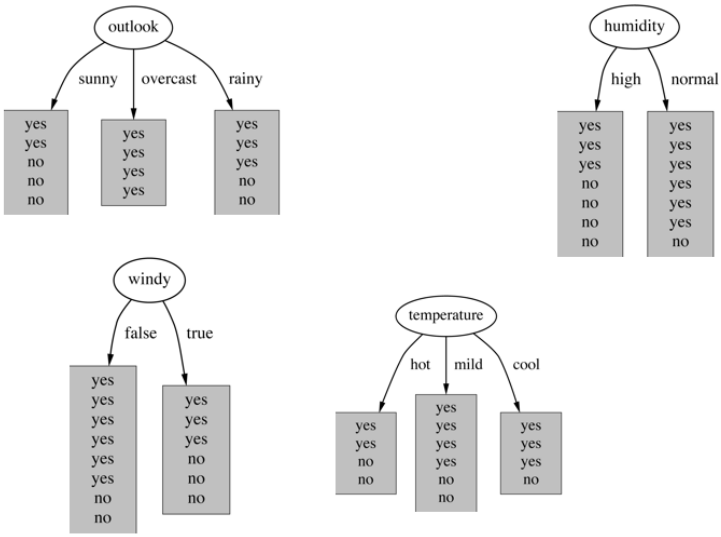
Data Mining, CSCI 347, Fall 2019
Decision Trees, Sept. 18

A simple decision tree algorithm is to calculate the information gained when splitting on each attribute, split on the attribute that yields the greatest gain, and recursively repeat this process for each branch.

Consider the weather dataset and splitting on the outlook attribute.



1. Calculate the information of the top node.

   info([9,5]) = entropy(9/14, 5/14)
   = -9/14 ($\log_2$ 9/14) – 5/14 ($\log_2$ 5/14)
   = -9/14 ($\log_{10}$ 9/14)/ ($\log_{10}$ 2) – 5/14 ($\log_{10}$ 5/14)/ ($\log_2$ 2)
   = 0.940 bits

2. Give the formula to calculate the information gained by splitting on Outlook.

   gain(Outlook) = info([9,5]) – info([2,3], [4,0], [3,2])

3. Calculate the purity of each of the three branches.

Outlook = sunny
info([2,3]) = entropy(2/5, 3/5)
= -2/5 ($\log_2$ 2/5) – 3/5 ($\log_2$ 3/5)
= -2/5 ($\log_{10}$ 2/5)/ ($\log_{10}$ 2) – 3/5 ($\log_{10}$ 3/5)/ ($\log_{10}$ 2)
= 0.971 bits

Outlook = overcast
info([4,0]) = entropy(4/4, 0/4)
= entropy(1,0)
= -1 ($\log_2$ 1) – 0 ($\log_2$ 0)
Ordinarily $\log_2$ 0 is not defined, ok since times 0
= -1 ($\log_{10}$ 1)/ ($\log_{10}$ 2) – 0
= -1 (0/($\log_{10}$ 2)) – 0
= 0 bits

Outlook = rainy
info([3,2]) = entropy(3/5, 2/5)
= entropy(2/5, 3/5)
= -2/5 ($\log_2$ 2/5) – 3/5 ($\log_2$ 3/5)
= -2/5 ($\log_{10}$ 2/5)/ ($\log_{10}$ 2) – 3/5 ($\log_{10}$ 3/5)/ ($\log_{10}$ 2)
= 0.971 bits

4. Calculate the weighted average information value for splitting on Outlook

info([3,2], [4,0], [3,2]))
= 5/14 * 0.971 + 4/14 * 0 + 5/14 * 0.971
= 0.693 bits

5. Information gain from splitting on Outlook.

info([9,5]) - info([3,2], [4,0], [3,2]))
= 0.940 – 0.693
= 0.247 bits

6. Calculate the intrinsic information value of the split

info([5,4,5]) = entropy(5/14, 4/14, 5/14)
= -5/14 ($\log_2$ 5/14) – 4/14 ($\log_2$ 4/14) – 5/14 ($\log_2$ 5/14
= -5/14 ($\log_{10}$ 5/14)/($\log_{10}$ 2) – 4/14 ($\log_{10}$ 4/14)/ ($\log_{10}$ 2)
– 5/14 ($\log_{10}$ 5/14)/($\log_{10}$ 2)
= 1.577 bits

7. Calculate the ratio by dividing the attribute gain by the intrinsic information value

Gain Ratio = Gain from Attribute / Intrinsic Value of Split
= 0.247/1.577
= 0.157

8. Repeat the process the other attributes.

Gain ratio = information gain of attribute split / intrinsic information of attribute split
= (information at root – average information for attribute split)
/ intrinsic information of attribute split

Gain ratio for splitting on Outlook
= (information at root – average information for attribute split)
/ intrinsic information of attribute split
= (info([9,5]) – info([2,3],[4,0],[3,2])) / info([5,4,5])
= (info([9,5]) – (5/14*info([2,3]) + 4/14*info([4,0]) + 5/14*info([3,2])
/ info([5,4,5])
= (entropy(9/14,5/14) –
(5/14*entropy(2/5,3/5) + 4/14*entropy(4/4,0/4) +
5/14*entropy(3/5,2/5)
/ entropy(5/14,4/14,5/14)
= (entropy(9/14,5/14) –
(5/14*entropy(2/5,3/5) + 4/14*entropy(4/4,0/4) +
5/14*entropy(3/5,2/5)
/ entropy(5/14,4/14,5/14)
= ($-9/14*\log_2(9/14)-5/14*\log_2(5/14)$) –
($5/14*[-2/5*\log_2(2/5)-3/5*\log_2(3/5)]$
$+ 4/14*[-4/4*\log_2(4/4)-0/4*\log_2(0/4)]$
$+ 5/14*[-3/5*\log_2(3/5)-2/5*\log_2(2/5)]$
/ $(-5/14*\log_2(5/14)-4/14*\log_2(4/14)-5/14*\log_2(5/14))$
= (0.940 bits –
(5/14*0.971 bits + 4/14* 0 bits + 5/14 * 0.971 bits)
/ 1.577 bits
= (0.940 bits –
(0.347 bits + 0 bits + 0.347 bits)
/ 1.577 bits
= (0.940 bits – 0.694 bits) / 1.577 bits
= 0.246 bits / 1.577 bits
= 0.156

Gain ratio for splitting on Temperature

= (information at root – average information for attribute split)
/ intrinsic information of attribute split

= (info([9,5]) – info([2,2],[4,2],[3,1])) / info([4,6,4])

= (info([9,5]) – (4/14*info([2,2]) + 6/14*info([4,2]) + 4/14*info([3,1]))
/ info([4,6,4])

= (entropy(9/14,5/14) –
(4/14*entropy(2/4,2/4) + 6/14*entropy(4/6,2/6) +
4/14*entropy(3/4,1/4))
/ entropy(4/14,6/14,4/14)

= (entropy(9/14,5/14) –
(4/14*entropy(2/4,2/4) + 6/14*entropy(4/6,2/6) +
4/14*entropy(3/4,1/4))
/ entropy(4/14,6/14,4/14)

=(0.940 bits – (4/14*1 + 6/14*0.438 + 4/14*0.791) / 1.362 bits

=(0.940 bits – 0.911 bits) / 1.362 bits

= 0.029 bits / 1.577 bits

= 0.021

Gain ratio for splitting on Humidity

= (information at root – average information for attribute split)
/ intrinsic information of attribute split

= (info([9,5]) – info([3,4],[6,1])) / info([7,7])

= (info([9,5]) – (7/14*info([3,4]) + 7/14*info([6,1]))
/ info([7,7])

= (entropy(9/14,5/14) –
(7/14*entropy(3/7,4/7) + 7/14*entropy(6/7,1/7))
/ entropy(7/14,7/14)

= (entropy(9/14,5/14) –
(7/14*entropy(3/7,4/7) + 7/14*entropy(6/7,1/7))
/ entropy(7/14,7/14)

=(0.940 bits – 0.788 bits) / 1 bit

= 0.152 bits / 1 bit

= 0.152

Gain ratio for splitting on Windy

= (information at root – average information for attribute split)
/ intrinsic information of attribute split

= (info([9,5]) – info([6,2],[3,3])) / info([8,6])

= (info([9,5]) – (8/14*info([6,2]) + 6/14*info([3,3]))
/ info([8,6])

= (entropy(9/14,5/14) –
(8/14*entropy(6/8,2/8) + 6/14*entropy(3/6,3/6))
/ entropy(8/14,6/14)

= (entropy(9/14,5/14) –
(8/14*entropy(6/8,2/8) + 6/14*entropy(6/7,1/7))

/ entropy(8/14,6/14)
=(0.940 bits – 0.892 bits) / 0.985 bits
= 0.048 bits / 0.985 bits
= 0.049

9. Based on these calculations, what attribute would appear as the root of the tree?

Outlook because it creates the greatest gain ratio.