

SODAS SOFTWARE

The **Tree** Procedure

Yves Lechevallier

INRIA-Rocquencourt

Domaine de Voluceau BP 105, Rocquencourt

78153 Le Chesnay Cedex, France

tel : 01 39 63 54 34

fax : 01 39 63 58 92

Yves.Lechevallier@inria.fr

Emmanuel Périnel

ENSAR - Unité de Mathématiques Appliquées

65, rue de St-Brieuc - CS 84215

35042 Rennes cedex

tel : 02 23 48 58 88

fax : 02 23 48 58 71

perinel@agrorennes.educagri.fr

# 1 Example : Sensor Analysis

- 6 products : stewed apple (“compotes de pommes”)  
tasted by 30 judges.
- 2 classes of products
  1. branded products (“ produit de marque ”)
  2. standard makes products (“ marque de distributeur ”)
- sensory descriptors : sweet, acid, bitter, astringent, cooked apple flavor, bright aspect, granular texture, etc.
- integer scale :  $0 \rightarrow 5$

judge	product	rank	uncooked apple	cooked apple	... sweet	acidity	bitterness	class
1	1	5	2.00	2.00	4.00	1.00	0.00	1
...	1	6	3.00	3.00	1.00	1.00	4.00	1
30	1	3	1.00	4.00	1.00	0.00	0.00	1
.								
.								
1	2	5	2.00	1.00	4.00	2.00	0.00	2
...	2	2	4.00	2.00	4.00	0.00	0.00	2
30	2	5	1.00	2.00	2.00	0.00	0.00	2
.								
.								
1	3	1	2.00	2.00	1.00	4.00	0.00	1
...	3	3	4.00	2.00	1.00	2.00	3.00	1
30	3	3	4.00	2.00	1.00	2.00	3.00	1
.								
.								
1	6	6	0.00	1.00	4.00	2.00	0.00	1
...	6	4	2.00	1.00	1.00	3.00	0.00	1
30	6	3	4.00	2.00	2.00	4.00	0.00	1

The “stewed apple ” data table

	uncooked apple					bitter					sweet								
	0	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5	
P1	0,00	0,10	0,26	0,35	0,26	0,03	0,26	0,52	0,13	0,10	0,00	0,00	0,32	0,42	0,10	0,13	0,03	0,00	...
P2	0,03	0,10	0,16	0,32	0,29	0,10	0,03	0,29	0,45	0,13	0,06	0,03	0,19	0,35	0,19	0,16	0,03	0,06	...
P3	0,00	0,10	0,26	0,35	0,19	0,10	0,06	0,35	0,16	0,32	0,06	0,03	0,23	0,39	0,19	0,13	0,06	0,00	...
P4	0,03	0,03	0,16	0,52	0,23	0,23	0,19	0,26	0,29	0,16	0,06	0,03	0,29	0,48	0,13	0,03	0,06	0,00	...
P5	0,00	0,00	0,19	0,35	0,26	0,19	0,10	0,26	0,16	0,29	0,16	0,03	0,29	0,45	0,19	0,03	0,03	0,00	...
P6	0,00	0,06	0,16	0,16	0,39	0,23	0,06	0,16	0,19	0,23	0,29	0,06	0,32	0,42	0,16	0,10	0,00	0,00	...

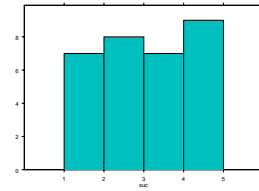
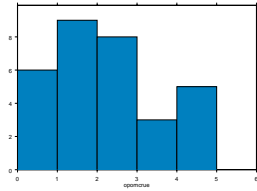
- Each object (a product) is described by a frequency distribution
- A distribution  $\rightarrow$  variability of the judgments

uncooked apple ...

...

sweet

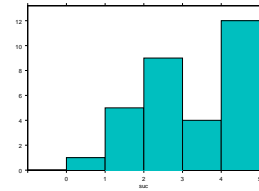
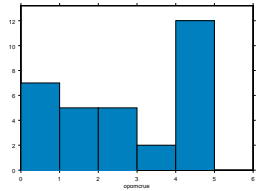
Produit1



...

...

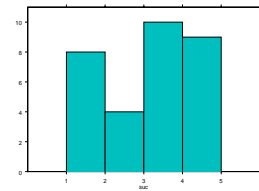
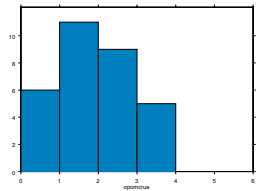
Produit2



...

...

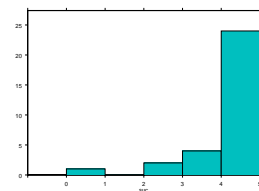
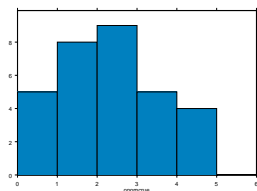
Produit3



...

...

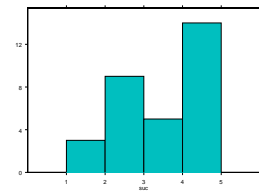
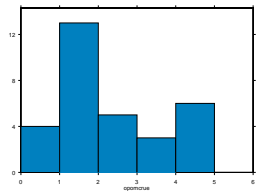
Produit4



...

...

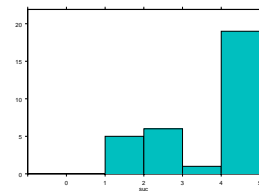
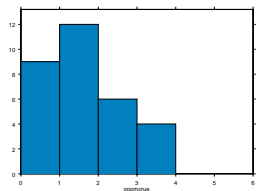
Produit5



...

...

Produit6



...

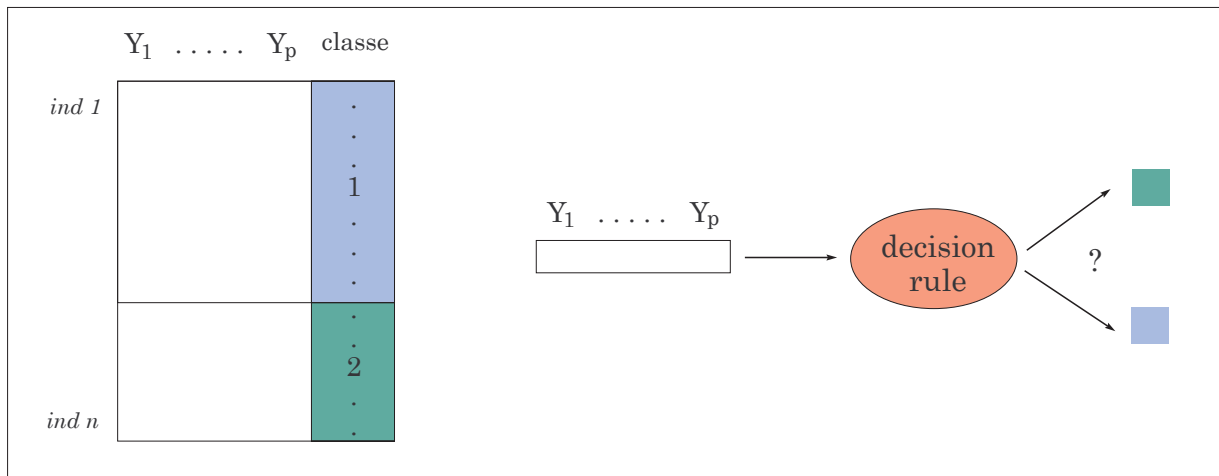
...

# Objective

What are the more specific sensory descriptors of each of the two classes of products ?

## DISCRIMINANT ANALYSIS :

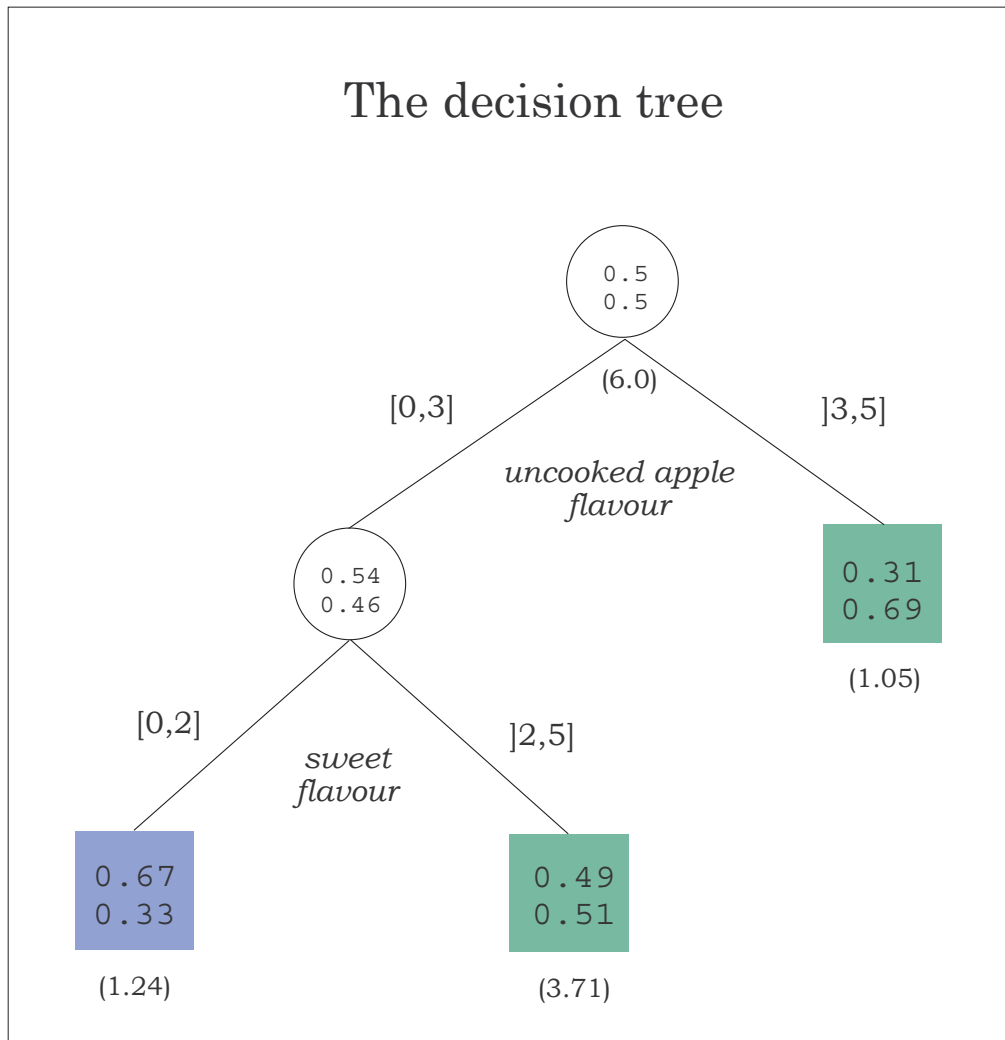
Build a mathematical rule able to discriminate the two classes



- Linear - Quadratic discriminant analysis
- Kernel discriminant analysis
- Factorial discriminant analysis
- **DECISION TREE**

= build "logical" decision rules in the form of a binary tree

## The decision tree



## Description of the two classes

branded products

[ uncooked apple < 3 ] and [ sweet < 2 ]

standard make products

[ uncooked apple > 3 ]

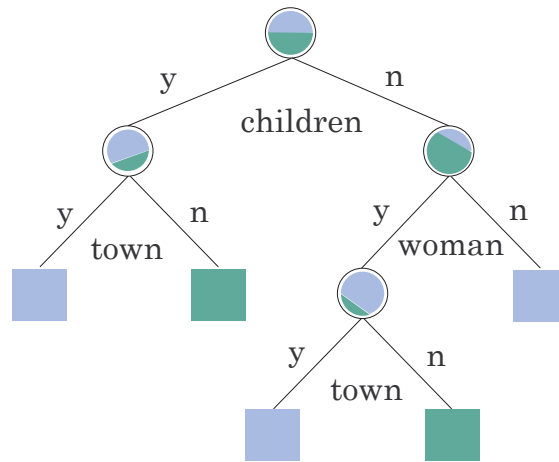
[ uncooked apple < 3 ] and [ sweet > 2 ]

# Decision tree methodology

Decision tree = Segmentation  
 = Tree growing = Recursive partition

- ① Building specific descriptions of the classes  
 on the basis of objects which class is known

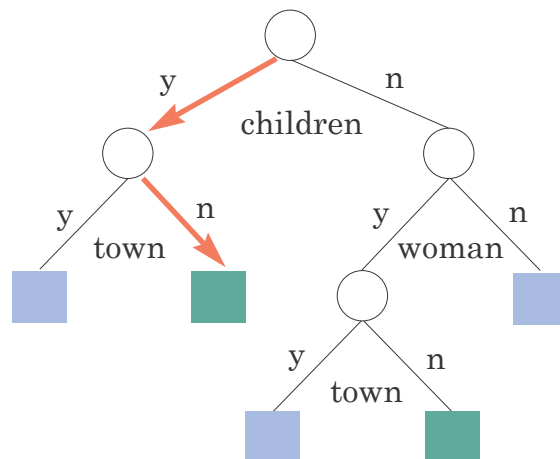
	$Y_1$	.....	$Y_p$	classe
<i>ind 1</i>				· · 1 · · ·
<i>ind n</i>				· · 2 · ·



- ② Using the decision rule for new objects  
 which class is unknown

$Y_1$	.....	$Y_p$

?



# Problem

How to extend standard algorithms  
of tree growing to symbolic data ?

## Contents

1. Description of the symbolic data table
2. The general algorithm
  - Set of binary questions
  - How to select the best split ?
  - The stopping rules
  - Assigning a label
  - Computing the misclassification rate
3. Running the `Tree` procedure on an simple example
  - Description of the input (data set)
  - Parameters of `Tree`
  - Description of the output (report)

## The symbolic data table

- A set of  $n$  objects :  $1, \dots, k, \dots, n$  described by  $p$  variables
- The objects are partitionned into  $m$  disjoint classes

	$Y_1$	⋯	$Y_j$	⋯	$Y_p$	$C$
1						
⋮						
$k$	$Y_1(k) \sim f_{k1}$	⋯	$Y_j(k) \sim f_{kj}$	⋯	$Y_p(k) \sim f_{kp}$	$C = c_k$
⋮						
$n$						

### Notations

- $C$  : Class variable
- $Y_1, Y_2, \dots, Y_p$  : predictors
- **Description** of object  $k$  for variable  $j$

→  $f_{kj}$  : distribution of frequencies or probabilities

Example 2 :

	color	size	petals	class
1	1(yellow)	[10,12]	1(12)	1
2	$\frac{1}{2}$ (blue), $\frac{1}{2}$ (yellow)	[30, 60]	$\frac{1}{6}$ (15), $\frac{1}{6}$ (16), ..., $\frac{1}{6}$ (20)	1
3	$\frac{1}{2}$ (blue), $\frac{1}{2}$ (red)	[27, 35]	1(2)	1
4	$\frac{3}{4}$ (blue), $\frac{1}{4}$ (red)	[18, 30]	$\frac{1}{2}$ (7), $\frac{1}{3}$ (8), $\frac{1}{6}$ (9)	2
5	$\frac{1}{2}$ (yellow), $\frac{1}{4}$ (blue), $\frac{1}{4}$ (red)	[5, 10]	$\frac{1}{3}$ (5), $\frac{1}{3}$ (6), $\frac{1}{3}$ (7)	2

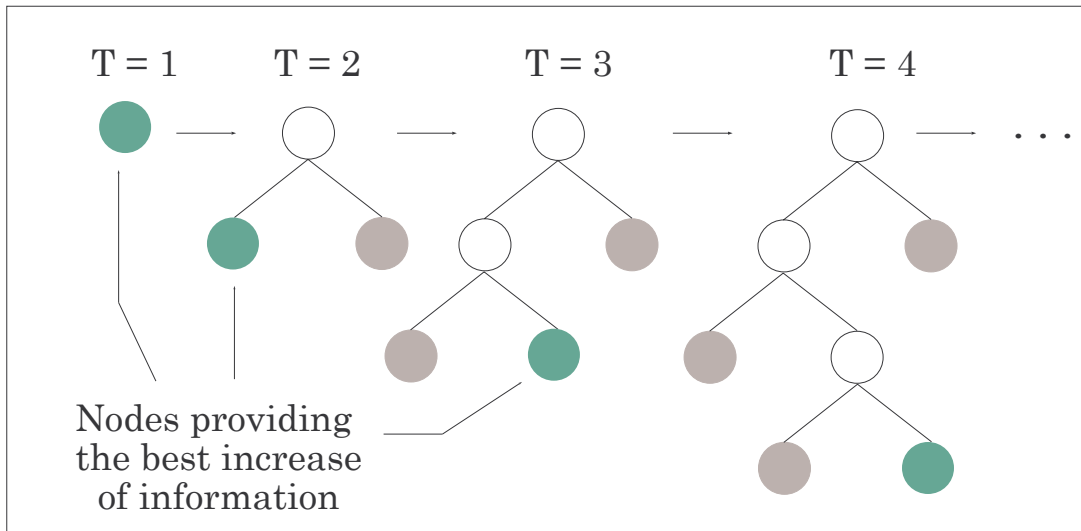
In this version of **Tree** :

- the mixed data table is not treated :
  - predictors are only numerical (interval descriptions)
    - size  $\sim$  [10, 12],
    - age  $\sim$  [45, 60],
    - rate of interest  $\sim$  [0.08, 0.12], ...
  - predictors are only nominal (frequency descriptions)
    - sexe  $\sim$  0.65(men), 0.35(women),
    - color  $\sim$   $\frac{3}{4}$ (blue),  $\frac{1}{4}$ (red),
    - student  $\sim$  0.34(Science), 0.45(Letter), 0.21(other)...
- The class variable is as in the standard case (nominal)

$$C = 1, \quad C = 2, \dots$$

**and not**  $C \sim 0.26(1), 0.74(2) \quad !$

## The general algorithm



## The set of binary questions

- **Continuous variables**

$$[Y_j < s] \quad \text{or} \quad [Y_j \geq s] \quad ?$$

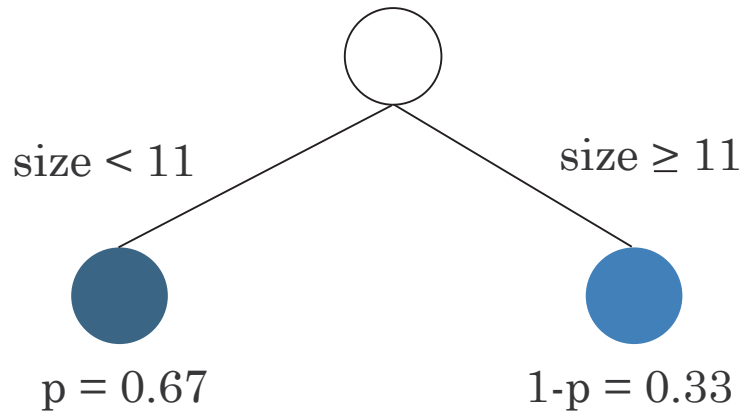
Example :  $[\text{size} < 8]$  or  $[\text{size} \geq 8]$  ?

- **Nominal variables**

$$[Y_j \in M] \quad \text{or} \quad [Y_j \in \overline{M}] \quad ?$$

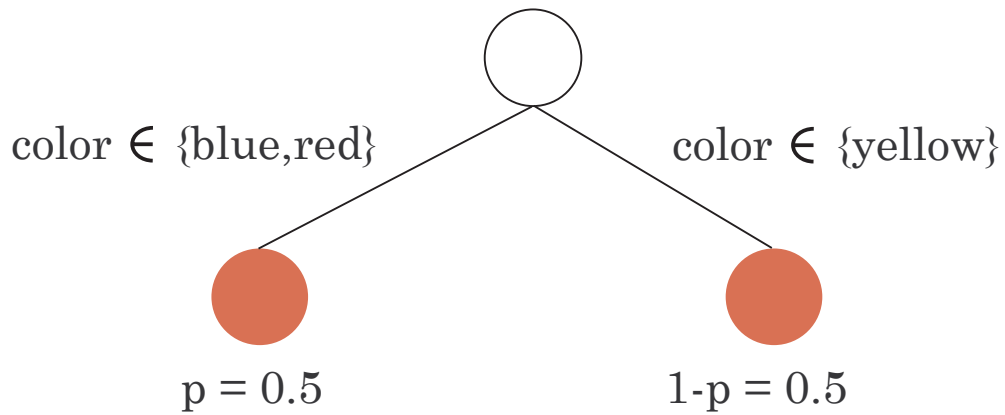
Example :  $[\text{color} \in \{\text{yellow}\}]$  or  $[\text{color} \in \{\text{red}, \text{blue}\}]$  ?

[ size  $\in$  [10,13] ]



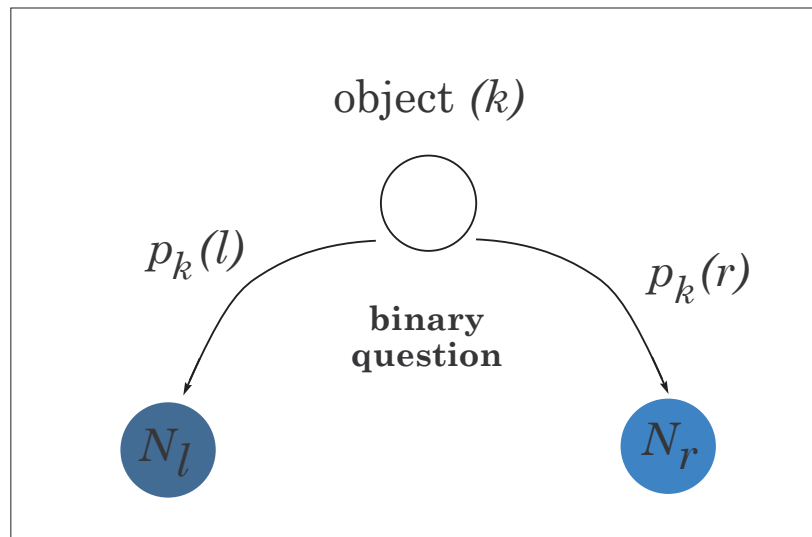
$$p = (13-11)/(13-10) = 0.67$$
$$1-p = (11-10)/(13-10) = 0.33$$

[ color  $\sim \frac{1}{2}$ (yellow),  $\frac{1}{2}$ (red) ]



$$p = P(\text{blue or red}) = P(\text{red}) = 0.5$$
$$1-p = P(\text{yellow}) = 0.5$$

## Probabilistic or fuzzy assignment



$p_k(r)$  = probabilistic membership of  $k$  to  $N_r$

$p_k(l)$  = probabilistic membership of  $k$  to  $N_l$

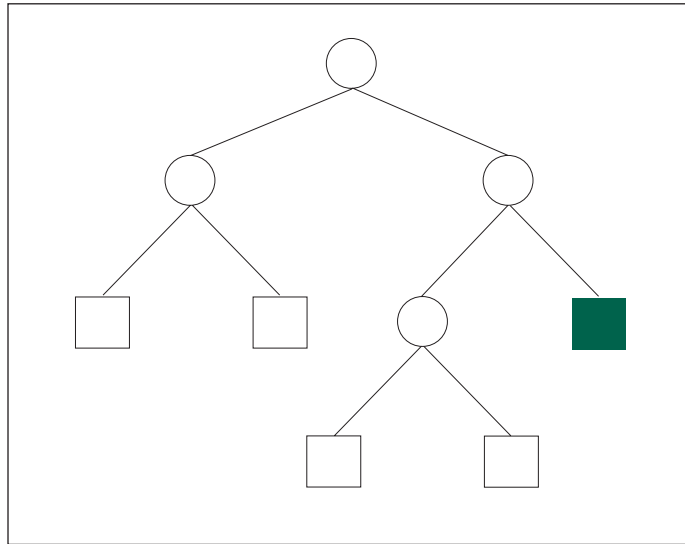
### Consequence

- Usual decision tree
  - each split creates a “fuzzy” partition of the objects
- Decision tree with symbolic data
  - each split creates disjoint subclasses of objects

In the Tree version

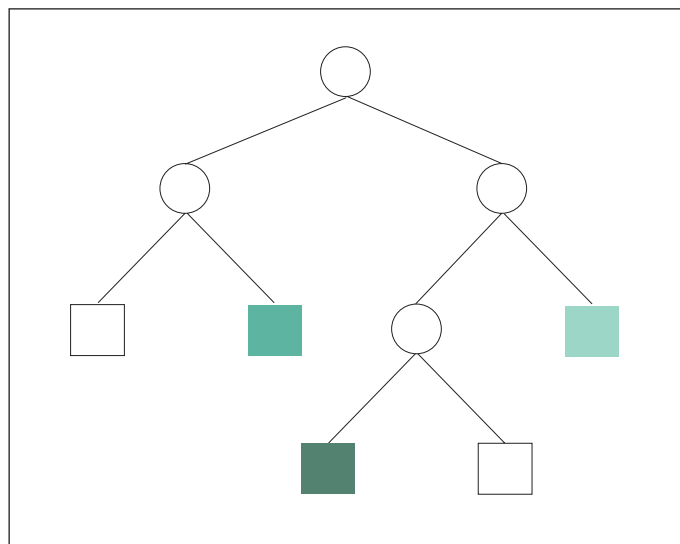
- **“fuzzy”** assignment, OR
- **“pure”** assignment

## Usual decision tree



→ An object belongs exclusively to one terminal node

## Decision tree with symbolic data



→ An object may belong to various leaves of the tree

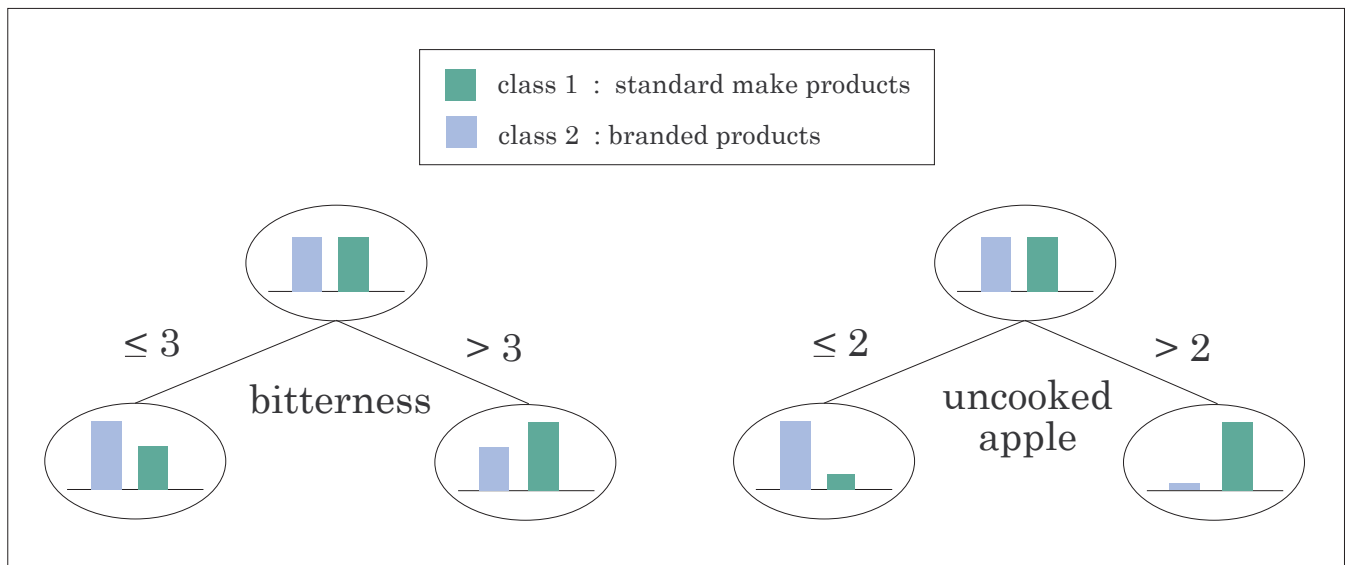
To build the tree :

1. Consider all binary questions
2. Select the best one

How to select the best split ?

- **Goal** : select the question which leads to the best prediction of the class variable
- **Best prediction** : the left and right nodes are as “pure” as possible with respect to the class variable

### Comparison of two binary question



**Impurity of a node** ( $t$ ) : many measures in the literature

In the Tree version

With the “fuzzy” assignment :

- *Log-likelihood*

$$\text{Imp}(N_t) = \log \prod_k P_t(c_k)$$

With the “pure” assignment :

- *Gini* :

$$\text{Imp}(N_t) = \sum_{m \neq n} P_t(m) \cdot P_t(n)$$

- *Information - Entropy* :

$$\text{Imp}(N_t) = \sum_m P_t(m) \cdot \log P_t(m)$$

→ does not change a lot the selection of the split...

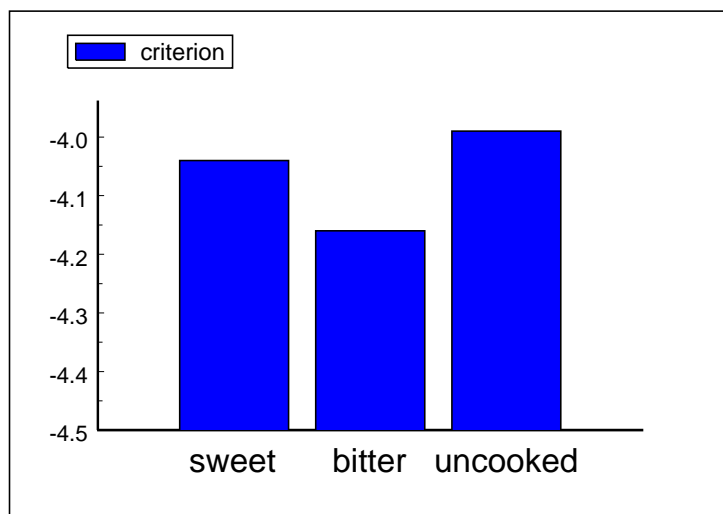
**Global quality of a split** : average impurity

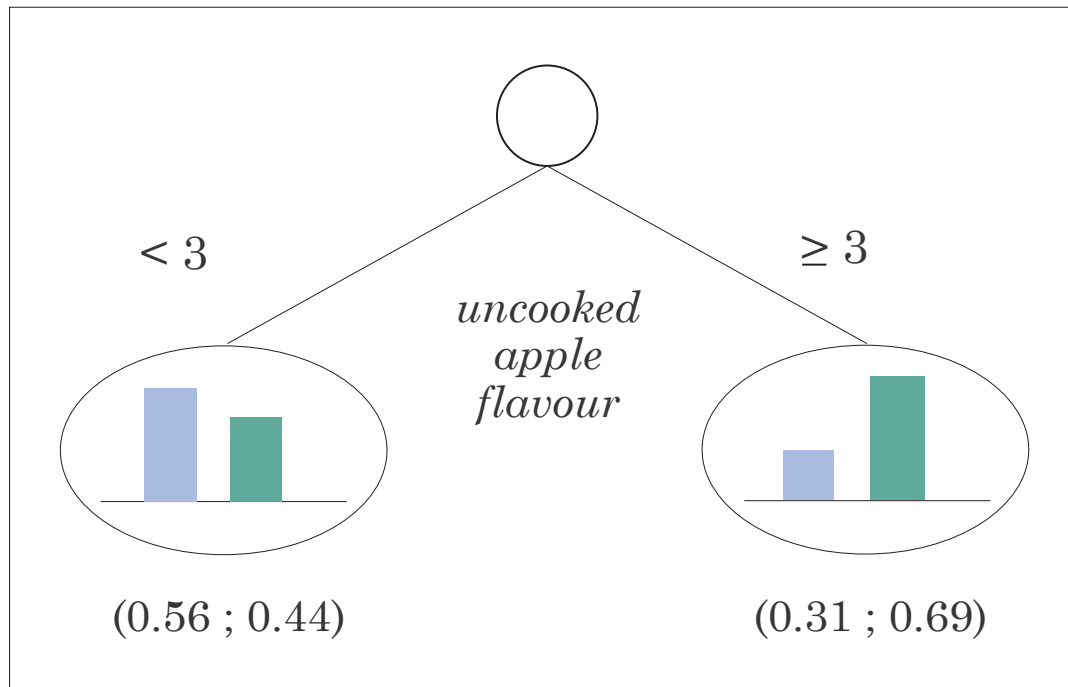
$$\text{Quality (split)} = p(r) \cdot \text{Imp}(N_r) + p(l) \cdot \text{Imp}(N_l)$$

# Application

Results for the first split

descriptor	binary split	criterion value
sweet	[0; 1] / ]1; 5]	-4.04
	[0; 2] / ]2; 5]	-4.14
	[0; 3] / ]3; 5]	-4.15
	[0; 4] / ]4; 5]	-4.08
bitter	[0; 1] / ]1; 5]	-4.16
	[0; 2] / ]2; 5]	-4.22
	[0; 3] / ]3; 5]	-4.36
	⋮	⋮
uncooked apple	[0; 1] / ]1; 5]	-4.19
	[0; 2] / ]2; 5]	-4.12
	[0; 3] / ]3; 5]	-3.99
	[0; 4] / ]4; 5]	-4.02





- $P[\text{branded mark} \mid \text{uncooked apple} < 3] = 0.56$
- $P[\text{standard make} \mid \text{uncooked apple} < 3] = 0.44$
- $P[\text{branded mark} \mid \text{uncooked apple} \geq 3] = 0.31$
- $P[\text{standard make} \mid \text{uncooked apple} \geq 3] = 0.69$

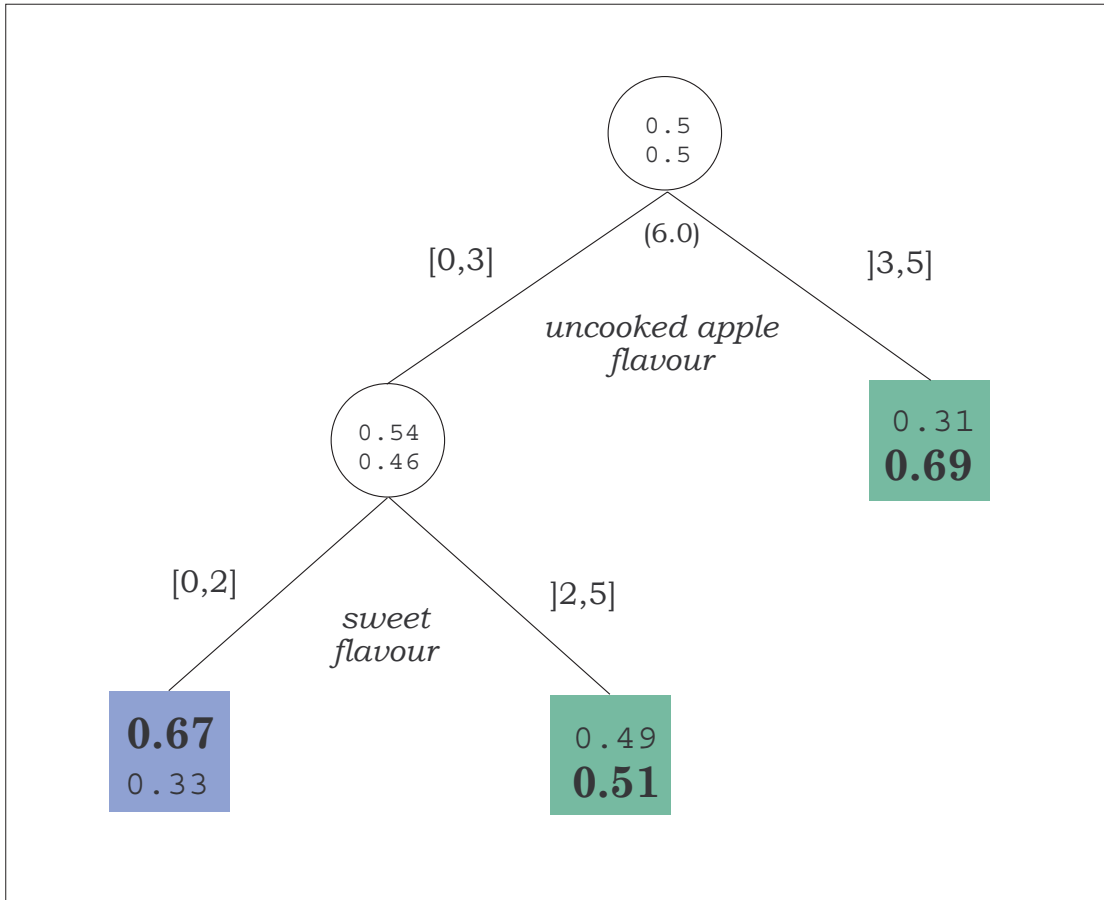
## Determining when to stop splitting

**A node is considered as terminal** (a leaf) if

1. its size is not large enough  
(default value = 5)
2. the number of objects who dont belong to the majority class is less than a given threshold  
(default value = 2)
3. it creates two son nodes of too small size  
(default value = 1)

# Assigning a label to a leaf

= according to the **majority rule**



label (leaf  $t$ ) =  $m$

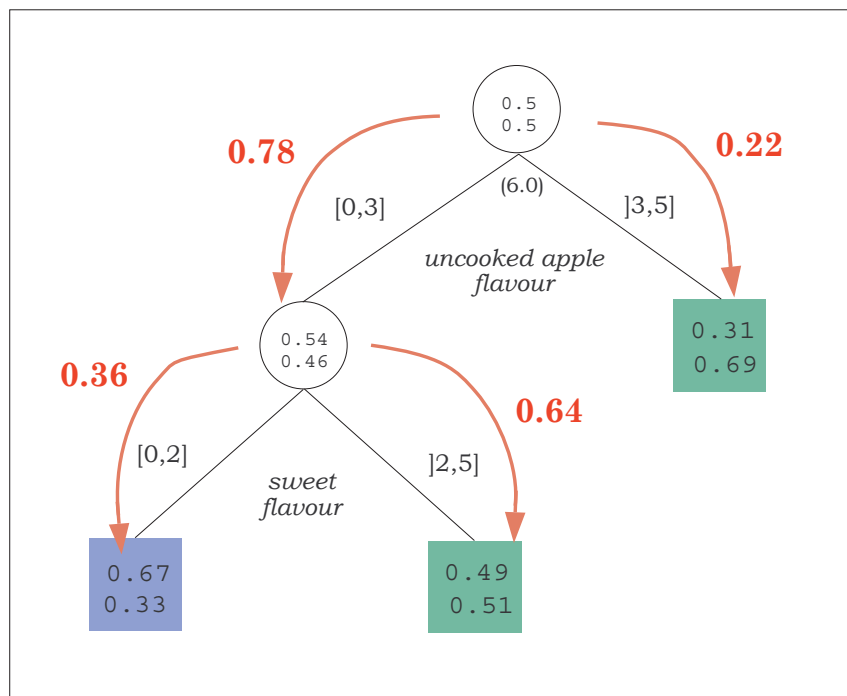
$\Leftrightarrow$

$$P[\text{class } m \mid t] > P[\text{class } m' \mid t] \quad \forall m'$$

# The misclassification rate

- Is the tree a good classifier ?
- Are the 6 objects well classified by the tree ?

We compute :  $P[\text{class } m \mid \text{object } k] \quad \forall k, m$



Example :

$$\begin{aligned}
 P[\text{class 1} \mid \text{object 1}] &= 0.67 \times (\mathbf{0.78} \times \mathbf{0.26}) \\
 &+ 0.49 \times (\mathbf{0.78} \times \mathbf{0.64}) \\
 &+ 0.31 \times \mathbf{0.22} \\
 &= \boxed{0.53} \\
 P[\text{class 2} \mid \text{object 1}] &= 0.47
 \end{aligned}$$

- **Membership probabilities**  $P[\text{class } m \mid \text{object } k]$  of the 6 objects to the 2 prior classes

	Prior classes		decision
	branded	standard make	
P1	<b>0.53</b>	0.47	1
P2	<b>0.51</b>	0.49	1
P3	<b>0.55</b>	0.45	1
P4	0.46	<b>0.54</b>	2
P5	0.49	<b>0.51</b>	2
P6	0.46	<b>0.54</b>	2

- **Confusion matrix**

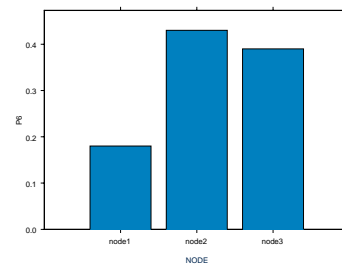
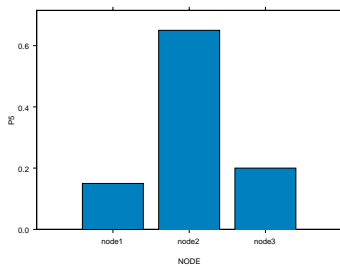
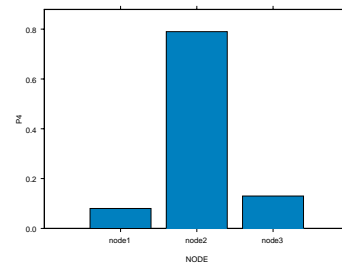
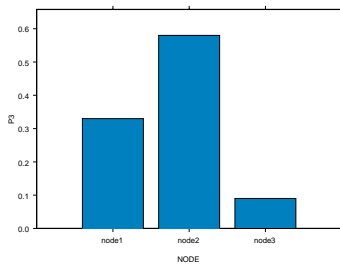
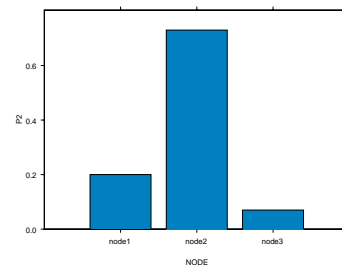
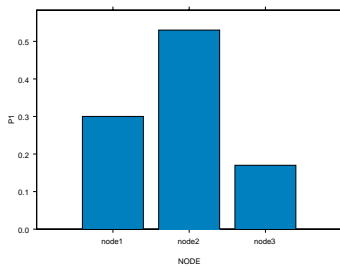
		prior class	
		class 1	class 2
decision	class 1	3	0
	class 2	0	3

$$\text{Misclassification rate} = \frac{\text{number of misclassified objects}}{\text{total number of objects}}$$

# Further information

- Membership probabilities of the 6 products to the 3 leaves

	Terminal nodes			
	1	2	3	
P1	0.30	<b>0.53</b>	0.17	1.00
P2	0.20	<b>0.73</b>	0.07	1.00
P3	0.33	<b>0.58</b>	0.09	1.00
P4	0.08	<b>0.79</b>	0.13	1.00
P5	0.15	<b>0.65</b>	0.20	1.00
P6	0.18	<b>0.43</b>	0.39	1.00

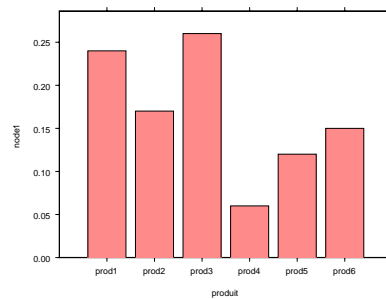


- Relative frequency of each product in the three leaves

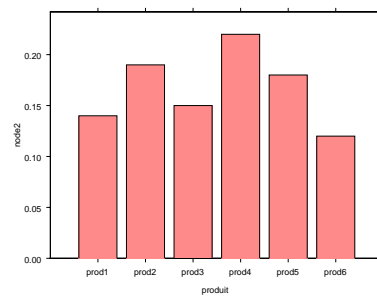
	Terminal nodes		
	1	2	3
P1	0.24	0.14	0.15
P2	0.17	0.19	0.07
P3	0.26	0.15	0.10
P4	0.06	0.22	0.12
P5	0.12	0.18	0.18
P6	0.15	0.12	0.38

1.00	1.00	1.00
------	------	------

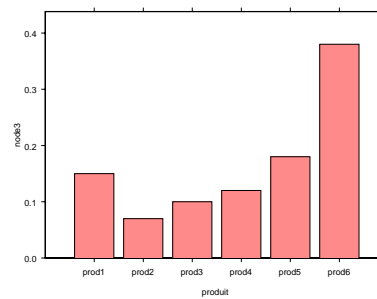
node 1



node 2

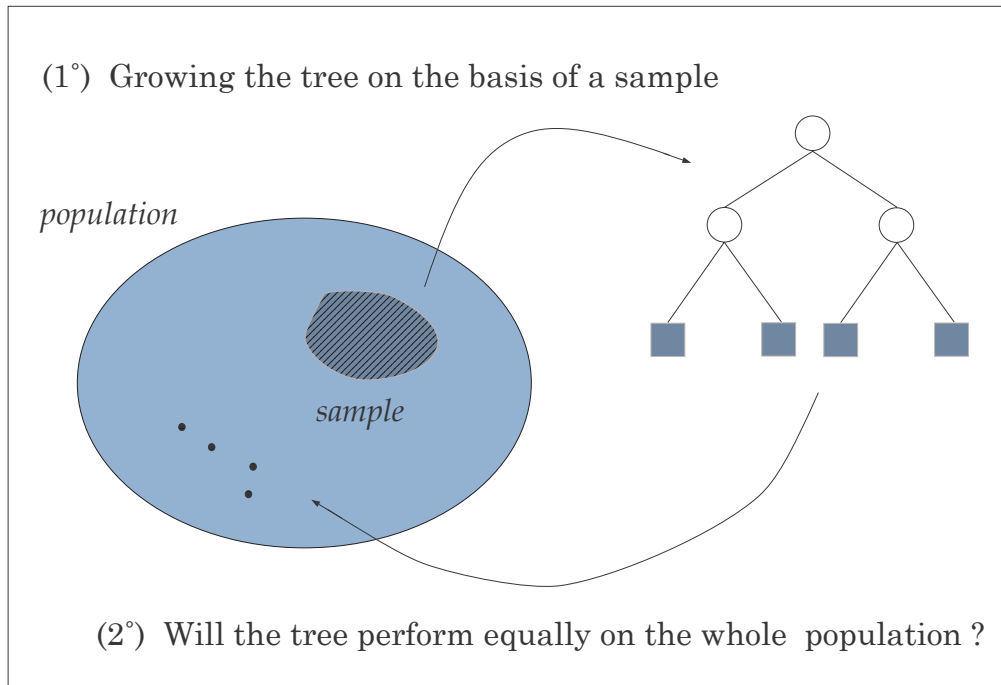


node 3



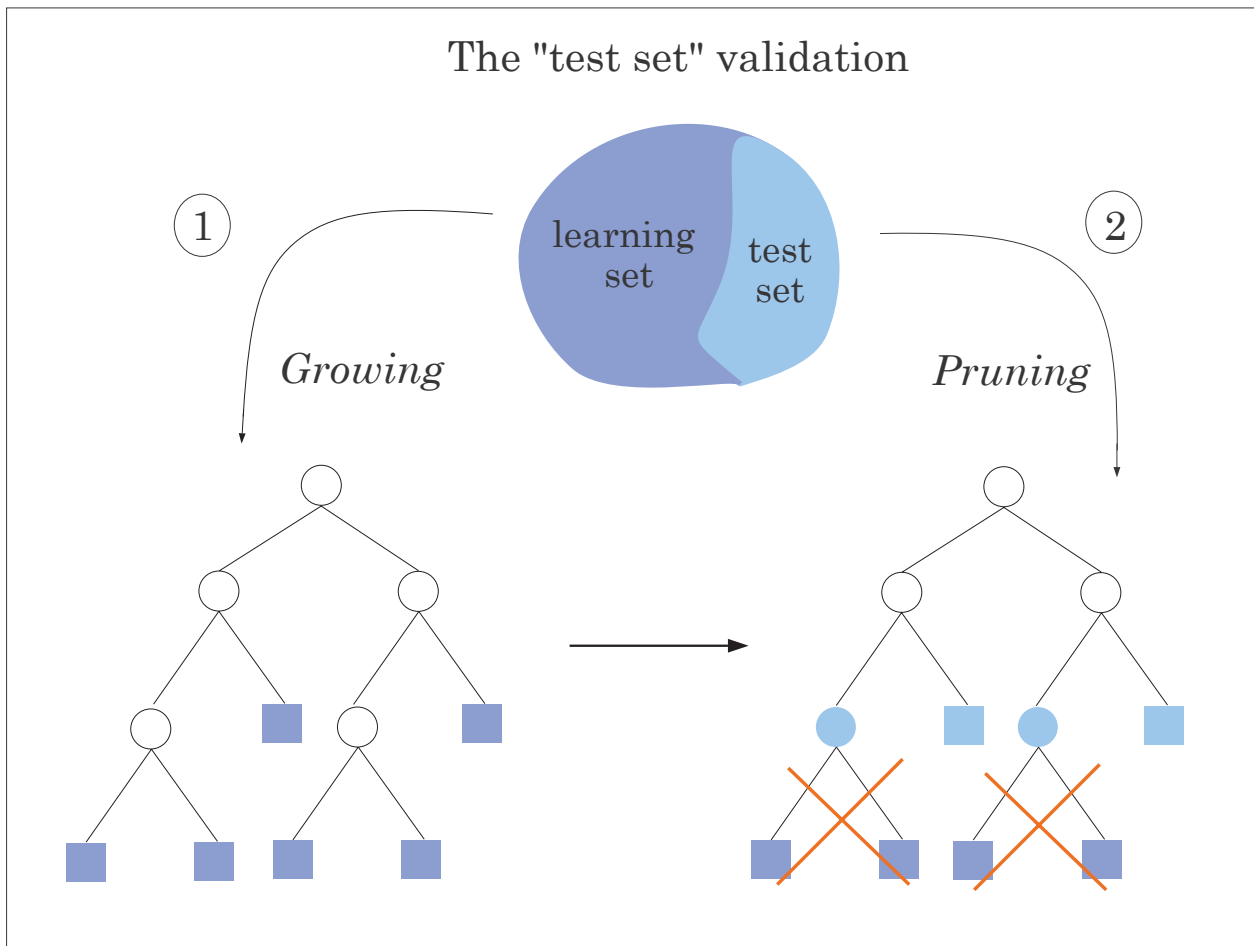
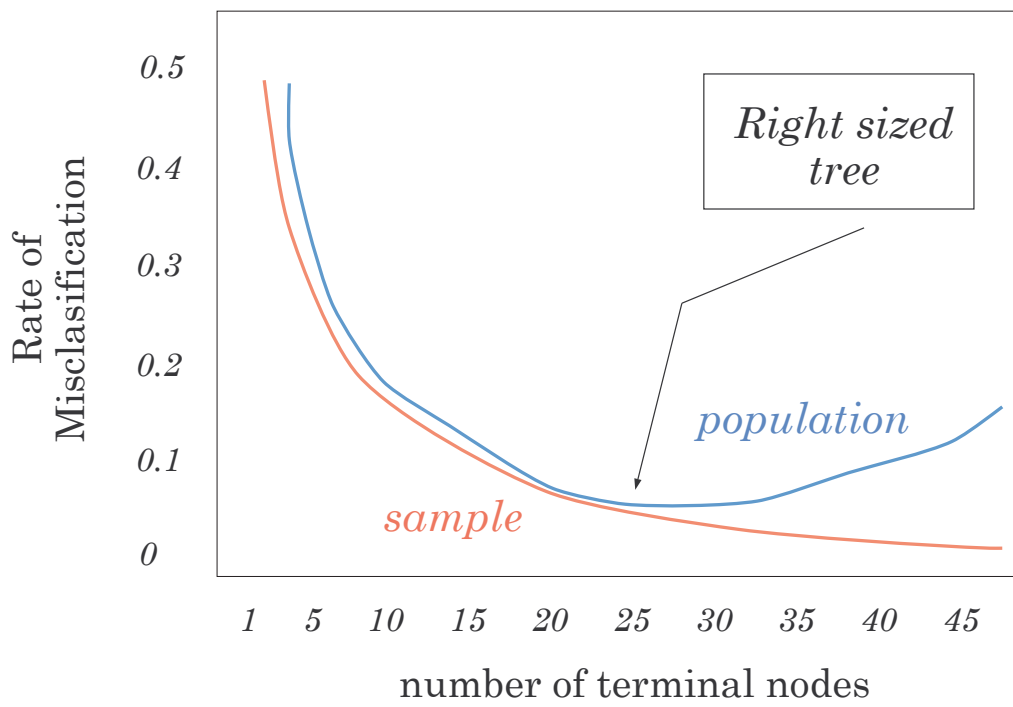
# Selecting the “right-sized” tree

## Problem



- Does the tree perform equally on the population ?
- What about the rate of misclassification (RM) for the whole population ?

= statistical validation of the tree...



## Statistical validation = *Not always necessary !*

- The studied set of objects corresponds to the whole population
- Statistical official data : regions, countries, towns, etc.

In the Tree version :

- Choose the percentage of objects in the test set  
(default value = 0)
- In many algorithms : test set  $\sim \frac{1}{3}$

## Running the **Tree** procedure

### Data :

- 12 fishes

1. "Ageneiosusbrevifili"
2. "Cynodongibbus"
3. "Hopliasaimara"
4. "Potamotrygonhystrix"
5. "Leporinusfasciatus"
6. "Leporinusfrederici"
7. "Dorasmicropoeus"
8. "Platydorascostatus"
9. "Pseudoancistrusbarbatus"
10. "Semaprochilodusvari"
11. "Acnodonoligacanthus"
12. "Myleusrubripinis"

- 4 classes (régimes)

1. Carnivores
2. Détritivores
3. Omnivores
4. Herbivores

- 13 predictors

1. LONG	8. REIN
2. POID	9. Foie/Muscle
3. MUSC	10. Reins/Muscle
4. INTE	11. Branchies/Muscle
5. ESTO	12. Intestins/Muscle
6. BRAN	13. Estomac/Muscle
7. FOIE	

# The output of `Tree` : report

## The prior classes

CLASS	SIZE	LEARNING	TEST
1	4	4	0
2	4	4	0
3	2	2	0
4	2	2	0
TOTAL	12	12	0

## Split of node 1

```
=====
| SPLIT OF A NODE           :      1 |
=====
```

### LEARNING SET

```
=====
|                               | N(k/t) | N(k) | P(k/t) | P(t/k) |
=====
| Carnivores                   |    4.00 |    4.00 |    33.33 |   100.00 |
| Détritivores                 |    4.00 |    4.00 |    33.33 |   100.00 |
| Omnivores                    |    2.00 |    2.00 |    16.67 |   100.00 |
| Herbivores                   |    2.00 |    2.00 |    16.67 |   100.00 |
=====
```

## Selecting the best split

```

=====
| Ord | variable | value | criterion |
=====
| 1 | ( 7) FOIE | 1124.9900 | 4.1917 |
| 2 | ( 9) Foie/Muscle | 4.9200 | 4.6892 |
| 3 | ( 4) INTE | 261.2000 | 4.7882 |
| 4 | ( 3) MUSC | 93.0000 | 4.8074 |
| 5 | ( 1) LONG | 18.8000 | 4.9636 |
=====

```

## The final tree

```

          +----- < 4 >Herbivores (0.00 0.57 0.00 2.00 )
          !
        !----2[ MUSC <= 255.000000]
          !
          !      +----- < 5 >Omnivores (0.13 0.00 2.00 0.00 )
          !
        !----1[ FOIE <= 1124.989990]
          !
          !      +----- < 6 >Détritivores (1.00 3.43 0.00 0.00 )
          !
          !----3[ BRAN <= 270.000000]
          !
          !      +----- < 7 >Carnivores (2.87 0.00 0.00 0.00 )

```

## The confusion matrix

CONFUSION MATRIX FOR TRAINING SET  
CONFUSION MATRIX FOR TRAINING SET

```
=====
|           | Carnivores| Détritivor| Omnivores | Herbivores| Total   |
=====
| Carnivores |          3 |          1 |          0 |          0 |          4 |
| Détritivores |          0 |          3 |          0 |          1 |          4 |
| Omnivores   |          0 |          0 |          2 |          0 |          2 |
| Herbivores  |          0 |          0 |          0 |          2 |          2 |
=====
| Total      |          3 |          4 |          2 |          3 |          12 |
=====
```

## Misclassification rate

MISCLASSIFICATION RATE BY CLASS

TRUE CLASS	( ERROR /SIZE )	FREQUENCY
Carnivores	( 1 / 4 )	25.00
Détritivores	( 1 / 4 )	25.00
Omnivores	( 0 / 2 )	0.00
Herbivores	( 0 / 2 )	0.00
TOTAL	( 2 / 12 )	16.67