CIS 572 Assignment 1

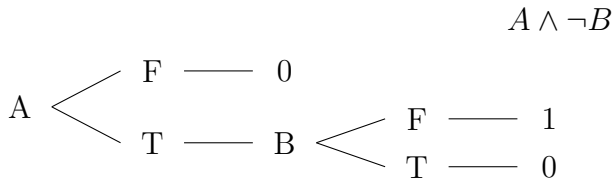
Steven Walton

April 7, 2019

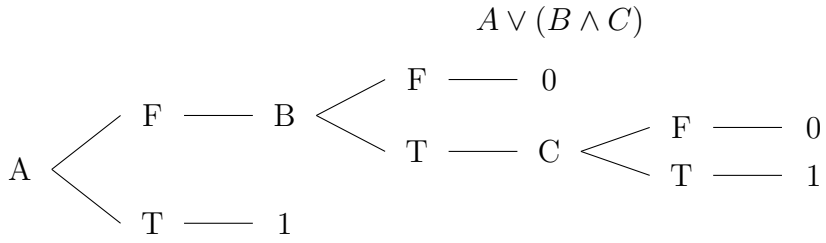
Problem 1

Answer Exercise 3.1 from Chapter 3 of Mitchell's machine learning book.

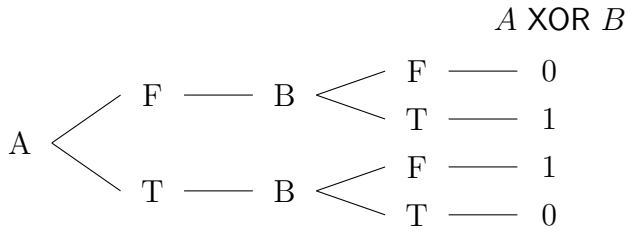
Part a)



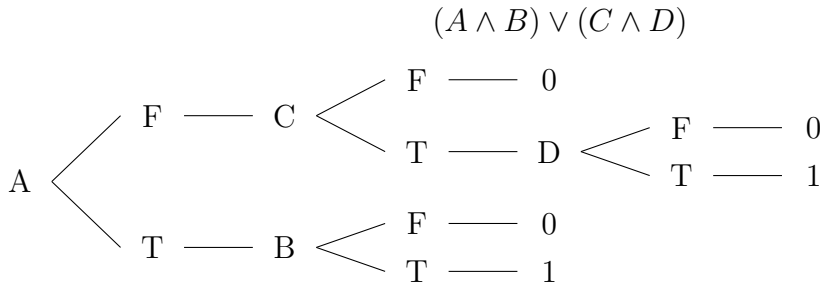
Part b)



Part c)



Part d)

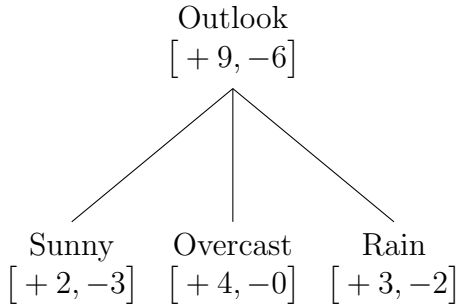


Problem 2

Consider the samples in the Play-tennis dataset from Table 3.2 in Mitchell's textbook. If you calculate the information-gain for all of the attributes of this set, you will observe that the attribute "Outlook" has the largest information- gain, which is equal to 0.246. Therefore, the attribute "Outlook" is the best heuristic choice for the root node.
(a) List the labels of the new tree branches below the root node.
(b) Which partition of the data will be assigned to each branch by ID3? Please list the sample IDs that will be assigned to each branch.
(c) Calculate the information gain for the remaining attributes in each branch, and determine which attribute will be chosen as the root of the sub-tree in each branch.

Part a)

Now that we know that Overcast has the highest information gain, of 0.246, we will use it as the root node. We will then create the next nodes and show their corresponding values for playing tennis or not, in the form of $\big[+yes, -no\big]$



Part b)

To determine how we will partition the data we will look at the entropy of each node.
Given sunny, we have:
Humidity: High $\big[+0, -3\big]$, Normal $\big[+2, -0\big]$
Temperature: High $\big[+0, -2\big]$, Mild $\big[+1, -1\big]$, Cold $\big[+1, -0\big]$

Wind: Strong $\big[+1,-1\big]$, Weak $\big[+1,-2\big]$

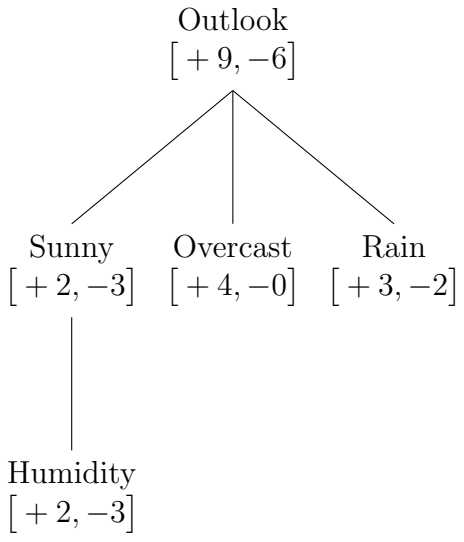
$$\begin{aligned} S(Sunny, Humidity) &= S_{sunny} - S_{sunny,high} - S_{sunny,low} \\ &= S_{sunny} - \frac{3}{5}(-1\log_2 1) - \frac{2}{5}(-1\log_2 1) \\ &= 0.97 - \frac{3}{5}0 - \frac{2}{5}0 \\ &= 0.97 \end{aligned}$$

We'll shorten steps from now on. We note that any $\big[+a,-a\big] = 1$ and $\big[+a,0\big] = \big[0,-a\big] = 0$

$$\begin{aligned} S(Sunny, Temperature) &= S_{sunny} - S_{sunny,hot} - S_{sunny,mild} - S_{sunny,cold} \\ &= 0.97 - \frac{2}{5}0 - \frac{2}{5}1 - \frac{2}{5}0 \\ &= 0.57 \end{aligned}$$

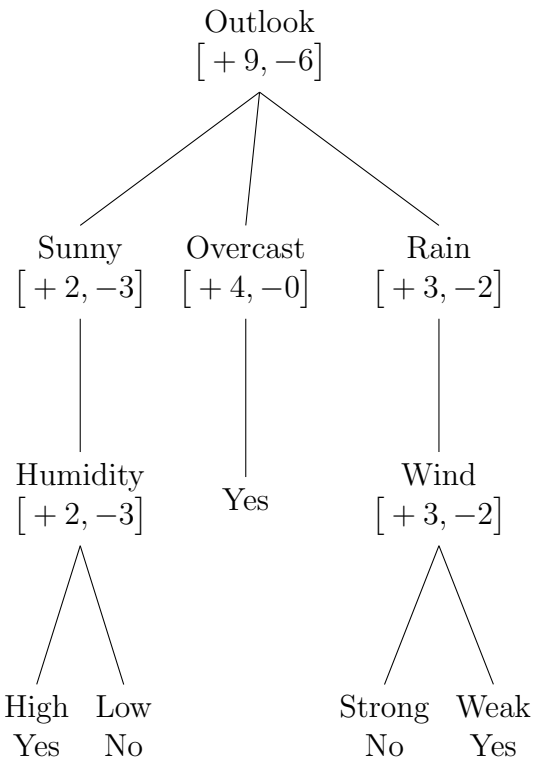
$$\begin{aligned} S(Sunny, Wind) &= S_{sunny} - S_{sunny,strong} - S_{sunny,weak} \\ &= 0.97 - \frac{2}{5}1 - \frac{3}{5}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) \\ &= 0.019 \end{aligned}$$

From here we can see that the best thing to split on is humidity and the worst is wind. That gives us the following update to our tree.



Part c)

Continuing with this process we get the following tree



Problem 3

Suppose a bank makes loan decisions using two decision trees, one that uses attributes related to credit history and one that uses other demographic attributes. Each decision tree separately classifies a loan application as "High Risk" or "Low Risk". The bank only offers a loan when both decision trees predict "Low Risk"

(a) Describe an algorithm for converting this pair of decision trees into a single decision tree that makes the same predictions (that is, it predicts non-risky only when both of the original decision trees would have predicted non-risky).

(b) Let n_1 and n_2 be the number of leaves in the first and second decision trees, respectively. Provide an upper bound on n, the number of leaves in the single equivalent decision tree, expressed as a function of n_1 and n_2.