Some enhancements in Decision Trees

Djamel A. Zighed

ERIC Laboratory
University of Lyon 2 (France)
zighed@univ-lyon2.fr

Outline

1. Decision tree / some recalls
2. Insensitivity of the criterion to the sample size
3. Entropy measure sensitive to the sample size
4. Lattice for mining small sample
5. Example on brandy control
6. Conclusion

1. Some recalls / Decision tree

Classification Task

Let's consider a learning sample Ω;

The size of Ω : $|\Omega| = n$

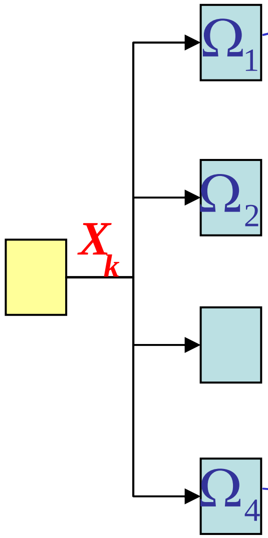
$\forall \omega \in \Omega; X(\omega) = \left(X_1(\omega), \cdots, X_p(\omega)\right)$: Predictive attributes

$\forall \omega \in \Omega; C(\omega)$: Predicted attribute; $C(\omega) \in \{c_1, c_2, \cdots, c_m\}$

$$\boxed{\Omega} \longrightarrow \begin{array}{|c|c|} c_1 = & n_1 \\ \vdots & \vdots \\ c_j = & n_j \\ \vdots & \vdots \\ c_m = & n_m \end{array}$$

X_k : Brings about a partition Π_1 on Ω

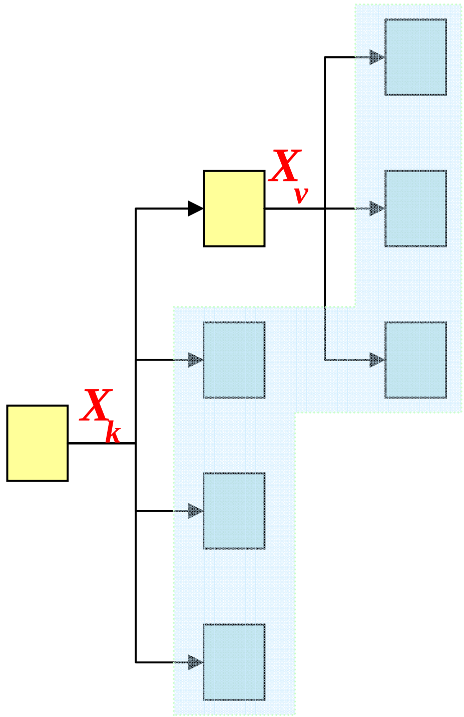
$$\begin{array}{|c|c|} \hline c_1/\Omega_1 = & n_{11} \\ \vdots & \vdots \\ c_j/\Omega_1 = & n_{1j} \\ \vdots & \vdots \\ c_m/\Omega_1 = & n_{1m} \\ \hline \end{array}$$



$\theta(\Pi_1)$ is the quality of the partition Π_1 brought about by X_k

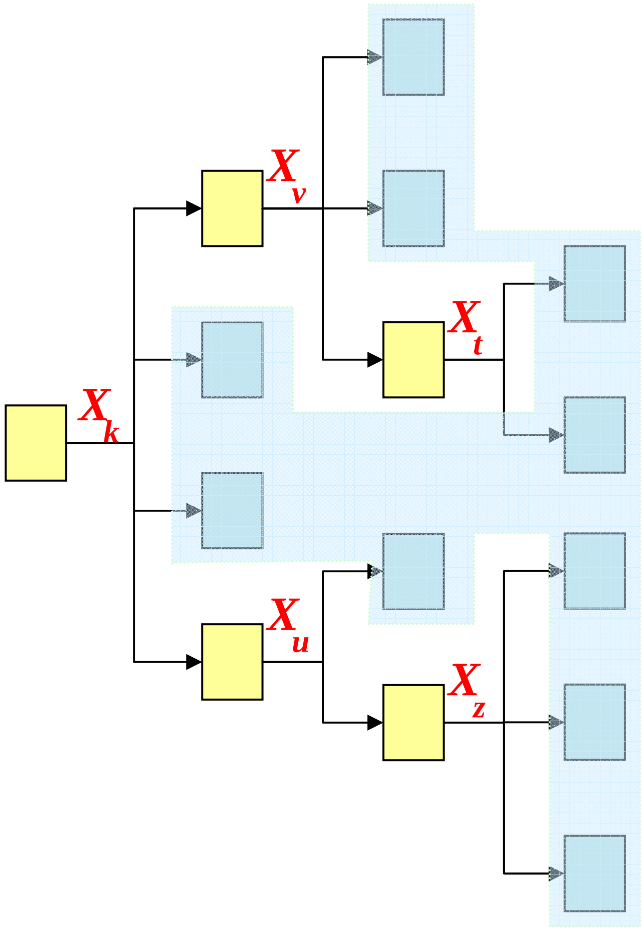
$$\begin{array}{|c|c|} \hline c_1/\Omega_4 = & n_{41} \\ \vdots & \vdots \\ c_j/\Omega_4 = & n_{4j} \\ \vdots & \vdots \\ c_m/\Omega_4 = & n_{4m} \\ \hline \end{array}$$

$X_k ; X_v$: Bring about a partition Π_2 on Ω



$\theta(\Pi_2)$

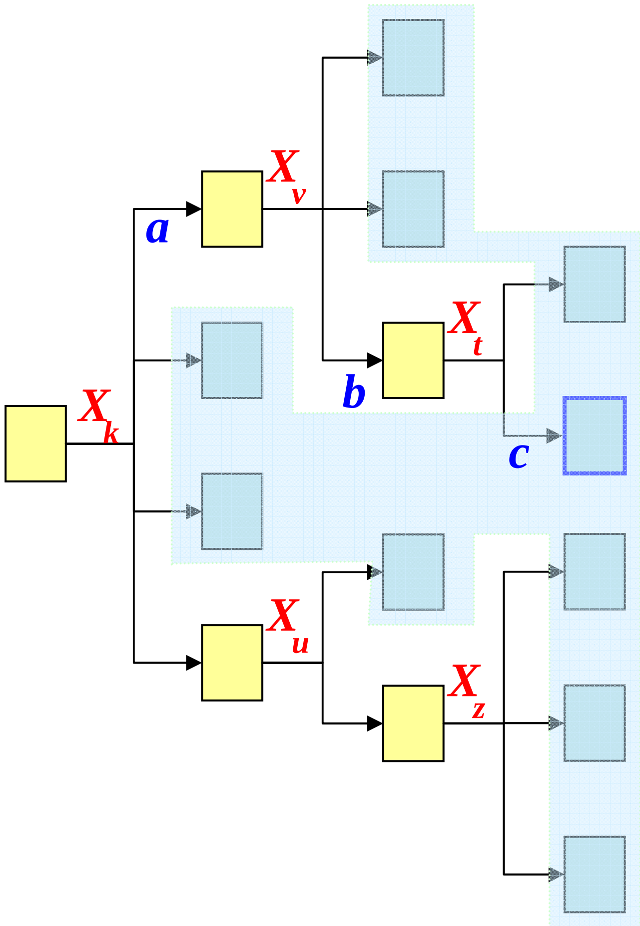
$X_k; X_v, X_t, X_u, X_z$: Bring about a partition Π_s on Ω



The size : $|\Pi_s| = \pi = 10$

$\theta(\Pi_s)$

Criterion for Growing the tree



$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

were

w_s is the weight of s

and

$h(s)$ i s the purity of s

$$h(s) = -\sum_{i=1}^{m} p(c_i / s) \log p(c_i / s)$$

$$h(s) = \sum_{i=1}^{m} p(c_i / s)(1 - p(c_i / s))$$

w_s = The probability of the path s :

$$w_s = P(X_k = x; X_v = y, \cdots) = \frac{|\Pi_s|}{|\Pi|}$$

$$p(c_i / s) = \frac{n_{is}}{n_{.s}}$$

Entropy measure :

$$h : \Pi \to IR^{+}$$

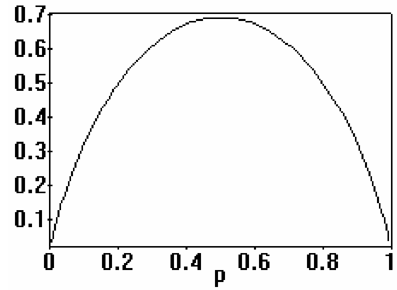
$$\Pi = \bigcup_{n} s_n$$

$$s_n = \left\{ \gamma \in R^n / \gamma_i \geq 0 (i = 1,..., n) \text{ and } \sum_{i=1}^{n} \gamma_i = 1 \right\}$$

- **Symetry** $h(\gamma_1, \gamma_2, ..., \gamma_n) = h(\gamma_{\sigma_1}, \gamma_{\sigma_2}, ..., \gamma_{\sigma_n})$
- **Minimality** $h(\gamma_1, \gamma_2, ..., \gamma_n) = 0 \Leftrightarrow \exists i / \gamma_i = 1 \ \forall j \neq i \ ; \gamma_j = 0$
- **Maximality** $h(\gamma_1, \gamma_2, ..., \gamma_n) = Max \Leftrightarrow \forall i \ (i = 1, ..., n) \ ; \gamma_i = \dfrac{1}{n}$
- **Continuity**
- **Concavity**

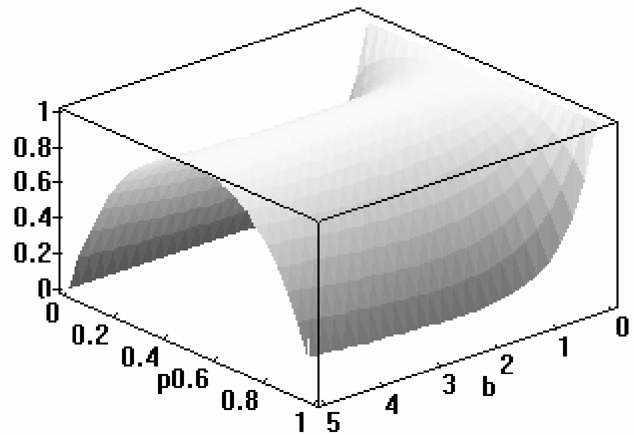
$$h(s) = h(p_1, p_2)$$

$$h(s) = -\sum_{i=1}^{2} p_i \log p_i$$



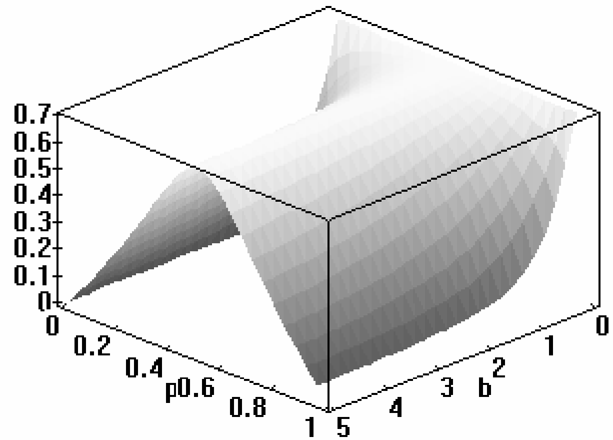
$$h(s) = \frac{1}{2^{\beta-1}-1}\left[\left(\sum_{i=1}^{2} p_i^{\beta} - 1\right)\right]$$

$$\beta > 0;\ \beta \neq 1$$



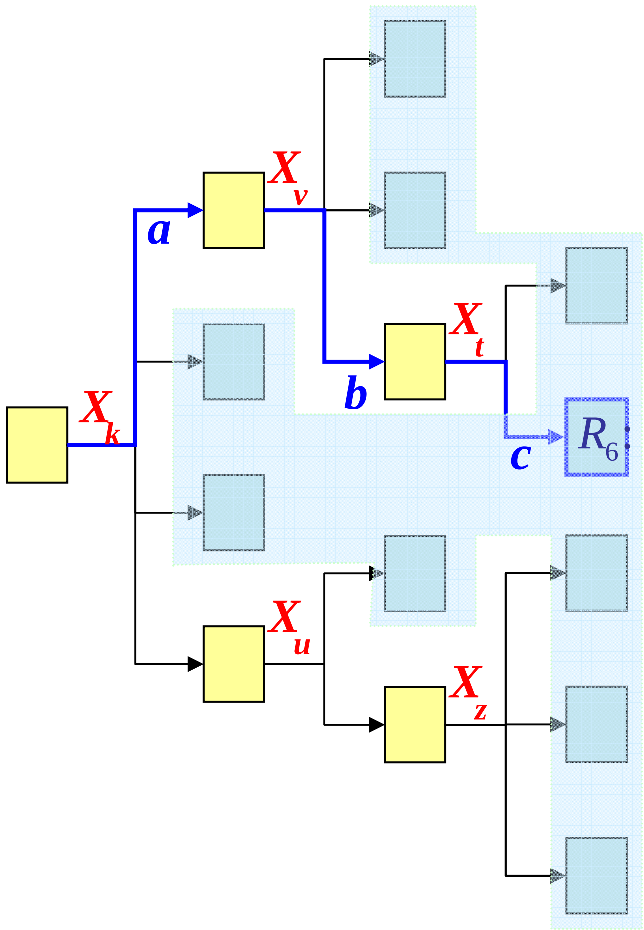
$$h(s) = \frac{1}{1-\beta}\log\left(\sum_{i=1}^{2} p_i^{\beta}\right)$$

$$\beta > 0;\ \beta \neq 1$$



Each partition Π on Ω is described by a model φ : a set of rules

R_i: If (condition) Then (Conclusion) (ε)



R_6: If $\big((X_k = a) \,\&\, (X_v = b) \,\&\, (X_t = c)\big)$ Then $(C = c_j)$ $(\varepsilon_6 = .10)$

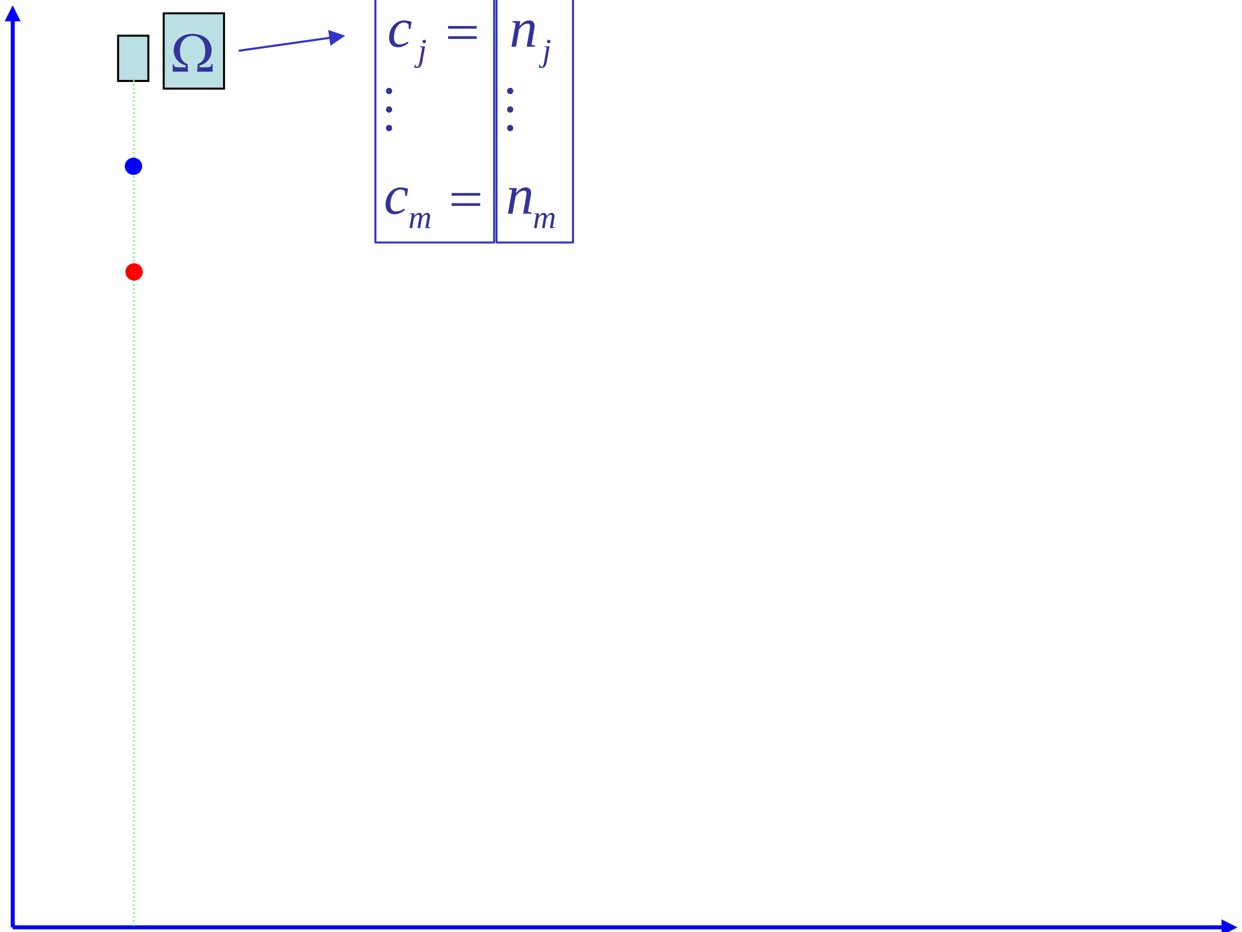
S is the set of terminal nodes

$$\varepsilon_s = 1 - f_{\max/s}$$

$$\mathrm{E}_{\Pi} = \sum_{s \in S} f_s \varepsilon_s$$

$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

$$E_{\Pi} = \sum_{s \in S} f_s \varepsilon_s$$



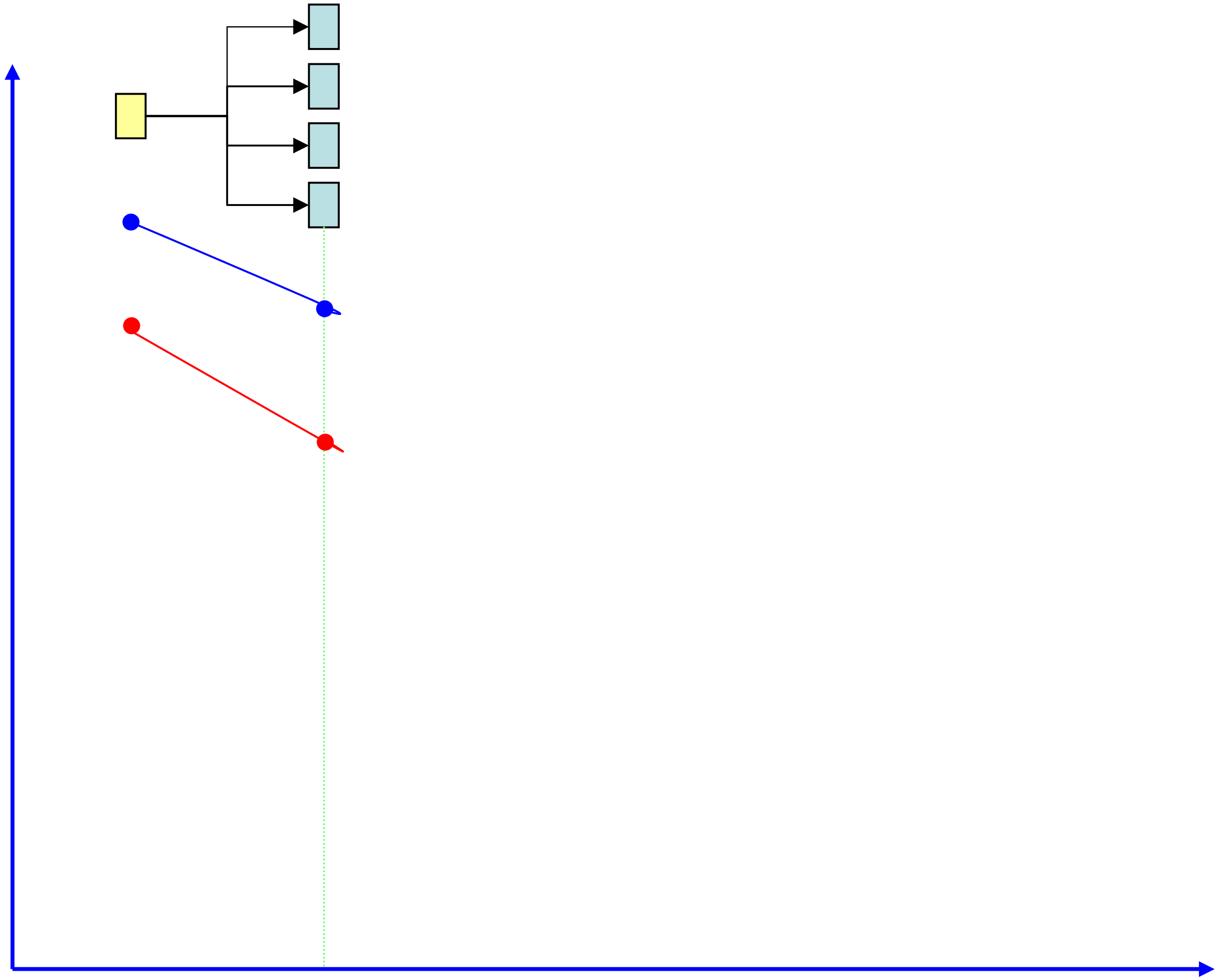
$$\begin{array}{|c|c|} c_1 = & n_1 \\ \vdots & \vdots \\ c_j = & n_j \\ \vdots & \vdots \\ c_m = & n_m \end{array}$$

$$|\Pi_1| = 1$$

Size of the partitions
Number of Terminal Nodes
Number of rules

$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

$$\mathrm{E}_{\Pi} = \sum_{s \in S} f_s \varepsilon_s$$

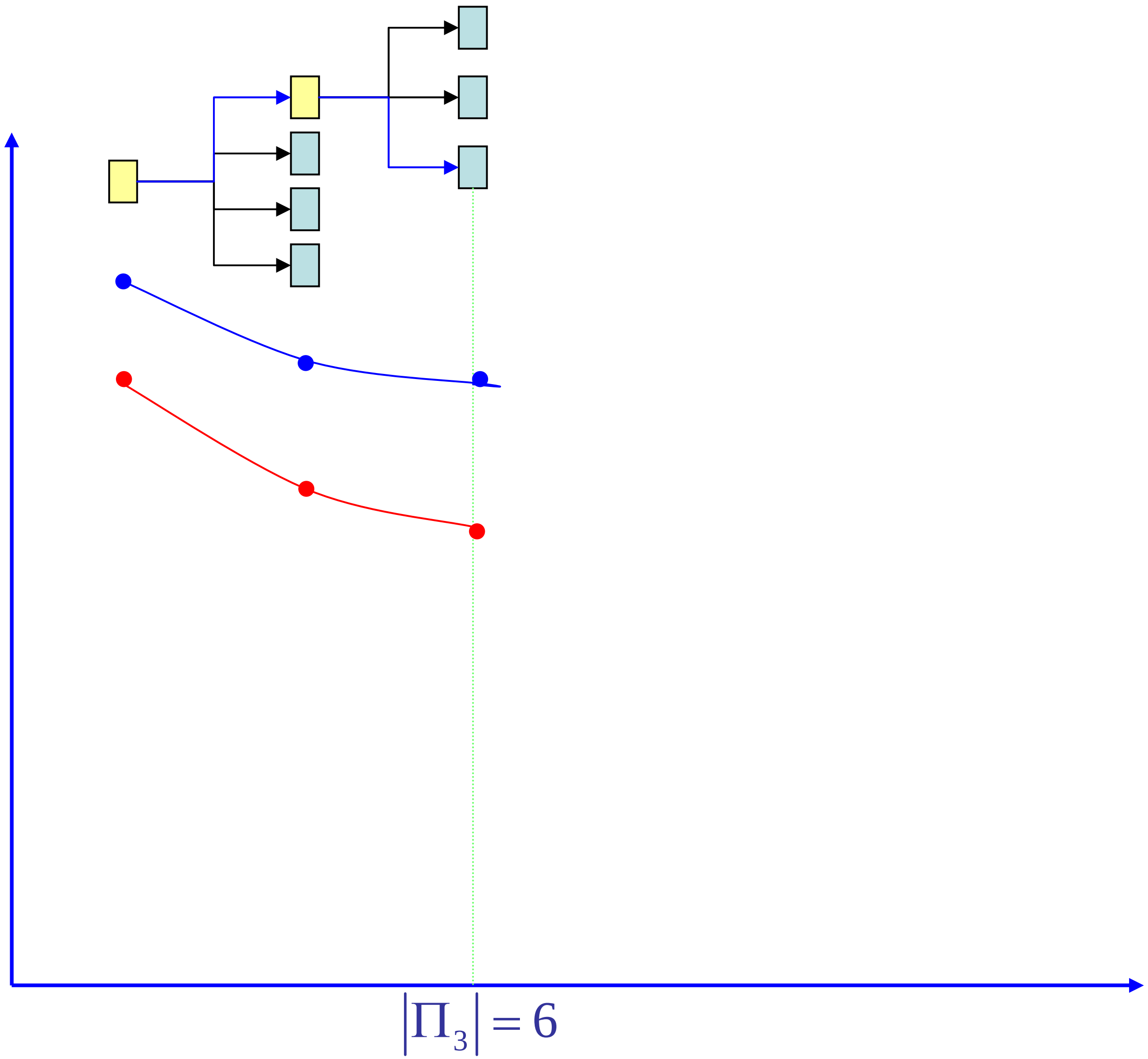


$$|\Pi_2| = 4$$

Size of the partitions
Number of Terminal Nodes
Number of rules

$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

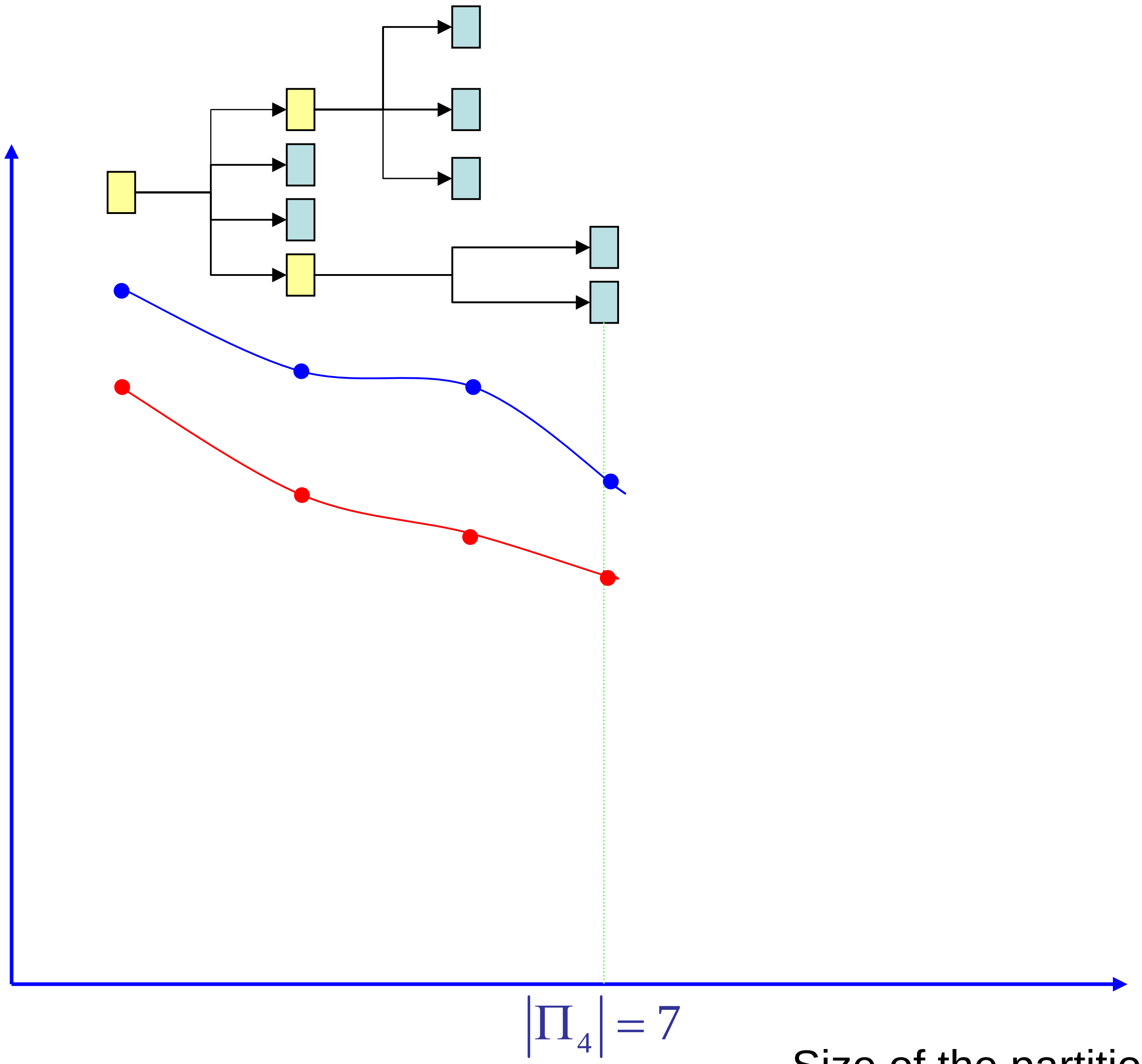
$$\mathrm{E}_{\Pi} = \sum_{s \in S} f_s \varepsilon_s$$



Size of the partitions
Number of Terminal Nodes
Number of rules

$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

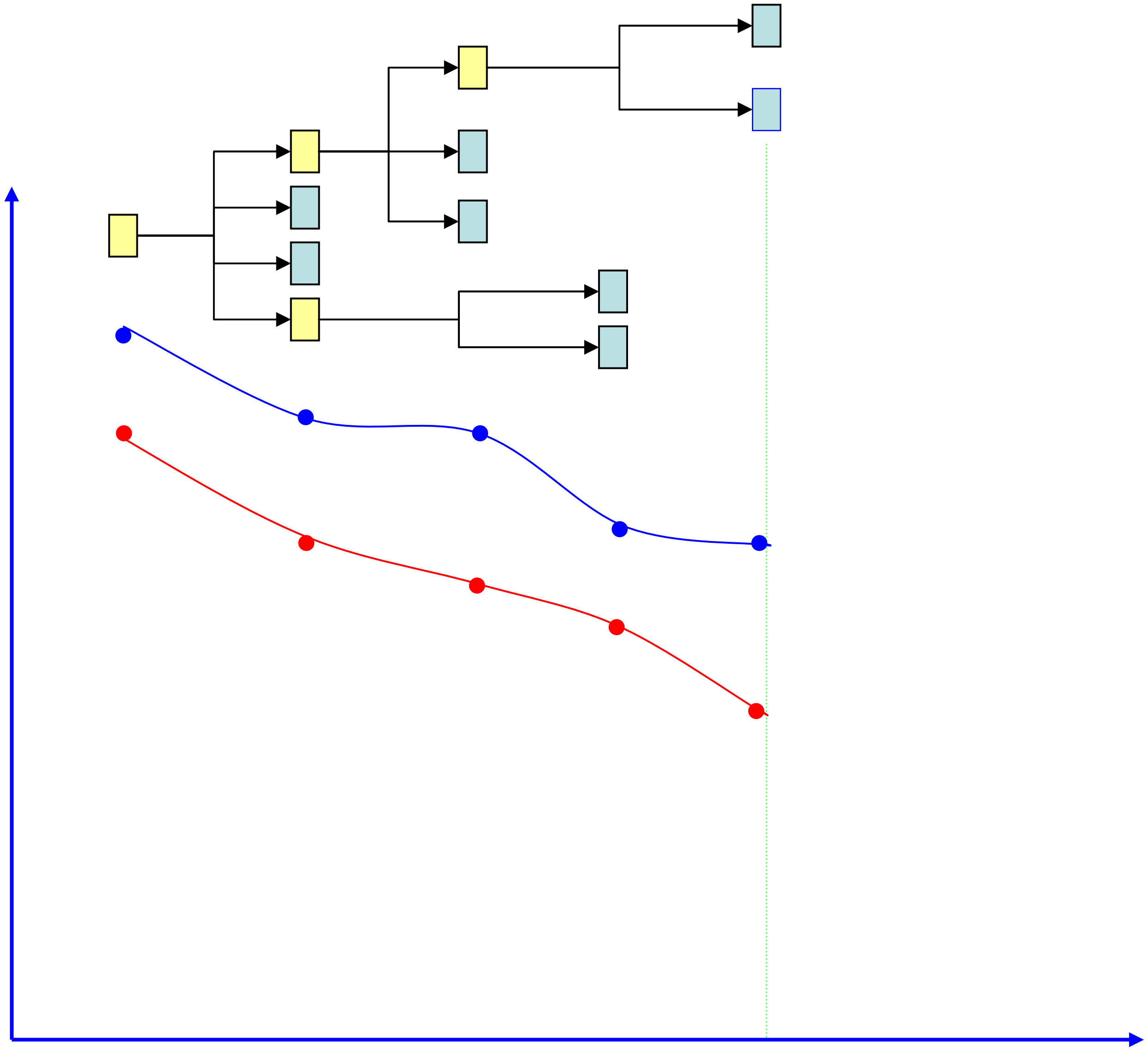
$$E_\Pi = \sum_{s \in S} f_s \varepsilon_s$$



Size of the partitions
Number of Terminal Nodes
Number of rules

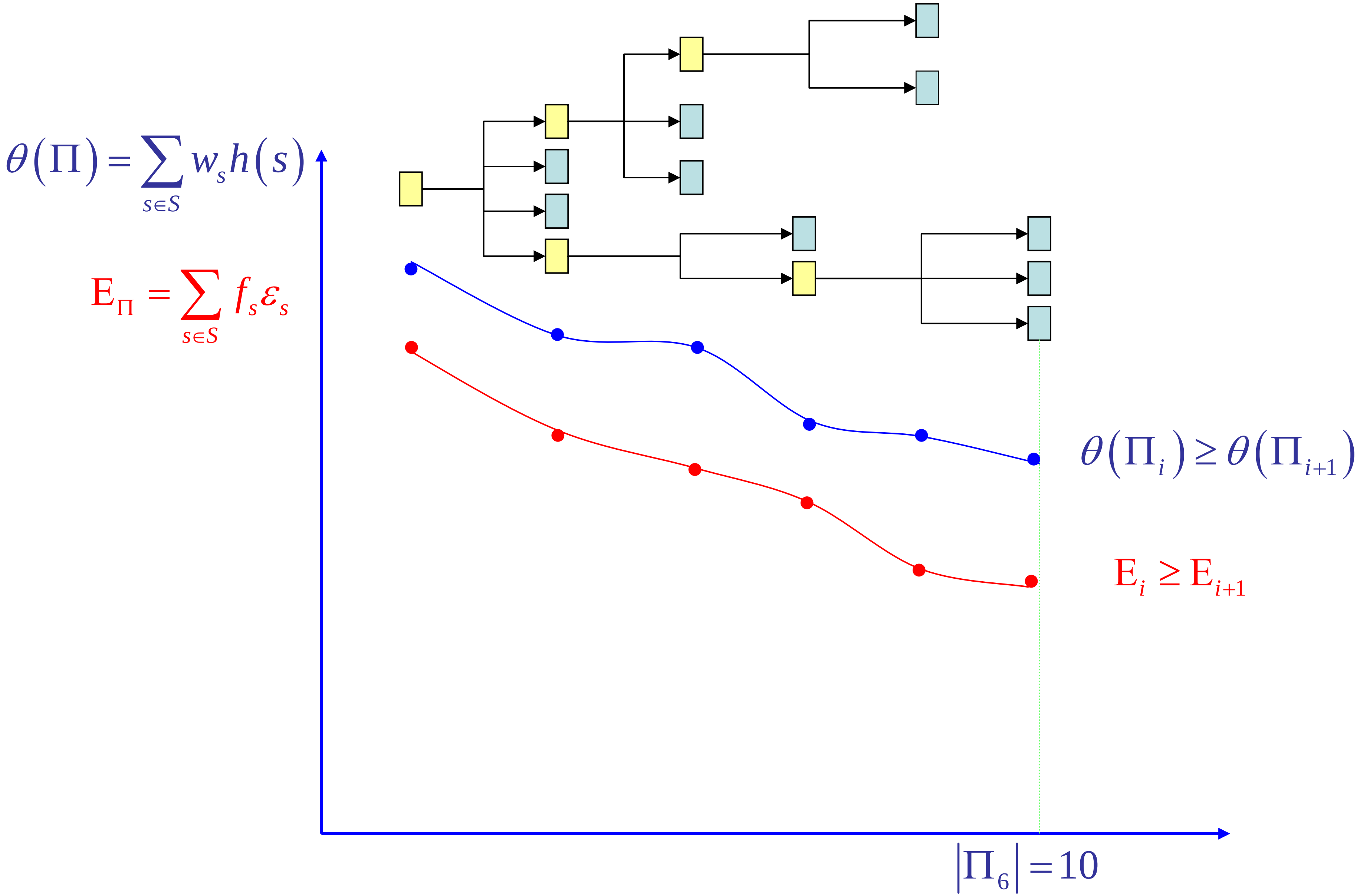
$$\theta(\Pi_s) = \sum_{s \in S} w_s h(s)$$

$$E_\Pi = \sum_{s \in S} f_s \varepsilon_s$$



$$|\Pi_5| = 8$$

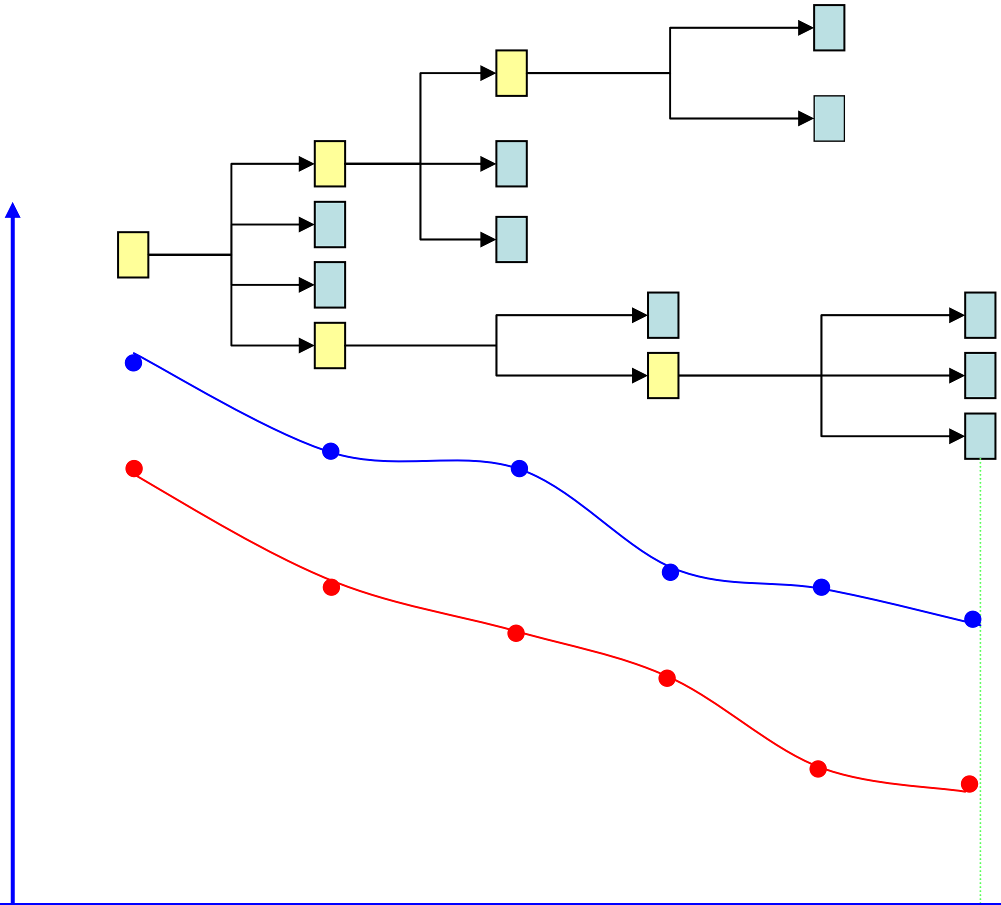
Size of the partitions
Number of Terminal Nodes
Number of rules



Size of the partitions
Number of Terminal Nodes
Number of rules

$$\theta(\Pi) = \sum_{s \in S} w_s h(s)$$

$$E_{\Pi} = \sum_{s \in S} f_s \varepsilon_s$$



$$\theta(\Pi_i) \geq \theta(\Pi_{i+1})$$

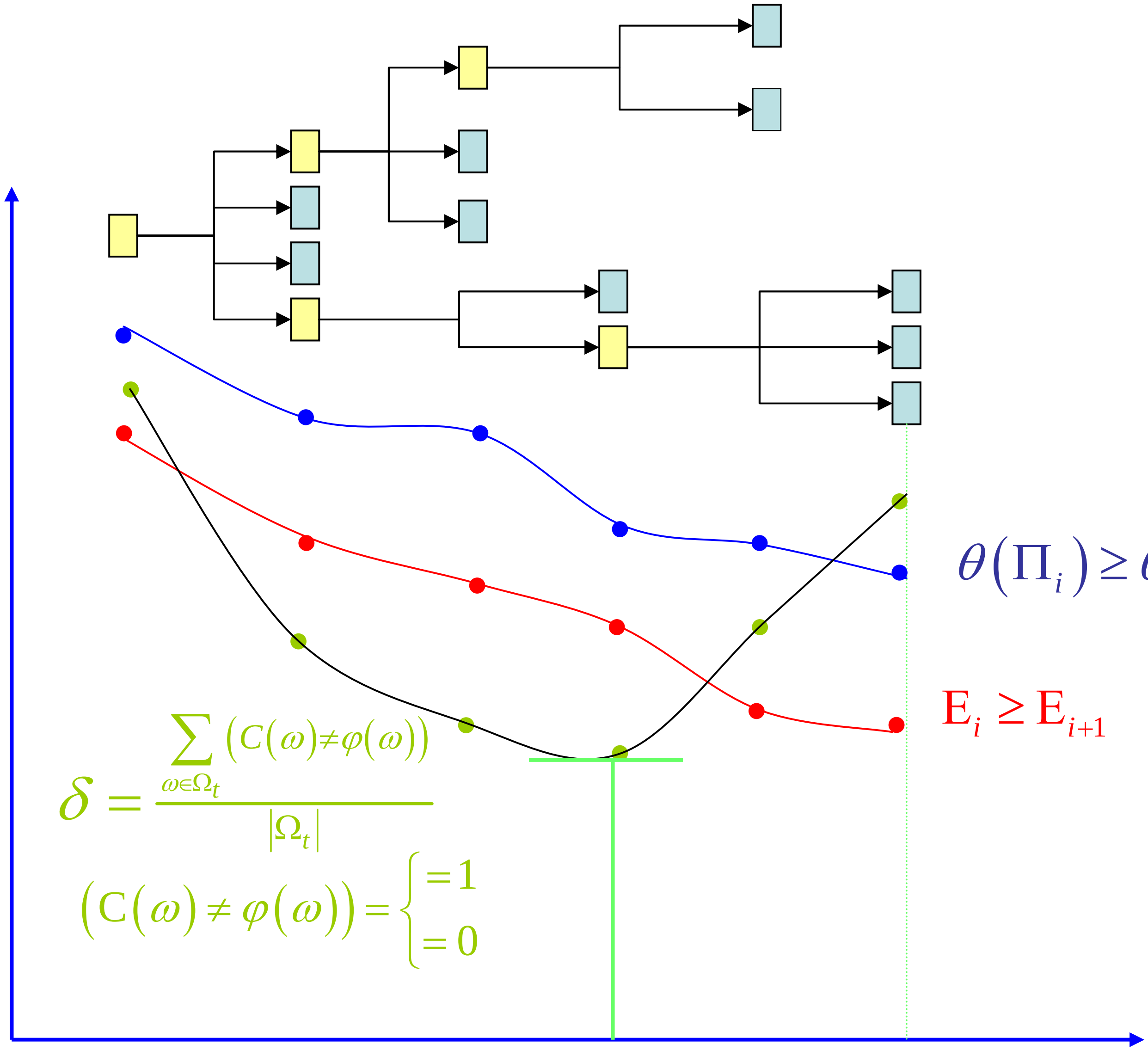
$$E_i \geq E_{i+1}$$

The largest tree will have the lowest value of θ and error E,
Shall we consider the predictive model associated to the largest tree as the best one ?

$$\theta(\Pi) = \sum_{s \in S} w_s h(s)$$

$$E_\Pi = \sum_{s \in S} f_s \varepsilon_s$$

Ω_t sample test



$$\delta = \frac{\sum_{\omega \in \Omega_t} \left(C(\omega) \neq \varphi(\omega)\right)}{|\Omega_t|}$$

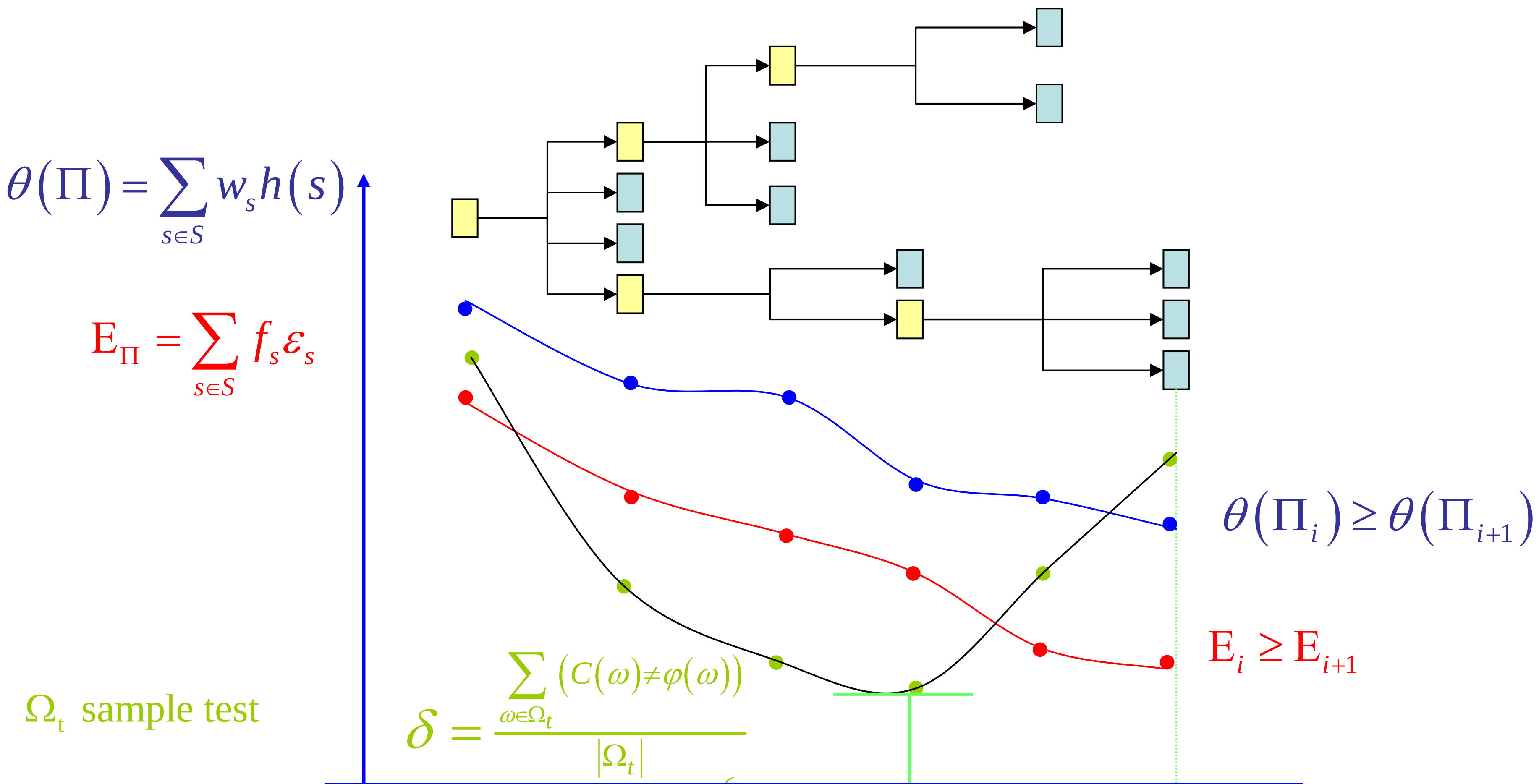
$$\left(C(\omega) \neq \varphi(\omega)\right) = \begin{cases} = 1 \\ = 0 \end{cases}$$

$$\theta(\Pi_i) \geq \theta(\Pi_{i+1})$$

$$E_i \geq E_{i+1}$$

Π_*

Size of the partitions
Number of Terminal Nodes
Number of rules

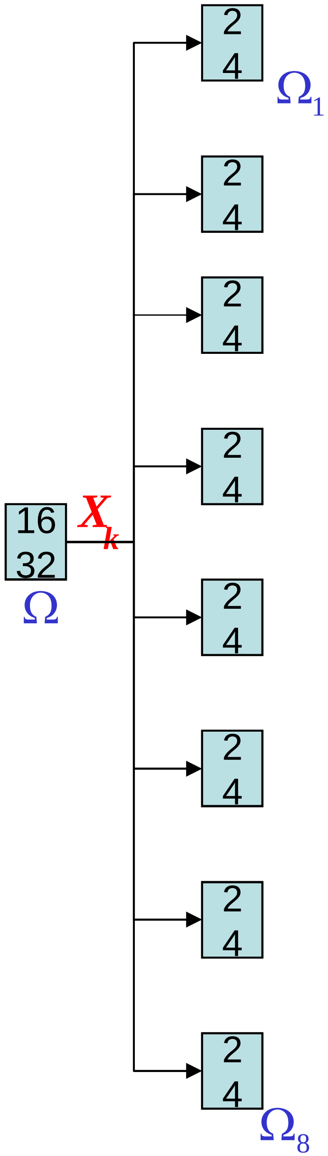


$$\theta(\Pi) = \sum_{s\in S} w_s h(s)$$

$$E_\Pi = \sum_{s\in S} f_s \varepsilon_s$$

aren't suitable for building the right tree

2. Insensitivity of the criterion to the sample size

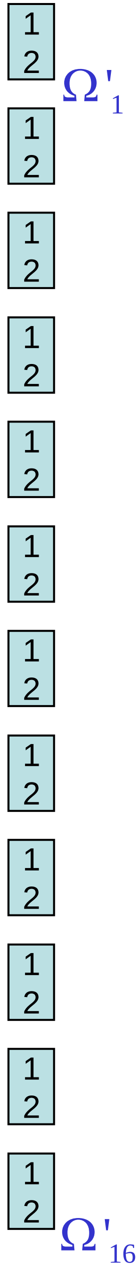


$$h(\Omega) = \sum_{i=1}^{8} w_i h(\Omega_i) = \sum_{i=1}^{16} w'_i \, h(\Omega'_i)$$

The probabilities used for the calculation of the criteria are estimated locally by the frequencies without taking into account the fact that the sample size at each node is decreasing.

In the other hand, we assume that

$$\left(\frac{16}{48};\frac{32}{48}\right) \equiv \left(\frac{2}{6};\frac{4}{6}\right) \equiv \left(\frac{1}{3};\frac{2}{3}\right)$$



The insensitivity of the criteria to the sample size may lead to different assumptions and solutions :

- We don't need any criteria nor algorithm for growing the tree, develop the largest one, we assume that there is a large data set in each node to make the estimates of the probabilities reliable ;
- Suppose that the predictive attributes are independent then, we can get a reliable estimates of the probabilities at each node by applying Bayes Formula;
- Start from the largest tree and apply one of the pruning techniques to find the right tree ;
- Fix a minimum size in each node before splitting;
- Introduce a penalty parameter for the complexity of the tree (complexity is the size of current partition), and allow merging modalities of predictive attributes, thus we better control the size of the current partition, we better use the sample;
- Use statistical criteria to stop the growing process at the right deep;

In all of these cases, the criteria for growing the tree remains insensitive to the sample size.

We aim to formulate a new criteria which depends on both the frequencies vector and the size of the sample on which the frequencies are estimated, thus we may avoid the over fitting without post treatment nor external parameterization.

3. Entropy measure sensitive to the sample size

$$\hat{h} : \Pi \rightarrow IR^{+} \qquad \Pi = \bigcup_n s_n$$

$$s_n = \left\{ \gamma \in R^n / \gamma_i \geq 0 (i = 1,..., n) \text{ and } \sum_{i=1}^{n} \gamma_i = 1 \right\}$$

- **Symetry** $\hat{h}(\gamma_1, \gamma_2, ..., \gamma_n) = \hat{h}(\gamma_{\sigma_1}, \gamma_{\sigma_2}, ..., \gamma_{\sigma_n})$
- **Minimality** $\hat{h}(\gamma_1, \gamma_2, ..., \gamma_n) = Min \Leftrightarrow \exists i / \gamma_i = 1 \forall j \neq i ; \gamma_j = 0$
- **Maximality** $\hat{h}(\gamma_1, \gamma_2, ..., \gamma_n) = Max \Leftrightarrow \forall i (i = 1, ..., n) ; \gamma_i = \frac{1}{n}$
- **Sensitivité to the sample size** $\hat{h}(\gamma_1, \gamma_2, ..., \gamma_n, n_1) > \hat{h}(\gamma_1, \gamma_2, ..., \gamma_n, n_2)$ if $n_1 < n_2$
- **Convergence** $\lim_{n \mapsto \infty} \hat{h}(\gamma_1, \gamma_2, ..., \gamma_n, n) \rightarrow h(p_1, p_2, ..., p_n)$
- **Concavity**

$$\hat{h}(\Omega) = \hat{h}(\gamma_1, \ldots, \gamma_i, \ldots, \gamma_m, n) = h(\hat{\gamma}_1, \ldots, \hat{\gamma}_i, \ldots, \hat{\gamma}_m)$$

were

h is any entropy measure

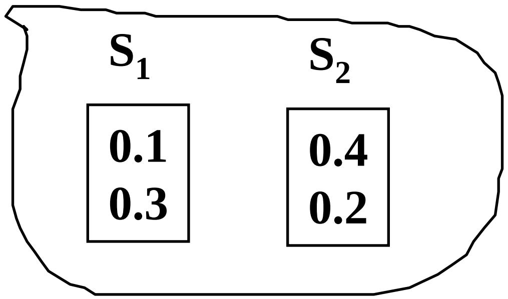
$$\hat{\gamma}_i = \frac{n_i + \lambda}{n + m.\lambda}$$

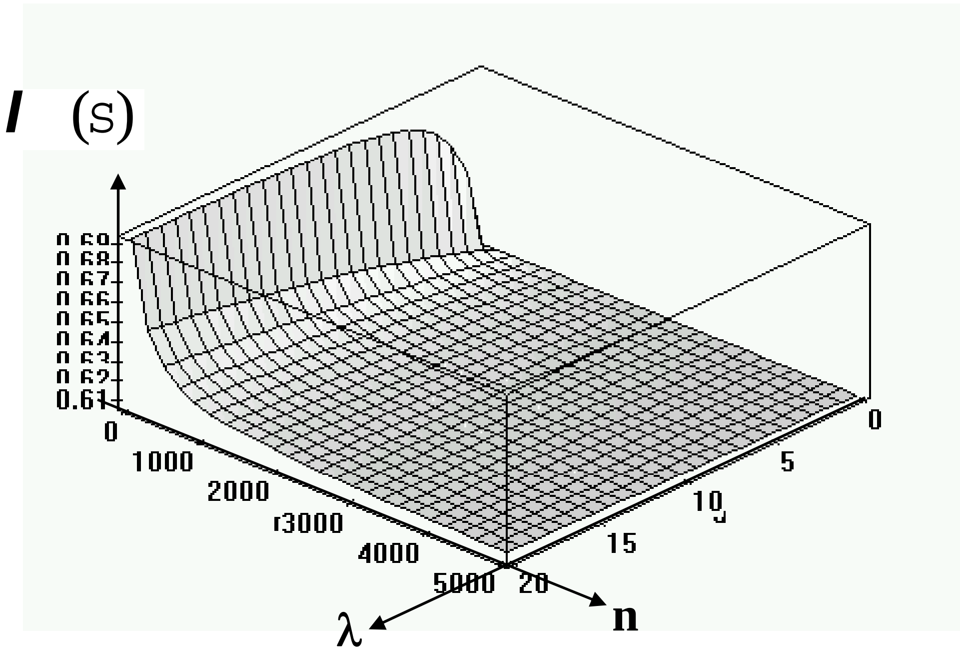
were $\lambda > 0$

$$\hat{h}(\Omega) = -\sum_{i=1}^{m} \frac{n_i + \lambda}{n + m.\lambda} \log \frac{n_i + \lambda}{n + m.\lambda}$$

$$\hat{h}(\Omega) = \sum_{i=1}^{m} \frac{n_i + \lambda}{n + m.\lambda} \left(1 - \frac{n_i + \lambda}{n + m.\lambda} \right)$$

$$\hat{h}(\Omega)=\sum_{j=1}^{k}\frac{n_{.j}}{n}\left(-\sum_{i=1}^{m}\frac{n_{ij}+\lambda}{n_{.j}+m\lambda}Log\left(\frac{n_{ij}+\lambda}{n_{.j}+m\lambda}\right)\right)$$





How to fix λ

τ ***soft constraint***

$$v = \begin{bmatrix} \tau+1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \qquad u = \begin{bmatrix} \tau \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

$$n_{.v} = \tau+1 \qquad n_{.u} = \tau$$

$$\lambda^* : \hat{h}\left(u_{\lambda^*}\right) - \hat{h}\left(v_{\lambda^*}\right) = \max_{\lambda}\left(\hat{h}\left(u_{\lambda}\right) - \hat{h}\left(v_{\lambda}\right)\right)$$

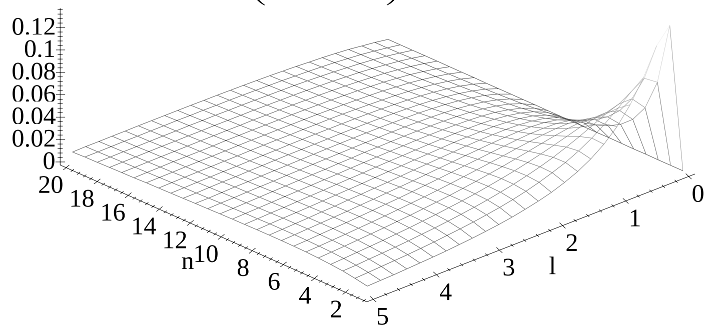
$$\begin{aligned} f(\lambda) &= \left(\hat{h}\left(u_{\lambda}\right) - \hat{h}\left(v_{\lambda}\right)\right) \\ &= (m-1)\frac{2\tau^2 + 2\tau + m\lambda(1+2\tau)}{(\tau+m\lambda)^2(\tau+1+m\lambda)^2} \end{aligned}$$

$$\text{Find } \lambda^* : f\left(\lambda^*\right) \geq f(\lambda); \lambda \neq \lambda^*$$

$$\frac{\partial f(\lambda)}{\partial \lambda} = 0$$

$m = 3$

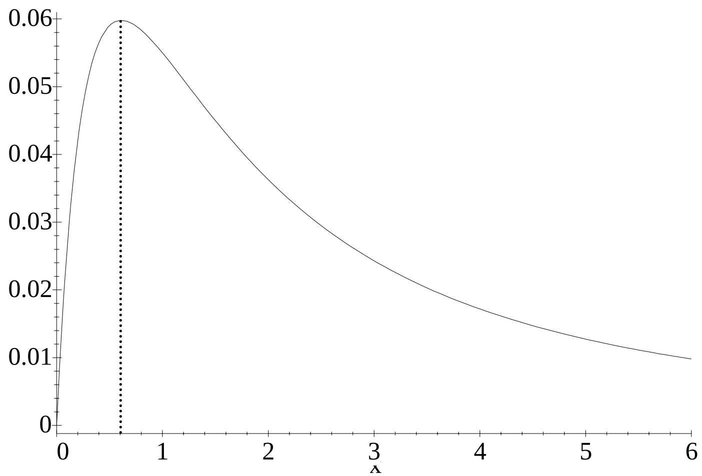
$$\varphi(\lambda, \tau)$$



$m = 2$

$\tau = 2$

$$f(\lambda) = \lambda \left(\frac{5\lambda + 6}{2\left(1 + \lambda^2 (3 + 2\lambda)^2\right)} \right)$$



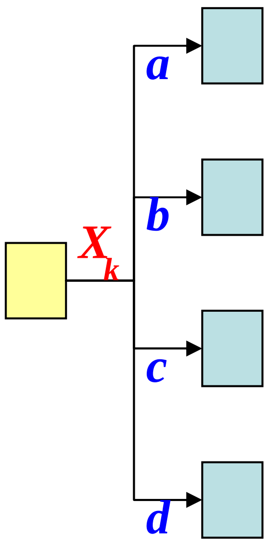
$\lambda^* = 0.61098$

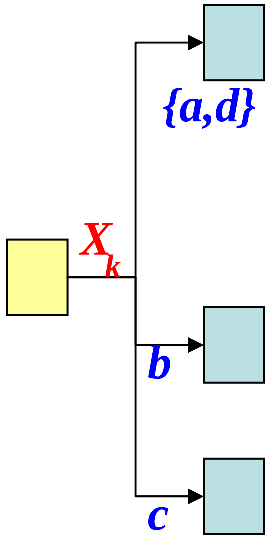
4. Lattice to learn from finite data set

Merging nodes reduces the size of the partition and allows deeper exploration

CHAID (Kass, 1980)

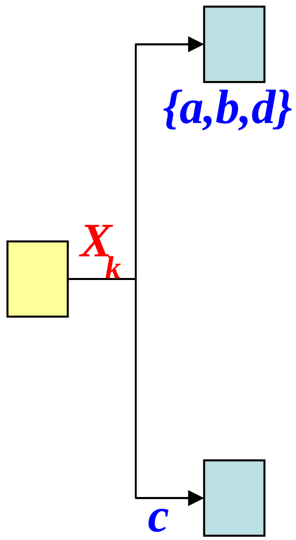
Looks for the best partition on the set of observed modalities of X_k in the current node





N-ary trees

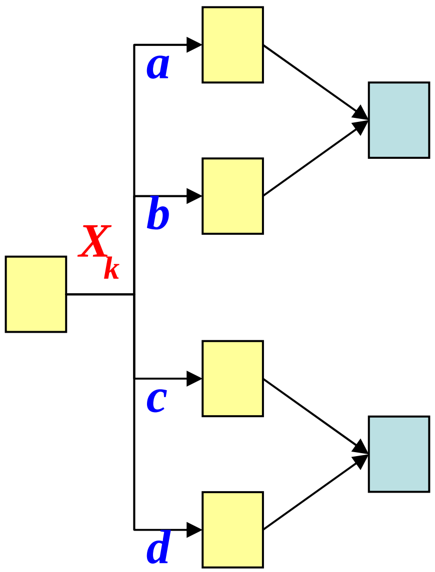
The merge is allowed only within brother nodes



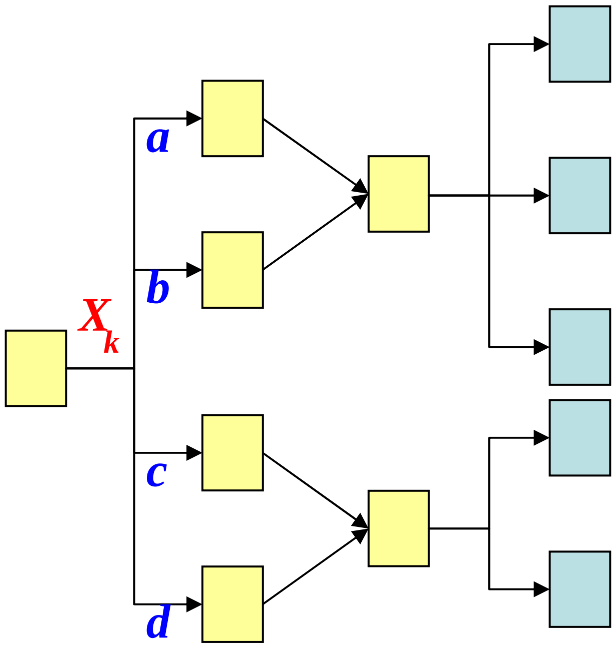
Binary trees

Looks for the the best bi-partition on the set of observed modalities of X_k in the current node

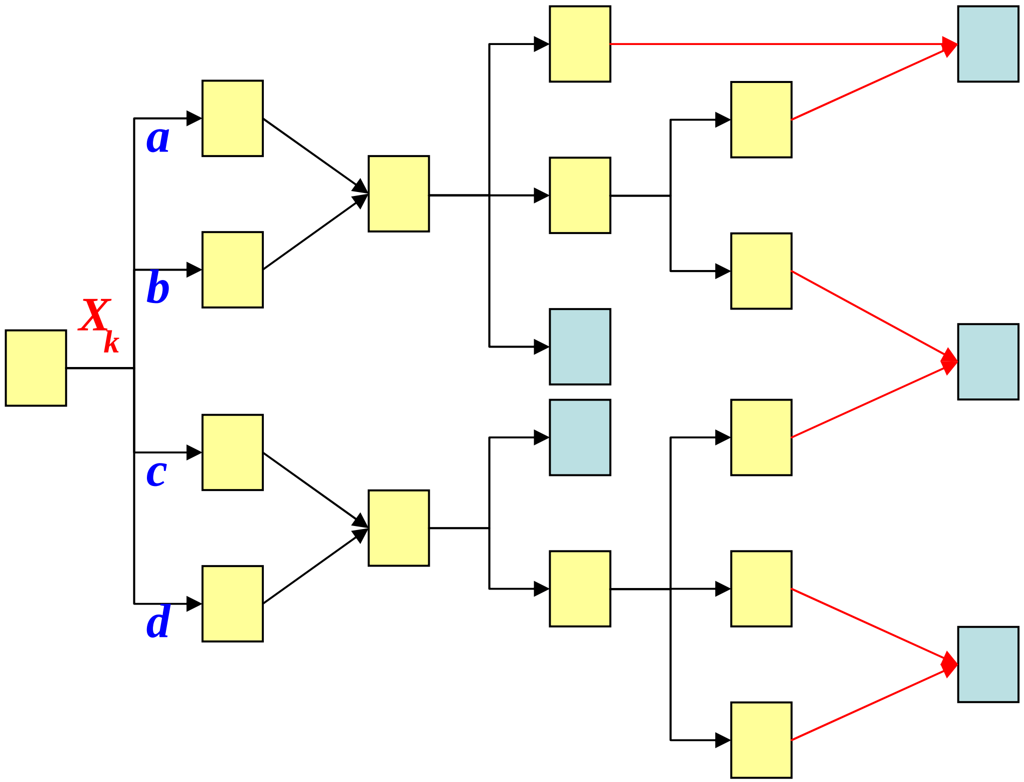
CART (Breiman et al. 84)



X_k
a
b
c
d

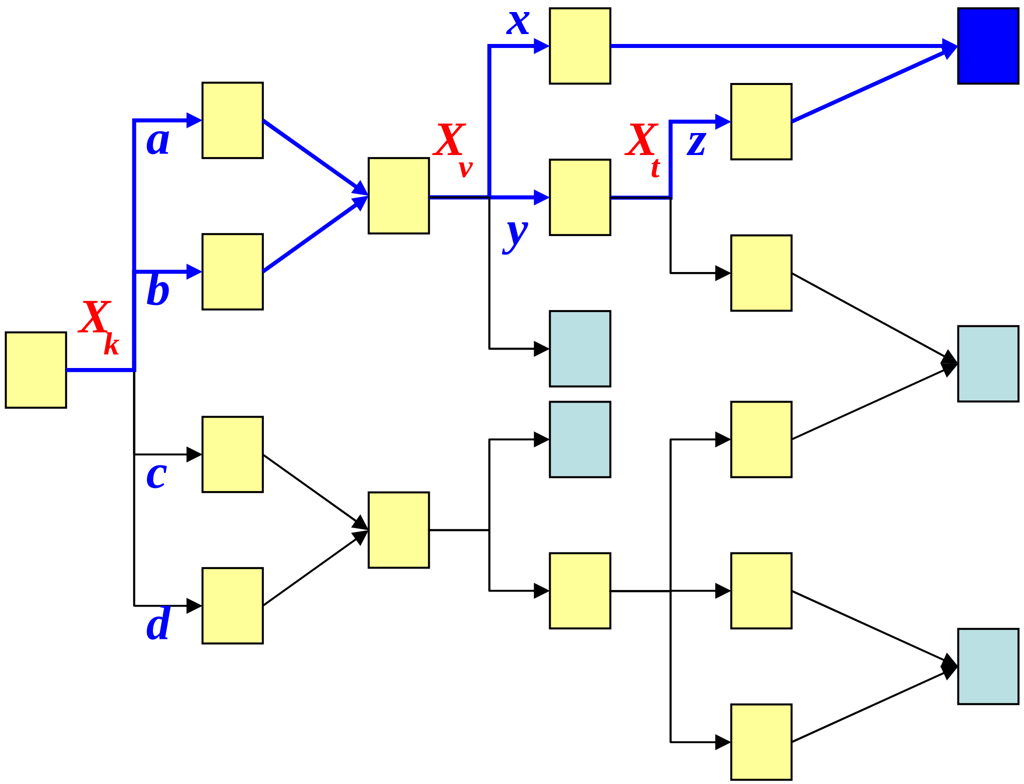


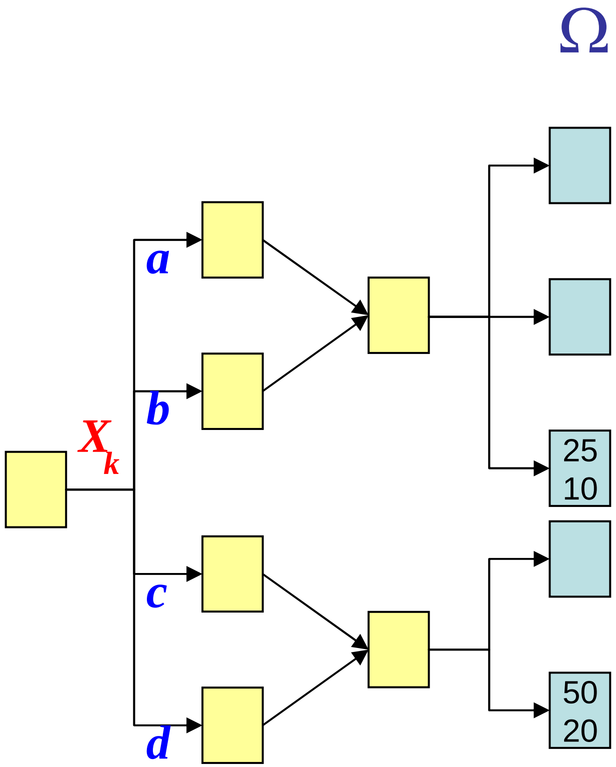
X_k
a
b
c
d



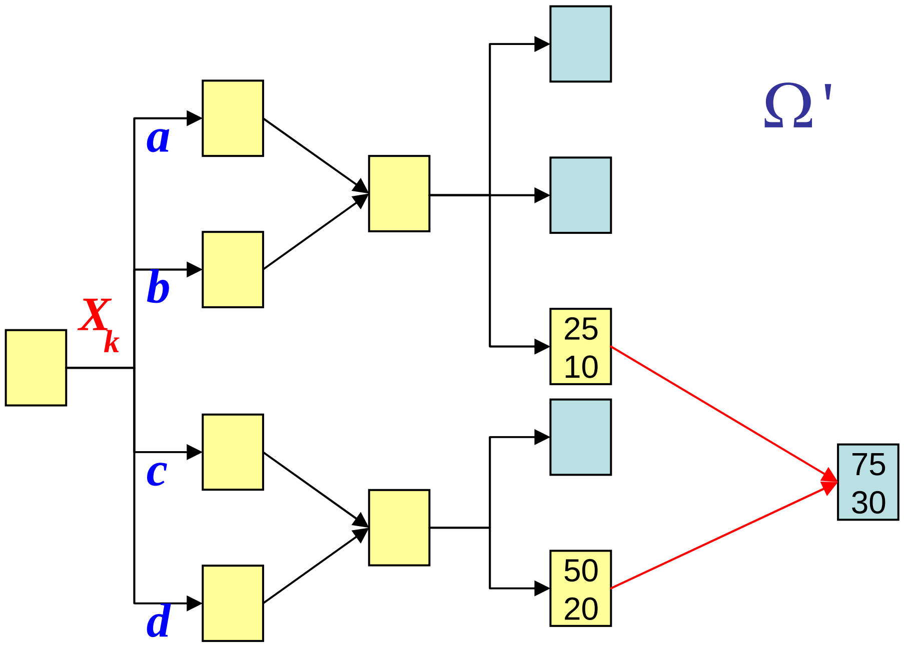
X_k
a
b
c
d

$$\text{If}\left[\left(X_k \in \{a,b\}\right) \wedge \left(\left(X_v = x\right) \vee \left(\left(X_v = y\right) \wedge \left(X_t = z\right)\right)\right)\right] \text{ Then } C = c_i$$





$$\hat{h}(\Omega) = \sum_{i=1}^{5} w_i \hat{h}(\Omega_i)$$



Ω'

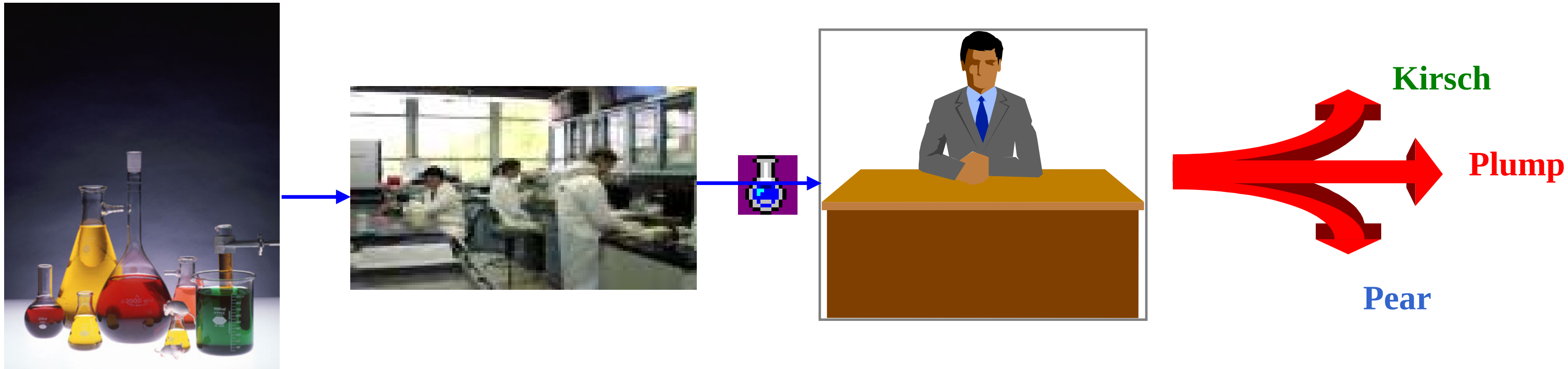
$$\hat{h}(\Omega') = \sum_{i=1}^{4} w_i \hat{h}(\Omega_i)$$

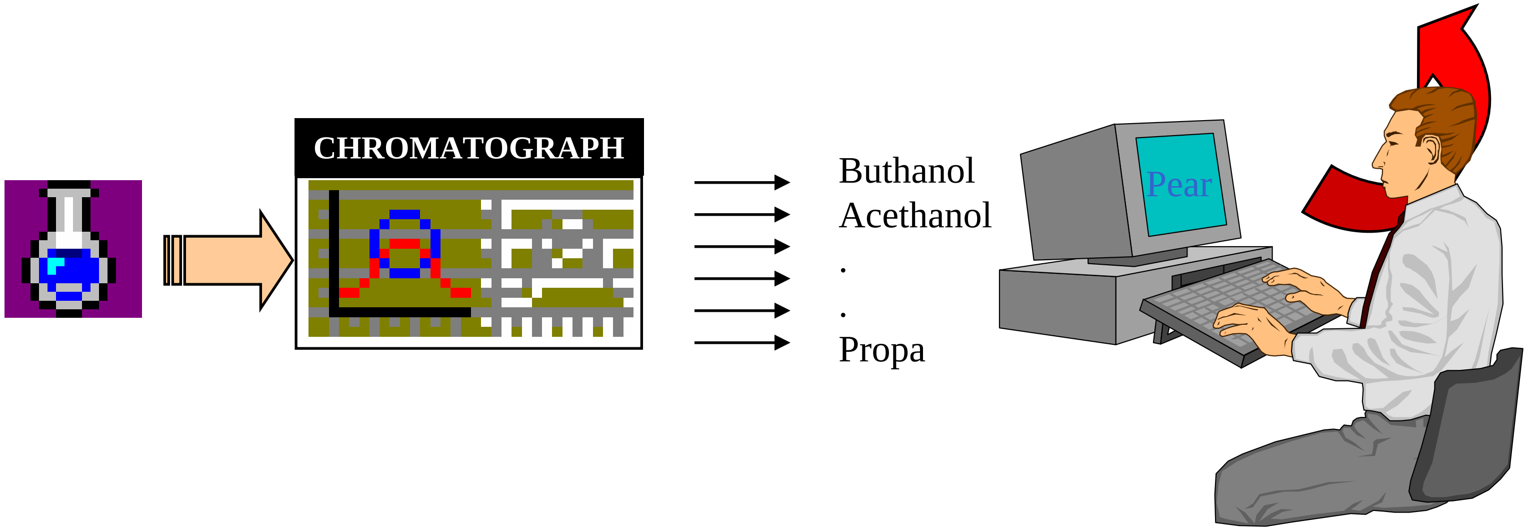
Merging such nodes must decreases the entropy

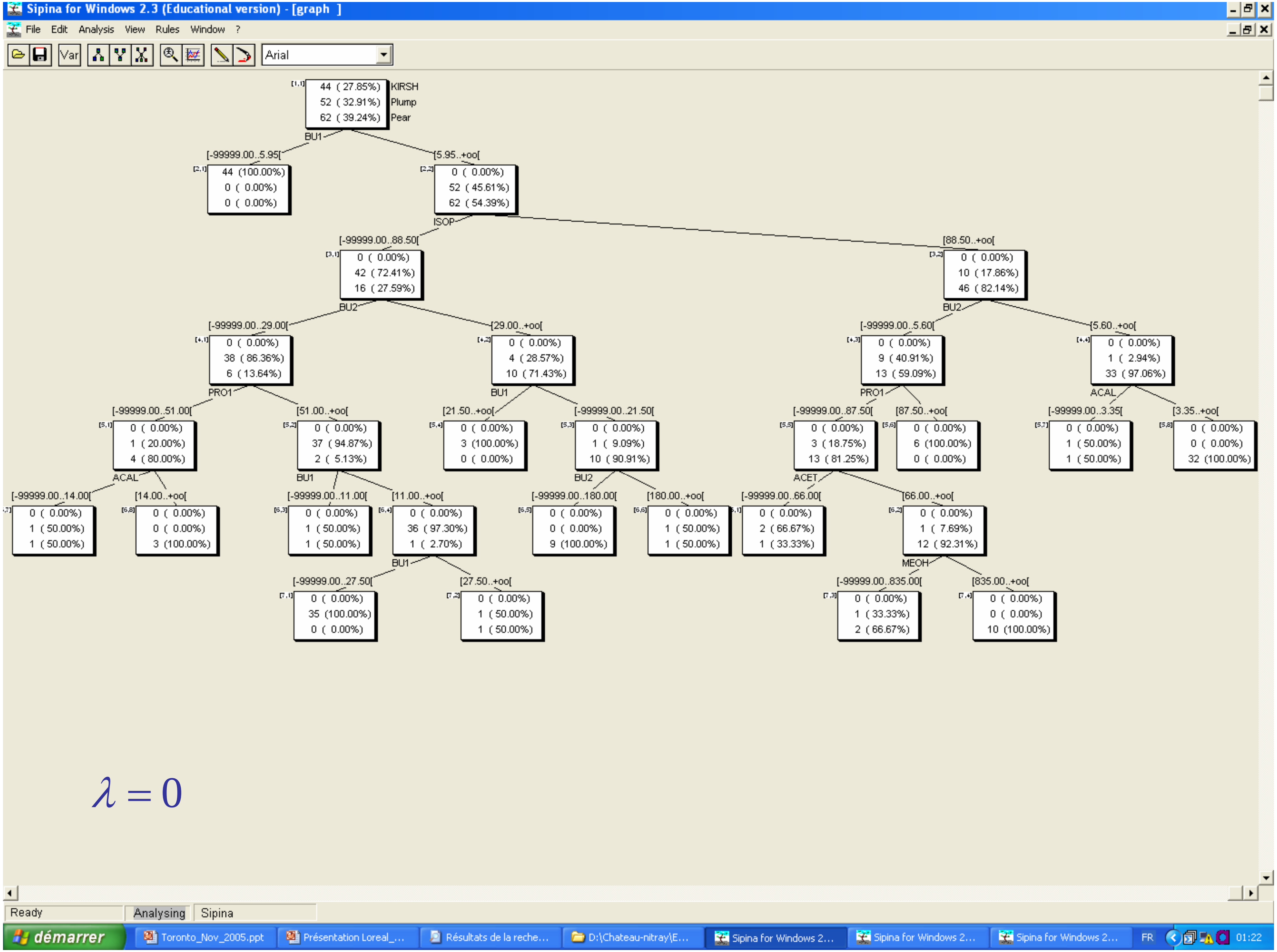
$$\hat{h}(\Omega') < \hat{h}(\Omega)$$

5. Example

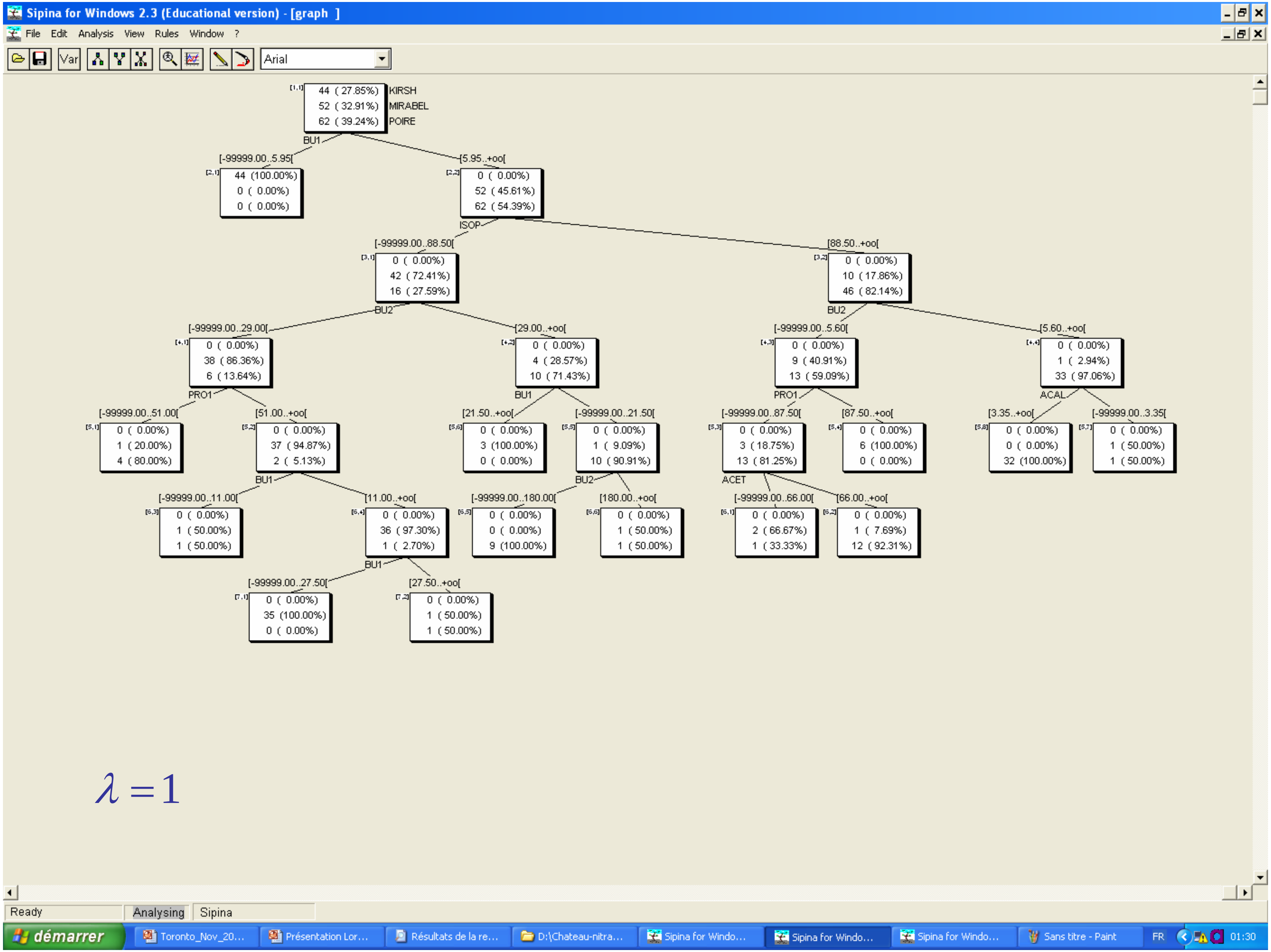
Quality Control of brandy



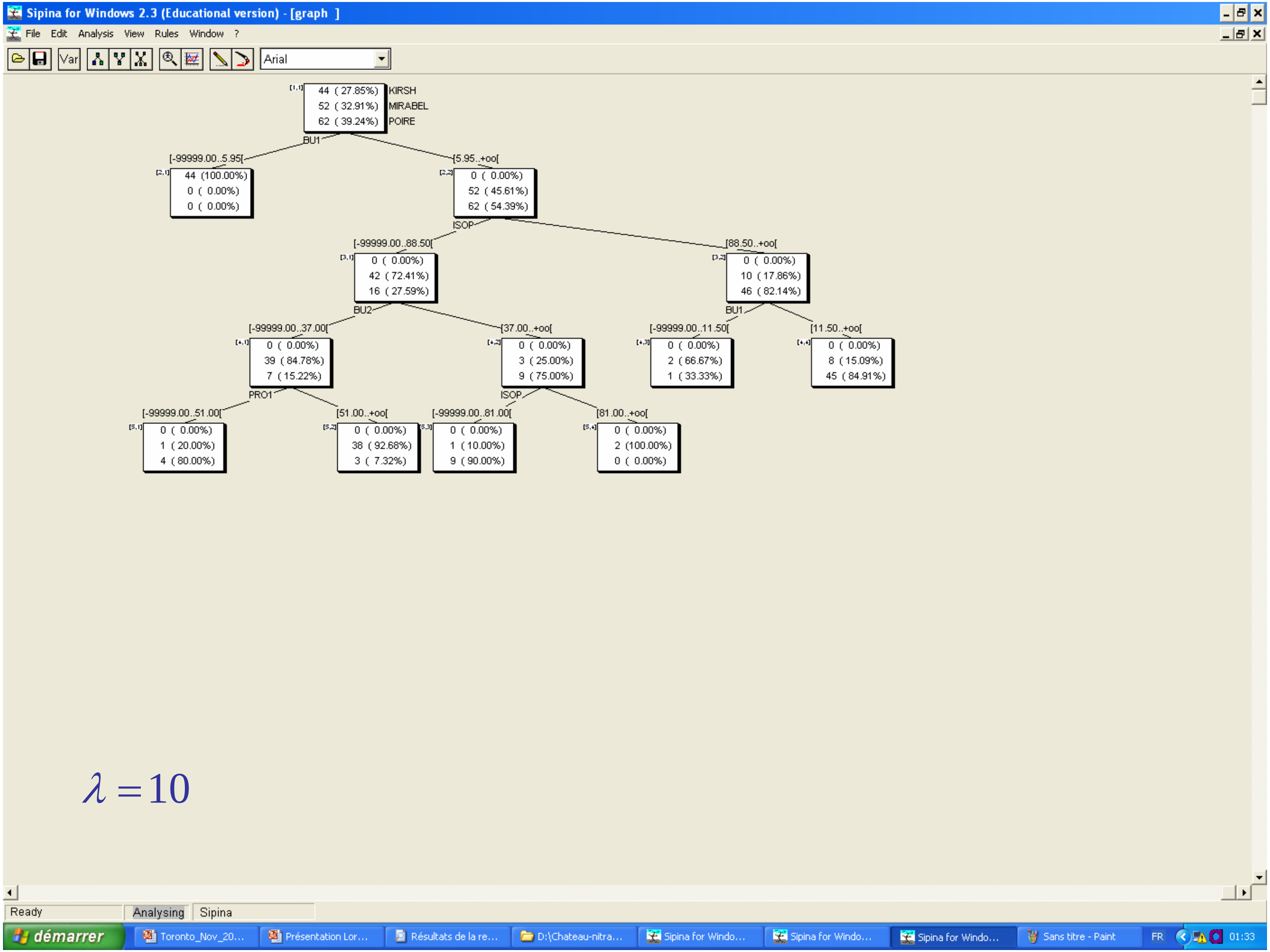



Sipina for Windows 2.3 (Educational version) - [graph]
File Edit Analysis View Rules Window ?
Arial
44 (27.85%) KIRSH
52 (32.91%) Plump
62 (39.24%) Pear
BU1
[-99999.00..5.95[
[5.95..+oo[
ISOP
[-99999.00..88.50[
[88.50..+oo[
BU2
PRO1
ACAL
BU1
ACET
MEOH
Ready
Analysing
Sipina
démarrer
Toronto_Nov_2005.ppt
Présentation Loreal_...
Résultats de la reche...
D:\Chateau-nitray\E...
Sipina for Windows 2...
FR
01:22

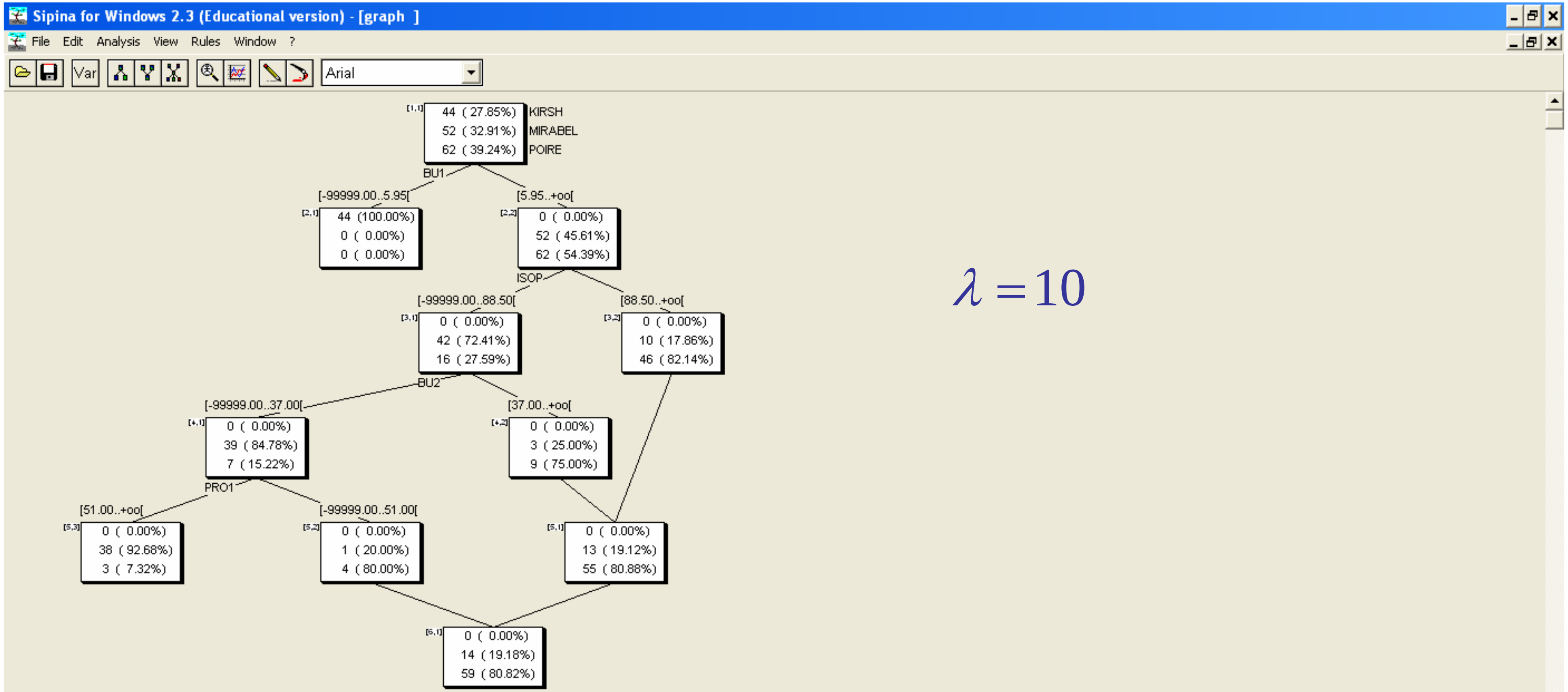
$$\lambda = 0$$


Sipina for Windows 2.3 (Educational version) - [graph]
File Edit Analysis View Rules Window ?
44 (27.85%) KIRSH
52 (32.91%) MIRABEL
62 (39.24%) POIRE
BU1
[-99999.00..5.95[
[5.95..+oo[
ISOP
[-99999.00..88.50[
[88.50..+oo[
BU2
PRO1
BU1
ACAL
ACET
Ready
Analysing
Sipina

$\lambda = 1$


Sipina for Windows 2.3 (Educational version) - [graph]
File Edit Analysis View Rules Window ?
44 (27.85%) KIRSH
52 (32.91%) MIRABEL
62 (39.24%) POIRE
BU1
[-99999.00..5.95[
[5.95..+oo[
44 (100.00%)
0 (0.00%)
0 (0.00%)
0 (0.00%)
52 (45.61%)
62 (54.39%)
ISOP
[-99999.00..88.50[
[88.50..+oo[
0 (0.00%)
42 (72.41%)
16 (27.59%)
0 (0.00%)
10 (17.86%)
46 (82.14%)
BU2
BU1
[-99999.00..37.00[
[37.00..+oo[
[-99999.00..11.50[
[11.50..+oo[
0 (0.00%)
39 (84.78%)
7 (15.22%)
0 (0.00%)
3 (25.00%)
9 (75.00%)
0 (0.00%)
2 (66.67%)
1 (33.33%)
0 (0.00%)
8 (15.09%)
45 (84.91%)
PRO1
ISOP
[-99999.00..51.00[
[51.00..+oo[
[-99999.00..81.00[
[81.00..+oo[
0 (0.00%)
1 (20.00%)
4 (80.00%)
0 (0.00%)
38 (92.68%)
3 (7.32%)
0 (0.00%)
1 (10.00%)
9 (90.00%)
0 (0.00%)
2 (100.00%)
0 (0.00%)
Ready
Analysing
Sipina
démarrer
Toronto_Nov_20...
Présentation Lor...
Résultats de la re...
D:\Chateau-nitra...
Sipina for Windo...
Sipina for Windo...
Sipina for Windo...
Sans titre - Paint
FR
01:33

$$\lambda = 10$$



$\lambda = 10$

If(BU1=[-oo..5.95[) then TYPE=KIRSH with <1.00>

If(BU1=[5.95..+oo[& BU2=[-OO..37[& ISOP=[-OO..88.5[& PRO1=[51..+oo[) then TYPE=Plump with <0.93>

If(BU1=[5.95..+oo[and ISOP=[88.50..+oo[) Or

(BU1=[5.95..+oo[and BU2=[37.00..+oo[and ISOP=[-OO..88.50[) Or

(BU1=[5.95..+oo[and BU2=[--OO..37.00[and ISOP=[-OO..88.50[and PRO1=[-OO..51.00[ ) then

TYPE=Pear with <0.80>

6. Conclusion