FMCS1 Lab 5
Solutions

Manuel Marques-Pita and Frank Keller

April 6, 2005

Task 1

Question 1

There are 1024 items. Each item is encoded using two bits, the size of the inventory is therefore $1024 \cdot 2 = 2048$ bits.

Question 2

A shorter code could be the following prefix-free code:

$$\begin{aligned} A &\rightarrow 1 \\ B &\rightarrow 10 \\ C &\rightarrow 110 \\ D &\rightarrow 111 \end{aligned}$$

The size of the inventory using this new code is then:

$$(512 \cdot 1) + (256 \cdot 2) + (2 \cdot 3 \cdot 128) = 1792$$

Question 3

The entropy of the set is:

$$\begin{aligned} H(512, 256, 128, 128) &= -(512/1024 \cdot \log 512/1024 + \\ &\quad 256/1024 \cdot \log 256/1024 + \\ &\quad 128/1024 \cdot \log 128/1024 + \\ &\quad 128/1024 \cdot \log 128/1024) \\ &= 1.75 \end{aligned}$$

Multiplying this number by the number of items in the inventory gives us a lower bound for the size the encoded inventory will require. In this case $1024 \cdot 1.75 = 1792$, which is already the size of the inventory with the new code above.

Question 4

First of all we need to calculate the entropy of the whole set, based on the decisions made (yes/no). There are 6 examples with outcome = yes and 4 with outcome = no:

$$H(6,4) = -(6/10 \cdot \log(6/10) + 4/10 \cdot \log(4/10)) = 0.9710$$

Question 5

For the split on attribute 'Hungry', we need to check, for each attribute value of 'Hungry', which examples are classified as positive and which ones as negative. 'Hungry' has two values (Yes, No):

For Hungry = Yes, E1, E5, E6 and E7 are positive (the frog will eat when 'Hungry'); E2 is negative (the frog won't eat even though it is 'Hungry'). For Hungry = No: E4 and E10 are positive; E3, E8 and E9 are negative.

Notice that the partition classifies a proportion: $P(\text{Hungry} = \text{Yes}) = 5/10$ for value 'Yes' and $P(\text{Hungry} = \text{No}) = 5/10$ for value 'No'.

For the value 'Yes', the entropy of the subset would be: $H(4,1)$ (four positives, one negative):

$$H(4,1) = -(4/5 \cdot \log(4/5) + 1/5 \cdot \log(1/5)) = 0.7219$$

For the value 'No' the entropy of the subset would be: $H(2,3)$ (two positives and 3 negatives), i.e., 0.9710. Using the formulae in the handout, we need:

1. $H(X)$, which is 0.9710 (calculated in Q4);
2. $H(X|Y)$, the entropy of the set given this partition on 'Hungry':

$$\begin{aligned} H(\text{Set}|\text{Hungry}) &= P(\text{Hungry} = \text{Yes}) \cdot H(4,1) \; + \\ &\quad P(\text{Hungry} = \text{No}) \cdot H(2,3) \\ &= 0.8464 \end{aligned}$$

3. The mutual information:

$$I(\text{Set}; \text{Hungry}) = 0.9710 - 0.8464 = 0.1246$$

Then the same process is repeated for 'Competition':

$$I(\text{Set}; \text{Competition}) = 0.2955$$

and for 'Food_value':

$$I(\text{Set}; \text{Food\_value}) = 0.0464$$

The highest Mutual Information (also called Information Gain) Takes place when the set is partitioned on the attribute Competition. This means, that the first question the frog should ask (given the data it has) is the value for this attribute.