DECISION TREES - EXAMPLE

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

Consider the following table (adapted from [LS98])

EXAMPLE	CREDIT HISTORY	DEBT	COLLATERAL	INCOME	RISK
1	bad	low	none	0–15	high
2	unknown	high	none	15–35	high
3	unknown	low	none	15–35	moderate
4	bad	low	none	0–15	high
5	unknown	low	adequate	>35	low
6	unknown	low	none	>35	low
7	unknown	high	none	0–15	high
8	bad	low	adequate	>35	moderate
9	good	low	none	>35	low
10	good	high	adequate	>35	low
11	good	high	none	0–15	high
12	good	high	none	15–35	moderate
13	good	high	none	>35	low
14	bad	high	none	15–35	high

Constructing Decision Trees
Solution (By C. Ruiz)

I. Selecting the first attribute:

```
          RISK: Low               Moderate          High
          x/14*[-y/x log2(y/x)     -z/x log2(z/x)    -w/x log2(w/x) ]
-------------------------------------------------------------------------
CREDIT  HISTORY
Entropy(Credit,S)
= Entropy([0,1,3],[2,1,2],[3,1,1])
= (4/14)*Entropy([0,1,3]) + (5/14)*Entropy([2,1,2]) + (5/14)*Entropy([3,1,1])

Bad       4/14*[-0                -1/4 log2(1/4)    -3/4 log2(3/4) ]
+Unknown  5/14*[-2/5 log2(2/5)     -1/5 log2(1/5)    -2/5 log2(2/5) ]
+Good     5/14*[-3/5 log2(3/5)     -1/5 log2(1/5)    -1/5 log2(1/5) ]
= 1.265
-------------------------------------------------------------------------
DEBT
Entropy(Debt,S) = Entropy([3,3,2],[2,1,4])
= (7/14)*Entropy([3,3,2])+ (7/14)*Entropy([2,1,4])

Low       7/14*[-3/7 log2(3/7)     -2/7 log2(2/7)    -2/7 log2(2/7) ]
+ High    7/14*[-2/7 log2(2/7)     -1/7 log2(1/7)    -4/7 log2(4/7) ]
=1.467
-------------------------------------------------------------------------
COLLATERAL
Entropy(Collateral,S) = Entropy([3,2,6],[2,1,0])
= (11/14)* Entropy([3,2,6],[2,1,0]) + (3/14)* Entropy([3,2,6],[2,1,0])

None     11/14*[-3/11 log2(3/11)   -2/11 log2(2/11)  -6/11 log2(6/11)]
+Adequate 3/14*[-2/3 log2(2/3)     -1/3 log2(1/3)    -0             ]
=1.324
-------------------------------------------------------------------------
INCOME
Entropy(Income,S) = Entropy([0,0,4],[0,2,2],[5,1,0])
= (4/14)*Entropy([0,0,4]) +(4/14)*Entropy([0,2,2]) +(6/14)*Entropy([5,1,0])

[0-15]    4/14*[-0                -0                -4/4 log2(4/4) ]
+[15-35]  4/14*[-0                -2/4 log2(2/4)    -2/4 log2(2/4) ]
+[>35]    6/14*[-5/6 log2(5/6)     -1/6 log2(1/6)    -0             ]
=0.564
-------------------------------------------------------------------------
Hence, INCOME should be selected as the root of the decision tree.
The partial tree would look like:
```

```
                    INCOME
     0-15 /         15-35 |         \ >35
         HIGH
         1,4,7,11
```

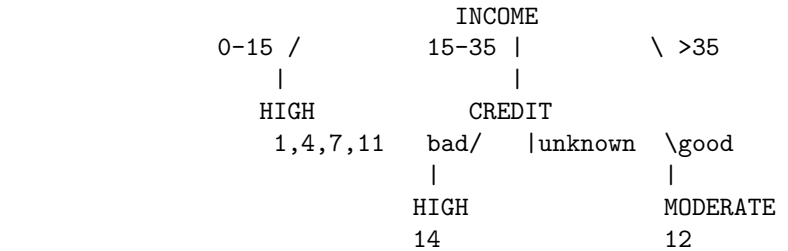
II. Now we need to select an attribute (among Credit History, Debt, and Collateral) to subdivide the case when income ranges from $15 to $35.
For this, we restrict our analysis to examples S' = 2, 3, 12, 14.
Hence n_t = 4.

```
          RISK: Low              Moderate          High
          x/4*[-y/x log2(y/x)     -z/x log2(z/x)    -w/x log2(w/x) ]
-------------------------------------------------------------------------
CREDIT HISTORY
Entropy(Credit,S') = Entropy([0,0,1],[0,1,1],[0,1,0])
= (1/4)*Entropy([0,0,1]) + (2/4)*Entropy([0,1,1]) + (1/4)*Entropy([0,1,0])
Bad       1/4*[-0                -0                -1/1 log2(1/1) ]
+Unknown  2/4*[-0                -1/2 log2(1/2)    -1/2 log2(1/2) ]
+Good     1/4*[-0                -1/1 log2(1/1)    -0             ]
=0.5
-------------------------------------------------------------------------
DEBT
Entropy(Debt,S') = Entropy([0,1,0],[0,1,2])
= (1/4)*Entropy([0,1,0]) + (3/4)*Entropy([0,1,2])

Low       1/4*[-0                -1/1 log2(1/1)    -0             ]
+ High    3/4*[-0                -1/3 log2(1/3)    -2/3 log2(2/3) ]
= 0.688
-------------------------------------------------------------------------
COLLATERAL
Entropy(Collateral,S') = Entropy([0,2,2],[0,0,0])
= (4/4)*Entropy([0,2,2]) + (0/4)*Entropy([0,0,0])

None      4/4*[-0                -2/4 log2(2/4)    -2/4 log2(2/4) ]
+Adequate 0/4*[-0                -0                -0             ]
=1
-------------------------------------------------------------------------
```

So Credit History is selected to subdivide Income $15-$35



Now we need to select an attribute (among Debt, and Collateral)
to subdivide the case when income ranges from $15 to $35 and the
Credit History is unknown.
For this, we restrict our analysis to examples S" = 2, 3.
Hence n_t = 2.

```
          RISK: Low               Moderate          High
          x/2*[-y/x log2(y/x)     -z/x log2(z/x)    -w/x log2(w/x) ]
-------------------------------------------------------------------------
DEBT
Entropy(Debt,S") = Entropy([0,1,0],[0,0,1])
= (1/2)*Entropy([0,1,0]) + (1/2)*Entropy([0,0,1])

Low       1/2*[-0                 -1/1 log2(1/1)    -0             ]
+ High    1/2*[-0                 -0                -1/1 log2(1/1) ]
= 0
-------------------------------------------------------------------------
COLLATERAL
Entropy(Collateral,S") = Entropy([0,1,1],[0,0,0])
= (2/2)*Entropy([0,1,1]) + (0/2)*Entropy([0,0,0])

None      2/2*[-0                 -1/2 log2(1/2)    -1/2 log2(1/2) ]
+Adequate 0/2*[-0                 -0                -0             ]
=1
-------------------------------------------------------------------------
```

Hence, Debt is selected as the attribute to subdivide the case when
income ranges from $15 to $35 and the Credit History is unknown.

III. Now we need to select an attribute (among Credit History,
Debt, and Collateral) to subdivide the case when income >$35.
For this, we restrict our analysis to examples S"' = 5, 6, 8, 9, 10, 13.
Hence n_t = 6.

```
          RISK: Low               Moderate          High
          x/6*[-y/x log2(y/x)     -z/x log2(z/x)    -w/x log2(w/x) ]
-------------------------------------------------------------------------
CREDIT HISTORY
Entropy(Credit,S"')= Entropy([0,1,0],[2,0,0],[3,0,0])
= (1/6)*Entropy([0,1,0]) + (2/6)*Entropy([2,0,0]) + (3/6)*Entropy([3,0,0])

Bad       1/6*[-0                 -1/1 log2(1/1)    -0             ]
+Unknown  2/6*[-2/2 log2(2/2)     -0                -0             ]
+Good     3/6*[-3/3 log2(3/3)     -0                -0             ]
=0
-------------------------------------------------------------------------
```

Since the average disorder for Credit History is 0, we don't even
need to consider the other two attributes.
Then the final tree is the following: