Developing a Decision Tree

Problem: Suppose you wish to develop a decision tree that can be used to answer the question "Is today a good day to fish"?

Observations (training data):

Data	**Wind**	**Water**	**Air**	**Forecast**	**Oracle**
1	Strong	Warm	Warm	Sunny	Yes
2	Weak	Warm	Warm	Sunny	No
3	Strong	Warm	Warm	Cloudy	Yes
4	Strong	Moderate	Warm	Rainy	Yes
5	Strong	Cold	Cool	Rainy	No
6	Weak	Cold	Cool	Rainy	No
7	Weak	Cold	Cool	Sunny	No
8	Strong	Moderate	Warm	Sunny	Yes
9	Strong	Cold	Cool	Sunny	Yes
10	Strong	Moderate	Cool	Rainy	No
11	Weak	Moderate	Cool	Sunny	Yes
12	Weak	Moderate	Warm	Sunny	Yes
13	Strong	Warm	Cool	Sunny	Yes
14	Weak	Moderate	Warm	Rainy	No

The ID3 algorithm

1. Determine the root node of the decision tree by choosing the attribute of the training data that maximizes the information gain.

    - Use the formula for Entropy:

    $$Entropy(S) \equiv -\sum_{i=1}^{k} p_i \log_2 p_i$$

    where S is the collection of examples, k is the number of categories, and p_i is the ratio of the cardinality of category i to the cardinality of S, as in $\left(p_i = \frac{N_i}{N} \right)$.

    - Then use the formula for Information Gain:

    $$Gain(S,a) = Entropy(S) - \sum_{v=values(a)} \frac{|S_v|}{|S|} Entropy(S_v)$$

    where *values(a)* is the set of all possible values for attribute a, and S_v is the subset of set S for which attribute a has value v.

1. Begin with the entire set of training examples, $S = [D_1 .. D_{14}]$

Set	$\oplus$	Θ	Entropy
S	8	6	.985

a) v (Wind) = {Weak, Strong}

Set	$\oplus$	Θ	Entropy
S_{Weak}	2	4	.918
S_{Strong}	6	2	.811

Gain (S, Wind) = .985 – (6/14)(.918) – (8/14)(.811) = .128

b) v (Water) = {Cold, Moderate, Warm}

Set	$\oplus$	Θ	Entropy
S_{Cold}	1	3	.811
$S_{Moderate}$	4	2	.918
S_{Warm}	3	1	.811

Gain (S, Water) = .985 – (4/14)(.811) – (6/14)(.918) – (4/14)(.811) = .128

c) v (Air) = {Cool, Warm}

Set	$\oplus$	Θ	Entropy
S_{Cool}	3	4	.985
S_{Warm}	5	2	.863

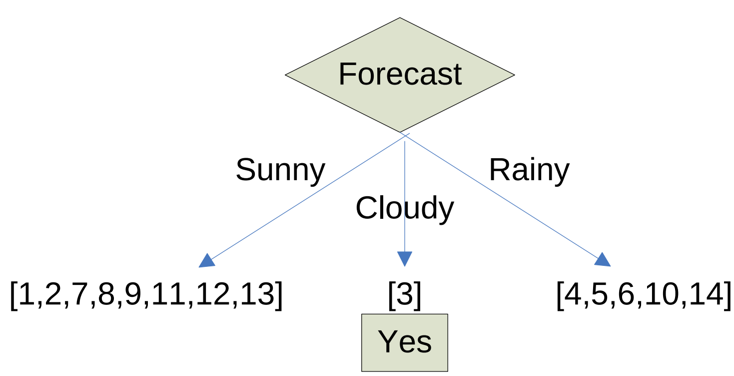
Gain (S, Air) = .985 – (7/14)(.985) – (7/14)(.863) = .061

d) v (Forecast) = {Rainy, Cloudy, Sunny}

Set	$\oplus$	Θ	Entropy
S_{Rainy}	1	4	.722
S_{Cloudy}	1	0	0
S_{Sunny}	6	2	.811

Gain (S, Forecast) = .985 – (5/14)(.722) – (1/14)(0) – (8/14)(.811) = $\boxed{.264}$

The "Forecast" attribute maximizes Information Gain and is chosen as the root node, leading to the following initial Decision Tree. Examples are then "sorted" down the tree accordingly.



This process continues recursively down each subtree, until each attribute appears once on a path, or until a leaf node is created.

2. At the 2nd level of the tree, continue with:

a) the "Sunny" training examples, S = [1,2,7,8,9,11,12,13]

$$\text{Gain}(S_{Sunny}, \text{Wind}) = .811 - (4/8)(0) - (4/8)(1) = \boxed{.311}$$
$$\text{Gain}(S_{Sunny}, \text{Water}) = .811 - (2/8)(1) - (3/8)(0) - (3/8)(.918) = .217$$
$$\text{Gain}(S_{Sunny}, \text{Air}) = .811 - (4/8)(.811) - (4/8)(.811) = 0$$

b) and the "Rainy" training examples, S = [4,5,6,10,14]

$$\text{Gain}(S_{Rainy}, \text{Wind}) = .722 - (3/5)(.918) - (2/5)(0) = .171$$
$$\text{Gain}(S_{Rainy}, \text{Water}) = .722 - (2/5)(0) - (3/5)(.918) = .171$$
$$\text{Gain}(S_{Rainy}, \text{Air}) = .722 - (3/5)(0) - (2/5)(1) = \boxed{.322}$$

3. Finally, at the 3rd level of the tree, continue with:

a) the "Sunny day, Weak wind" training examples, S = [2,7,11,12]

$$\text{Gain}(S_{Sunny,Weak}, \text{Water}) = 1 - (1/4)(0) - (2/4)(0) - (1/4)(0) = \boxed{1}$$
$$\text{Gain}(S_{Sunny,Weak}, \text{Air}) = 1 - (2/4)(1) - (2/4)(1) = 0$$

b) and the "Rainy day, Warm air" training examples, S = [4,14]

$$\text{Gain}(S_{Rainy,Warm}, \text{Wind}) = 1 - (1/2)(0) - (1/2)(0) = \boxed{1}$$
$$\text{Gain}(S_{Rainy,Warm}, \text{Water}) = 1 - (0/2)(1) - (2/2)(1) - (0/2)(1) = 0$$

Leaving us with the final Decision Tree for this collection of training data: