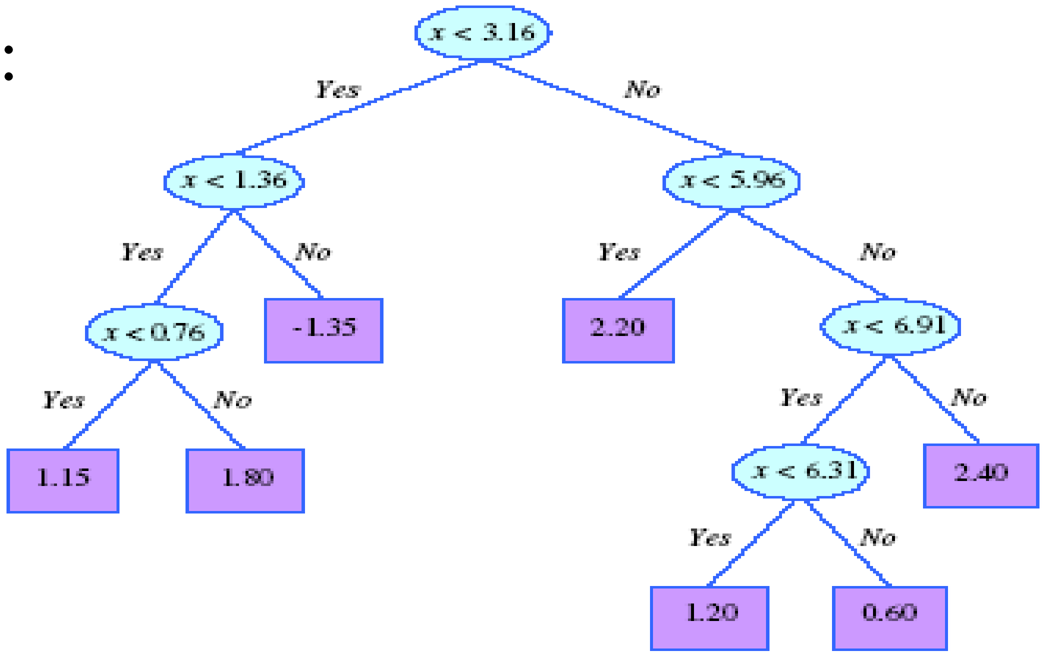
CH. 9: Decision Trees

9.1 Introduction

- A decision tree (DT) consists of internal (decision) and leaf (outcome) nodes. It can be built from a labeled training sample, i.e., supervised learning.

- Example:



- Each decision node goes with a test function with discrete outcomes for labeling branches.

  Each leaf node has a class label (for classification) or a numeric value (for regression).

9.2 Univariate Trees

- In a univariate tree, each internal node makes a decision based on only one attribute.

  In a multivariate tree, each internal node makes a decision based on multiple attributes.

In this chapter, we focus on univariate trees.

- A tree learning algorithm starts at the root with the complete training data. At each step, it looks for the best split of the subset of training data corresponding to the node under consideration based on a chosen attribute. It continues until no split is needed and a leaf node is created.

  Refer to Appendix for the classification and regression tree (**CART**) algorithm.

- The goodness of a split is measured by a criterion.

- The goodness of a split is quantified by the measure of **information gain** G

$$G(X,a) = H(X) - \sum_{v \in V(a)} \frac{|X_v|}{|X|} H(X_v),$$

where a : an attribute

$V(a)$: the set of all possible values of a

$X_v = \{x \in X \mid a(x) = v\}$

X : a set of training examples

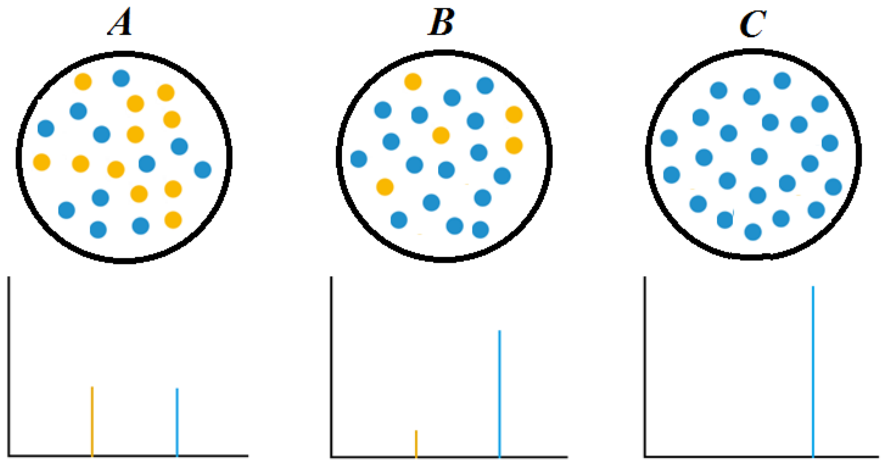
p_i : the proportion of X belonging to class i

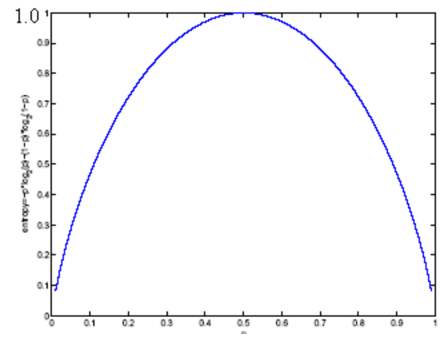
$H(X) = -\sum_{i=1}^{c} p_i \log_2 p_i$,: entropy:

Examples: 2-classes case

Entropy: $H(p_1, p_2) = H(p, 1-p)$

$$= -p\log_2 p - (1-p)\log_2(1-p)$$





$$H_A > H_B > H_C$$

Entropy reflects degree of uncertainty.

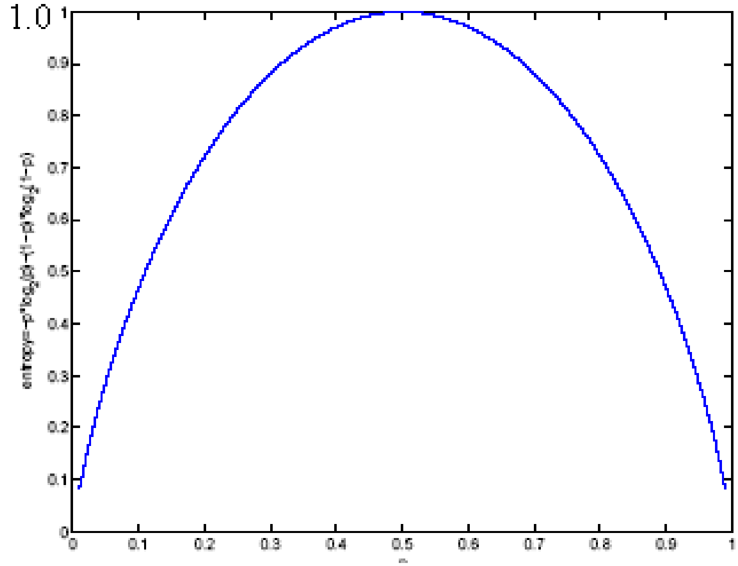
For K-class case, $H(p_1, \cdots, p_K) = -\sum_{i=1}^{K} p_i \log_2 p_i$

When $\forall i,\ p_i = 1/K$, H is largest.

Other examples of measures.

1. Gini index: $\phi(p_1, p_2) = 2p(1-p)$

2. Misclassification error: $\phi(p_1, p_2) = 1 - \max\{p, 1-p\}$



$$\sum_{v \in V(a)} \frac{|X_v|}{|X|} H(X_v)$$ impurity: the sum of the entropies of each subset X_v, weighted by the fraction of $|X_v|/|X|$, i.e., the expected value of entropy after X is partitioned using attribute a

$G(X,a)$: measures the expected reduction in entropy caused by partitioning set S according to attribute a.

Choose a, if $G(X,a) = \max_{a'} G(X,a')$

Example:

年紀	收入	是否為學生	購買筆電與否
<=30	高	否	否
31…40	高	否	是
>40	中	否	是
>40	低	是	否
31…40	低	是	是
<=30	中	否	否
<=30	低	是	是
<=30	中	是	是
31…40	中	否	是
31…40	高	是	是
>40	中	是	否

$X = \{7 \text{ 購買}, 4 \text{ 未購}\}$

Entropy:

$$H(X) = H(\{7,4\})$$

$$= -\sum_{i=1}^{c} p_i \log_2 p_i$$

$$= -\left(\frac{7}{11}\right)\log_2\left(\frac{7}{11}\right) - \left(\frac{4}{11}\right)\log_2\left(\frac{4}{11}\right)$$

$$= 0.415 + 0.531 = 0.956$$

$N = 11$

$X = \{7 \text{ 購買}, 4 \text{ 未購}\}$

$H(X) = H(\{7,4\}) = 0.958$

購買筆電與否	出現次數	出現機率	H 熵
是	7	7/11	$H(X) = H(\{7,4\})$ $= -\left(\frac{7}{11}\right)\log_2\left(\frac{7}{11}\right) - \left(\frac{4}{11}\right)\log_2\left(\frac{4}{11}\right)$ $= 0.415 + 0.531 = 0.956$
否	4	4/11	

以"年紀"為特徵值下購買筆電與否的impurity

Impurity: $I(\text{age}) = \sum_{v \in V(\text{age})} \frac{|X_v|}{|X|} H(X_v)$

$= \frac{|X_{\le 30}|}{|X|} H(X_{\le 30}) + \frac{|X_{30-40}|}{|X|} H(X_{30-40})$

$+ \frac{|X_{>40}|}{|X|} H(X_{>40})$

年紀	購買筆電	未購買筆電
<=30 (4)	2	2
31…40 (4)	4	0
>40 (3)	1	2

$= \frac{4}{11} \boldsymbol{H}(2,2) + \frac{4}{11} \boldsymbol{H}(4,0) + \frac{3}{11} \boldsymbol{H}(1,2) = \frac{4}{11} (-(\frac{2}{4}) \log_2 (\frac{2}{4}) - (\frac{2}{4}) \log_2 (\frac{2}{4}))$

$+ \frac{4}{11} (-(\frac{4}{4}) \log_2 (\frac{4}{4}) - (\frac{0}{4}) \log_2 (\frac{0}{4})) + \frac{3}{11} (-(\frac{1}{3}) \log_2 (\frac{1}{3}) - (\frac{2}{3}) \log_2 (\frac{2}{3}))$

$= 0.364 + 0 + 0.250 = 0.614$

年紀	購買筆電	未購買筆電	I impurity
<=30 (4)	2	2	$I(\text{年紀})$ $= \frac{4}{11}\boldsymbol{H}(2,2) + \frac{4}{11}\boldsymbol{H}(4,0) + \frac{3}{11}\boldsymbol{H}(1,2)$
31…40 (4)	4	0	$= \frac{4}{11}(-(\frac{2}{4})\log_2(\frac{2}{4})-(\frac{2}{4})\log_2(\frac{2}{4}))$ $+ \frac{4}{11}(-(\frac{4}{4})\log_2(\frac{4}{4})-(\frac{0}{4})\log_2(\frac{0}{4}))$
>40 (3)	1	2	$+ \frac{3}{11}(-(\frac{1}{3})\log_2(\frac{1}{3})-(\frac{2}{3})\log_2(\frac{2}{3}))$ $=0.364+0+0.250=0.614$

以"收入"為特徵值下購買筆電與否的impurity

$$\text{Impurity: } I(\text{income}) = \sum_{v \in V(\text{income})} \frac{|X_v|}{|X|} H(X_v)$$

$$= \frac{|X_{\text{high}}|}{|X|} H(X_{\text{high}}) + \frac{|X_{\text{middle}}|}{|X|} H(X_{\text{middle}})$$

$$+ \frac{|X_{\text{low}}|}{|X|} H(X_{\text{low}})$$

收入	購買筆電	未購買筆電
高 (3)	2	1
中 (5)	3	2
低 (3)	2	1

$$= \frac{3}{11} \boldsymbol{H}(2,1) + \frac{5}{11} \boldsymbol{H}(3,2) + \frac{3}{11} \boldsymbol{H}(2,1) \quad = \frac{3}{11} \left(-\left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right)\right)$$

$$+ \frac{5}{11} \left(-\left(\frac{3}{5}\right) \log_2 \left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right)\right) + \frac{3}{11} \left(-\left(\frac{2}{3}\right) \log_2 \left(\frac{2}{3}\right) - \left(\frac{1}{3}\right) \log_2 \left(\frac{1}{3}\right)\right)$$

$$= 0.250 + 0.441 + 0.250 = 0.941$$

<table>
<tr><th>收入</th><th>購買筆電</th><th>未購買筆電</th><th>I impurity</th></tr>
<tr><td>高
(3)</td><td>2</td><td>1</td><td rowspan="3">I(收入)
$= \frac{3}{11}\boldsymbol{H}(2,1) + \frac{5}{11}\boldsymbol{H}(3,2) + \frac{3}{11}\boldsymbol{H}(2,1)$
$= \frac{3}{11}(-(\frac{2}{3})\log_2(\frac{2}{3}) - (\frac{1}{3})\log_2(\frac{1}{3}))$
$+ \frac{5}{11}(-(\frac{3}{5})\log_2(\frac{3}{5}) - (\frac{2}{5})\log_2(\frac{2}{5}))$
$+ \frac{3}{11}(-(\frac{2}{3})\log_2(\frac{2}{3}) - (\frac{1}{3})\log_2(\frac{1}{3}))$
$= 0.250 + 0.441 + 0.250 = 0.941$</td></tr>
<tr><td>中
(5)</td><td>3</td><td>2</td></tr>
<tr><td>低
(3)</td><td>2</td><td>1</td></tr>
</table>

以"是否為學生"為特徵值下購買筆電與否的impurity

$$\text{Impurity: } I(\text{student?}) = \sum_{v \in V(\text{student?})} \frac{|X_v|}{|X|} H(X_v)$$

$$= \frac{|X_{\text{yes}}|}{|X|} H(X_{\text{yes}}) + \frac{|X_{\text{no}}|}{|X|} H(X_{\text{no}})$$

學生	購買筆電	未購買筆電
是 (6)	4	2
否 (5)	3	2

$$= \frac{6}{11} \boldsymbol{H}(4,2) + \frac{5}{11} \boldsymbol{H}(3,2)$$

$$= \frac{6}{11} \left(-\left(\frac{4}{6}\right) \log_2 \left(\frac{4}{6}\right) - \left(\frac{2}{6}\right) \log_2 \left(\frac{2}{6}\right)\right)$$

$$+ \frac{5}{11} \left(-\left(\frac{3}{5}\right) \log_2 \left(\frac{3}{5}\right) - \left(\frac{2}{5}\right) \log_2 \left(\frac{2}{5}\right)\right)$$

$$= 0.501 + 0.441 = 0.942$$

<table>
<tr><th>學生</th><th>購買筆電</th><th>未購買筆電</th><th>I impurity</th></tr>
<tr><td>是
(6)</td><td>4</td><td>2</td><td rowspan="2">I (是否為學生)
$= \frac{6}{11} \boldsymbol{H}(4,2) + \frac{5}{11} \boldsymbol{H}(3,2)$
$= \frac{6}{11} (-(\frac{4}{6}) \log_2 (\frac{4}{6}) - (\frac{2}{6}) \log_2 (\frac{2}{6}))$
$+ \frac{5}{11} (-(\frac{3}{5}) \log_2 (\frac{3}{5}) - (\frac{2}{5}) \log_2 (\frac{2}{5}))$
$= 0.501 + 0.441 = 0.942$</td></tr>
<tr><td>否
(5)</td><td>3</td><td>2</td></tr>
</table>

Information gain: $$G(X,a)=H(X)-\sum_{v\in V(a)}\frac{|X_v|}{|X|}H(X_v)$$

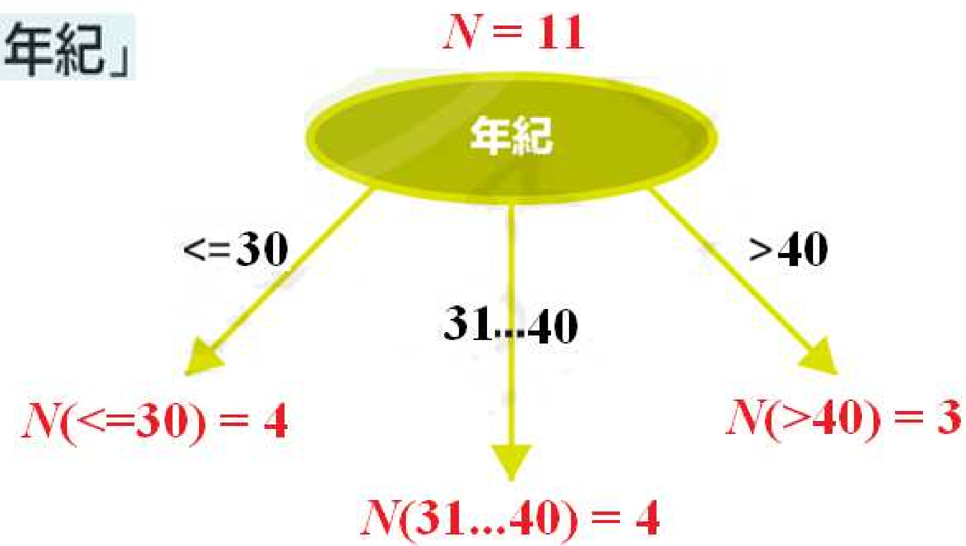
特徵值「年紀」的資訊獲利＝0.956－0.614＝0.342

特徵值「收入」的資訊獲利＝0.956－0.941＝0.015

特徵值「是否為學生」的資訊獲利＝0.956－0.942＝0.014

Choose a, if $G(X,a)=\max_{a'} G(X,a')$

Select attribute「年紀」



將年紀<=30歲 的資料擷取出來

年紀	收入	是否為學生	購買筆電與否
<=30	高	否	否
<=30	中	否	否
<=30	低	是	是
<=30	中	是	是

$X = (2\ \text{購買}, 2\ \text{未購})$ Entropy: $H(X) = -\sum_{i=1}^{c} p_i \log_2 p_i$

$$= -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 0.5 + 0.5 = 1$$

購買筆電與否	出現次數	出現機率	H 熵
是	2	2/4	$H(X) = H(2,2)$ $= -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)$ $= 0.5 + 0.5 = 1$
否	2	2/4	

以"收入"為特徵值下購買筆電與否的impurity

$$\text{Impurity: } I(\text{income}) = \sum_{v \in V(\text{income})} \frac{|X_v|}{|X|} H(X_v)$$

$$= \frac{|X_{\text{high}}|}{|X|} H(X_{\text{high}})$$

$$+ \frac{|X_{\text{middle}}|}{|X|} H(X_{\text{middle}}) + \frac{|X_{\text{low}}|}{|X|} H(X_{\text{low}})$$

收入	購買筆電	未購買筆電
高 (1)	0	1
中 (2)	1	1
低 (1)	1	0

$$= \frac{1}{4} \boldsymbol{H}(0,1) + \frac{2}{4} \boldsymbol{H}(1,1) + \frac{1}{4} \boldsymbol{H}(1,0)$$

$$= \frac{1}{4}\left(-\left(\frac{0}{1}\right) \log_2\left(\frac{0}{1}\right) - \left(\frac{1}{1}\right) \log_2\left(\frac{1}{1}\right)\right) + \frac{2}{4}\left(-\left(\frac{1}{2}\right) \log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2\left(\frac{1}{2}\right)\right)$$

$$+ \frac{1}{4}\left(-\left(\frac{1}{1}\right) \log_2\left(\frac{1}{1}\right) - \left(\frac{0}{1}\right) \log_2\left(\frac{0}{1}\right)\right) = 0 + 0.5 + 0 = 0.5$$

收入	購買筆電	未購買筆電	I imnpurity
高 (1)	0	1	
中 (2)	1	1	
低 (1)	1	0	

$$I(\text{收入})$$

$$=\frac{1}{4}\boldsymbol{H}(0,1)+\frac{2}{4}\boldsymbol{H}(1,1)+\frac{1}{4}\boldsymbol{H}(1,0)$$

$$=\frac{1}{4}(-(\frac{0}{1})\log_2(\frac{0}{1})-(\frac{1}{1})\log_2(\frac{1}{1}))$$

$$+\frac{2}{4}(-(\frac{1}{2})\log_2(\frac{1}{2})-(\frac{1}{2})\log_2(\frac{1}{2}))$$

$$+\frac{1}{4}(-(\frac{1}{1})\log_2(\frac{1}{1})-(\frac{0}{1})\log_2(\frac{0}{1}))$$

$$=0+0.5+0=0.5$$

以"是否為學生"為特徵值下購買筆電與否的impurity

Impurity: $I(\text{student?}) = \sum_{v \in V(\text{student?})} \frac{|X_v|}{|X|} H(X_v)$

$$= \frac{|X_{\text{yes}}|}{|X|} H(X_{\text{yes}}) + \frac{|X_{\text{no}}|}{|X|} H(X_{\text{no}})$$

學生	購買筆電	未購買筆電
是 (2)	2	0
否 (2)	0	2

$$= \frac{2}{4} \boldsymbol{H}(2,0) + \frac{2}{4} \boldsymbol{H}(0,2)$$

$$= \frac{2}{4} \left(-\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) - \left(\frac{0}{2}\right) \log_2 \left(\frac{0}{2}\right)\right)$$

$$+ \frac{2}{4} \left(-\left(\frac{0}{2}\right) \log_2 \left(\frac{0}{2}\right) - \left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right)\right) = 0 + 0 = 0$$

<table>
<tr><th>學生</th><th>購買筆電</th><th>未購買筆電</th><th>I imnpurity</th></tr>
<tr><td>是
(2)</td><td>2</td><td>0</td><td rowspan="2">I(是否為學生)
$=\frac{2}{4}\boldsymbol{H}(2,0)+\frac{2}{4}\boldsymbol{H}(0,2)$
$=\frac{2}{4}(-(\frac{2}{2})\log_2(\frac{2}{2})-(\frac{0}{2})\log_2(\frac{0}{2}))$
$+\frac{2}{4}(-(\frac{0}{2})\log_2(\frac{0}{2})-(\frac{2}{2})\log_2(\frac{2}{2}))$
$=0+0=0$</td></tr>
<tr><td>否
(2)</td><td>0</td><td>2</td></tr>
</table>

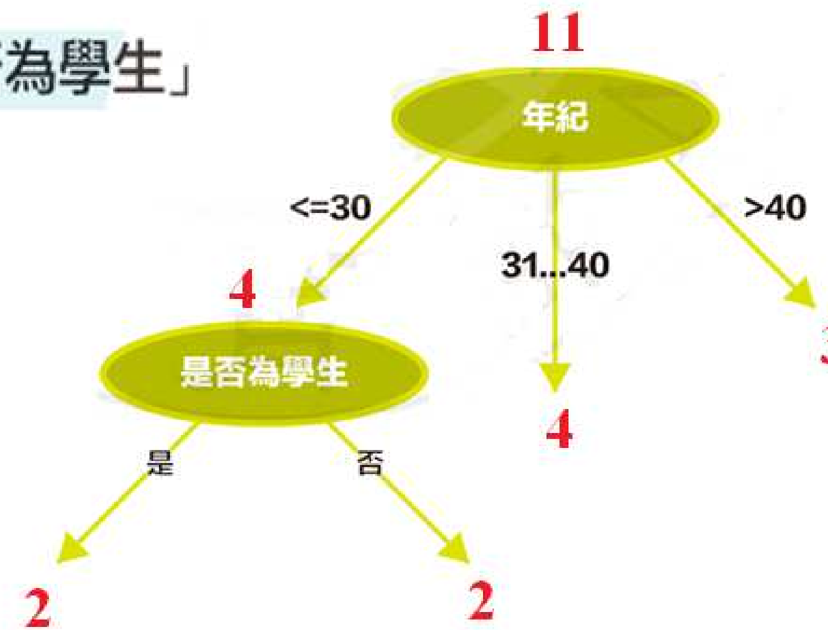
Information gain: $$G(X,a) = H(X) - \sum_{v \in V(a)} \frac{|X_v|}{|X|} H(X_v)$$

特徵值「收入」的資訊獲利 = **1.0 - 0.5 = 0.5**

特徵值「是否為學生」的資訊獲利 = **1.0 - 0 = 1.0**

Choose a, if $G(X,a) = \max_{a'} G(X,a')$

Select attribute「是否為學生」



年紀<=30且是學生 的資料表

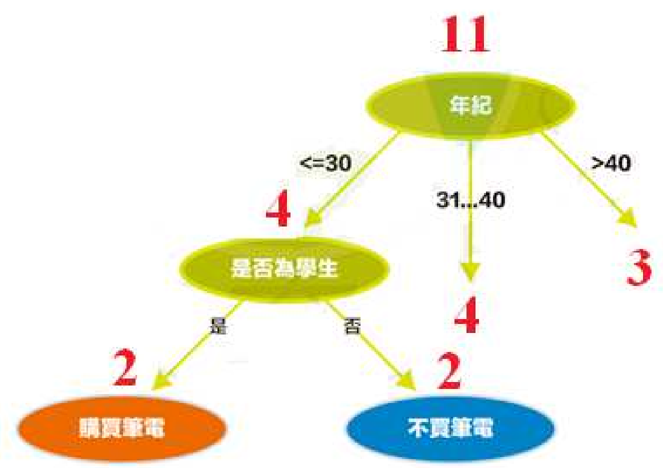
年紀	收入	是否為學生	購買筆電與否
<=30	低	是	是
<=30	中	是	是

All 是

年紀<=30且不是學生 的資料表

年紀	收入	是否為學生	購買筆電與否
<=30	高	否	否
<=30	中	否	否

All 否



將年紀介於31…40歲 的資料擷取出來

年紀	收入	是否為學生	購買筆電與否
31…40	高	否	是
31…40	低	是	是
31…40	中	否	是
31…40	高	是	是

All 是

將年紀>40歲的資料擷取出來

年紀	收入	是否為學生	購買筆電與否
>40	中	否	是
>40	低	是	否
>40	中	是	否

$X = (1\ \text{購買}, 2\ \text{未購})$ Entropy: $H(X) = -\sum_{i=1}^{c} p_i \log_2 p_i$

$$= -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = ?$$

購買筆電與否	出現次數	出現機率	H 熵
是	1	1/3	$H(X) = H(1,2)$ $= -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right)$ $= ?$
否	2	2/3	

以"收入"為特徵值下購買筆電與否的impurity

收入	購買筆電	未購買筆電	I imnpurity
中	1	1	I(收入) = ?
低	0	1	

Impurity:

$$I(\text{income}) = \sum_{v \in V(\text{income})} \frac{|X_v|}{|X|} H(X_v)$$

以"是否為學生"為特徵值下購買筆電與否的impurity

是否為學生	購買筆電	未購買筆電	I imnpurity
是	0	2	I(是否為學生) = ?
否	1	0	

Impurity:

$$I(\text{sudent?}) = \sum_{v \in V(\text{sudent?})} \frac{|X_v|}{|X|} H(X_v)$$

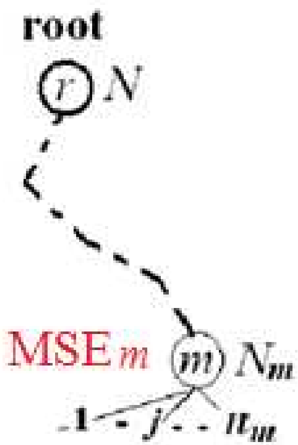
9.2.2 Regression Trees

Given a training sample

$$X = \{(\boldsymbol{x}^1, r^1), \cdots (\boldsymbol{x}^t, r^t), \cdots, (\boldsymbol{x}^n, r^n)\},$$

during constructing a tree, the goodness of a split of a node is measured by the mean square error.

Considering node m, let X_m: the subset of X reaching node m, $N_m = |X_m| = \sum_t b_m(\boldsymbol{x}^t)$,

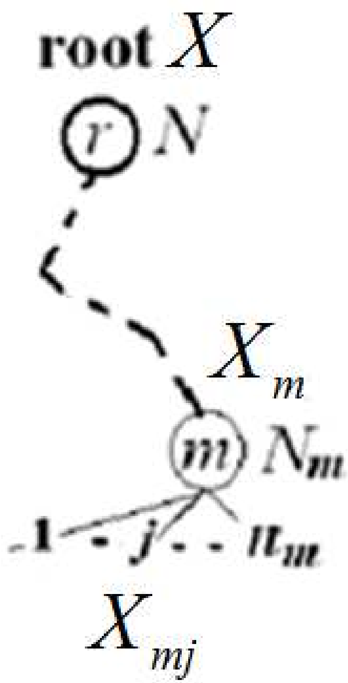


$$b_m(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in X_m \\ 0 & \text{otherwise} \end{cases}, \quad g_m = \frac{\sum_t b_m(\boldsymbol{x}^t)\, r^t}{\sum_t b_m(\boldsymbol{x}^t)}$$

The mean square error: $E_m = \dfrac{1}{N_m} \sum_t b_m(\boldsymbol{x}^t)(r^t - g_m)^2$

If the error is acceptable, i.e. $E_m < \theta_r$, a leaf node is created and g_m value is stored.

Otherwise, data reaching node m is split further such that the sum of the errors in the branches is minimized. Let



X_{mj} : the subset of X_m taking branch j

n_m : # branched at node m

$$b_{mj}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in X_{mj} \\ 0 & \text{otherwise} \end{cases},$$

$$g_{mj} = \frac{\sum_t b_m j\left(\boldsymbol{x}^t\right)\, r^t}{\sum_t b_{mj}\left(\boldsymbol{x}^t\right)}$$: the mean value in branch j of node m

The total MSE after splitting:

$$E'_m = \frac{1}{N_m} \sum_j \sum_t \left(r^t - g_{mj}\right)^2 b_{mj}\left(\boldsymbol{x}^t\right).$$

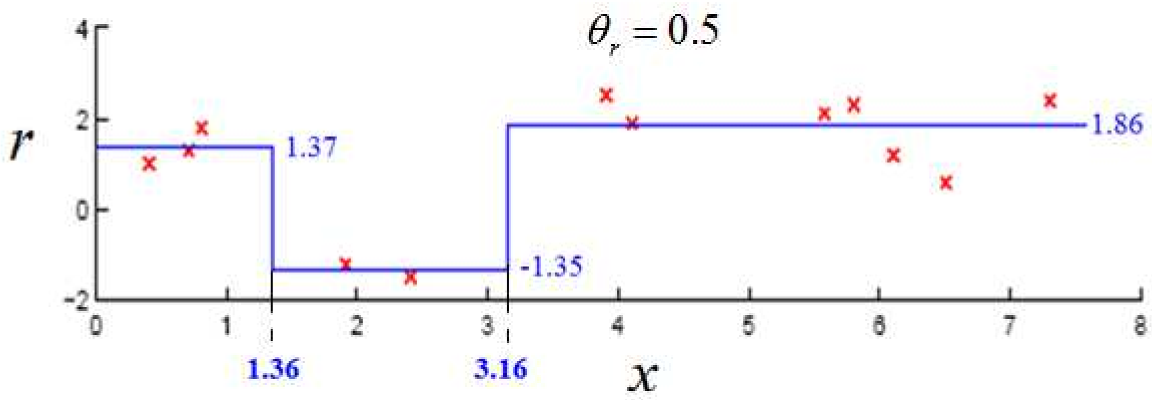
We look for the split that E'_m is minimum.

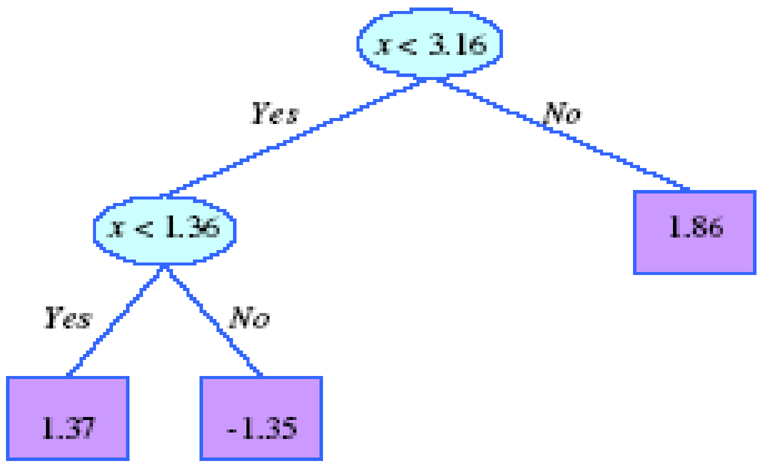
- The CRDT algorithm for classification can be modified to training a regression tree by replacing (i) entropy with mean square error and (ii) class labels with averages.
- Another possible error function:

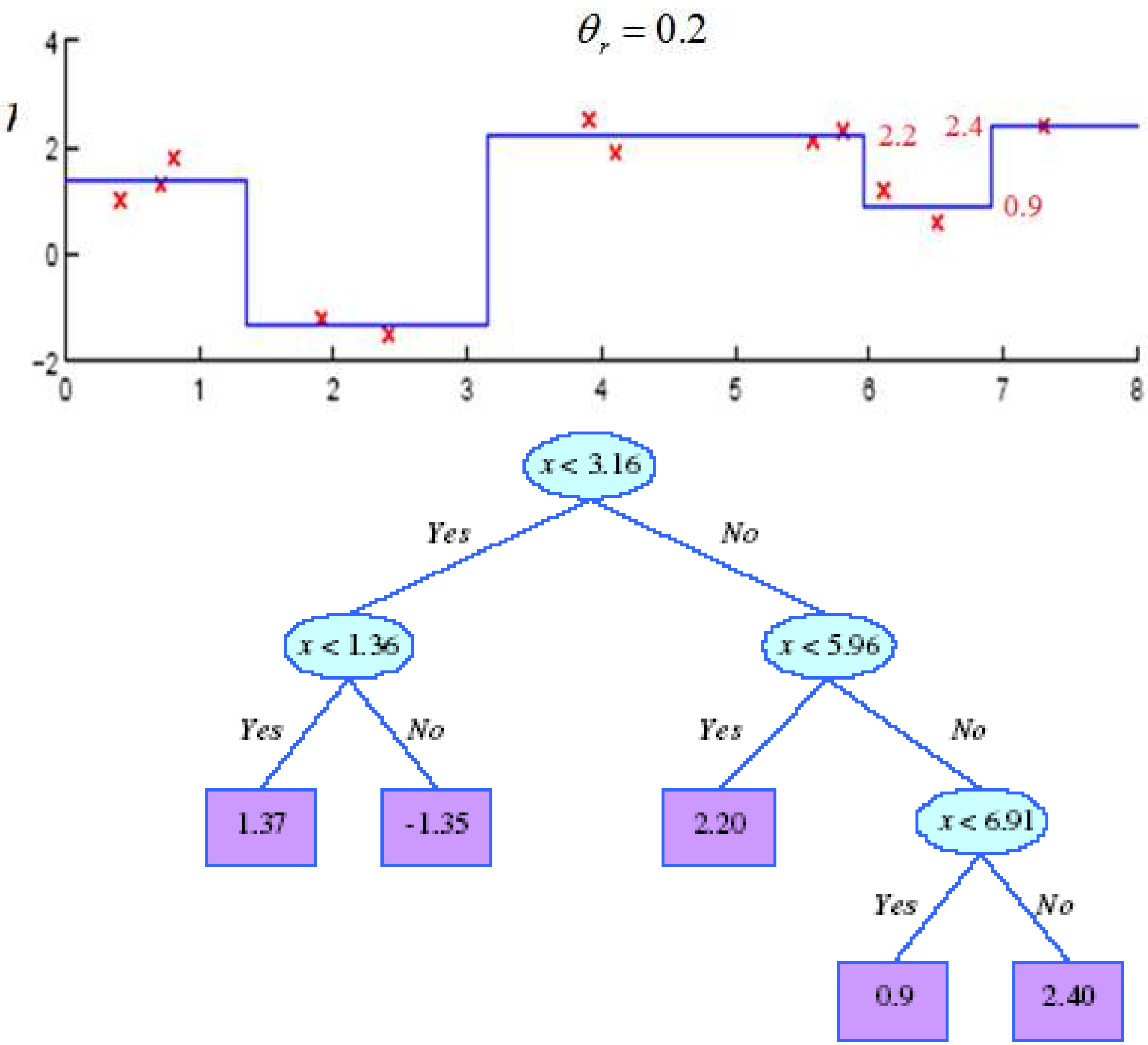
Worst possible error $E_m = \max_j \max_t \frac{1}{N_m} \left| r^t - g_{mj} \right| b_{mj}(\boldsymbol{x}^t)$

- A small error threshold value θ_r leads to a large tree and overfit.

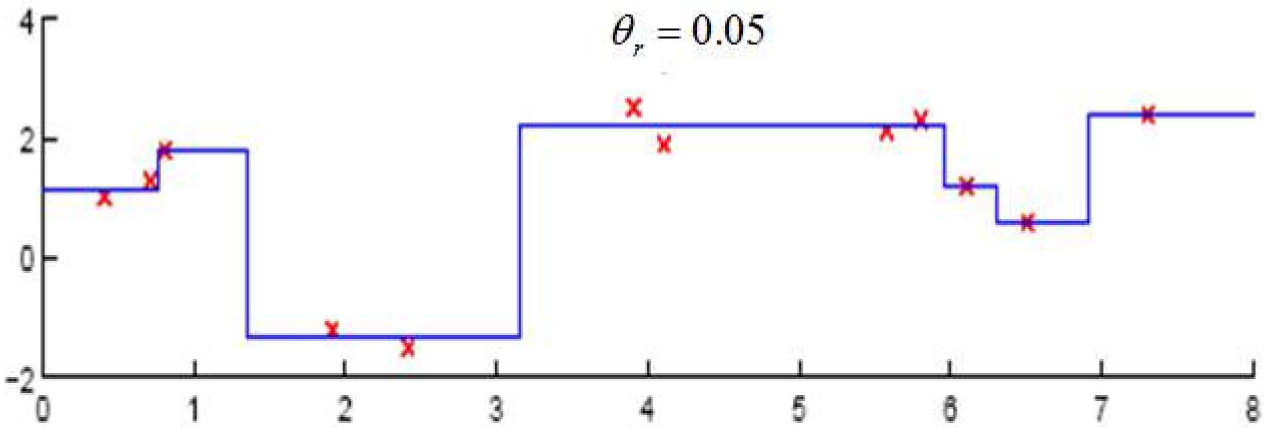
Example:

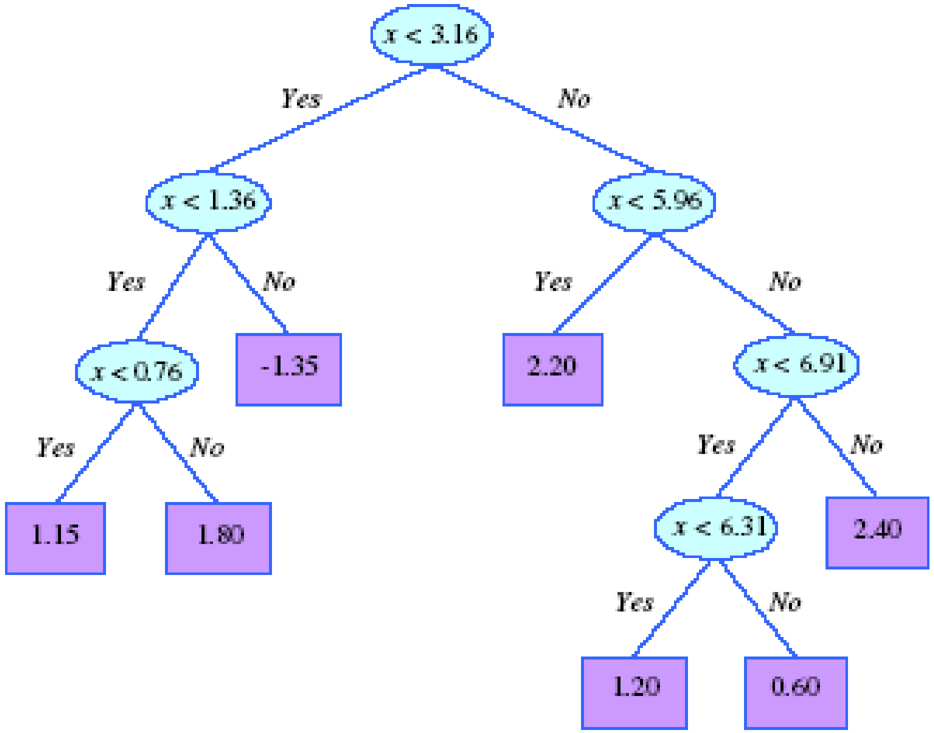






$\theta_r = 0.2$
2.2
2.4
0.9
x < 3.16
Yes
No
x < 1.36
x < 5.96
1.37
-1.35
2.20
x < 6.91
0.9
2.40


θr = 0.05
4
2
0
-2
0
1
2
3
4
5
6
7
8


x < 3.16
Yes
No
x < 1.36
x < 5.96
Yes
No
Yes
No
x < 0.76
-1.35
2.20
x < 6.91
Yes
No
Yes
No
1.15
1.80
x < 6.31
2.40
Yes
No
1.20
0.60

If the error is acceptable, a leaf node is created and the mean value g_m is stored.

We may store a linear function $g_m(\boldsymbol{x}) = \boldsymbol{w}_m^T\boldsymbol{x} + w_{m0}$.

9.3 Pruning

- Two types of pruning:

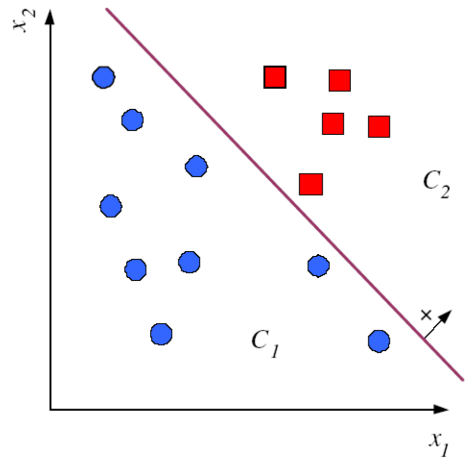
  Prepruning : Early stopping, e.g., small number of examples reaching a node

  Postpruning: Grow the whole tree then prune unnecessary subtrees that cause overfitting

* Prepruning is faster, postpruning is more accurate

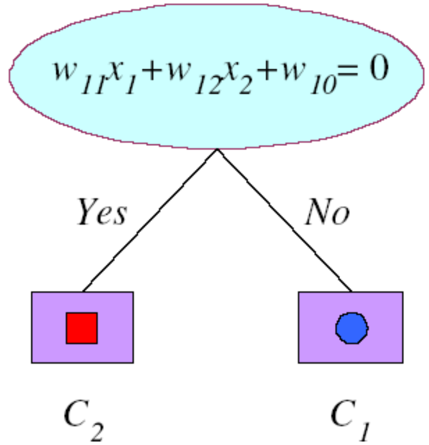
9.6 Multivariate Trees

At a decision node *m*, all input dimensions can be used to split the node.



Linear multivariate node:

$$f_m(\boldsymbol{x}):\ \boldsymbol{w}_m^T\boldsymbol{x} + w_{m0} > 0$$



Quadratic multivariate node:

$$f_m(\boldsymbol{x}):\ \boldsymbol{x}^T W_m \boldsymbol{x} + \boldsymbol{w}_m^T\boldsymbol{x} + w_{m0} > 0$$

Sphere node:

$$f_m(\boldsymbol{x}):\ \left\|\boldsymbol{x} - \boldsymbol{c}_m\right\| \le \alpha_m$$

Appendix: CART Algorithm

- N : # training instances

  N_m : # instances reaching node m

  f_m : test function at m

  n_m : # branched from m

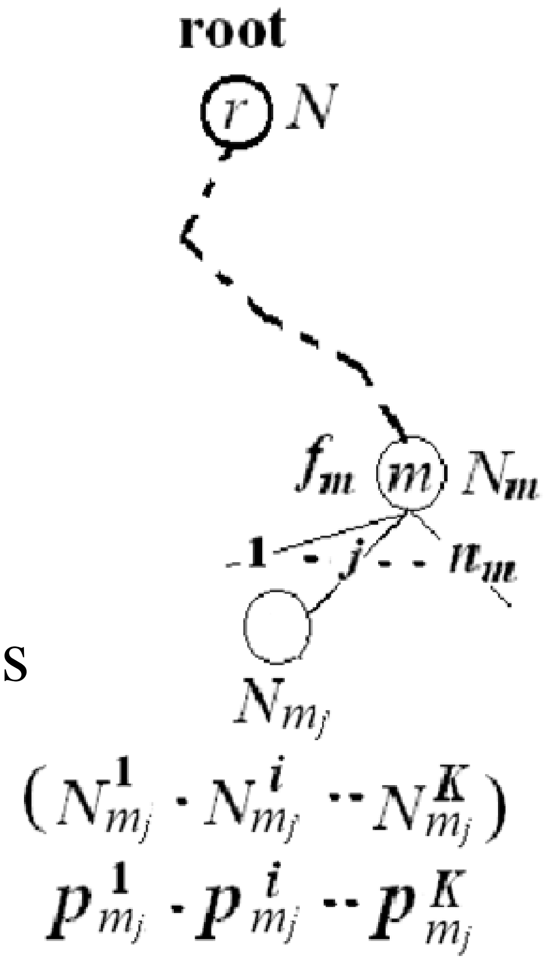
  N_{mj} : # instances along branch j

  N^i_{mj} : # instances belonging to class

  K : # classes

  $p^i_{mj} = N^i_{mj} / N_{mj}$: probability of

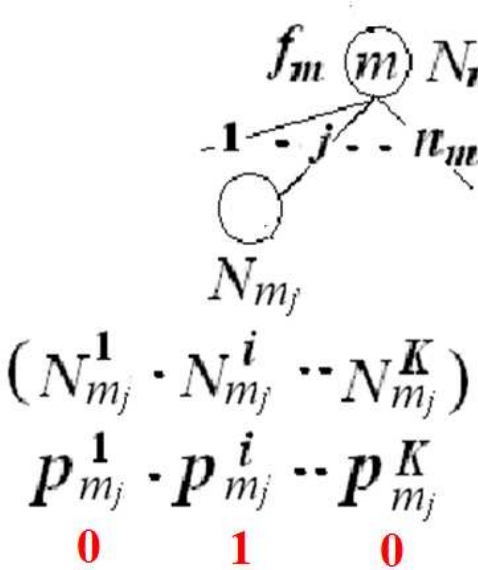
  an instance reaching m belonging to C_j.



- Node m_j is pure if $\forall i \;\; p^i_{m_j}$ are either 0 or 1.

  ($\because \; \sum_{i=1}^{K} p^i_{mj} = 1$, only one of the probabilities is 1
  and the others are all 0)

  A leaf node for which $p^i_{m_j} = 1$
  can be added.



CART Algorithm:

- If node m is pure, generate a leaf and stop;
  Otherwise, split and continue recursively
- Impurity after split: $I = -\sum_{j=1}^{n_m} \frac{N_{mj}}{N_m} \sum_{i=1}^{K} p^i_{mj} \log_2 p^i_{mj}$

- For all attributes, calculate their split impurity and choose the one with the minimum impurity.

Difficulties with the CART Algorithm:

1. Splitting favors attributes with many values
   $\because$ many values → many branches → less impurity
2. Noise may lead to a very large tree if highly pure tree is desired.
   Strategy: introducing thresholds for impurity measures I of nodes and probabilities p_{mj}^{i} of creating leaf nodes.