

CIS 572 Assignment 1

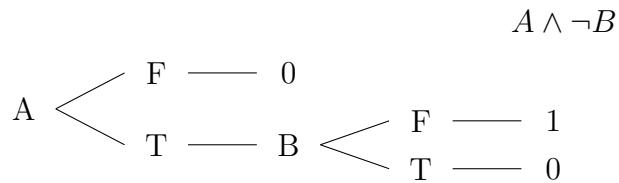
Steven Walton

April 7, 2019

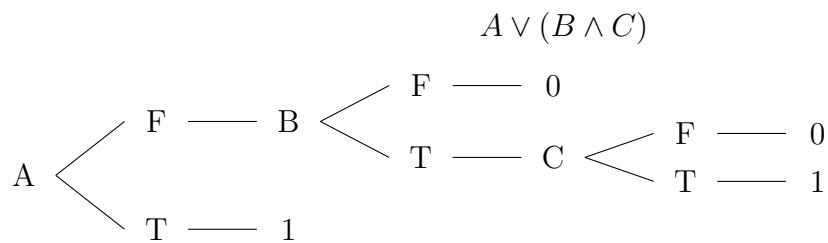
Problem 1

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

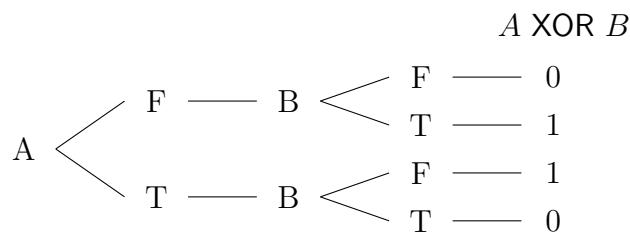
Part a)



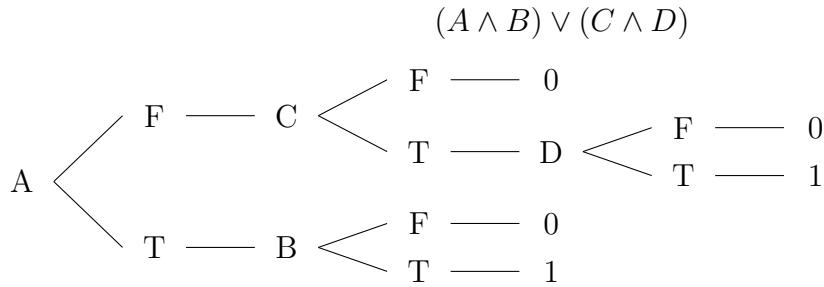
Part b)



Part c)



Part d)



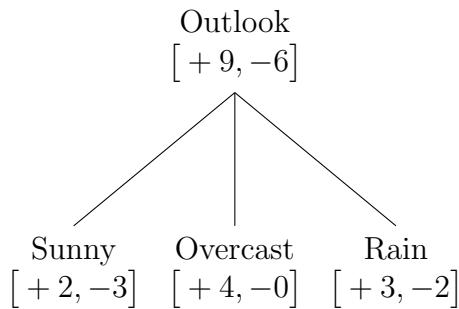
Problem 2

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 $[+yes, -no]$



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 $[+ 0, -3]$, Normal $[+ 2, -0]$

Temperature: High $[+ 0, -2]$, Mild $[+ 1, -1]$, Cold $[+ 1, -0]$

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

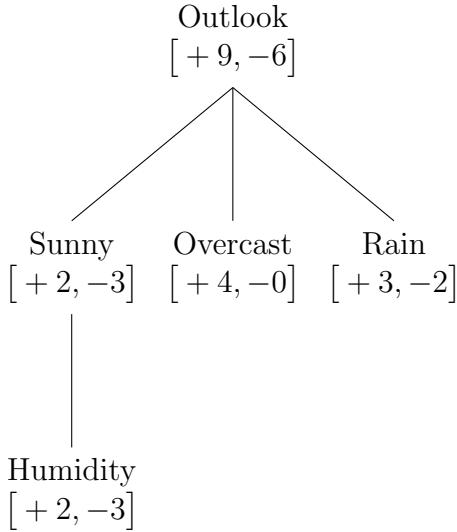
$$\begin{aligned}
 S(Sunny, Humidity) &= S_{sunny} - S_{sunny,high} - S_{sunny,low} \\
 &= S_{sunny} - \frac{3}{5}(-1 \log_2 1) - \frac{2}{5}(-1 \log_2 1) \\
 &= 0.97 - \frac{3}{5}0 - \frac{2}{5}0 \\
 &= 0.97
 \end{aligned}$$

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

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

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

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

