T-79.230 **Kevät 2003**
Agenttipohjaisen tietojenkäsittelyn perusteet
Laskuharjoitus 5
Solutions

1. We construct a decision tree from sample data by adding each variable to it one at a time. We will use information theoretic choice criterion: on each step we add the variable that adds most information (= that decreases the entropy the most) about the distribution.

   We can compute the information (entropu)) I of a probability distribution P as follows: laskemalla

$$I(P(v_1), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$$

   In the sample data x_1, x_2, x_6, and x_7 are birds while x_3, x_4, x_5, x_8, and x_9 are not. Thus, the initial information is:

$$I(\frac{4}{9}, \frac{5}{9}) = \frac{4}{9} \log_2 \frac{9}{4} + \frac{5}{9} \log_2 \frac{9}{5} = 0.991$$

   Next, we go through all variables and compute which one of them is the best choise. The first variable is “flies?”. When we examine the sample data we notice that two birds (x_1 and x_2) fly and two (x_6, x_7) do not. Of the other animals only, x_4 and x_5 fly. Thus, the entropy is:

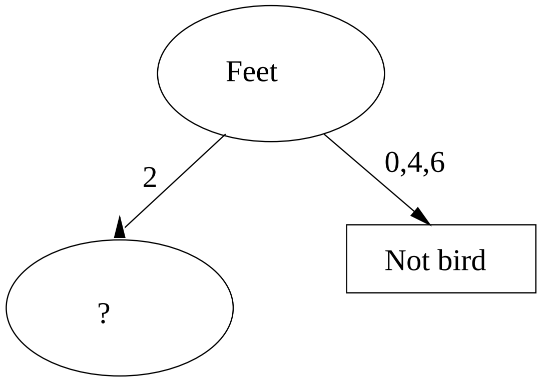
$$I(\text{Flies?}) = \frac{4}{9} I(\frac{2}{4}, \frac{2}{4}) + \frac{5}{9} I(\frac{2}{5}, \frac{3}{5}) = 0.984$$

   Here the left term is the case of flying animals, and the right one the case of non-flying ones. We notice that the node “flies?” does not help us much.

   On the other hand, the number of legs is much better choice, since all birds are two-legged:

$$I(\text{Legs}) = \frac{5}{9} I(\frac{4}{5}, \frac{1}{5}) + \frac{4}{9} I(0, 1) = 0.401$$

   When we compute the entropies of the rest choices, we notice that “legs” is the best choice for the root of the tree.



Next we try to find the best decision variable for the node with '?'.

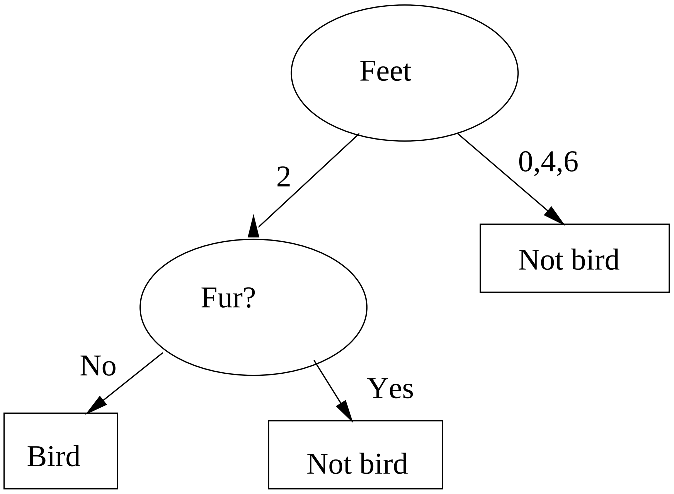
For the variable "flies?" the rest of samples divide so that x_1, x_2, and x_5 fly, while x_6 and x_7 do not, so:

$$I(\text{Flies?}) = \frac{3}{5}I(\frac{2}{3}, 13) + \frac{2}{5}I(1,0) = 0.551$$

Also in this case "flies?" is not the best one, since by choosing "fur?" the birds can be completely separated from the other animals:

$$I(fur?) = \frac{1}{5}I(0,1) + \frac{4}{5}I(1,0) = 0$$

Thus, we have constructed a decision tree that works correctly with the sample data.



The main problem with learning decision trees is that they are only as good as the sample data. If the sample data is not complete enough, the tree may give completely wrong answers. Also, if there is too much data and too many variables, the algorigthm can find correlations that are actually only statistical oddities.

The tree above has too little sample data. Since a human has two legs and no fur, the tree classifies a person as a bird.