Data-X HW 7 Decision Trees

October 16, 2018

```
In [1]: from IPython.display import display, Latex, Markdown
```

1.

Firstly, we have $P(Defaulter = 1) = 0.5$, $P(Defaulter = 0) = 0.5$.
So $H(Defaulter) = 0.5 * log_2(2) + 0.5 * log_2(2) = 1$
For variable 'HasJob',
we have $P(Defaulter = 1|HasJob = 1) = 2/5$, $P(Defaulter = 1|HasJob = 0) = 2/3$.
So,
$H(Defaulter|HasJob = 1) = \frac{2}{5} * log_2(\frac{5}{2}) + \frac{3}{5} * log_2(\frac{5}{3}) = 0.97$
$H(Defaulter|HasJob = 0) = \frac{2}{3} * log_2(\frac{3}{2}) + \frac{1}{3} * log_2(3) = 0.92$
$H(Defaulter|HasJob) = 0.97 * \frac{5}{8} + 0.92 * \frac{3}{8} = 0.95125$
Info Gained of 'HasJob' $= H(Defaulter)H(Defaulter|HasJob) = 1 - 0.95125 = 0.04875$
For variable 'HasFamily',
we have $P(Defaulter = 1|HasFamily = 1) = 1/4$, $P(Defaulter = 1|HasFamily = 0) = 3/4$.
So,
$H(Defaulter|HasFamily = 1) = \frac{1}{4} * log_2(4) + \frac{3}{4} * log_2(\frac{4}{3}) = 0.81$
$H(Defaulter|HasFamily = 0) = \frac{3}{4} * log_2(\frac{4}{3}) + \frac{1}{4} * log_2(4) = 0.81$
$H(Defaulter|HasFamily) = 0.81 * \frac{1}{2} + 0.81 * \frac{1}{2} = 0.81$
Info Gained of 'HasFamily' $= H(Defaulter)H(Defaulter|HasFamily) = 1 - 0.81 = 0.19$
For variable 'IsAbove30years',
we have $P(Defaulter = 1|IsAbove30years = 1) = 1/2$, $P(Defaulter = 1|HasFamily = 0) = 1/2$.
So,
$H(Defaulter|IsAbove30years = 1) = \frac{1}{2} * log_2(2) + \frac{1}{2} * log_2(2) = 1$
$H(Defaulter|IsAbove30years = 0) = \frac{1}{2} * log_2(2) + \frac{1}{2} * log_2(2) = 1$
$H(Defaulter|IsAbove30years) = \frac{6}{8} * 1 + \frac{2}{8} * 1 = 1$
Info Gained of 'IsAbove30years' $= H(Defaulter)H(Defaulter|IsAbove30years) = 1 - 1 = 0$
As a result, the best feature to do the first split in a binary decision tree in order to maximize the information gain in the next split is 'HasFamily'.

2.

$h(A) = log_2(10/7) = 0.5146bit$
$h(B) = log_2(5) = 2.32bits$
$h(C) = log_2(10) = 3.32bits$
$H(S) = \frac{7}{10} * log_2(10/7) + \frac{1}{5} * log_2(5) + \frac{1}{10} * log_2(10) = 1.157bits$
According to Source Coding Theorem:

The H(S) is the smallest codeword length that is theoretically possible for signal ‘S’, which means that theoretically the smallest codeword length of S is 1.157 bits per symbol