Lab 9 : Decision tree from scratch

1. Consider the following instances, S:

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Let's create a decision tree from scratch without a programming tool.

2. First we need to work out which attribute will be the root node for the decision tree.

Given an arbitrary categorisation, C into categories c_1, ..., c_n, and a set of examples, S, for which the proportion of examples in c_i is p_i, then the entropy of S is:

$$Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t)$$

P_x = probability for x to occur.

Entropy(Decision) = $-P_{Cinema}*\log_2(P_{Cinema}) - P_{Tennis}*\log_2(P_{Tennis}) - P_{Stay\text{-}in}*\log_2(P_{Stay\text{-}in}) - P_{Shopping}*\log_2(P_{Shopping})$
= $-[\ (0.6)*\log_2(0.6) + (0.2)*\log_2(0.2) + (0.1)*\log_2(0.1) + (0.1)*\log_2(0.1)\ ]$
= **1.57095**

3. Next we need to calculate the best of the following formula:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Note:
- Entropy(Decision, Weather), k ∈ { Sunny, Windy, Rainy }
*- Entropy(Decision, Weather) = P_{Sunny}*Entropy(Sunny) + P_{Windy}*Entropy(Windy) + P_{Rainy}*Entropy(Rainy)*
*= (3/10)*Entropy(Sunny) + (4/10)*Entropy(Windy) + (3/10)*Entropy(Rainy)*

Weather	**Decision**
Sunny	Cinema
Sunny	Tennis
Sunny	Tennis

*Entropy(Sunny) = -[P_{Cinema}*$log_2(P_{Cinema})$ + P_{Tennis}*$log_2(P_{Tennis})$ + $P_{stay\text{-}in}$*$log_2(P_{stay\text{-}in})$ + $P_{Shopping}$*$log_2(P_{Shopping})$]*
= -[(1/3)log_2(1/3) + (2/3)*log_2(2/3) + 0 + 0]*
= 0.918

E(Cinema, Tennis, Stay-in, Shopping) will be used as abbreviation in this section.

Gain(Decision, Weather)
= Entropy(Decision) – Entropy(Weather)
= 1.57095 - [P_{Sunny}*Entropy(Sunny) + P_{Windy}*Entropy(Windy) + P_{Rainy}*Entropy(Rainy)]
= 1.57095 - [(3/10)*E(1, 2, 0, 0) + (4/10)*E(3, 0, 1, 0) + (3/10)*E(2, 0, 0, 1)]
= 1.57095 - [(0.3)*-((1/3)*log_2(1/3) + (2/3)*log_2(2/3)) + (0.4)*-((3/4)*log_2(3/4) + (1/4)*log_2(1/4)) + (0.3)*-((2/3)*log_2(2/3) + (1/3)*log_2(1/3))]
= **0.6955**

Gain(Decision, Parents)
= 1.57095 - [(5/10)*E(5, 0, 0, 0) + (5/10)*E(1, 2, 1, 1)]
= 1.57095 - [0 + (0.5)*-((1/5)*log_2(1/5) + (2/5)*log_2(2/5) + (1/5)*log_2(1/5) + (1/5)*log_2(1/5))]
= **0.61**

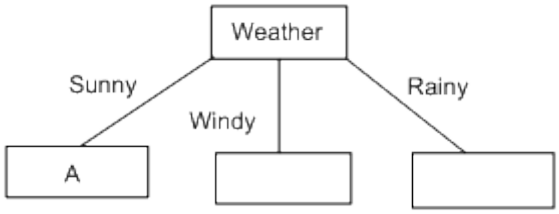
Gain(Decision, Money)
= 1.57095 - [(7/10)*E(3, 2, 1, 1) + (3/10)*E(3, 0, 0, 0)]
= 1.57095 - [(0.7)*-((3/7)*log_2(3/7) + (2/7)*log_2(2/7) + (1/7)*log_2(1/7) + (1/7)*log_2(1/7)) + 0]
= **0.2813**

The feature with the highest gain will be the root node, Weather (0.6955).

4. Draw your resulting decision tree.

5. Once you have decided on the tree, look at the activities based on weeks. For instance, assume you have chosen Weather, henceforth we investigate Sunny. For Sunny weather, the instances are W1,W2, W10 for which the activities are Cinema, Cinema, Tennis. As Cinema, Cinema and Tennis is present as a set and consists of non-Empty set, we will establish an attribute node, A. Do the same for the rest of your tree.

You may end up with something like the following



6. Now we have to fill in the choice of attribute A, which we know cannot be weather, because we've already removed that from the list of attributes to use. So, we need to calculate the values for **Gain(S_{Sunny}, parents) and Gain(S_{Sunny}, money)**. Firstly, Entropy(S_{Sunny}) = 0.918. Next, we set S to be S_{Sunny} = {W1,W2,W10} (and, for this part of the branch, we will ignore all the other examples). In effect, we are interested only in this part of the table:

Weekend (Example)	Weather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Calculate the following

Gain(S_{Sunny}, parents) and Gain(S_{Sunny}, money). Which has the higher gain, and what impact does that have on the selection of attribute A?

Entropy(Sunny) = 0.918

Gain(Sunny, Parents)
= Entropy(Sunny) – Entropy(Sunny, Parents)
= 0.918 - [P_{Yes}*Entropy(Yes) + P_{No}*Entropy(No)]
= 0.918 - [(1/3)*E(1, 0, 0, 0) + (2/3)*E(0, 2, 0, 0)]
= 0.918 - [(1/3)*0 + (2/3)*0]
= **0.918**

Gain(Sunny, Money)
= Entropy(Sunny) – Entropy(Sunny, Money)
= 0.918 - [P_{Rich}*Entropy(Rich) + P_{Poor}*Entropy(Poor)]
= 0.918 - [(3/3)*E(1, 2, 0, 0) + 0]
= 0.918 - [(1)*-((1/3)*$\log_2$(1/3) + (2/3)*$\log_2$(2/3) + 0 + 0) + 0]
= **0 … No difference**

Therefore, Parents as the branching node.

Tree, Class = Decision. *Rounded off to 2 decimal places.*

1. Root Node.

Entropy(Decision)
$= -P_{Cinema} * \log_2(P_{Cinema}) - P_{Tennis} * \log_2(P_{Tennis}) - P_{Stay\text{-}in} * \log_2(P_{Stay\text{-}in}) - P_{Shopping} * \log_2(P_{Shopping})$
$= -[\ (0.6) * \log_2(0.6) + (0.2) * \log_2(0.2) + (0.1) * \log_2(0.1) + (0.1) * \log_2(0.1)\ ]$
= 1.57

Entropy($Weather_{Sunny}$)
$= -((1/3) * \log_2(1/3) + (2/3) * \log_2(2/3) + 0 + 0)$
$= 0.92$

Entropy($Weather_{Windy}$)
$= -((3/4) * \log_2(3/4) + (1/4) * \log_2(1/4) + 0 + 0)$
$= 0.81$

Entropy($Weather_{Rainy}$)
$= -((2/3) * \log_2(2/3) + 0 + (1/3) * \log_2(1/3) + 0)$
$= 0.92$

Entropy(Decision, Weather)
$= (3/10)(0.92) + (4/10)(0.81) + (3/10)(0.92)$
$= 0.88$

Gain(Decision, Weather)
$= 1.57 - 0.88$
= 0.69

Entropy($Parents_{Yes}$)
$= -((5/5) * \log_2(5/5) + 0 + 0 + 0)$
$= 0$

Entropy($Parents_{No}$)
$= -((1/5) * \log_2(1/5) + (2/5) * \log_2(2/5) + (1/5) * \log_2(1/5) + (1/5) * \log_2(1/5))$
$= 1.92$

Entropy(Decision, Parents)
$= (5/10)(0) + (5/10)(1.92)$
$= 0.96$

Gain(Decision, Parents)
$= 1.57 - 0.96$
= 0.61

Entropy($Money_{Rich}$)
$= -((3/7) * \log_2(3/7) + (2/7) * \log_2(2/7) + (1/7) * \log_2(1/7) + (1/7) * \log_2(1/7))$
$= 1.84$

Entropy($Money_{Poor}$)
$= -((3/3) * \log_2(3/3) + 0 + 0 + 0)$

= 0

Entropy(Decision, Money)
= (7/10)(1.84) + (3/10)(0)
= 1.29

Gain(Decision, Money)
= 1.57 – 1.29
= **0.28**

Therefore, first split will be Weather.

2. Sunny Node.

Entropy(Sunny) = 0.92

Entropy(Sunny, $Parents_{Yes}$)
= $-((1/1)*\log_2(1/1) + 0 + 0 + 0)$
= 0

Entropy(Sunny, $Parents_{No}$)
= $-(0 + (2/2)*\log_2(2/2) + 0 + 0)$
= 0

Entropy(Sunny, Parents)
= (1/3)(0) + (2/3)(0)
= 0

Gain(Sunny, Parents)
= 0.92 – 0
= **0.92**

Entropy(Sunny, $Money_{Rich}$)
= $-((1/3)*\log_2(1/3) + (2/3)*\log_2(2/3) + 0 + 0)$
= 0.92

Entropy(Sunny, $Money_{Poor}$)
= 0

Entropy(Sunny, Money)
= (3/3)(0.92) + 0
= 0.92

Gain(Sunny, Parents)
= 0.92 – 0
= **0**

Therefore, Sunny node split will be Parents.

3. Windy Node.

Entropy(Windy) = 0.81

Entropy(Windy, $Parents_{Yes}$)
= $-((2/2)*\log_2(2/2) + 0 + 0 + 0)$
= 0

Entropy(Windy, $Parents_{No}$)
= $-((1/2)*\log_2(1/2) + 0 + 0 + (1/2)*\log_2(1/2))$
= 1

Entropy(Windy, Parents)
= (2/4)(0) + (2/4)(1)
= 0.5

Gain(Windy, Parents)
= 0.81 – 0.5
= **0.31**

Entropy(Windy, $Money_{Rich}$)
= $-((2/3)*\log_2(2/3) + 0 + 0 + (1/3)*\log_2(1/3))$
= 0.92

Entropy(Windy, $Money_{Poor}$)
= $-((1/1)*\log_2(1/1) + 0 + 0 + 0)$
= 0

Entropy(Windy, Money)
= (3/4)(0.92) + (1/4)(0)
= 0.69

Gain(Windy, Money)
= 0.81 – 0.69
= **0.12**

Therefore, Windy node split will be Parents.

4. Rainy Node.

Entropy(Rainy) = 0.92

Entropy(Rainy, $Parents_{Yes}$)
= $-((2/2)*\log_2(2/2) + 0 + 0 + 0)$
= 0

Entropy(Rainy, $Parents_{No}$)
= $-(0 + 0 + (1/1)*\log_2(1/1) + 0)$
= 0

Entropy(Rainy, Parents)

= 0

Gain(Rainy, Parents)
= 0.92 – 0
= **0.92**

Therefore, Windy node split will be Parents.

Tree Visualization. *Without Pruning*.