REMISE EN CAUSE DES DIVISIONS DANS UN ALGORITHME DE CLASSIFICATION DESCENDANTE

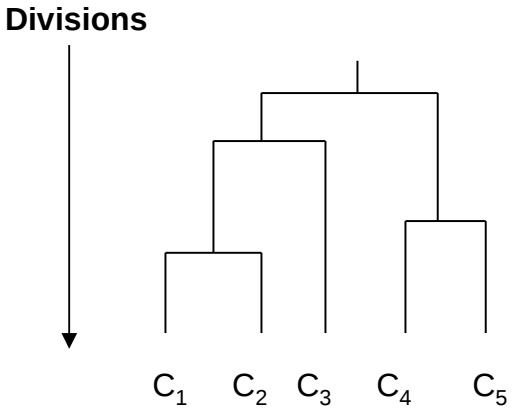
Marie CHAVENT
MAB, Université Bordeaux 1
351 cours de la libération
33405 Bordeaux cedex

Chavent@math.u-bordeaux.fr

Méthodes divisives

Classification :

Approche hiérarchique descendante

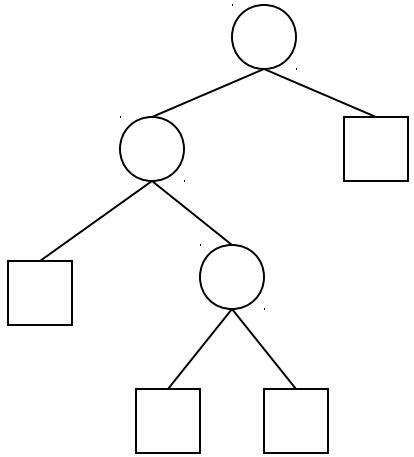


- Comment diviser une classe
- Choix de la classe à diviser

Recherche de classes homogènes et/ou bien séparées sur toutes les variables

Discrimination :

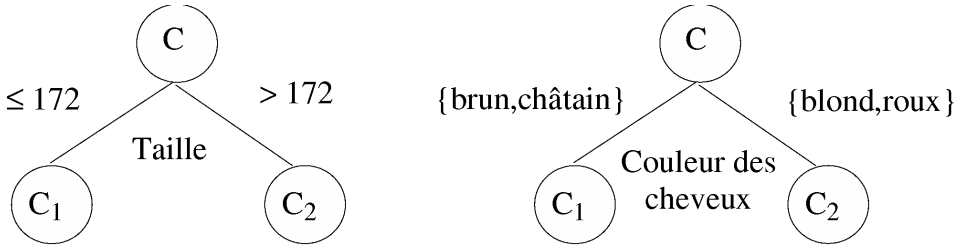
Méthodes de segmentation



Recherche de classes homogènes sur la variable à expliquer

Classification descendante hiérarchique

Division monothétique d'une classe



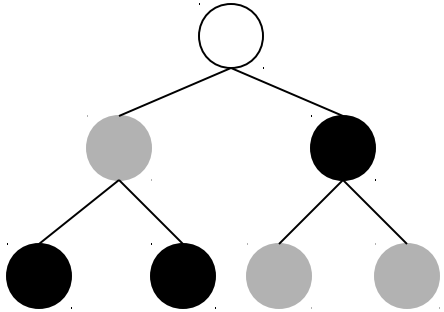
- Division en fonction d 'une variable et de sa transformation binaire
- Critère d 'évaluation des bi-partitions : inertie intra-classe

Autres approches :

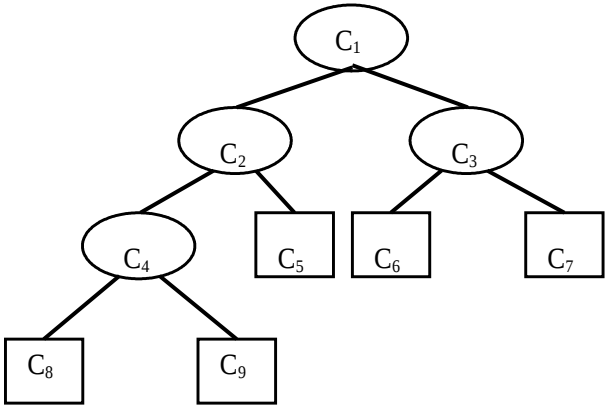
- énumération complète : Edward et Cavalli-Sforza (1965)
- heuristiques : Roux (1985), MacNaughton-Smith (1964)
- optimisation du diamètre : Guénoche, Hansen et Jaumard (1991)

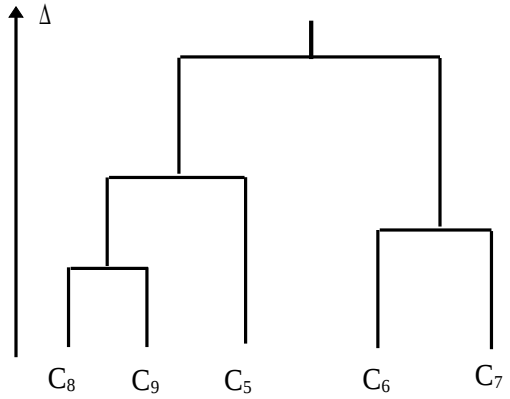
Classification descendante hiérarchique

Choix de la classe à diviser

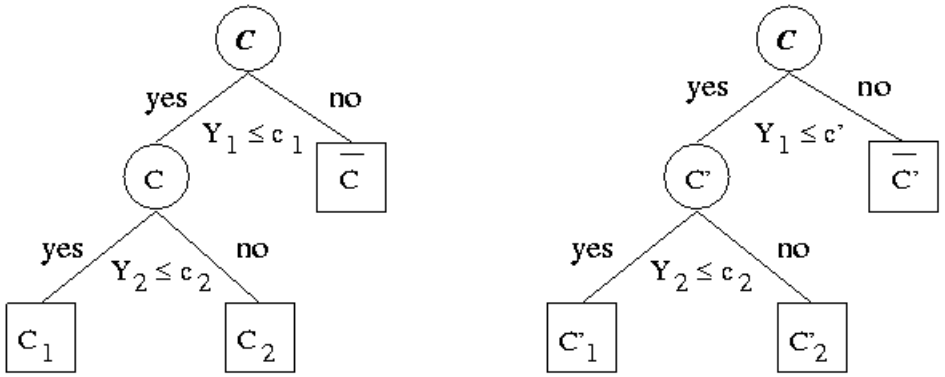


- choix de la classe qui induit la partition de plus petite inertie intra-classe
- choix de la classe C qui maximise $\Delta(C)=I(C)-I(C_1)-I(C_2)$



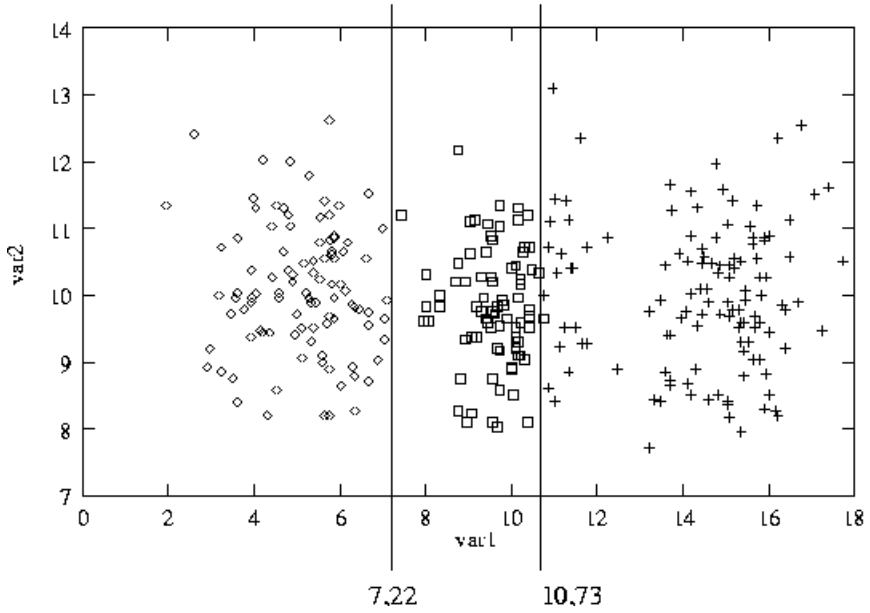


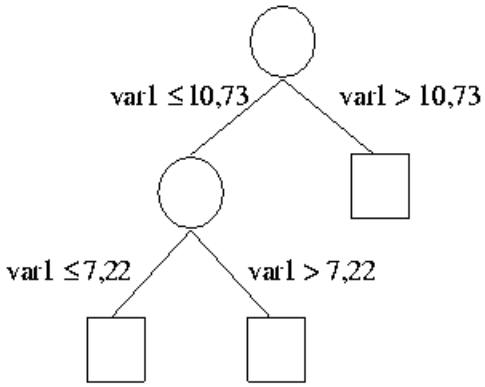
Remise en cause d ’une division



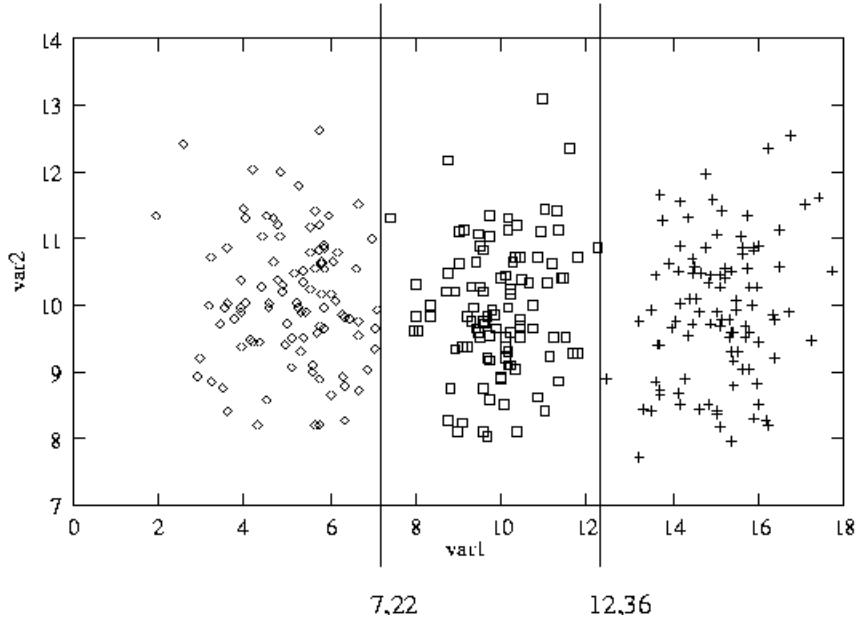
- Pour toutes les transformations binaires de Y_1, on évalue les partitions en trois classes induites par cette nouvelle variable binaire, la transformation binaire de Y_2 étant fixée.
- On retient la nouvelle transformation binaire de Y_1 qui induite la meilleure partition en trois classes.

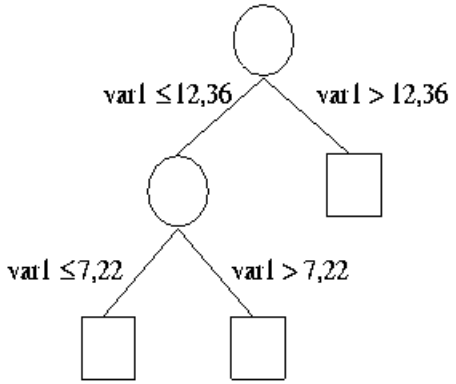
Avant :


var2
var1
7,22
10,73

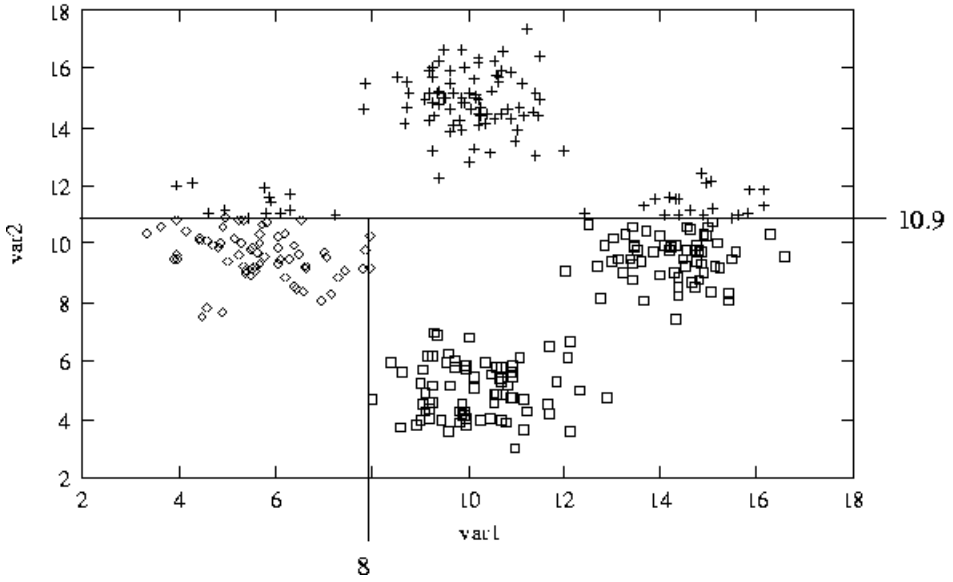

var1 ≤ 10,73
var1 > 10,73
var1 ≤ 7,22
var1 > 7,22

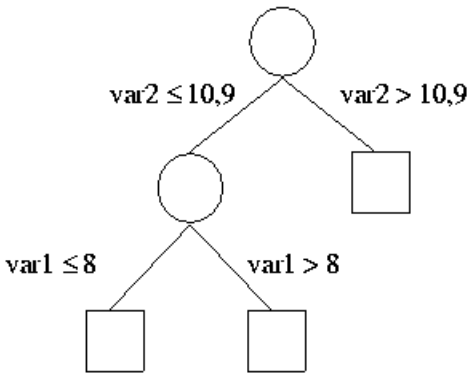
Après :


var2
var1
7,22
12,36

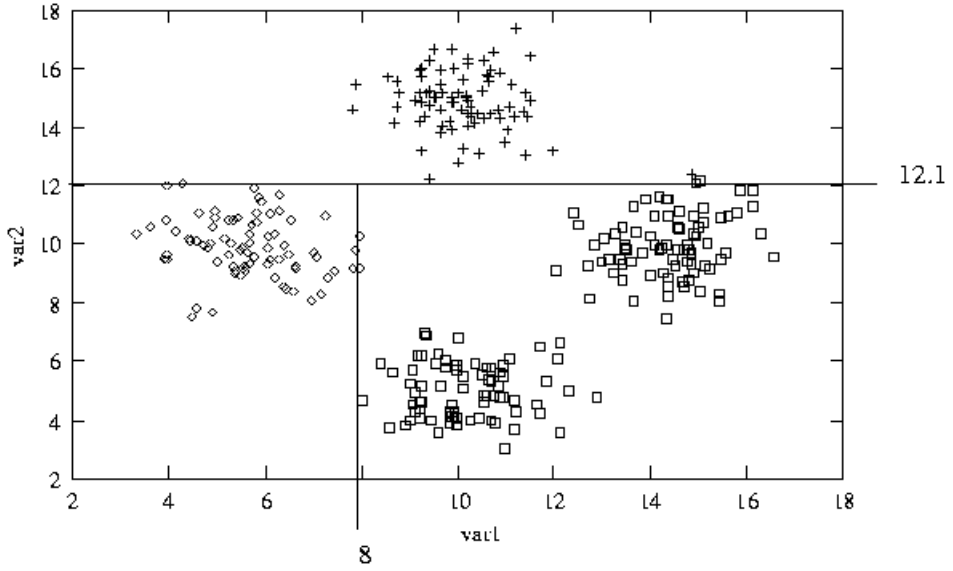

var1 ≤ 12,36
var1 > 12,36
var1 ≤ 7,22
var1 > 7,22

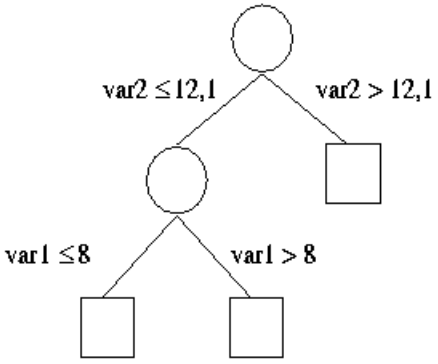
Avant :


var2
var1
10.9
8


var2 ≤ 10,9
var2 > 10,9
var1 ≤ 8
var1 > 8

Après :


var2
var1
12.1
8

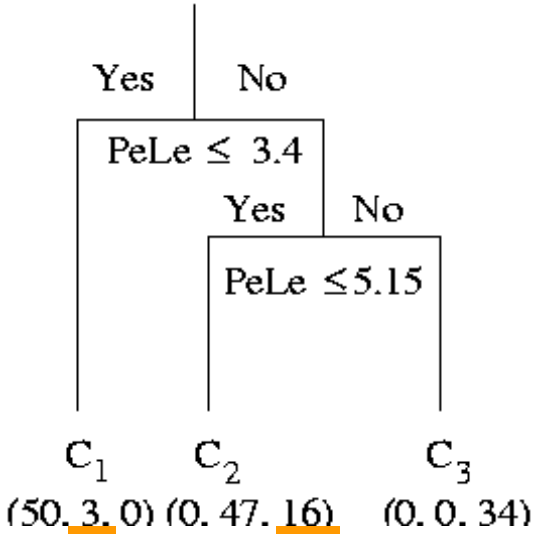

var2 ≤ 12,1
var2 > 12,1
var1 ≤ 8
var1 > 8

Les Iris de Fisher

	PeLe	PeWi	SeLe	SeWi	Varieté
1					Setoa
⋮					⋮
51					Versicolor
⋮					⋮
101					Virginia
⋮					⋮
150					

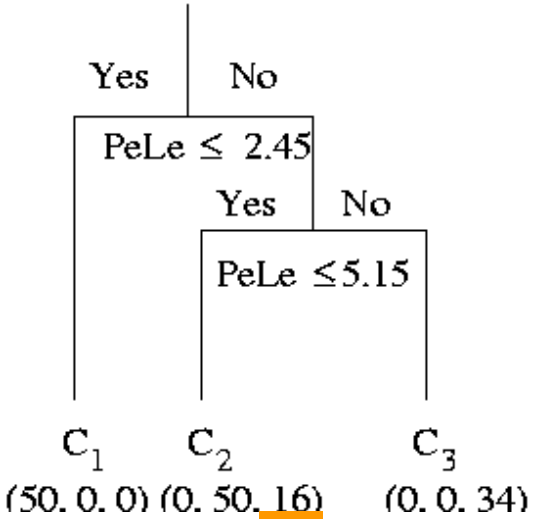
1. Distance euclidienne d_M avec $M=I$

Avant :



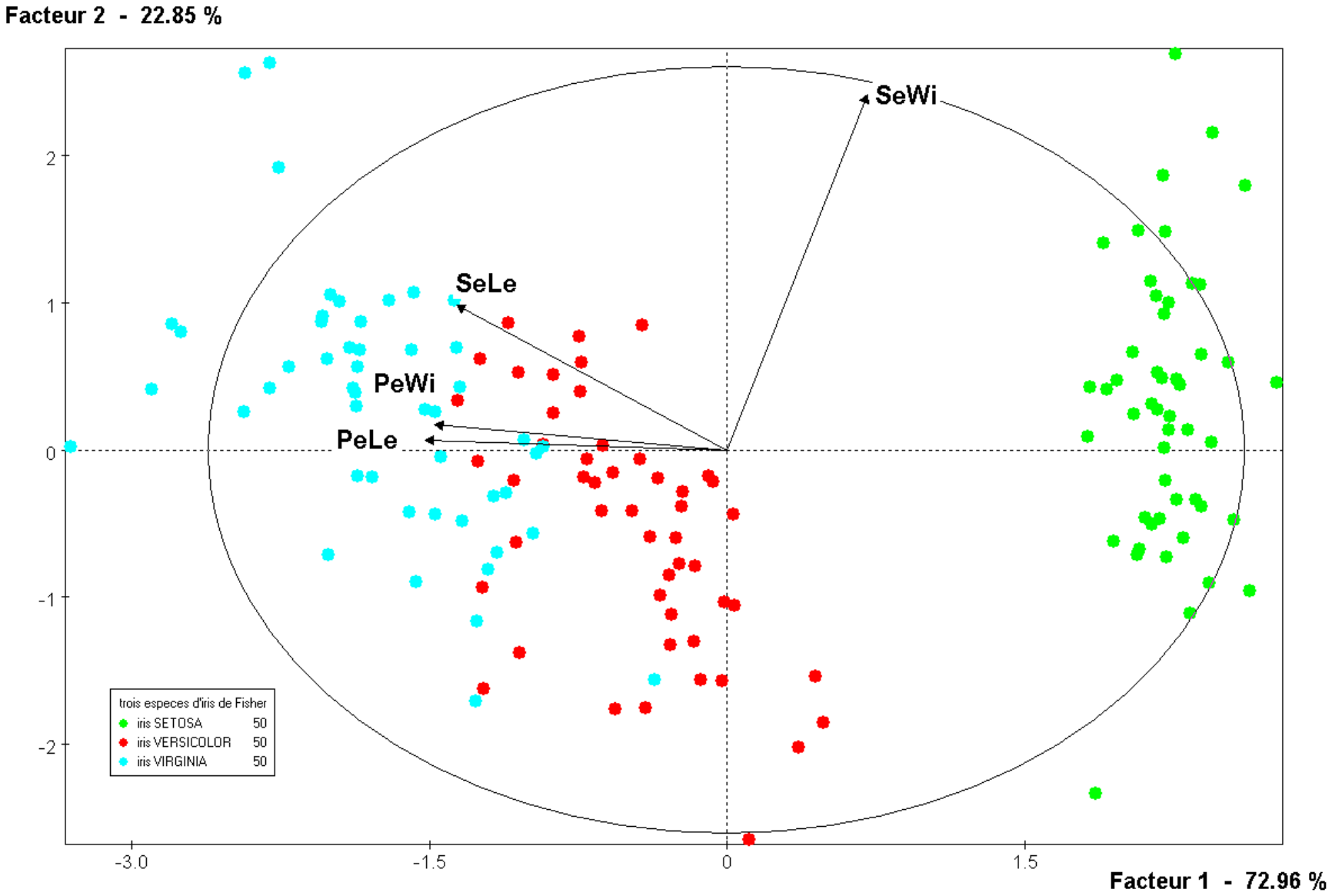
⇨ 19 iris mal classés

Après :

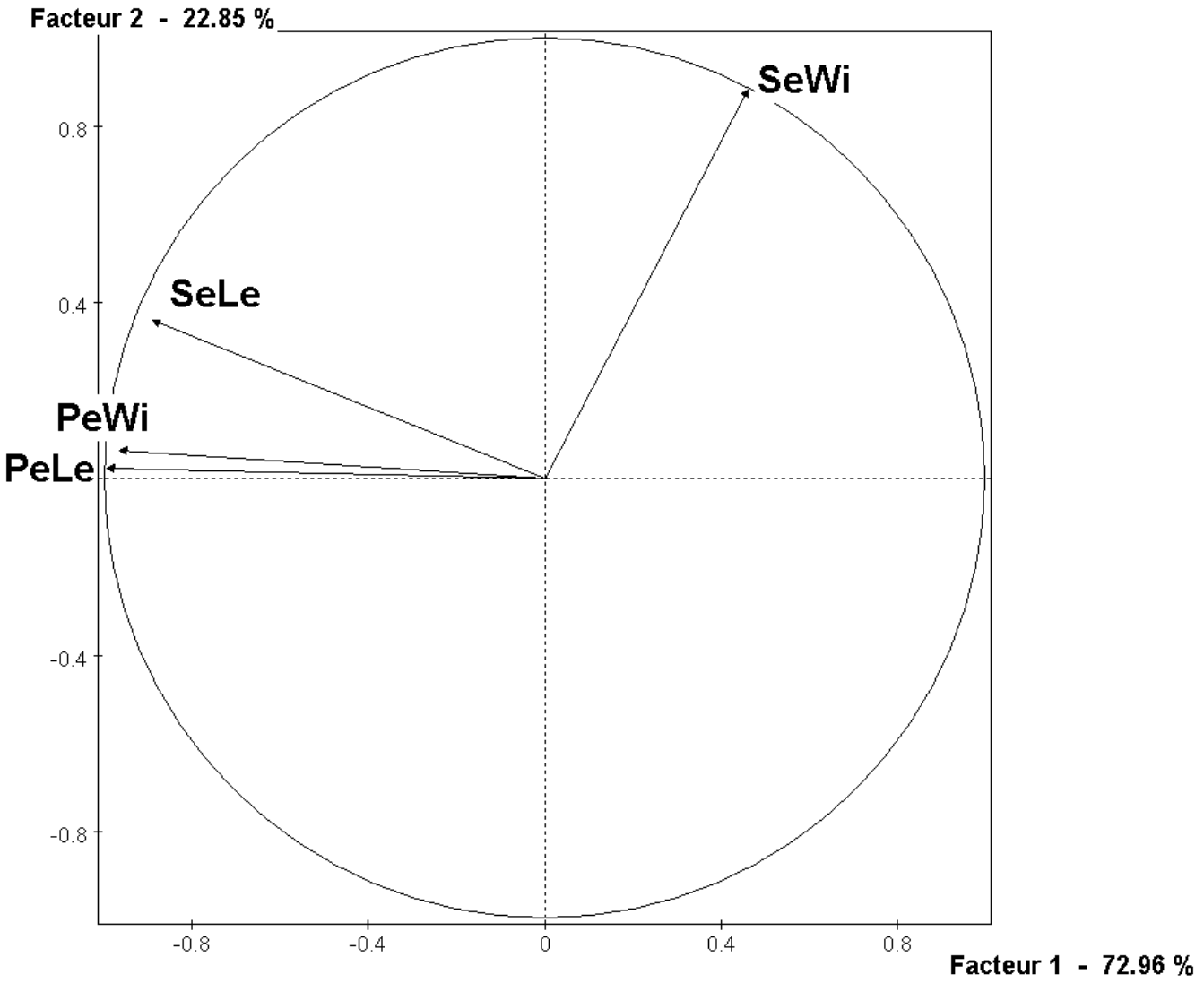


⇨ 16 iris mal classés

ACP sur les 4 variables continues



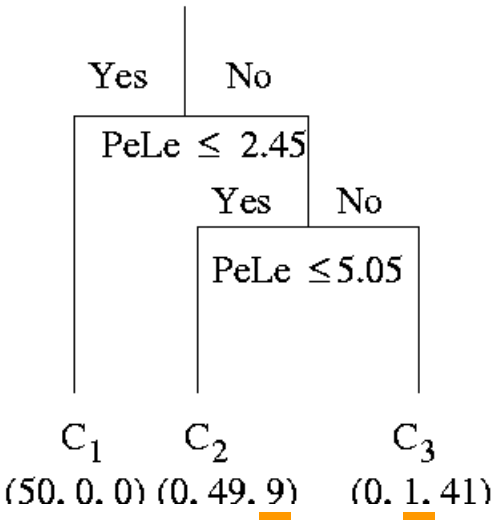
ACP : cercle des corrélations



Les Iris de Fisher

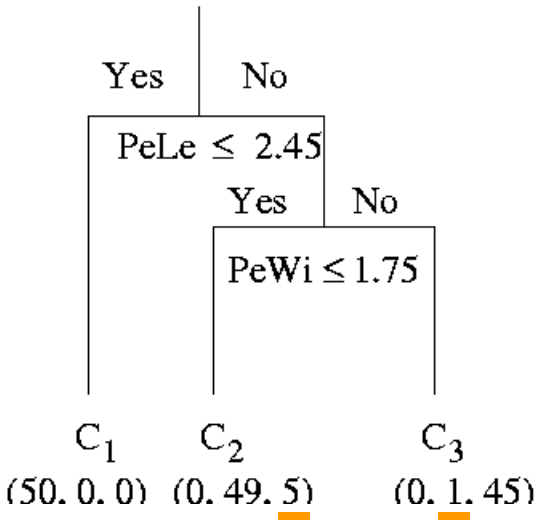
2. Distance euclidienne d_M normalisée par l 'inverse de l 'écart maximum, $M=D_{1/U_i^2}$

Globale :



➡ 10 iris mal classés

Locale :

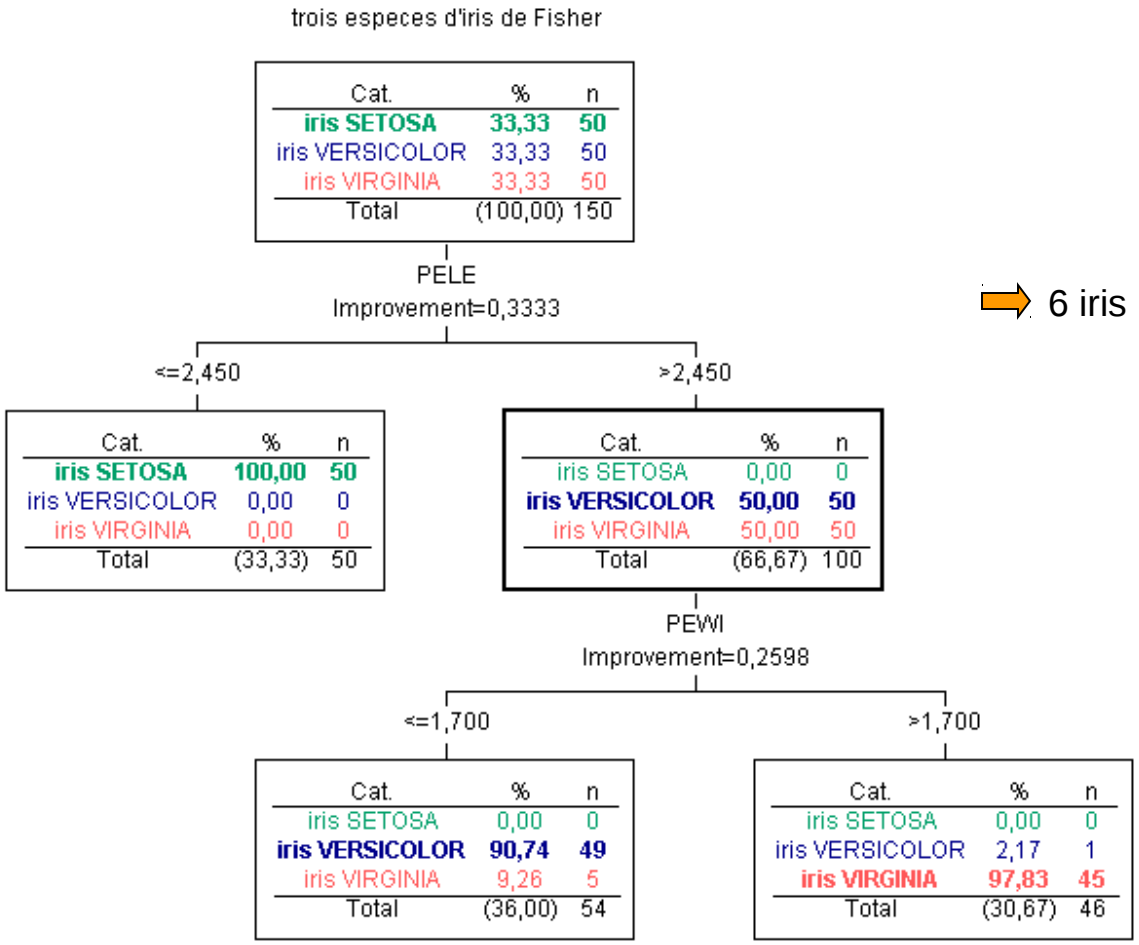


➡ 6 iris mal classés

Segmentation non supervisée ➡ 4%

	Setoa	Versicolor	Virginia
C_1	50	0	0
C_2	0	49	5
C_3	0	1	45

Classification And Regression Tree (CART, Breiman, Friedman Olshen and Stone, 1984)



⇨ 6 iris mal classés

Segmentation supervisée ⇨ 4%

Conclusion et perspectives

- Les Iris de Ficher : segmentation non supervisée et supervisée

- Règles d'affectation permettent

  ⇨ Interprétation monothétique des classes

  ⇨ Réviser les coupures

  ⇨ Échantillonner et affecter

- Données symboliques

- Divisions non binaires, affectation floues, validation, règles d 'arrêt et élagage...