Exercise

Use the first 7 training examples to construct a decision tree.
In case of ties between features F_i and F_j such that i<j, favor F_i:

	F_1	F_2	F_3	F_4	F_5	Class
Example 1	T	T	F	F	F	T
Example 2	F	F	T	T	F	T
Example 3	T	F	F	T	F	T
Example 4	T	F	T	F	T	T
Example 5	F	T	F	F	F	F
Example 6	T	T	F	T	T	F
Example 7	F	T	T	T	T	F
Example 8	F	F	F	T	T	?

How would the constructed decision tree classify the 8th training example?

Exercise

- First, choose from {F_1, F_2, F_3, F_4, F_5} to become the root.

$H(\text{class}) - H(\text{class}/F_1)$ =
-4/7log(4/7)-3/7log(3/7)-[4/7(-3/4log(3/4)-1/4log(1/4))+3/7(2/3log(2/3)-1/3log(1/3))] = **0.128**

$H(\text{class}) - H(\text{class}/F_2)$ =
-4/7log(4/7)-3/7log(3/7)-[4/7(-3/4log(3/4)-1/4log(1/4))+3/7(-3/3log(3/3))] = **0.522**

$H(\text{class}) - H(\text{class}/F_3)$ =
-4/7log(4/7)-3/7log(3/7)-[3/7(-2/3log(2/3)-1/3log(1/3))+4/7(-2/4log(2/4)-2/4log(2/4))] = **0.02**

$H(\text{class}) - H(\text{class}/F_4)$ =
-4/7log(4/7)-3/7log(3/7)-[4/7(-2/4log(2/4)-2/4log(2/4))+3/7(-2/3log(2/3)-1/3log(1/3))] = **0.02**

$H(\text{class}) - H(\text{class}/F_5)$ =
-4/7log(4/7)-3/7log(3/7)-[3/7(-1/3log(1/3)-2/3log(2/3))+4/7(-3/4log(3/4)-1/4log(1/4))]= **0.128**

Exercise

- Since F_2 has the maximum gain, F_2 becomes the root.
- Then, choose from {F_1, F_3, F_4, F_5} to be F_2's F-child.
  - Since all examples are 'In Class', it becomes F2's F-child
- Next, choose from {F_1, F_3, F_4, F_5} to be F_2's T-child.

	F_1	F_3	F_4	F_5	Class
Example 1	T	F	F	F	T
Example 5	F	F	F	F	F
Example 6	T	F	T	T	F
Example 7	F	T	T	T	F

H(class) – H(class/F1) =
-1/4log(1/4)-3/4log(3/4)-[2/4(-1/2log(1/2)-1/2log(1/2))+2/4(-2/2log(2/2))]
≈ **0.311**

H(class) – H(class/F3) =
-1/4log(1/4)-3/4log(3/4)-[1/4(-1/1log(1/1))+3/4(-1/3log(1/3)-2/3log(2/3))]
≈ **0.122**

Exercise

$H(\text{class}) - H(\text{class}/F_4) =$

$-1/4\log(1/4)-3/4\log(3/4)-[2/4(-2/2\log(2/2))+2/4(-1/2\log(1/2)-1/2\log(1/2))]$

$\approx \mathbf{0.311}$

$H(\text{class}) - H(\text{class}/F_5) =$

$-1/4\log(1/4)-3/4\log(3/4)-[2/4(-2/2\log(2/2))+2/4(-1/2\log(1/2)-1/2\log(1/2))]$

$\approx \mathbf{0.311}$

- F_1, F_4 and F_5 have the maximum gain, we break ties in favor of F_1 to be F_2's T-child.
- Then, determine F_1's F-child.
  - Since all examples are 'Not in Class', it becomes F_1's F-child
- Then, choose from $\{F_3, F_4, F_5\}$ to be F_1's T-child.

	F_3	F_4	F_5	Class
Example 1	F	F	F	T
Example 6	F	T	T	F

Exercise

$H(\text{class}) - H(\text{class}/F_3) =$

$-1/2\log(1/2)-1/2\log(1/2)-[2/2(-1/2\log(1/2)-1/2\log(1/2))] = \mathbf{0}$

$H(\text{class}) - H(\text{class}/F_4) =$

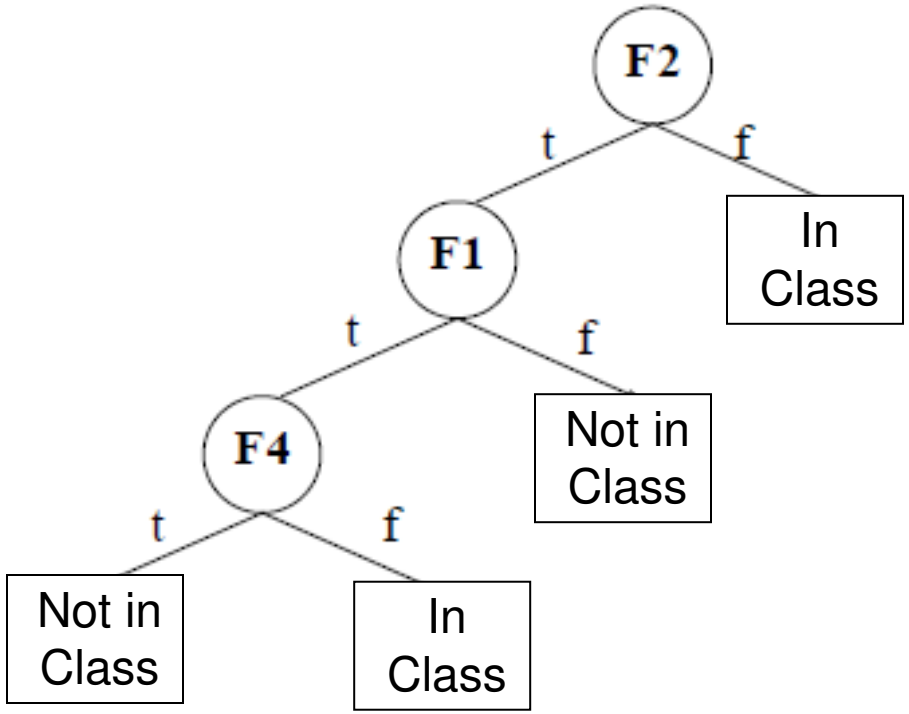
$-1/2\log(1/2)-1/2\log(1/2)-[1/2(-1/1\log(1/1))+1/2(-1/1\log(1/1)] = \mathbf{1}$

$H(\text{class}) - H(\text{class}/F_5) =$

$-1/2\log(1/2)-1/2\log(1/2)-[1/2(-1/1\log(1/1))+1/2(-1/1\log(1/1)] = \mathbf{1}$

- F_4 and F_5 have the maximum gain, we break the tie in favor of F_4 to be F_1's T-child.
- Then, determine F_4's F-child.
  - Since the only example is 'In Class', it becomes F_4's F-child
- Then, choose either F_3 or F_5 to be F_4's T-child.
  - Since the only example is 'Not in Class', it becomes F_4's T-child

Exercise



Application Problems – Decision Trees

The initial entropy of the training sample:

$$E(S) = -(\frac{5}{14}log_2\frac{5}{14} + \frac{9}{14}log_2\frac{9}{14}) = 0.9403$$

$$InfoGain(S,T) = 0.9403 - \frac{4}{14} - \frac{6}{14}(-(\frac{2}{6}log_2\frac{2}{6} + \frac{4}{6}log_2\frac{4}{6})) - \frac{4}{14}(-(\frac{1}{4}log_2\frac{1}{4} + \frac{3}{4}log_2\frac{3}{4})) = 0.0292$$

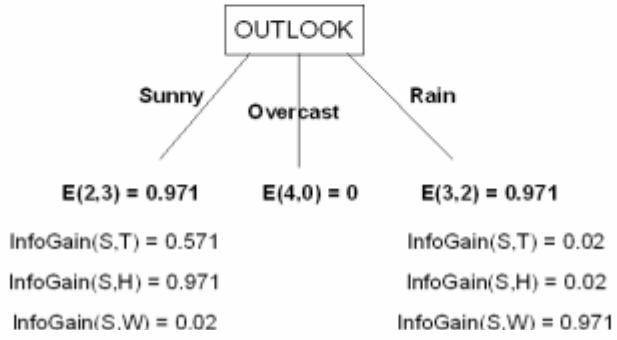
$$InfoGain(S,H) = 0.9403 - \frac{7}{14}(-(\frac{4}{7}log_2\frac{4}{7} + \frac{3}{7}log_2\frac{3}{7})) - \frac{7}{14}(-(\frac{6}{7}log_2\frac{6}{7} + \frac{1}{7}log_2\frac{1}{7})) = 0.1518$$

$$InfoGain(S,W) = 0.9403 - \frac{8}{14}(-(\frac{2}{8}log_2\frac{2}{8} + \frac{6}{8}log_2\frac{6}{8})) - \frac{6}{14} = 0.0481$$

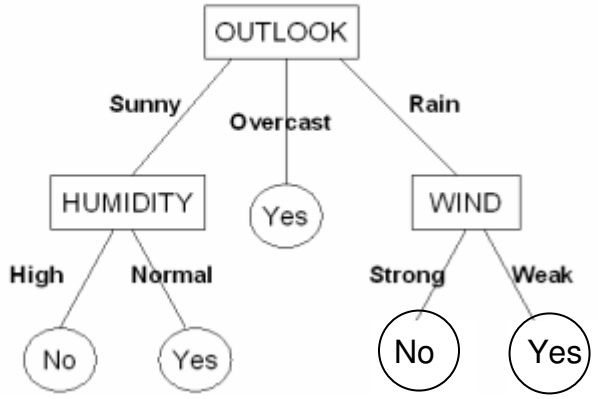
$$InfoGain(S,0\ ) = 0.9403 - \frac{5}{14}(-(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5})) - \frac{5}{14}(-(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}) - \frac{4}{14}0) = 0.2468$$

The first arrtribute to split on is therefore: OUTLOOK.

Next, we choose an attribute to split on in every leaf of the tree:



The fully developed tree is:



Application Problems – Decision Trees

2. For example : D1, D2, D4, D10, D11, D12

3.