AI2 Module 3
Tutorial 2: Sample Solutions

Jacques Fleuriot
School of Informatics (Amended David Talbot January 30, 2005.)

Part 1

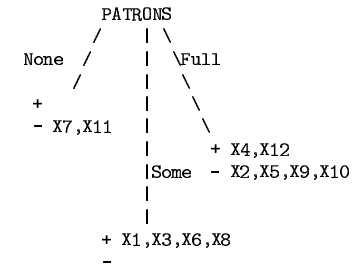
The main aim for this tutorial is to gain experience with the algorithm for learning decision trees.

1. We use the notation [number-positive, number-negative] to describe the splits. We start with all the examples and a split of [6,6]. Using the entropy formula from the slides:

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

We get $I(6,6) = 2(1/2)\log 2 = 1$.

If we split on Patrons we get the following:



that is, 3 partitionss with ([0,2], [4,0], [2,4]) splits. Now, recalling that $I(X,0) = I(0,X) = 0$ we have

$$\begin{aligned} \mathit{Remainder}(\mathit{Patrons}) &= (2/12)\cdot 0 + (4/12)\cdot 0 + (6/12)[(2/6)\log(6/2) + (4/6)\log(6/4)] \\ &= [(2/12)\log(6/2) + (4/12)\log(6/4)] = 0.459 \end{aligned}$$

Type yields (ordered (French, Italian, Thai, Burger)) 4 parts with ([1,1], [1,1], [2,2], [2,2]). Now, recalling that $I(X,X) = 1$ we have

$$\mathit{Remainder}(\mathit{Type}) = 2(2/12)1 + 2(4/12)1 = 1$$

Hungry yields (ordered (Yes, No)) 2 parts with ([4,2], [2,4]) and

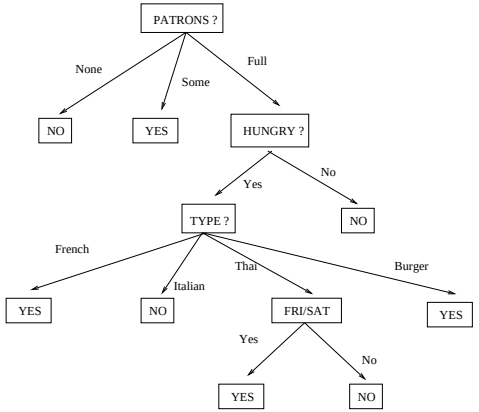
$$\mathit{Remainder}(\mathit{Hungry}) = 2(6/12)[(2/6)\log(6/2) + (4/6)\log(6/4)] = 0.918$$

Now $\mathit{Gain}() = I() - \mathit{Remainder}()$ is $0.541, 0, 0.082$ respectively, and clearly Patrons wins.

Note that the relative order that we get seems intuitively reasonable given the partitions (the more skewed the better). We base the next steps on this intuition alone.

2. After each outcome splits up the examples, the result is a new decision tree learning problem with fewer examples, and one fewer attribute. Recursively, choose the next most important attribute to classify the remaining set.

   In the next steps the attributes to choose are HUNGRY, followed by TYPE, followed by FRI/SAT.



3. If we accept this resulting tree as the solution, we notice that:

   The tree has never seen a case where the wait is 0-10 minutes but the restaurant is full — in this case if you were hungry you would most certainly wait, but the tree will answer NO.

   This indicates that our tree may not be so good. One cause for this is that we had a rather small number of examples. With more examples, *hopefully*, the statistics for each leaf will be reliable and such things will not happen.

   Another cause, of course, is that we only have a heuristic and in some cases it just indicates the wrong thing ...

Part 2

The average entropy *after* the split on A is given by

$$\mathit{Remainder}(A) = \sum_{i=1}^{3}\frac{p_i+n_i}{p+n}I(p_i,n_i)$$

where p_i and n_i are the number of positive and negative examples, respectively, in subset i after the split, and p and n are the number of positive and negative examples, respectively, before the split.

The information gain (improvement) after the split is given by

$$\mathit{Gain}(A) = I(p,n) - \mathit{Remainder}(A)$$

Thus, for given example, we have

$$Gain(A) = I(6,14) - (\frac{10}{20}I(2,8) + \frac{2}{20}I(0,2) + \frac{8}{20}I(4,4))$$

$$= (\frac{6}{20}\log\frac{20}{6} + \frac{14}{20}\log\frac{20}{14}) - (\frac{10}{20}(\frac{2}{10}\log\frac{10}{2} + \frac{8}{10}\log\frac{10}{8}) + \frac{2}{20}(0) + \frac{8}{20}(1))$$

$$= \frac{3}{10}\log\frac{10}{3} + \frac{7}{10}\log\frac{10}{7} - \frac{1}{10}\log 5 - \frac{2}{5}\log\frac{5}{4} - \frac{2}{5}$$