CS4600 - Introduction to Intelligent Systems

Homework 8 - Decision Trees - Sample Solution

Assume that you have the following training examples available:

	F1	F2	F3	F4	F5	Class
Example 1	t	t	f	f	f	p
Example 2	f	f	t	t	f	p
Example 3	t	f	f	t	f	p
Example 4	t	f	t	f	t	p
Example 5	f	t	f	f	f	n
Example 6	t	t	f	t	t	n
Example 7	f	t	t	t	t	n

Use all of the training examples to construct a decision tree. In case of ties between features, break ties in favor of features with smaller numbers (for example, favor F1 over F2, F2 over F3, and so on).

How does the resulting decision tree classify the following example:

	F1	F2	F3	F4	F5	Class
Example 8	f	f	f	t	t	?

Some formula:

$$\text{Gain}(A) = I(p,n) - E(A)$$

$$E(A) = \frac{p_t+n_t}{p+n} I(p_t+n_t) + \frac{p_f+n_f}{p+n} I(p_f+n_f)$$

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Pre-compute some I(x,y)

$I(0,x) = I(x,0) = -1 \log_2 1 - 0 \log_2 0 = 0$
$I(x,x) = - 1/2 \log_2 1/2 - 1/2 \log_2 1/2 = 1$
$I(1,2) = I(2,1) = - 2/3 \log_2 2/3 - 1/3 \log_2 1/3 = 0.918$
$I(1,3) = I(3,1) = - 3/4 \log_2 3/4 - 1/4 \log_2 1/4 = 0.811$
$I(3,4) = I(4,3) = - 4/7 \log_2 4/7 - 3/7 \log_2 3/7 = 0.985$

First, choose from {F1, F2, F3, F4, F5} to become the root.

	F1	F2	F3	F4	F5	Class
Example 1	t	t	f	f	f	p
Example 2	f	f	t	t	f	p
Example 3	t	f	f	t	f	p
Example 4	t	f	t	f	t	p
Example 5	f	t	f	f	f	n
Example 6	t	t	f	t	t	n
Example 7	f	t	t	t	t	n

E(F1) = 4/7 * I(3,1) + 3/7 * I(1,2) = 4/7 * 0.811 + 3/7 * 0.918 = 0.857
E(F2) = 4/7 * I(1,3) + 3/7 * I(3,0) = 4/7 * 0.811 + 3/7 * 0 = 0.463
E(F3) = 3/7 * I(2,1) + 4/7 * I(2,2) = 3/7 * 0.918 + 4/7 * 1 = 0.965
E(F4) = 4/7 * I(2,2) + 3/7 * I(2,1) = 4/7 * 1 + 3/7 * 0.918 = 0.965
E(F5) = 3/7 * I(1,2) + 4/7 * I(3,1) = 3/7 * 0.918 + 4/7 * 0.811 = 0.857

Gain(F1) = I(4,3) - E(F1) = 0.128
Gain(F2) = I(4,3) - E(F2) = 0.522
Gain(F3) = I(4,3) - E(F3) = 0.020
Gain(F4) = I(4,3) - E(F4) = 0.020
Gain(F5) = I(4,3) - E(F5) = 0.128

Since Gain(F2) is the highest, F2 becomes the root.

Then, choose from {F1, F3, F4, F5} to be F2’s f-child.

	F1	F2	F3	F4	F5	Class
Example 2	f	f	t	t	f	p
Example 3	t	f	f	t	f	p
Example 4	t	f	t	f	t	p

Since all examples are in class p, it becomes F2's f-child

Next, choose from {F1, F3, F4, F5} to be F2's t-child.

	F1	F2	F3	F4	F5	Class
Example 1	t	t	f	f	f	p
Example 5	f	t	f	f	f	n
Example 6	t	t	f	t	t	n
Example 7	f	t	t	t	t	n

$$
\begin{aligned}
E(F1) &= 2/4 * I(1,1) + 2/4 * I(0,2) = 2/4 * 1 + 2/4 * 0 = 0.5 \\
E(F3) &= 1/4 * I(0,1) + 3/4 * I(1,2) = 1/4 * 0 + 3/4 * 0.918 = 0.6885 \\
E(F4) &= 2/4 * I(0,2) + 2/4 * I(1,1) = 2/4 * 0 + 2/4 * 1 = 0.5 \\
E(F5) &= 2/4 * I(0,2) + 2/4 * I(1,1) = 2/4 * 0 + 2/4 * 1 = 0.5
\end{aligned}
$$

$$
\begin{aligned}
Gain(F1) &= I(1,3) - E(F1) = 0.311 \\
Gain(F3) &= I(1,3) - E(F3) = 0.1225 \\
Gain(F4) &= I(1,3) - E(F4) = 0.311 \\
Gain(F5) &= I(1,3) - E(F5) = 0.311
\end{aligned}
$$

F1, F4, and F5 have the maximum Gain(), we break ties in favor of features with smaller numbers and thus choose F1 to be F2's t-child.

Then, choose from {F3, F4, F5} to be F1's f-child.

	F1	F2	F3	F4	F5	Class
Example 5	f	t	f	f	f	n
Example 7	f	t	t	t	t	n

Therefore, class "n" becomes F1's f-child.

Then, choose from {F3, F4, F5} to be F1's t-child.

	F1	F2	F3	F4	F5	Class
Example 1	t	t	f	f	f	p
Example 6	t	t	f	t	t	n

E(F3) = 0/2 * I(0,0) + 2/2 * I(1,1) = 0/2 * 0 + 2/2 * 1 = 1
E(F4) = 1/2 * I(0,1) + 1/2 * I(1,0) = 1/2 * 0 + 1/2 * 0 = 0
E(F5) = 1/2 * I(0,1) + 1/2 * I(1,0) = 1/2 * 0 + 1/2 * 0 = 0

Gain(F3) = I(1,3) - E(F3) = 0
Gain(F4) = I(1,3) - E(F4) = 1
Gain(F5) = I(1,3) - E(F5) = 1

F4 and F5 have the highest Gain(). F4 are favored by the tie-breaking scheme and, thus, becomes F1's t-child.

Next, choose either F3 or F5 to be F4's f-child.

	F1	F2	F3	F4	F5	Class
Example 1	t	t	f	f	f	p

Since the only example has class p, "p" becomes F4's f-child.

Then, choose either F3 or F5 to be F4's t-child.

	F1	F2	F3	F4	F5	Class
Example 6	t	t	f	t	t	n

For similar reason as before, class n becomes F4's t-child.

The final tree below will classify example8 (f,f,f,t,t) as belonging to class p.