Decision Trees

Some exercises

Exemplifying the application of the ID3 algorithm on a toy *mushrooms* dataset

CMU, 2002(?) spring, Andrew Moore, midterm example questions, pr. 2

You are stranded on a deserted island. Mushrooms of various types grow widely all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider:

Example	*IsNotHeavy*	*IsSmelly*	*IsSpotted*	*IsSmooth*	*IsComestible*
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W.

For the a–d questions, consider only mushrooms A through H.

a. What is the entropy of *IsComestible*?

b. Which attribute should you choose as the root of a decision tree? *Hint*: You can figure this out by looking at the data without explicitly computing the information gain of all four attributes.

c. What is the information gain of the attribute you chose in the previous question?

d. Build a ID3 decision tree to classify mushrooms as poisonous or not.

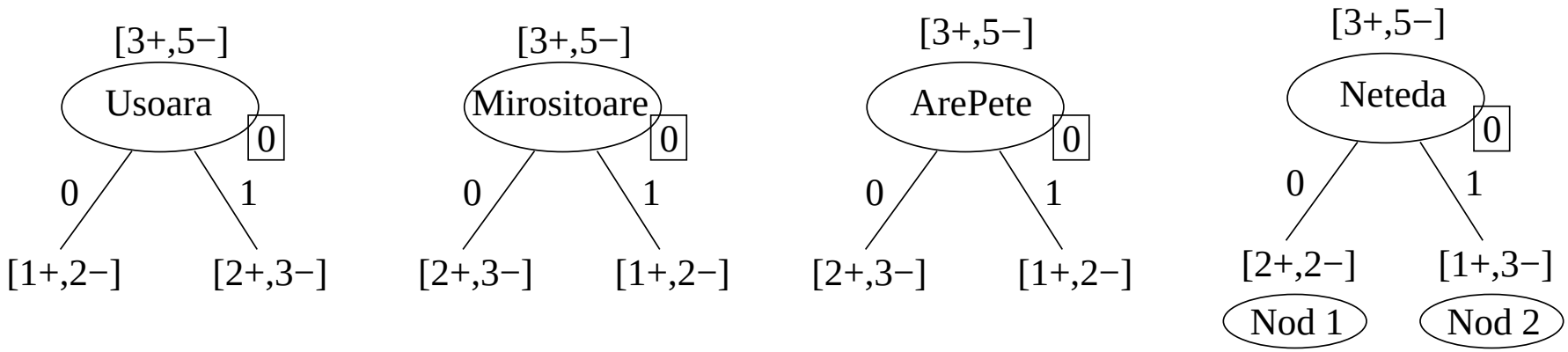
e. Classify mushrooms U, V and W using the decision tree as poisonous or not poisonous.

f. If the mushrooms A through H that you know are not poisonous suddenly became scarce, should you consider trying U, V and W? Which one(s) and why? Or if none of them, then why not?

a.

$$\begin{aligned} H_{Comestibilă} &= H[3+,5-] \stackrel{def.}{=} -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = \frac{3}{8}\log_2\frac{8}{3} + \frac{5}{8}\log_2\frac{8}{5} = \\ &= \frac{3}{8}3 - \frac{3}{8}\log_2 3 + \frac{5}{8}3 - \frac{5}{8}\log_2 5 = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5 \approx \\ &\approx 0.9544 \end{aligned}$$

b.


[3+,5−]
Usoara
0
0
1
[1+,2−]
[2+,3−]
[3+,5−]
Mirositoare
0
0
1
[2+,3−]
[1+,2−]
[3+,5−]
ArePete
0
0
1
[2+,3−]
[1+,2−]
[3+,5−]
Neteda
0
0
1
[2+,2−]
[1+,3−]
Nod 1
Nod 2

c.

$$
\begin{aligned}
H_{0/Netedă} &\overset{def.}{=} \frac{4}{8}H[2+,2-] + \frac{4}{8}H[1+,3-] = \frac{1}{2}\cdot 1 + \frac{1}{2}\left(\frac{1}{4}\log_2\frac{4}{1} + \frac{3}{4}\log_2\frac{4}{3}\right) \\
&= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{4}\cdot 2 + \frac{3}{4}\cdot 2 - \frac{3}{4}\log_2 3\right) = \frac{1}{2} + \frac{1}{2}\left(2 - \frac{3}{4}\log_2 3\right) \\
&= \frac{1}{2} + 1 - \frac{3}{8}\log_2 3 = \frac{3}{2} - \frac{3}{8}\log_2 3 \approx 0.9056
\end{aligned}
$$

$$
\begin{aligned}
IG_{0/Netedă} &\overset{def.}{=} H_{Comestibilă} - H_{0/Netedă} \\
&= 0.9544 - 0.9056 = 0.0488
\end{aligned}
$$

d.

$$
\begin{aligned}
H_{0/\,Uşoară} &\overset{def.}{=} \frac{3}{8}H[1+,2-] + \frac{5}{8}H[2+,3-] \\
&= \frac{3}{8}\left(\frac{1}{3}\log_2\frac{3}{1} + \frac{2}{3}\log_2\frac{3}{2}\right) + \frac{5}{8}\left(\frac{2}{5}\log_2\frac{5}{2} + \frac{3}{5}\log_2\frac{5}{3}\right) \\
&= \frac{3}{8}\left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 3 - \frac{2}{3}\cdot 1\right) + \frac{5}{8}\left(\frac{2}{5}\log_2 5 - \frac{2}{5}\cdot 1 + \frac{3}{5}\log_2 5 - \frac{3}{5}\log_2 3\right) \\
&= \frac{3}{8}\left(\log_2 3 - \frac{2}{3}\right) + \frac{5}{8}\left(\log_2 5 - \frac{3}{5}\log_2 3 - \frac{2}{5}\right) \\
&= \frac{3}{8}\log_2 3 - \frac{2}{8} + \frac{5}{8}\log_2 5 - \frac{3}{8}\log_2 3 - \frac{2}{8} \\
&= \frac{5}{8}\log_2 5 - \frac{4}{8} \approx 0.9512
\end{aligned}
$$

$$
\Rightarrow IG_{0/\,Uşoară} \overset{def.}{=} H_{Comestibilă} - H_{0/\,Uşoară} = 0.9544 - 0.9512 = 0.0032,
$$

$$
IG_{0/\,Uşoară} = IG_{0/Mirositoare} = IG_{0/ArePete} = 0.0032 < IG_{0/Netedă} = 0.0488
$$

Observaţie importantă

În loc să fi calculat efectiv aceste câştiguri de informaţie, pentru a determina atributul cel mai ,,bun“, ar fi fost suficient să comparăm valorile entropiilor condiţionale medii $H_{0/Netedă}$ şi $H_{0/Uşoară}$:

$$IG_{0/Netedă} > IG_{0/Uşoară} \Leftrightarrow H_{0/Netedă} < H_{0/Uşoară}$$

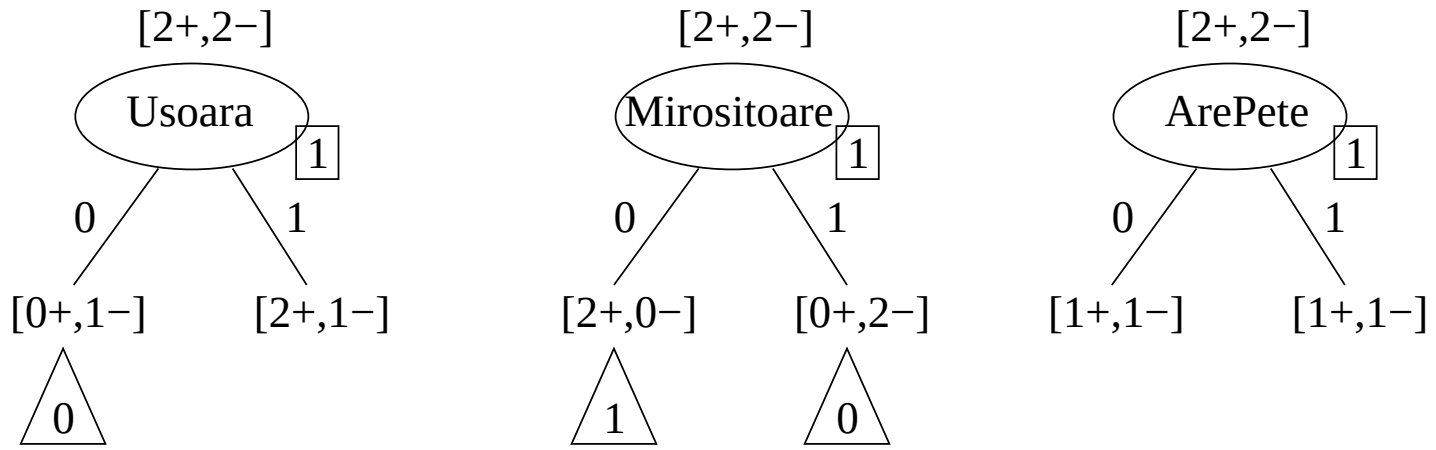
$$\Leftrightarrow \quad \frac{3}{2} - \frac{3}{8}\log_2 3 < \frac{5}{8}\log_2 5 - \frac{1}{2}$$

$$\Leftrightarrow \quad 12 - 3\log_2 3 < 5\log_2 5 - 4$$

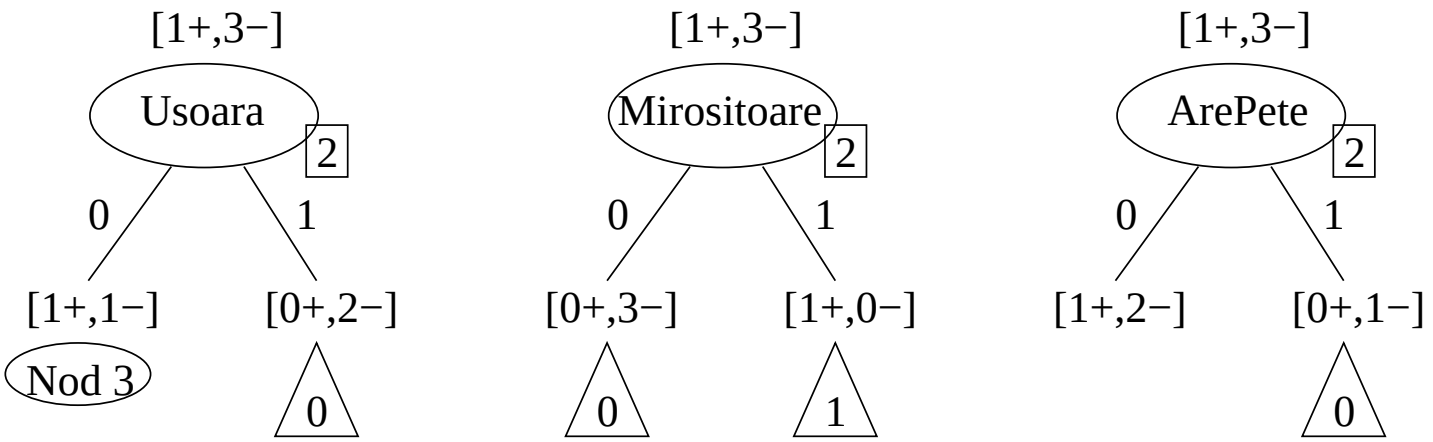
$$\Leftrightarrow \quad 16 < 5\log_2 5 + 3\log_2 3$$

$$\Leftrightarrow \quad 16 < 11.6096 + 4.7548 \ \textbf{(adev.)}$$

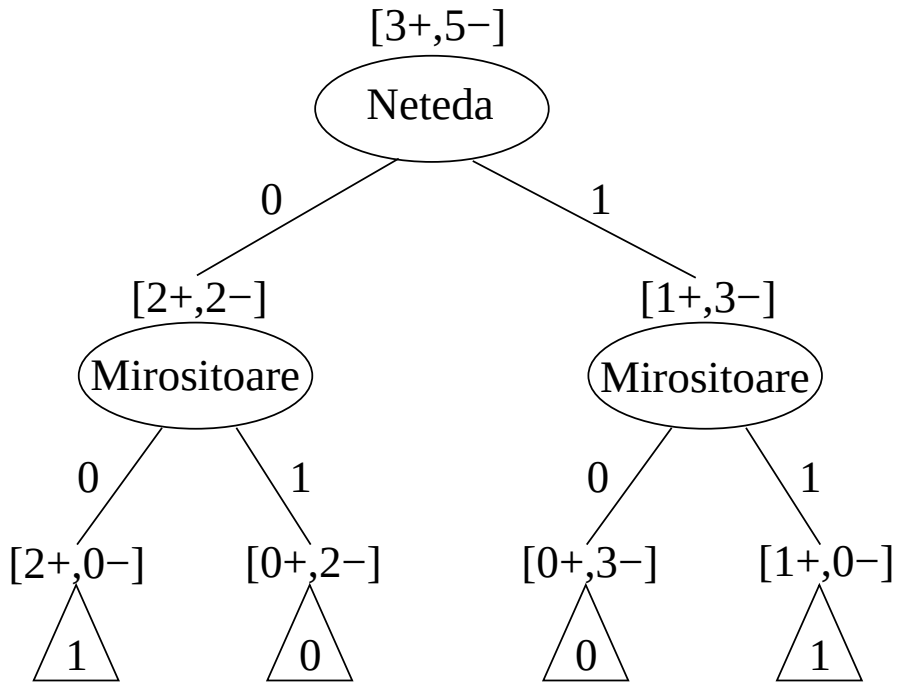
Nodul 1: Netedă = 0



Nodul 1: Netedă = 1



Arborele ID3



U	$\boldsymbol{Netedă = 1,\ Mirositoare = 1 \Rightarrow Comestibilă = 1}$
V	$\boldsymbol{Netedă = 1,\ Mirositoare = 1 \Rightarrow Comestibilă = 1}$
W	$\boldsymbol{Netedă = 0,\ Mirositoare = 1 \Rightarrow Comestibilă = 0}$

Exemplifying greedy character of the ID3 algorithm

CMU, 2003 fall, Andrew Moore, midterm, pr. 9.a

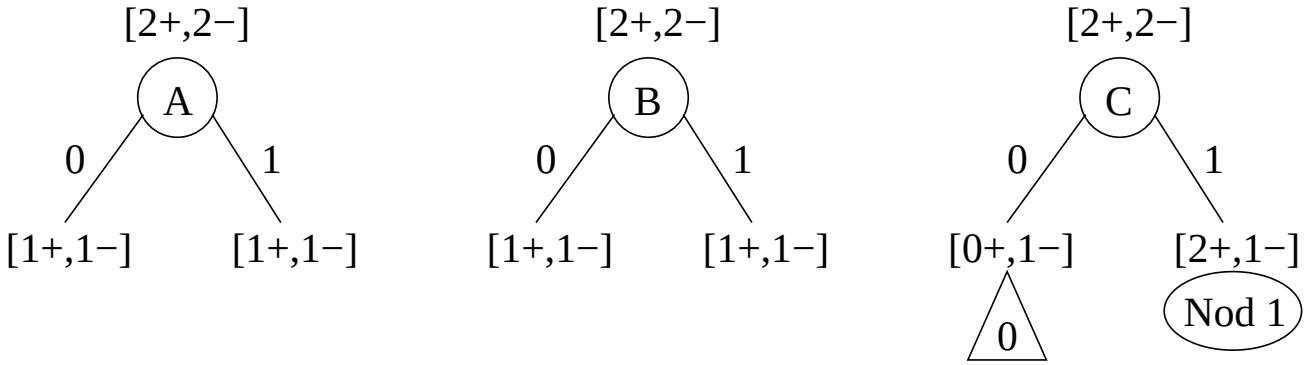
Fie atributele binare de intrare A, B, C, atributul de ieşire Y şi următoarele exemple de antrenament:

A	B	C	Y
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

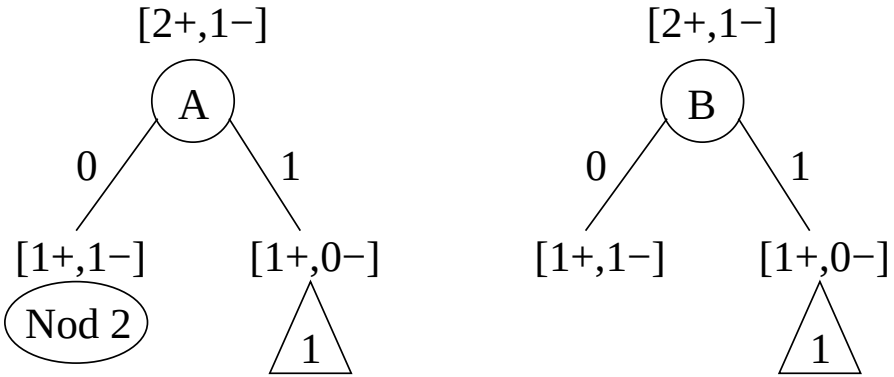
a. Determinaţi arborele de decizie calculat de algoritmul ID3. Este acest arbore de decizie consistent cu datele de antrenament?

Answer

***Nodul 0*: (rădăcina)**



***Nodul 1*: Avem de clasificat instanţele cu $C = 1$, deci alegerea se face între atributele A şi B.**

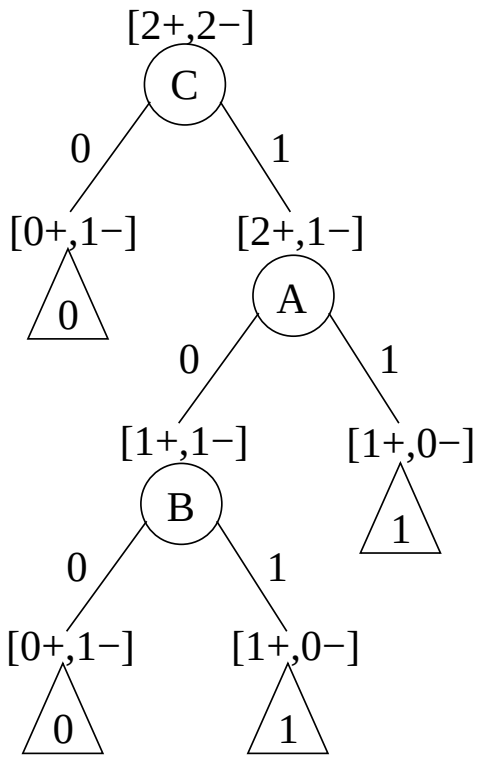


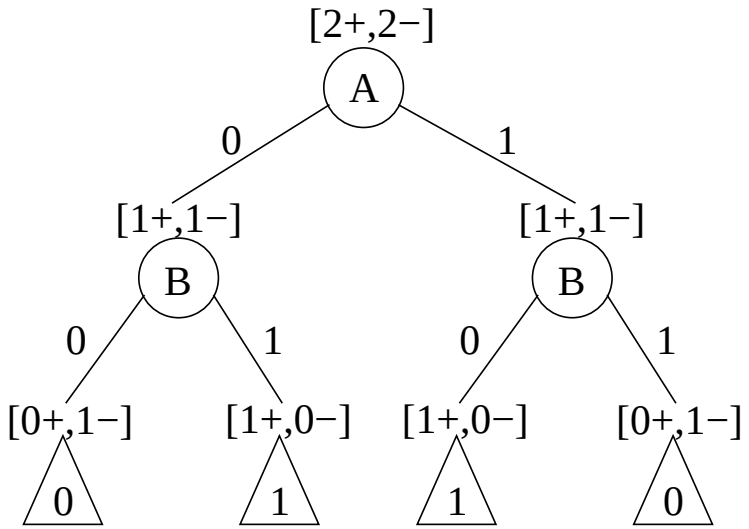
Cele două entropii condiţionale medii sunt egale:

$$H_{1/A} = H_{1/B} = \frac{2}{3}H[1+, 1-] + \frac{1}{3}H[1+, 0-]$$

Aşadar, putem alege oricare dintre cele două atribute. Pentru fixare, îl alegem pe A.

***Nodul 2*: La acest nod nu mai avem decât atributul B, deci îl vom pune pe acesta. Arborele complet este reprezentat în slide-ul următor în partea stângă:**





Prin construcţie, algoritmul ID3 este consistent cu datele de antrenament dacă acestea sunt consistente (i.e., necontradictorii). În cazul nostru, se verifică imediat că datele de antrenament sunt consistente.

b. Există un arbore de decizie de adâncime mai mică (decât cea a arborelui ID3) consistent cu datele de mai sus? Dacă da, ce concept (logic) reprezintă acest arbore?

Se observă că atributul de ieşire Y reprezintă de fapt funcţia logică A XOR B.
Reprezentând această funcţie ca arbore de decizie, vom obţine arborele din slide-ul precedent, în partea dreaptă.
Acest arbore are cu un nivel mai puţin decât arborele construit cu algoritmul ID3.
Prin urmare, arborele ID3 nu este optim din punctul de vedere al numărului de niveluri.

Exemplifying

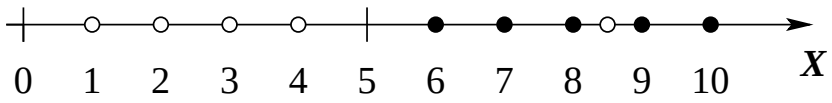
The application of the ID3 algorithm on continuous attributes;

Decision surfaces; decision boundaries;

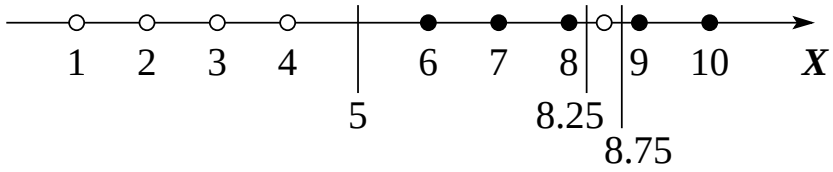
The computation of the CVLOO error

CMU, 2002 fall, Andrew Moore, midterm, pr. 3

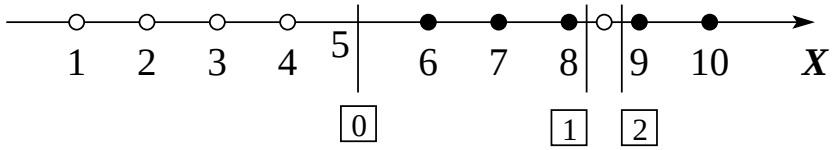
- **training data:**



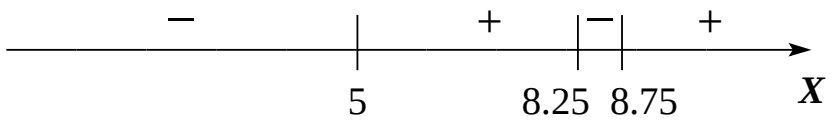
- **discretization / decision thresholds:**



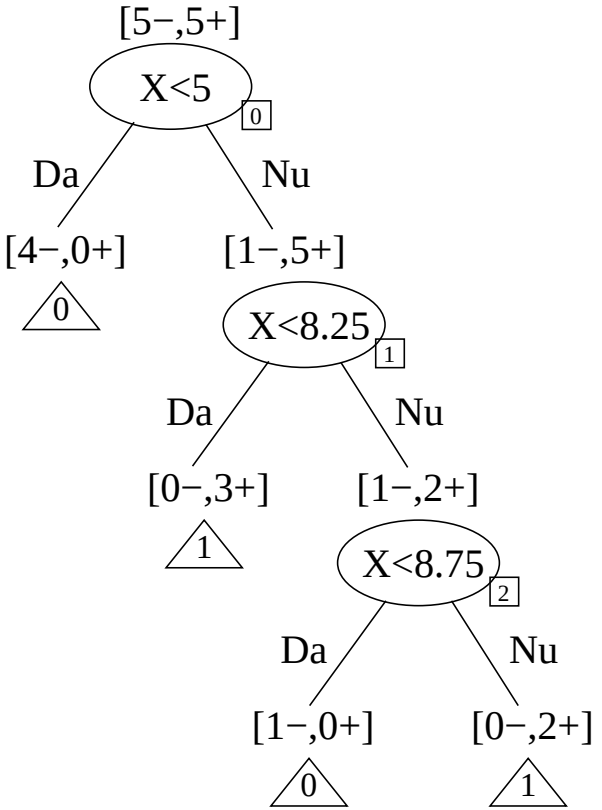
- **compact representation of the ID3 tree:**



- **decision "surfaces":**

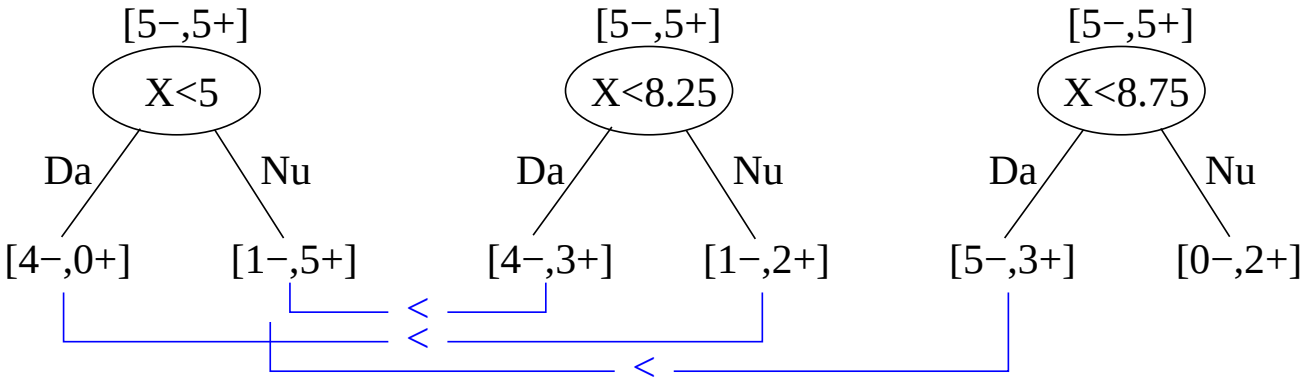


ID3 tree:

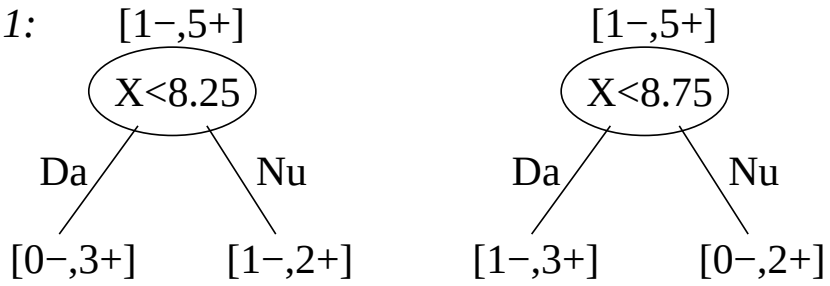


ID3:
IG computations

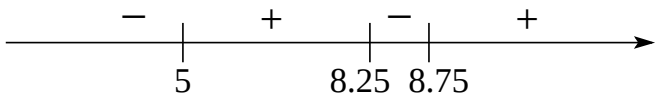
Level 0:



Level 1:

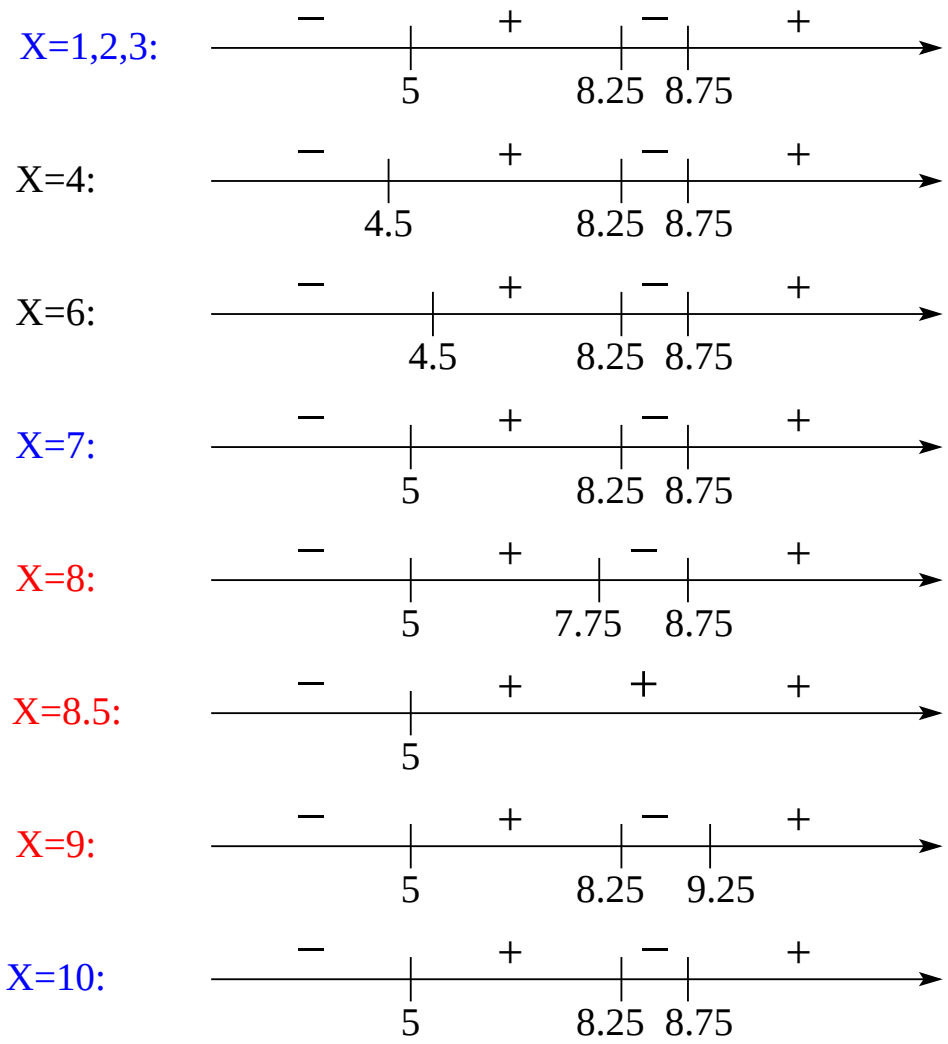


Decision "surfaces":

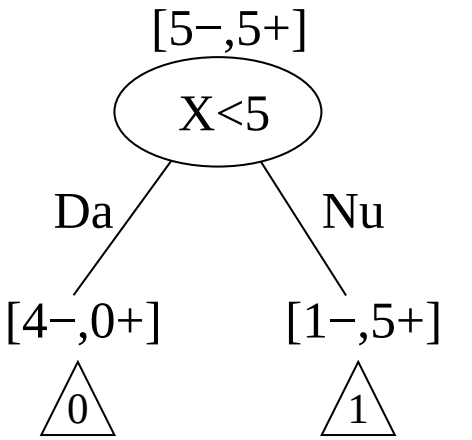


ID3, CVLOO:
Decision surfaces

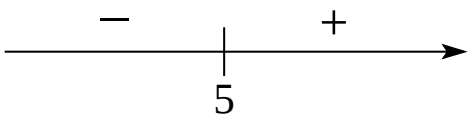
CVLOO error: 3/10



DT2



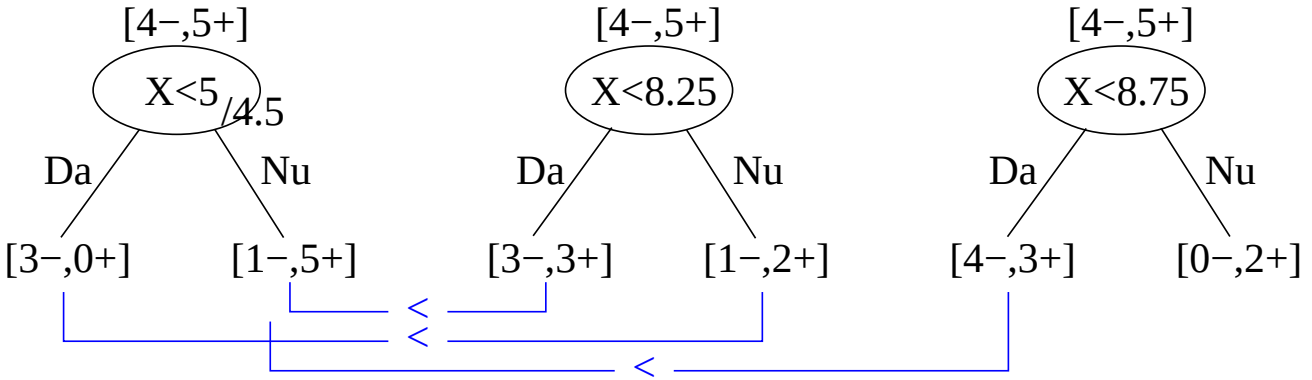
Decision "surfaces":



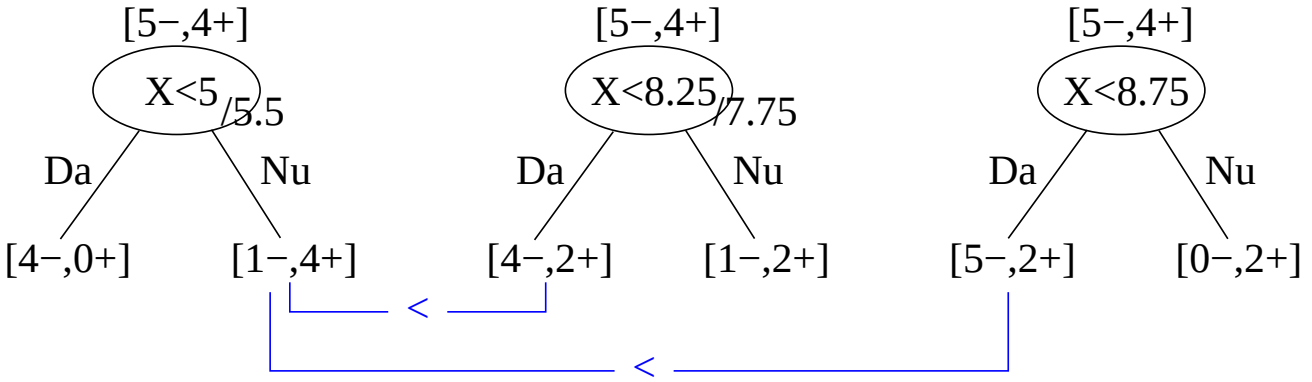
DT2, CVLOO

IG computations

Case 1: X=1, 2, 3, 4

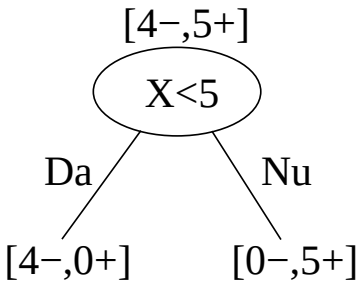


Case 2: X=6, 7, 8

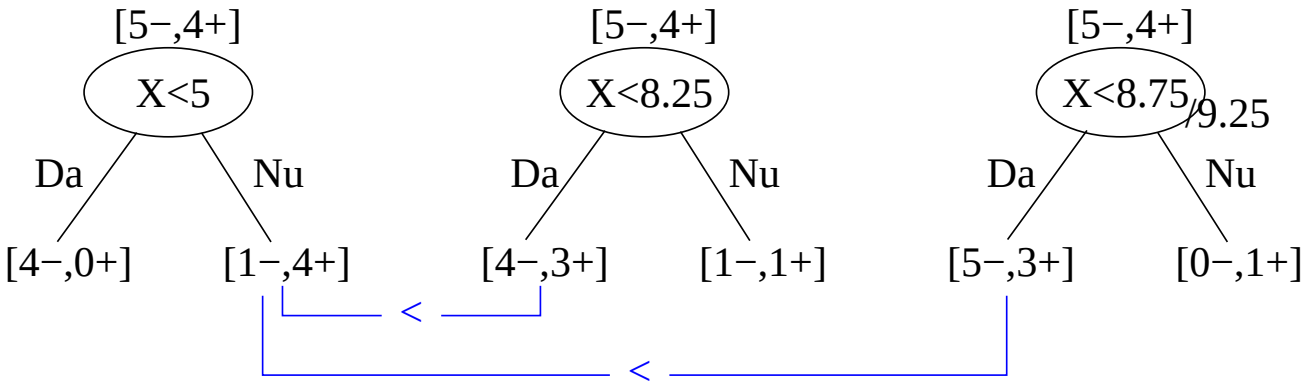


DT2, CVLOO
IG computations
(cont'd)

Case 3: X=8.5



Case 2: X=9, 10



CVLOO error: 1/10

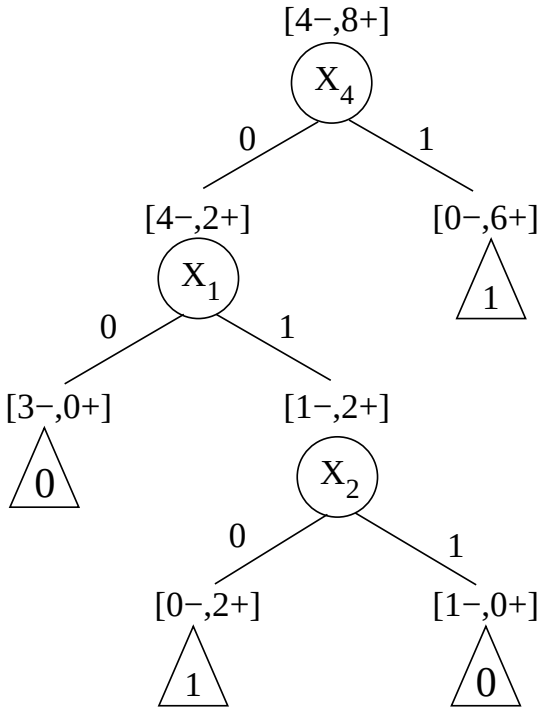
Exemplifying

χ^2-Based Pruning of Decision Trees

CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 2.1

Input:

X_1	X_2	X_3	X_4	***Class***
1	1	0	0	0
1	0	1	0	1
0	1	0	0	0
1	0	1	1	1
0	1	1	1	1
0	0	1	0	0
1	0	0	0	1
0	1	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	1	1
0	0	0	0	0



Idea

While traversing the ID3 tree [usually in bottom-up manner],
remove the nodes for which
there is not enough ("significant") statistical evidence that
there is a dependence between
the values of the input attribute tested in that node and the values of the output attribute (the labels),
supported by the set of instances assigned to that node.

Contingency tables

O_{X_4}	$X_4 = 0$	$X_4 = 1$
Class = 0	4	0
Class = 1	2	6

$\overset{N=12}{\Rightarrow}$

$$\begin{cases} P(\textbf{\textit{Class}} = 0) = \dfrac{4}{12} = \dfrac{1}{3},\ P(\textbf{\textit{Class}} = 1) = \dfrac{2}{3} \\ P(X_4 = 0) = \dfrac{6}{12} = \dfrac{1}{2},\ P(X_4 = 1) = \dfrac{1}{2} \end{cases}$$

$O_{X_1 \mid X_4=0}$	$X_1 = 0$	$X_1 = 1$
Class = 0	3	1
Class = 1	0	2

$\overset{N=6}{\Rightarrow}$

$$\begin{cases} P(\textbf{\textit{Class}} = 0 \mid X_4 = 0) = \dfrac{4}{6} = \dfrac{2}{3} \\ P(\textbf{\textit{Class}} = 1 \mid X_4 = 0) = \dfrac{1}{3} \\ P(X_1 = 0 \mid X_4 = 0) = \dfrac{3}{6} = \dfrac{1}{2} \\ P(X_1 = 1 \mid X_4 = 0) = \dfrac{1}{2} \end{cases}$$

$O_{X_2 \mid X_4=0, X_1=1}$	$X_2 = 0$	$X_2 = 1$
Class = 0	0	1
Class = 1	2	0

$\overset{N=3}{\Rightarrow}$

$$\begin{cases} P(\textbf{\textit{Class}} = 0 \mid X_4 = 0, X_1 = 1) = \dfrac{1}{3} \\ P(\textbf{\textit{Class}} = 1 \mid X_4 = 0, X_1 = 1) = \dfrac{2}{3} \\ P(X_2 = 0 \mid X_4 = 0, X_1 = 1) = \dfrac{2}{3} \\ P(X_2 = 1 \mid X_4 = 0, X_1 = 1) = \dfrac{1}{3} \end{cases}$$

The rationale behind the computation of the expected number of observations

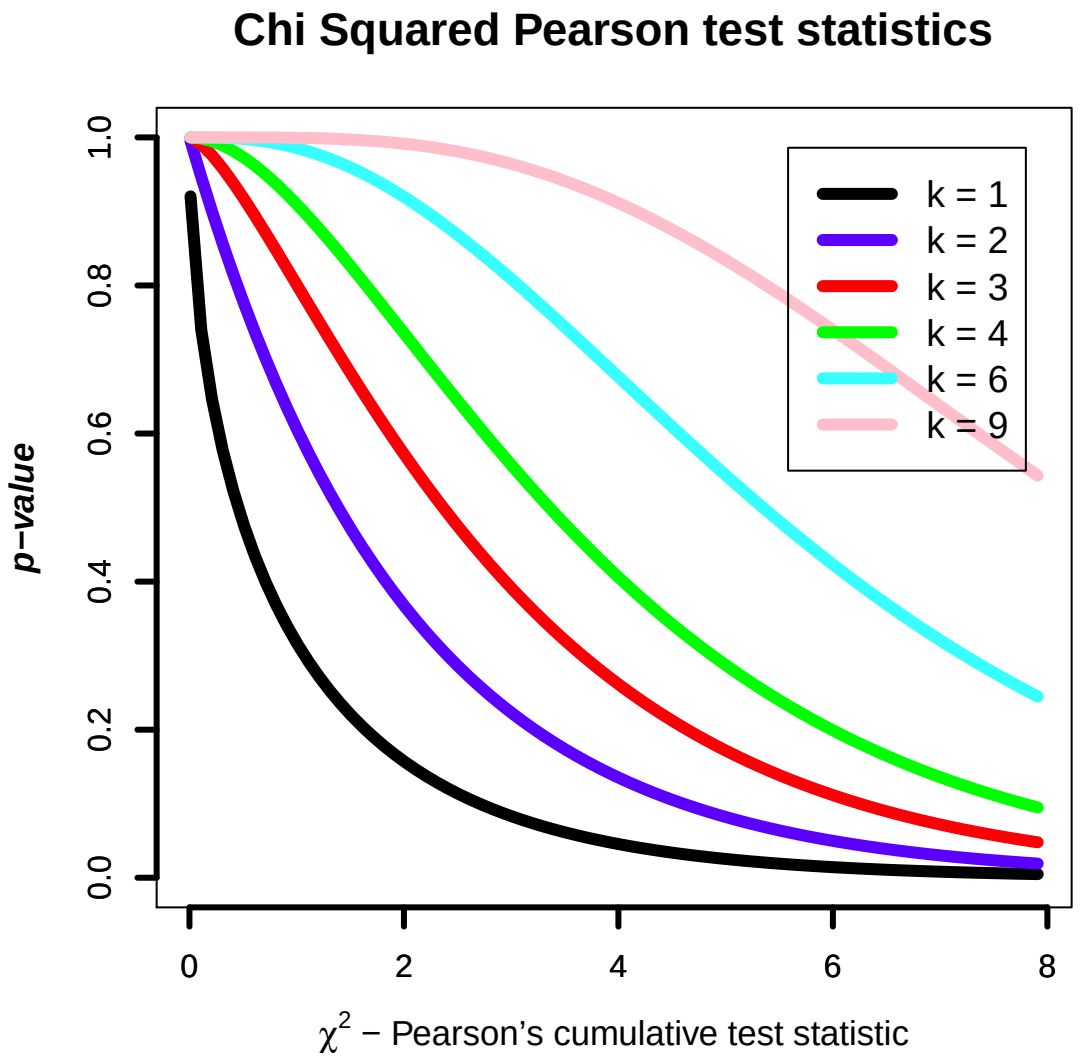
$$P(C = i, A = j) = P(C = i) \cdot P(A = j)$$

$$P(C = i) = \frac{\sum_{k=1}^{c} O_{i,k}}{N} \textbf{ and } P(A = j) = \frac{\sum_{k=1}^{r} O_{k,j}}{N}$$

$$P(C = i, A = j) \overset{\textbf{\textit{indep.}}}{=} \frac{\left(\sum_{k=1}^{c} O_{i,k}\right)\left(\sum_{k=1}^{r} O_{k,j}\right)}{N^2}$$

$$E[C = i, A = j] = N \cdot P(C = i, A = j)$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$


Chi Squared Pearson test statistics
p−value
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
k = 1
k = 2
k = 3
k = 4
k = 6
k = 9
χ² − Pearson's cumulative test statistic

Expected number of observations

E_{X_4}	$X_4 = 0$	$X_4 = 1$
$\boldsymbol{Class = 0}$	2	2
$\boldsymbol{Class = 1}$	4	4

$E_{X_1 \mid X_4}$	$X_1 = 0$	$X_1 = 1$
$\boldsymbol{Class = 0}$	2	2
$\boldsymbol{Class = 1}$	1	1

$E_{X_2 \mid X_4, X_1 = 1}$	$X_2 = 0$	$X_2 = 1$
$\boldsymbol{Class = 0}$	$\frac{2}{3}$	$\frac{1}{3}$
$\boldsymbol{Class = 1}$	$\frac{4}{3}$	$\frac{2}{3}$

$E_{X_4}(\boldsymbol{Class} = 0, X_4 = 0)$:

$N = 12, P(\boldsymbol{Class} = 0) = \frac{1}{3}$ **şi** $P(X_4 = 0) = \frac{1}{2} \Rightarrow$

$N \cdot P(\boldsymbol{Class} = 0, X_4 = 0) = N \cdot P(\boldsymbol{Class} = 0) \cdot P(X_4 = 0) = 12 \cdot \frac{1}{3} \cdot \frac{1}{2} = 2$

χ^2 Statistics

$$\chi^2_{X_4} = \frac{(4-2)^2}{2} + \frac{(0-2)^2}{2} + \frac{(2-4)^2}{4} + \frac{(6-4)^2}{4} = 2+2+1+1 = 6$$

$$\chi^2_{X_1|X_4=0} = \frac{(3-2)^2}{2} + \frac{(1-2)^2}{2} + \frac{(0-1)^2}{1} + \frac{(2-1)^2}{1} = 3$$

$$\chi^2_{X_2|X_4=0,X_1=1} = \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} + \frac{\left(1-\frac{1}{3}\right)^2}{\frac{1}{3}} + \frac{\left(2-\frac{4}{3}\right)^2}{\frac{4}{3}} + \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} = \frac{4}{9}\cdot\frac{27}{4} = 3$$

p-values: $0.0143,\ 0.0833$**, and respectively** 0.0833.

The first one of these p-values is smaller than ε, therefore the root node (X_4) cannot be prunned.

Output (pruned tree) for 95% confidence level