Solution to Decision Trees Prac Sheet

Last modified on Tuesday, 16 June 2020 by f.maire@qut.edu.au

Exercise 1

Consider the data table below

D =

X_1	X_2	X_3	X_4	C
F	F	F	F	P
F	F	T	T	P
F	T	F	T	P
T	T	T	F	P
T	F	F	F	N
T	T	T	T	N
T	T	T	F	N

$\mathbf{X} = \{X_1, X_2, X_3, X_4\}$

What is the entropy of D?

Write Pp for the probability of C=P, and Pn for the probability of C=N.

We have Pp = 4/7 and Pn = 3/7

H(D) = - Pn log2(Pn) -Pp log2(Pp)

Numerically

Pp = 0.5714

Pn = 0.4286

The entropy H(D) is 0.9852

What is the information gain of X_1 ?

H(C | X1=N) = H([0-,3+]) = 0

H(C | X1=P) = H([3-,1+]) = 0.8113

H(C | X1) = 3/7 H([0-,3+]) + 4/7 H([3-,1+]) = 0.4636

Information gain of X1 : H(D) - H(C | X1) = 0.5216

What is the information gain of X_2 ?

Similarly, Gain(X 2) = 0.9852-((4/7)*1+(3/7)*0.91829) = 0.02

Build a DT to a depth of 3.

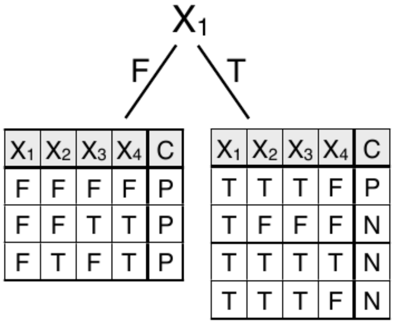
First split

$\mathbf{D}$ =

X_1	X_2	X_3	X_4	C
F	F	F	F	P
F	F	T	T	P
F	T	F	T	P
T	T	T	F	P
T	F	F	F	N
T	T	T	T	N
T	T	T	F	N

$\mathbf{X} = \{X_1, X_2, X_3, X_4\}$

$\text{Gain}(X_1) = 0.52$
$\text{Gain}(X_2) = 0.01$
$\text{Gain}(X_3) = 0.01$
$\text{Gain}(X_4) = 0.01$



X_1	X_2	X_3	X_4	C
F	F	F	F	P
F	F	T	T	P
F	T	F	T	P

X_1	X_2	X_3	X_4	C
T	T	T	F	P
T	F	F	F	N
T	T	T	T	N
T	T	T	F	N

Left subtree,

$\mathbf{D}$ =

X_1	X_2	X_3	X_4	C
F	F	F	F	P
F	F	T	T	P
F	T	F	T	P

$\mathbf{X} = \{X_2, X_3, X_4\}$

All instances have the same class.
Return class P.

Right subtree,

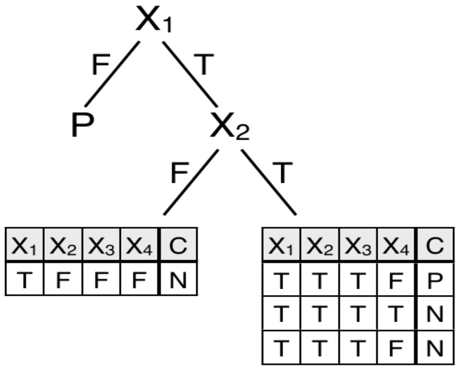
$\mathbf{D}$ =

X_1	X_2	X_3	X_4	C
T	T	T	F	P
T	F	F	F	N
T	T	T	T	N
T	T	T	F	N

$\mathbf{X} = \{X_2, X_3, X_4\}$

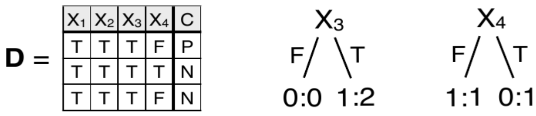
All attributes have same information gain.
Break ties arbitrarily.
Choose X_2

So far we have,



For the left subtree of the tree rooted at X2, all instances have the same class, we will return class N.

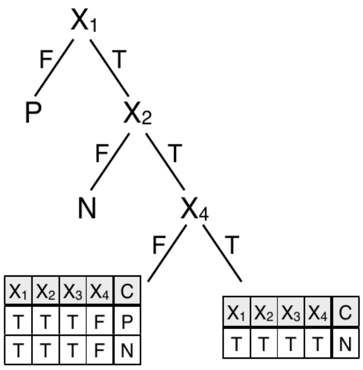
For the right subtree of the tree rooted at X2, we have



X = {X_3,X_4}

X_3 has zero information gain
X_4 has positive information gain
Choose X_4

That is, we now have

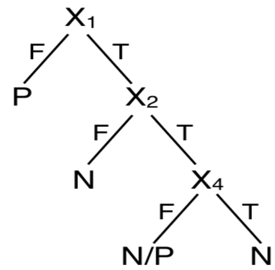


For the left subtree of X4, X3 has zero information gain. No suitable attribute for splitting.

Return most common class (break ties arbitrarily). Note: data is inconsistent!

For the right subtree of X4, All instances have the same class. Return N.

The final tree is



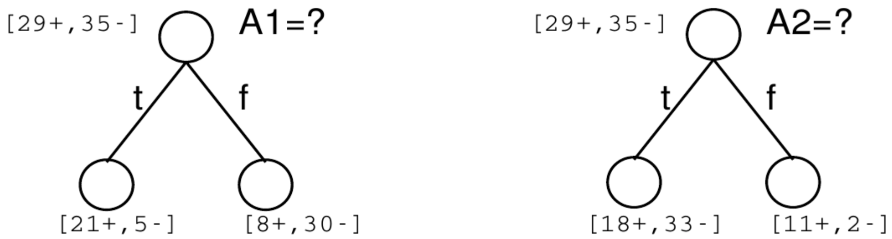
Exercise 2

Recall that

$Gain(S, A) =$ expected reduction in entropy due to sorting on A

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- *Compute the information gain of the attribute A1 and A2*

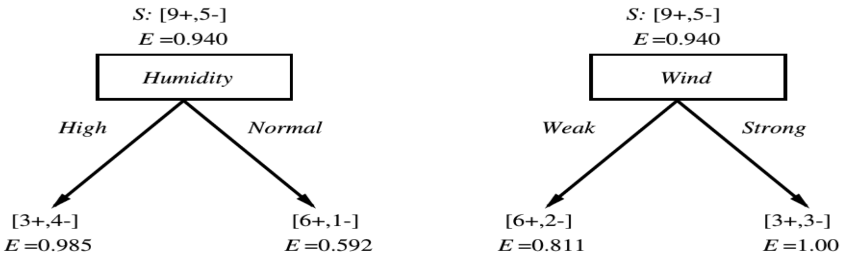


The information gain for A1 and A2 can be computed as follows

```
>>> from math import log
>>> ent = lambda p,n: -p/(p+n)*log(p/(p+n))-n/(p+n)*log(n/(p+n))
>>> IG1 = ent(29,35) - (21+5)/(29+35)*ent(21,5) - (8+30)/(29+35)*ent(8,30)
>>> IG1
0.18429041551228614
>>> IG2 = ent(29,35) - (18+33)/(29+35)*ent(18,33) - (11+2)/(29+35)*ent(11,2)
>>> IG2
0.08417016765262439
```

Exercise 3

- *Which attribute is the best classifier?*



We have

Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

Therefore Humidity is the most informative attribute with respect to class prediction.