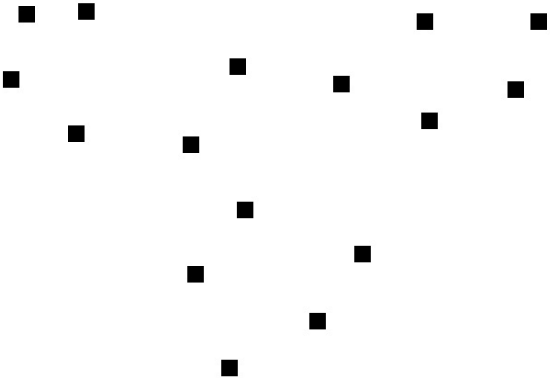
$$\mathcal{I}_{inter} = \sum_{k=1}^{K} \mu_k \, d_M^2 \left(G_k , G \right) \; .$$

Inertia decomposition

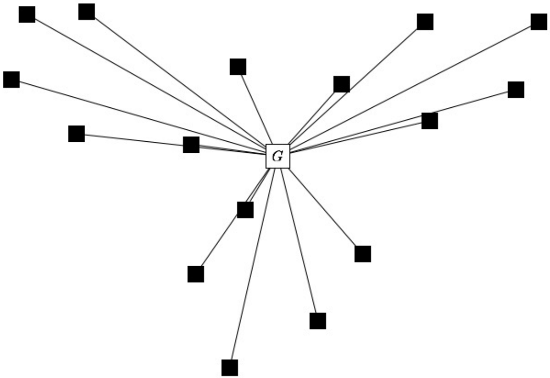
Huygens theorem gives:

$$\mathcal{I}_{tot} = \mathcal{I}_{intra} + \mathcal{I}_{inter} \quad .$$

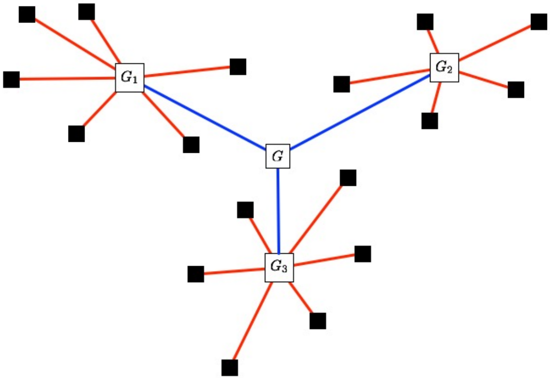
A dataset



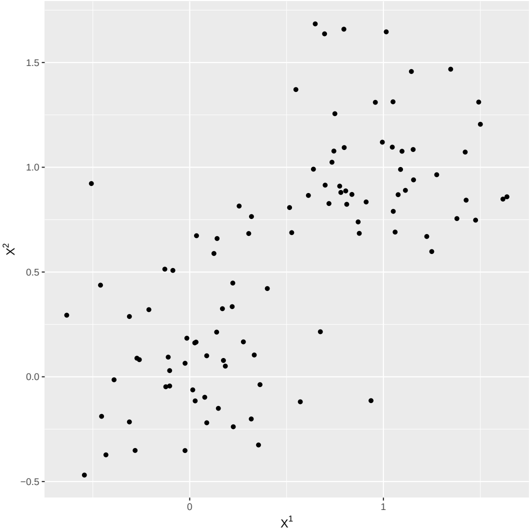
Total inertia



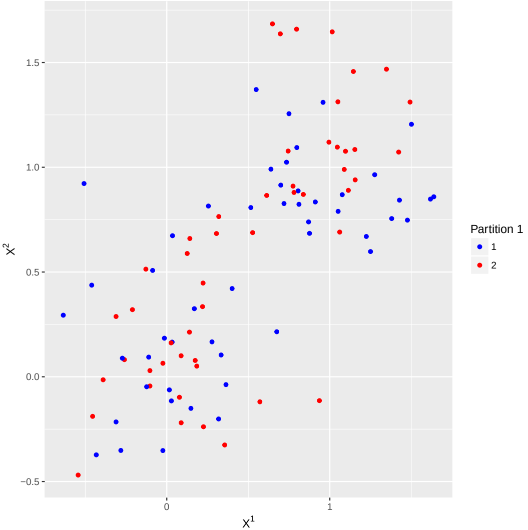
Intra-class and inter-class inertia ($K = 3$)



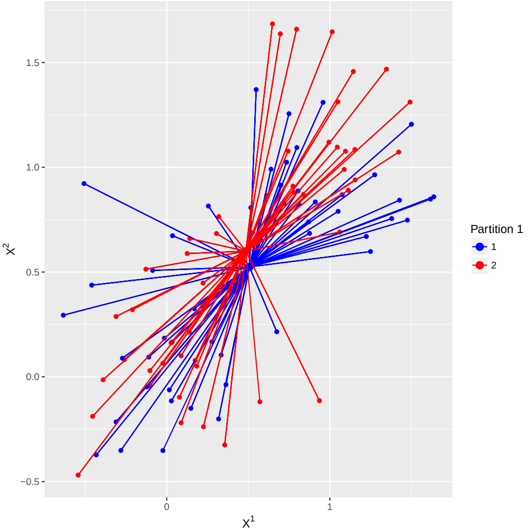
Example



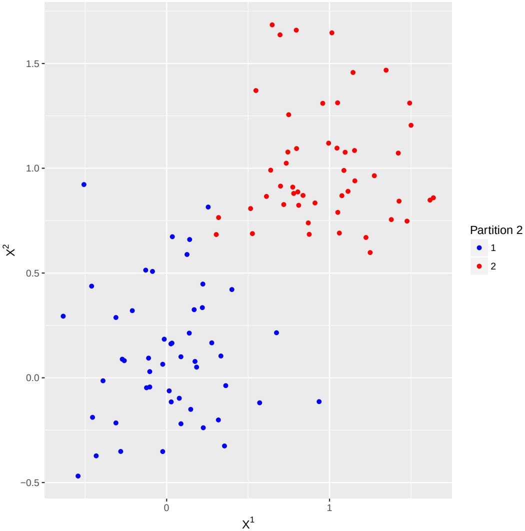
Example: partition 1 with $K = 2$ I



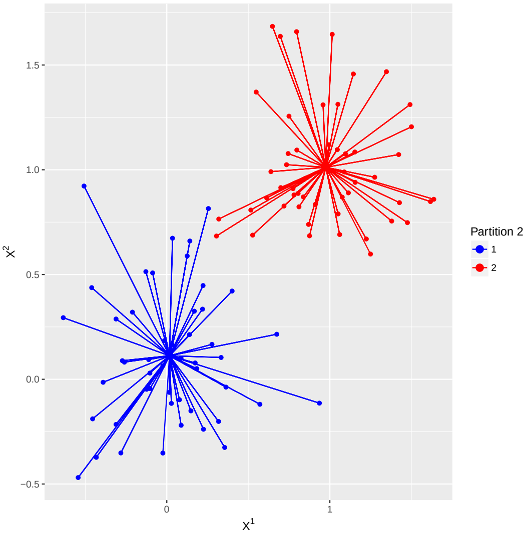
Example: partition 1 with $K = 2$ II



Example: partition 2 with $K = 2$ I



Example: partition 2 with $K = 2$ II



Clustering criterion

For K known, a good clustering minimizes intra-class inertia and maximizes inter-class inertia

The quality of a clustering can be measured by $\frac{\mathcal{I}_{inter}}{\mathcal{I}_{tot}}$.

Table of contents

Partition

$\mathcal{P} = \{C_1, \dots, C_K\}$ is a partition of Ω if:

- $\forall k \in \{1, \dots, K\} : C_k \neq \emptyset$,
- $\forall (k, k') \in \{1, \dots, K\}^2 \times \{1, \dots, K\} : C_k \cap C_{k'} = \emptyset$,
- $\bigcup_{k \in \{1,\dots,K\}} C_k = \Omega$.

Partitioning algorithms

- Partitioning algorithms principle: build a first partition and then improve it till convergence.
- K is chosen a priori.
- K-means is the most known of these methods.

K-means algorithm

To build a partition with K clusters:

1. Initialization
   Choose K seed points in the data set (a priori or random choice).
2. Points assignment
   Assign each of the n individuals to the cluster with the nearest seed point (with the chosen distance).
3. Seed points computing
   K seed points are replaced by the K barycenters of clusters of step 2.
4. Go back to step 2 and stop when there is no more new assignment.

Illustration ($K = 2$) I

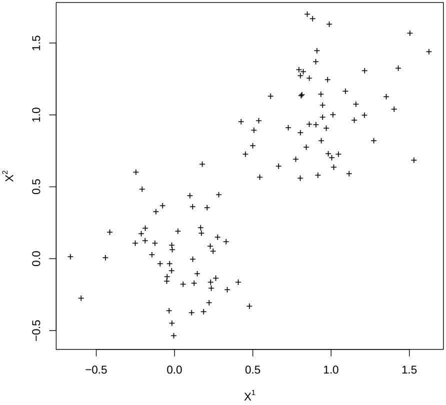


Illustration ($K = 2$) II

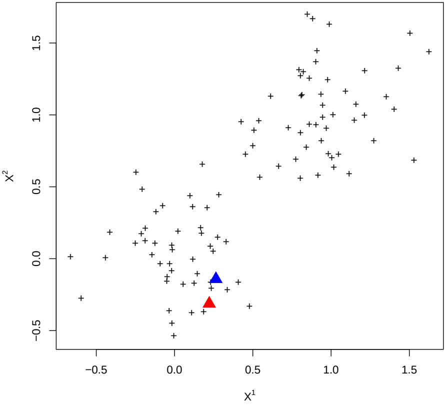


Illustration ($K = 2$) III

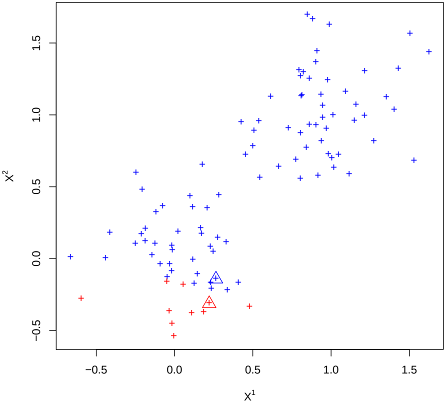


Illustration ($K = 2$) IV

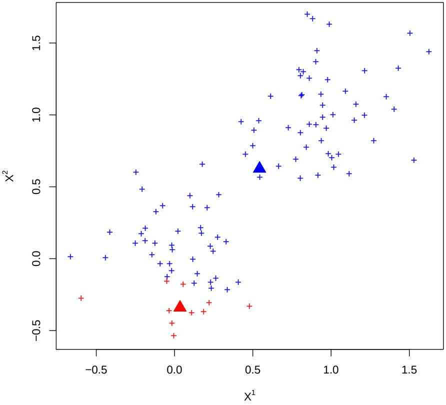


Illustration ($K = 2$) V

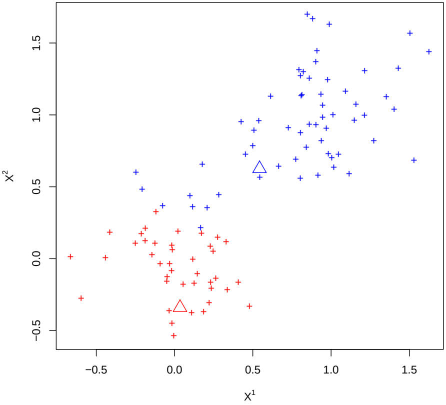


Illustration ($K = 2$) VI

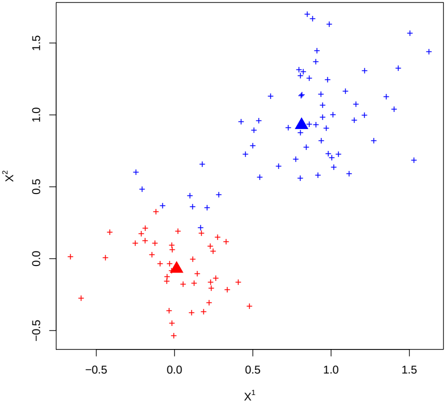
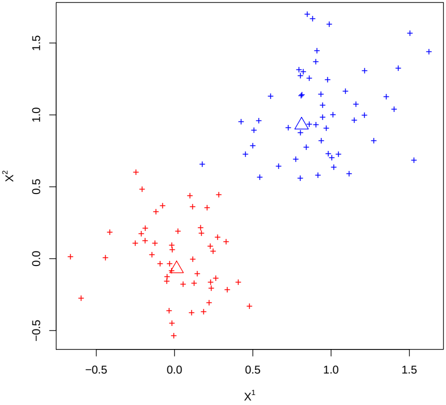


Illustration ($K = 2$) VII



Properties

- In K-means intra-class variance decreases with the number of iterations.
- K-means algorithm is sensitive to outliers.

Some alternatives

- K-medoids (Partitioning Around Medoids: PAM): barycenter is replaced by the medoid (the most centrally located point):

$$\text{medoid}\left(C_k\right) = \arg\min_{i \in C_k} \sum_{j / j \in C_k, j \neq i} \omega_i \omega_j d^2\left(j, i\right).$$

  PAM algorithm is less sensitive to outliers than K-means algorithm.

- The "real" K-means: barycenters are computed after each assignment (order of individuals isn't neutral).

Conclusion

Partitioning algorithms:

- converge fastly (towards a local minimum of intra-class inertia),
- but:
  - need to know the number of clusters K,
  - depend on seed points (another idea: compute K-means with a lot of different seed points).

Table of contents

Principle

- Agglomerative hierarchical clustering
  Initially each individual is a cluster and, iteratively, clusters are merged together.

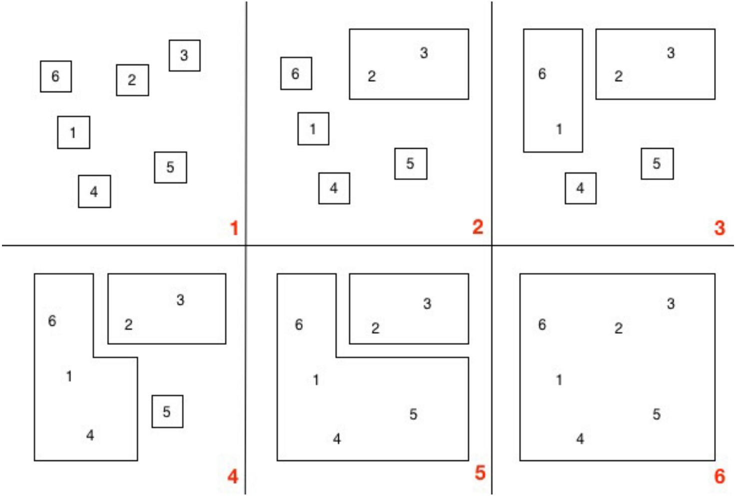
- Divisive hierarchical clustering
  Initially all individuals are in a cluster and, iteratively, clusters are divided.

Agglomerative hierarchical clustering

1. Initialization ($k = n$)
   The first partition has n clusters:
   $$\mathcal{P}_n = \{\{1\}, \ldots, \{n\}\} \ .$$
2. Agglomeration ($k \in \{n-1, \ldots, 2\}$)
   $\mathcal{P}_k$ is built by agglomerating the 2 closest clusters (in the sense of an ultrametric distance ∇) in $\mathcal{P}_{k-1}$.
3. End ($k = 1$)
   The last partition has only 1 cluster:
   $$\mathcal{P}_1 = \{1, \ldots, n\} \ .$$

We obtain a binary hierarchical tree.

Illustration



Hierarchy

$\mathcal{H}$ is a hierarchy on Ω if:

- $\Omega \in \mathcal{H}$,
- $\forall i \in \Omega : \{i\} \in \mathcal{H}$,
- $\forall (H_1, H_2) \in \Omega^2 : H_1 \cap H_2 \in \{\emptyset, H_1, H_2\}$.

Hierarchy height

Hierarchy height (level) of $\mathcal{H}$ is a function $h : \mathcal{H} \to \mathbb{R}^+$ such that:

- $\forall H \in \mathcal{H} : h(H) = 0 \Leftrightarrow H \in \{\{1\}, \ldots, \{n\}\}$,
- $\forall (H_1, H_2) \in \mathcal{H}^2 : H_1 \subset H_2 \Rightarrow h(H_1) \leq h(H_2)$.

One use dendrogram.

Linkage strategy

- The choice of the linkage strategy ∇ is sensitive.
- ∇ is an ultrametric distance.

Some linkage strategies

For 2 clusters A and B:

- Single linkage:

$$\nabla(A,B) = \min_{i \in A, j \in B} d(i,j) \ .$$

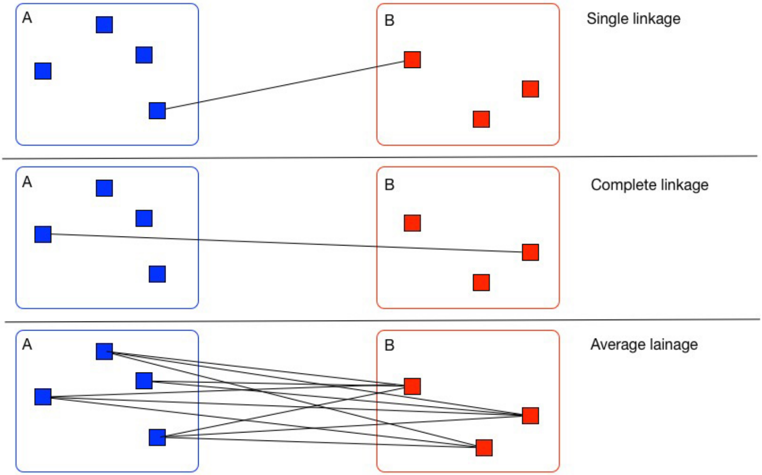
- Complete linkage:

$$\nabla(A,B) = \max_{i \in A, j \in B} d(i,j) \ .$$

- Average lainage:

$$\nabla(A,B) = \frac{1}{\mathrm{Card}(A)\,\mathrm{Card}(B)} \sum_{i \in A, j \in B} d(i,j) \ .$$

Example



Ward linkage

Ward linkage (Ward linkage):

$$\nabla(A, B) = \frac{\mu_A \, \mu_B}{\mu_A + \mu_B} d^2(G_A, G_B)$$

where:

- μ_A and μ_B are the weights of A and B,
- G_A and G_B are the barycenters of A and B,
- d is the euclidean distance.

Interpretation

If A and B are linked, the interclass inertia variation is:

$$\Delta\mathcal{I}_{inter} = \mu_A d^2\left(G_A, G\right) + \mu_B d^2\left(G_B, G\right) - \left(\mu_A + \mu_B\right) d^2\left(G_{A\cup B}, G\right).$$

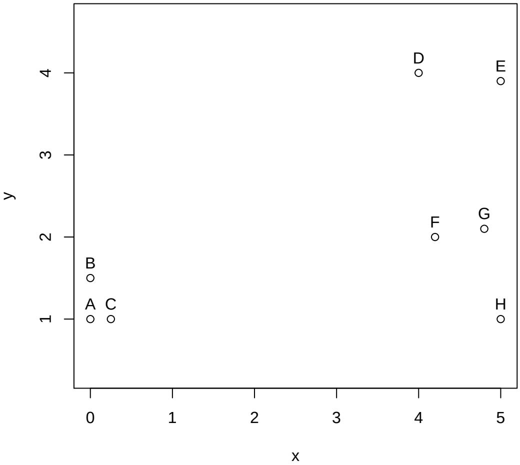
Huygens theorem gives:

$$\Delta\mathcal{I}_{inter} = \frac{\mu_A \mu_B}{\mu_A + \mu_B} d^2\left(G_A, G_B\right).$$

Dendrogram

- The dendrogram is a tree representation of the hierarchical clustering.
- Each level (height) shows the clusters considered.
- Leaves: individual clusters.
- Root: one cluster.

Exemple I



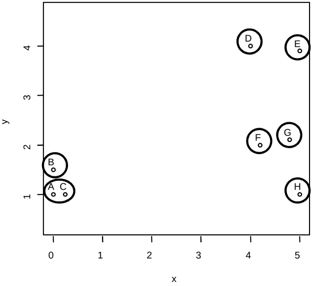
Exemple II

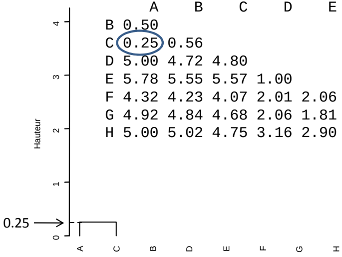
- Euclidean distance.
- Single linkage.

Distances matrix:

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

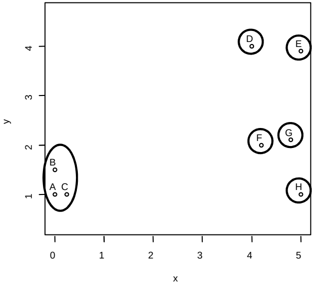
Exemple III

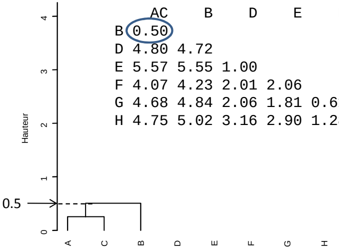




	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Exemple IV





	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

Exemple V

7e regroupement

	ABC
DEFGH	4.07

6e regroupement

	ABC	DE
DE	4.72	
FGH	4.07	1.81

5e regroupement

	ABC	DE	FG
DE	4.72		
FG	4.23	1.81	
H	4.07	2.90	1.12

4e regroupement

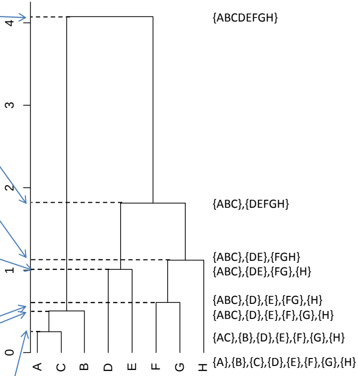
	ABC	D	E	FG
D	4.72			
E	5.55	1.00		
FG	4.07	2.01	1.81	
H	4.75	3.16	2.90	1.12

3e regroupement

	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

2e regroupement

	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12



1er regroupement

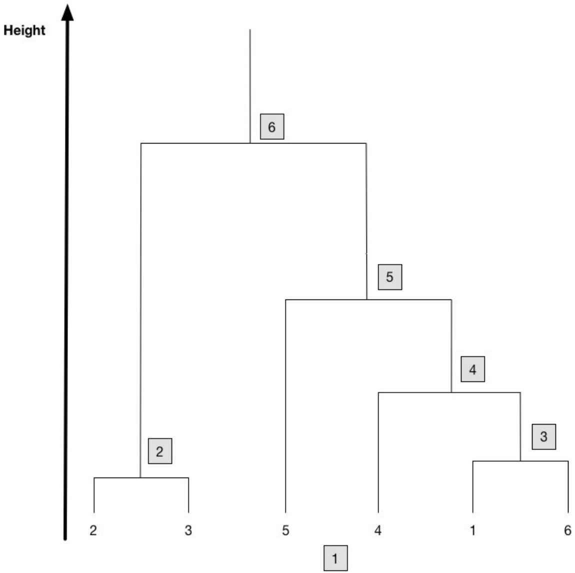
	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Choice of the number of clusters

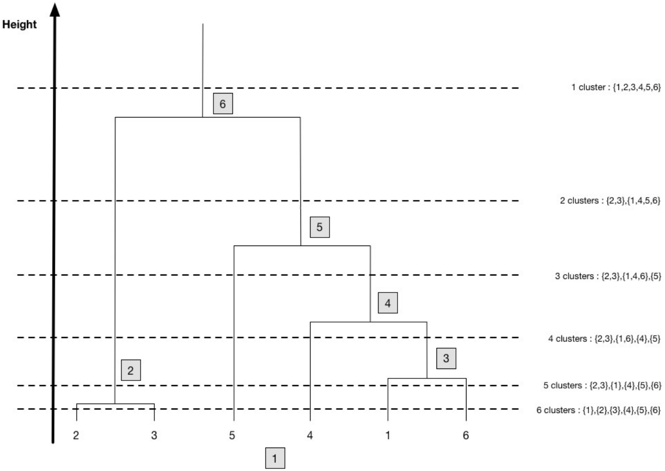
A classic method consists in finding an "elbow" in the plot of:

$$\forall k \in \{2, \ldots, n\} : \frac{\mathcal{I}_{inter}\left(C_{k+1}\right) - \mathcal{I}_{inter}\left(C_{k}\right)}{\mathcal{I}_{tot}} .$$

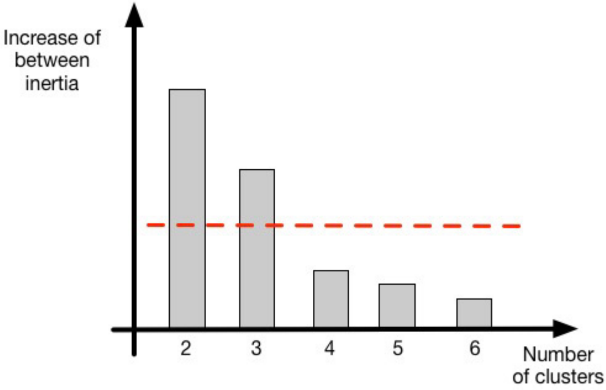
Illustration



Illustration



Illustration



Terminology

- AGNES: AGglomerative NESting.
- DIANA: DIvisive ANAlysis.

Conclusion

Hierarchical clustering algorithms:

- don't depend on seed points,
- but:
  - are computationally costly,
  - could be far from the optimum if the number of linkages $n - K$ is too high.

Table of contents

Conclusion

Interpret clusters:

- Using exploratory analysis of variables (active or supplementary) by clusters.
- Using medoids by clusters.

References

Tufféry, S. (2011). *Data mining and statistics for decision making*. Wiley series in Computational Statistics. Wiley.