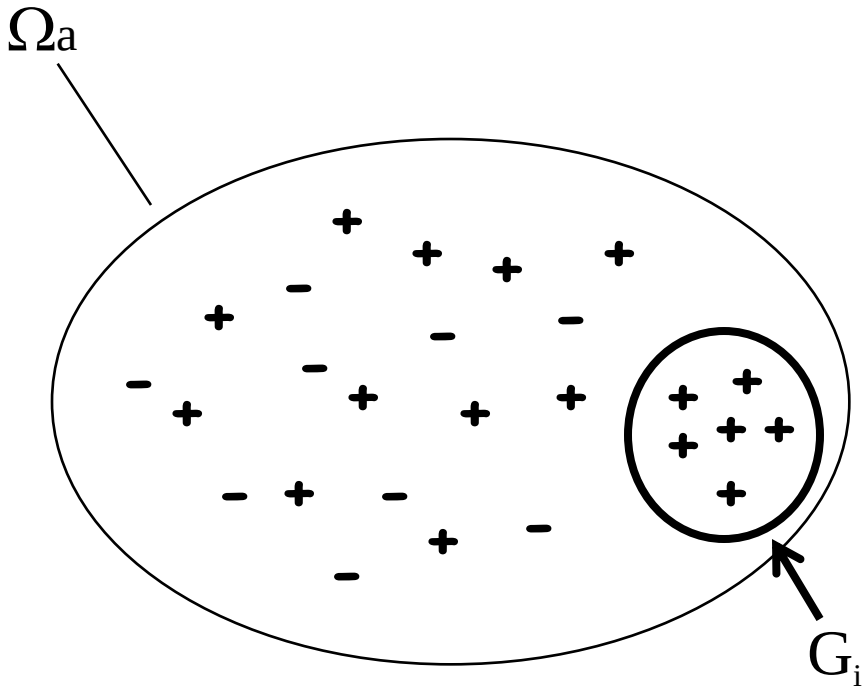
Introduction to Decision Trees

Ricco RAKOTOMALALA
Ricco.Rakotomalala@univ-lyon2.fr

Goal of the Decision Tree Learning (Classification Tree)

Goal: splitting the instances into subgroups with maximum of "purity" (homogeneity) regarding the target attribute



Binary target attribute Y with the values {+,-}
(Decision Trees Algorithms can handle multiclass problem)

Each subgroup G_i must be as homogenous as possible regarding Y
i.e. populated by instances with only the '+' (or the '-') label.

$$\text{IF}(\omega \in G_i)\,\text{THEN}(Y = +\,or\,-)$$

The goal is to obtain the most concise and accurate rule with the conditional probability P(Y=+/X) ≈1 [or P(Y=-/X) ≈1]

The description of the subgroups is based on :

☞ Logical Classification Rules

☞ With the most relevant descriptors

Example of Decision Tree

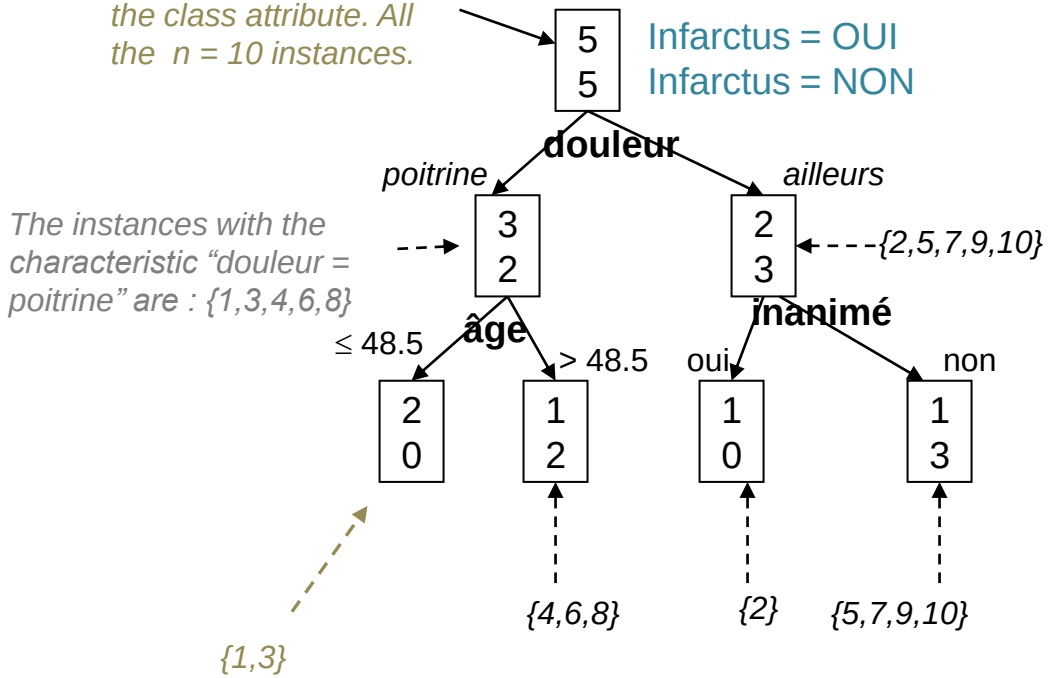
Numéro	Infarctus	Douleur	Age	Inanimé
1	oui	poitrine	45	oui
2	oui	ailleurs	25	oui
3	oui	poitrine	35	non
4	oui	poitrine	70	oui
5	oui	ailleurs	34	non
6	non	poitrine	60	non
7	non	ailleurs	67	non
8	non	poitrine	52	oui
9	non	ailleurs	58	non
10	non	ailleurs	34	non

Y X

Problems to solve:

- choosing the best splitting attribute at each node
- determining the best cut point when we handle continuous attribute
- what stopping rule for the decision tree growing (more generally, how to determine the right size of the tree)
- what is the best conclusion for a rule (leaf)

Absolute frequencies for the class attribute. All the n = 10 instances.



The instances with the characteristic "douleur = poitrine" are : {1,3,4,6,8}

This leaf (terminal node) is homogenous regarding the class attribute "Infarctus". All the instances are "Infarctus = oui". We can extract the prediction rule :

IF "douleur = poitrine" **AND** "age ≤ 48.5" **THEN** "Infarctus = oui"

Choosing the splitting attribute

Choosing the descriptor X* which is the most related to the target attribute Y.

Another point of view is "choosing the splitting attribute so that the induced subgroups are the most homogenous as possible on average"

- The chi-square (χ^2) statistic for contingency table can be used
- Actually, various measures of association can be used (based on Gini Impurity, Shannon entropy)

	$x_{i,1}$	...	x_{i,L_i}
Y_1			
⋮		$n_{k,l} = card(\{\omega \in \Omega_a / Y(\omega) = Y_k \text{ et } X_i(\omega) = X_{i,l}\})$	
Y_K			

Cross tabulation between Y and X

Selection process

$$X^* = \arg \max_{i=1,\ldots,p} \chi^2_{Y,X_i}$$

Improvement : χ^2 mechanically increases with
- n, number of instances into the node
- number of rows of the table
- number of columns of the table

these values are the same whatever the descriptors that we evaluate

the measure must not be biased in favor of the multi-way splits !

A possible solution : Tschuprow's t

(descriptors with a high number of values are penalized)

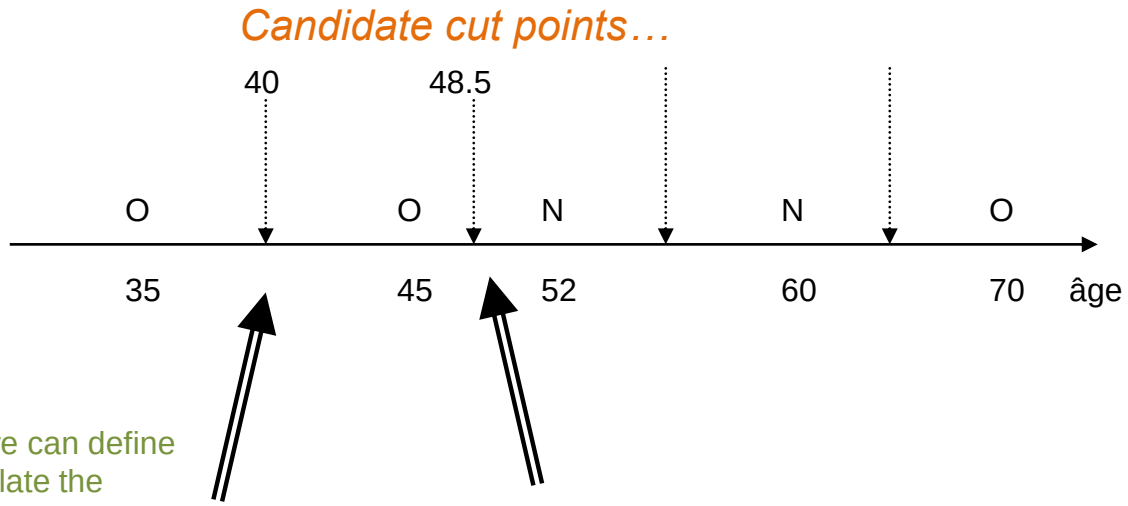
$$t^2_{Y,X_i} = \frac{\chi^2_{Y,X_i}}{n\sqrt{(K-1)(L_i-1)}}$$

(t ≥ 0 and t ≤ 1)

Continuous descriptors : determining the best “cut point”

How to choose the best "cut point" for the discretization of a continuous attribute ?

(e.g. how was determined the value 48.5 in the decision tree above ?)



For each possible cut point, we can define a contingency table and calculate the goodness of split

	$age < 40$	$age \geq 40$
$Inf. = oui$	1	2
$Inf. = non$	0	2

$\chi^2_{Infarctus,Age<40}$

	$age < 48.5$	$age \geq 48.5$
$Inf. = oui$	2	1
$Inf. = non$	0	2

$\chi^2_{Infarctus,Age<48.5}$

…

The “cut point” for the variable X

- must be located between two successive values of the descriptor
- enables to partition the data and defines a contingency table

The “best cut-point” maximizes the association between X and Y !

Stopping rule – Pre-pruning

Which reasons allow to stop the growing process?

Group homogeneity : confidence criterion

Confidence threshold (e.g. a node is considered homogenous if the relative frequency of one of the groups is higher than 98%)

Size of the nodes : support criterion

Min. size node to split (e.g. a node with less than 10 instances is not split)
Min. instances in leaves (e.g. a split is accepted if and only if each of the generated leaves contains at least 5 instances)

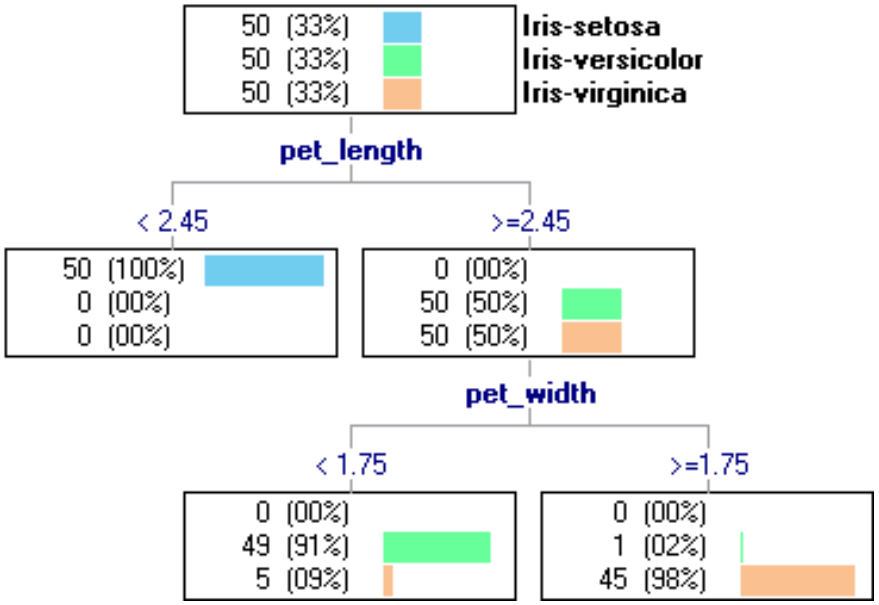
Chi-square test of independence: a statistical approach

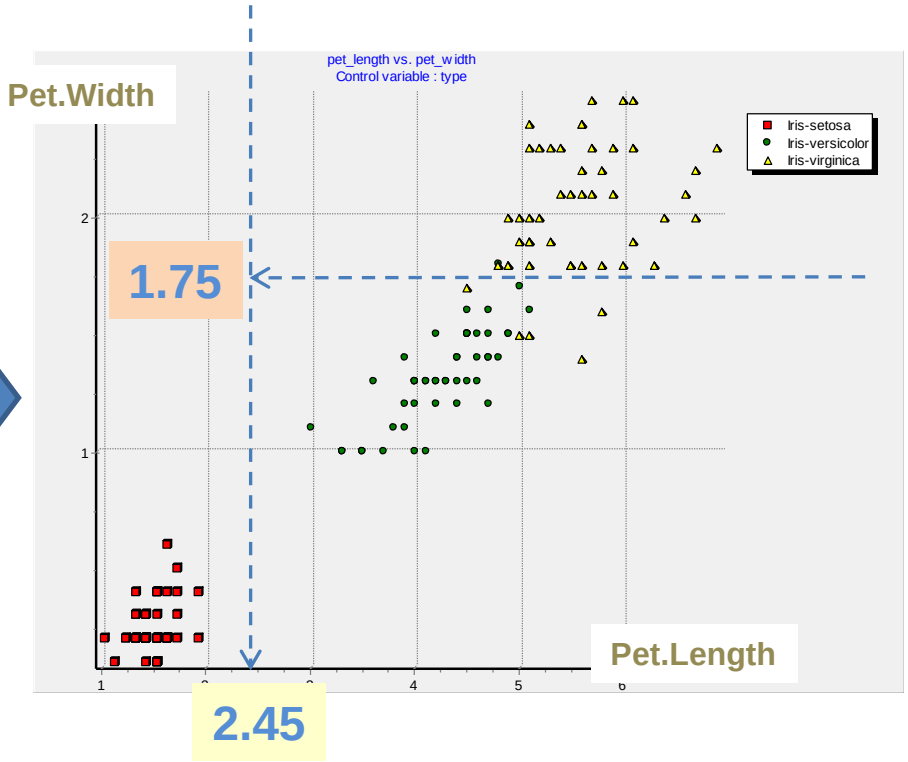
$$\begin{cases} H_0 : Y \text{ and } X^* \text{ are independent} \\ H_1 : Y \text{ and } X^* \text{ are not independent} \end{cases}$$

But the null hypothesis is very often rejected, especially when we deal with a large dataset. We must set a very low significance level.

The idea is above all to control the size of the tree and avoid the overfitting problem.

An example – Fisher's Iris Dataset (using Sipina Software)





Advantages and shortcomings of the Decision Trees

Advantages :

- Intelligible model - The domain expert can understand and evaluate.
- Direct transformation of the tree into a set of rules without loss of information.
- Automatic selection of the relevant variables.

- Nonparametric method.
- Handling both continuous and discrete attributes.
- Robust against outliers.
- Can handle large database.

- Interactive construction of the tree. Integration of domain knowledge.

Shortcomings :

- Data fragmentation on small dataset. High variance.
- Because its greedy characteristic, some interactions between variables can be missed (e.g. a tree can represent the XOR problem but no algorithm can find it).
- A compact representation of a complex underlying concept is sometimes difficult.

References

- "Classification and Regression Trees", L. Breiman, J. Friedman, R. Olshen and C. Stone, 1984.
- "C4.5: Programs for machine learning", R. Quinlan, Morgan Kaufman, 1993.
- "Induction graphs : machine learning and data mining", D. Zighed and R. Rakotomalala, Hermès, 2000 (in French).