Classification Tutorial- Decision tree solution

Decision tree for the following dataset

	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Dana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none

Not related to classification result

Nome	Hair	Height	Weight	Lotion	Result
Sarah	blonde	average	light	no	sunburned
Dana	blonde	tall	average	yes	none
Alex	brown	short	average	yes	none
Annie	blonde	short	average	no	sunburned
Emily	red	average	heavy	no	sunburned
Pete	brown	tall	heavy	no	none
John	brown	average	heavy	no	none
Katie	blonde	short	light	yes	none

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Attributes and values

- Hair : Blonde, Red, Brown
- Height : Average, Toll, Short
- Weight : Light, Average, Heavy
- Lotion Yes, No

‹

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

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

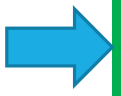
Entropy Calculation for whole dataset

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

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

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).

Result	
sunburned	none
3	5



Entropy(Result) Entropy(3,5)

Entropy(3/8 , 5/8)

Entropy(0.375,0.625)

= - (0.375 * log2 0.375) - (0.625 * log2 0.625)

= - (0.375 * -1.4150) — (0.625 * -0.6780)

=0.9525

STEP 2:

Calculate information gain for each attribute

Information Gain:

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

Parent node, p, is split into k partitions; n_i is number of records in partition i

Hair	Height	Weight	Lotion	Result
blonde	average	light	NO	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	No	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	No	none
brown	average	heavy	no	none
blonde	short	light	yes	none

ENTROPY of partitions based on hair attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Hair	blonde	2	2	4
	brown	0	3	3
	red	1	0	1
		3	5	8

Entropy(Result, Hair)

= P(blonde)" Entropy(2,2) + P(brown)* Entropy(0,3)+ P(red)* Entropy(1,0)

(4/8) * Entropy(2,2) + (3/8)* Entropy(0,3) + (1/8) * Entropy(1,0)

= (0.50 *1) + (0.38 *0) + (0.13 * 0)

=0.5

INFORMATION GAIN for splitting based on hair attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Information Gain:

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

Parent node, p, is split into k partitions; n_i is number of records in partition i

$$= 0.9525 - 0.50$$

$$= 0.4525$$

ENTROPY of partitions based on height attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Height	short	1	2	3
	average	2	1	3
	tall	0	2	2
		3	5	8

Entropy(Result,Height)

= P(short)* Entropy(1,2) + P(average)* Entropy(2,1)+ P(tall)* Entropy(0,2)

(3/8) * Entropy(1,2) + (3/8)* Entropy(2,1) + (2/8) * Entropy(0,2)

= (0.38 * 0.9182) + (0.38 * 0.9182) + (0.25 * 0)

=0.6930

INORMATION GAIN for splitting based on height attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Information Gain:

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

Parent node, p, is split into k partitions; n_i is number of records in partition i

= 0.9525 – 0.6930

= 0.2595

ENTROPY of partitions based on weight attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Weight	Light	1	1	2
	average	1	2	3
	heavy	1	2	3
		3	5	8

Entropy(Result,Weight)

= P(light)" Entropy(1,l) + P(average)* Entropy(1,2)+ P(heavy)* Entropy(1,2)

= (2/8) * Entropy(1,1) + (3/8)* Entropy(1,2) + (3/8) * Entropy(1,2)

= (0.25 * 1) + (0.38 * 0.9182) + (0.38 * 0.9182)

=0.9430

INFORMATION GAIN for splitting based on weight attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Information Gain:

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

Parent node, p, is split into k partitions; n_i is number of records in partition i

$$=0.9525-0.9430$$

$$= 0.0095$$

ENTROPY of partitions based on lotion attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Lotion	yes	0	3	3
	NO	3	2	5
		3	5	8

Entropy(Result,Lotion)

= P(yes)* Entropy(0,3) + P(no)* Entropy(3,2)

(3/8) * Entropy(0,3) + (5/8)* Entropy(3,2)

= (0.38 * 0) + (0.625 * 0.9709)

= 0.6068

INFORMATION GAIN for splitting based on lotion attribute

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
brown	short	average	yes	none
blonde	short	average	no	sunburned
red	average	heavy	no	sunburned
brown	tall	heavy	no	none
brown	average	heavy	no	none
blonde	short	light	yes	none

Information Gain:

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

Parent node, p, is split into k partitions; n_i is number of records in partition i

$$=0.9525-0.6068=0.3157$$

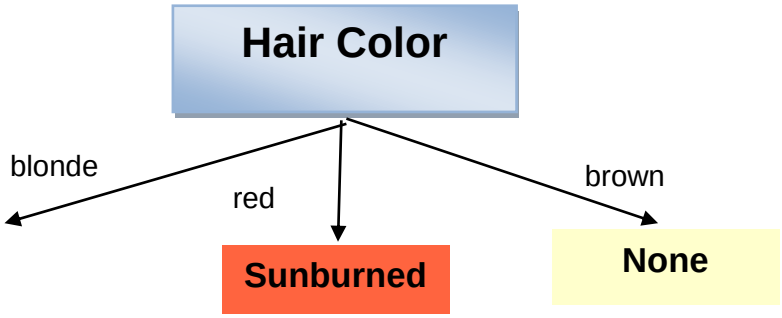
STEP 3:

Choose the attribute with highest information gain as the splitting node (decision node)

Entropy(Result, Hair)	= 0.9525 — 0.50	= 0.4525
Entropy(Result, Height)	= 0.9525 — 0.6930	= 0. 2595
Entropy(Result, Weight)	= 0.9525 — 0.9430	= 0.0095
Entropy(Result, Lotion)	= 0.9525 — 0.6068	= 0.3157

STEP 4:
DIVIDE THE DATASET BY ITS BRANCHES AND REPEAT THE SAME PROCESS ON EVERY BRANCH

All the nodes in branch Red belong to sunburned so it is leaf node with label=sunburned
All the nodes in branch Brown belong to None so it is leaf node with label=None
We should split branch blonde more because it is not pure yet



Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

ENTROPY for whole of this partition.

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

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

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).

Result	
sunburned	none
2	2

Entropy(Result) Entropy(2,2)

Entropy(2/4 , 2/4)

Entropy(0.5,0.5)

= - (0.5 * log2 0.5) - (0.5 * log2 0.5)

= - (0.5 * -1) — (0.5 * -1)

=1

ENTROPY calculation for partitioning based on attribute height

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Height	short	1	1	2
	average	1	o	1
	tall	0	1	1
		2	2	4

Entropy(Result,Height)

= P(short)* Entropy(l,1) + P(average)* Entropy(1,0)+ P(tall)* Entropy(0,1)

= (2/4) * Entropy(1,1) + (1 /4)* Entropy(1,0) + (1/4) * Entropy(0,1)

= ‹o.5*1) + (0.25 * o + 0.25*0)

=0.5

Information Gain= 1 — 0.5 =0.5

ENTROPY calculation for partitioning based on attribute weight

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

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

		Result		
		sunburned	none	
Weigt	light	1	1	2
	average	1	1	2
		2	2	4

Entropy(Result,Weight)

= P(light)" Entropy(1,l) + P(average)* Entropy(1,1)

= (2/4) * Entropy(1,1) + (2/4)* Entropy(1,1)

= (0.5 * 1) + (0.5 * 1)

= 1

Information Gain= 1 — 1 =0

ENTROPY calculation for partitioning based on attribute lotion

Hair	Height	Weight	Lotion	Result
blonde	average	light	no	sunburned
blonde	tall	average	yes	none
blonde	short	average	no	sunburned
blonde	short	light	yes	none

		Result		
		sunburned	none	
Lotion	yes	0	2	2
	no	2	0	2
		2	2	4

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

Entropy(Result,Lotion)

= P(yes)* Entropy(0,2) + P(no)* Entropy(2,0)

= (2/4) * Entropy(0,2) + (2/4)* Entropy(2,0)

= (0.5 * 0) + (0.5 * 0)

=0

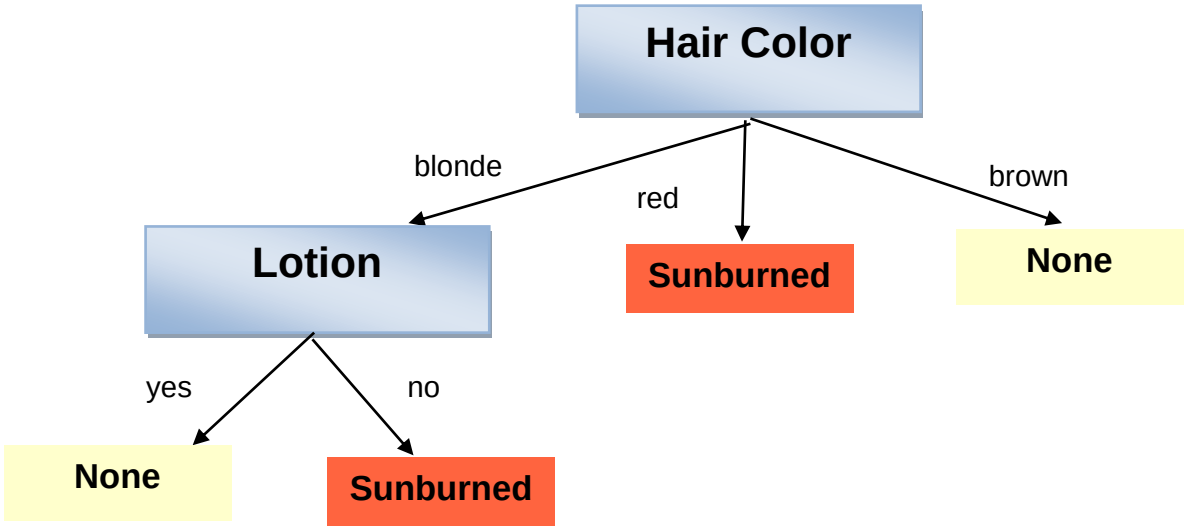
Information Gain= 1 — 0= 1

Choose the attribute with highest information gain as the next decision node

Entropy(Result, Height) = 1 — 0.5 = 0.5

Entropy(Result, Weight) = 1 — 1 = 0

Entropy(Result, Lotion) = 1 — 0 = 1


Hair Color
blonde
red
brown
Lotion
Sunburned
None
yes
no
None
Sunburned