Decision Tree Construction:

Step 1:

	Play	
	Yes	No
	9	5
Pr.	9/14	5/14

$\longrightarrow$ $$H(\text{Play?}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Entropy is a measure of uncertainity

Goal: Find the most simple decision tree

Split the table (tree) according to the "attribute" with the highest "information gain".

$$G(X) \triangleq H(\text{Play?}) - H(X) = H(\underset{\uparrow}{S}) - H(X)$$

Source, original table

where X represent an attribute.

Step 2: Calculate the Gain of each attribute.

2a) Let X = Temperature. (which takes 3 values cool, mild, hot)

		Play		Bernoulli:	
		Yes	No		
T	C	3	1	3/4	1/4
	M	4	2	4/6	2/6
	H	2	2		

$$H(C) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}$$
$$H(M) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}$$
$$H(Hot) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$H(T) = \frac{4}{14}H(C) + \frac{6}{14}H(M) + \frac{4}{14}(H(Hot)) = 0.911$$

Weighted average of H(C), H(M) & H(Hot)

$$G(T) = 0.94 - 0.911 = 0.029$$ (small gain)

Because H(T) is high (uncertainity) | G(Outlook) = 0.246

2b) Let X = Outlook. $H(\text{Outlook}) = \frac{5}{14}H(S) + 0 + \frac{5}{14}H(R) = 0.6\ldots$

		Play		Pr.
		Yes	No	
Outlook	S	2	3	2/5, 3/5
	O	4	0	1, 0
	R	3	2	3/5, 2/5

$\longrightarrow$ $H(S) = -\frac{2}{5}\log\frac{2}{5} + \left(-\frac{3}{5}\log\frac{3}{5}\right)$

$\longrightarrow$ $H(O) = 0$ overcast = 0

$\longrightarrow$ $H(R) = H(S)_{\text{any}} = \left(-\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}\right)$

2 c) Let X= Humidity.

$G(\text{Humidity}) = 0.151$

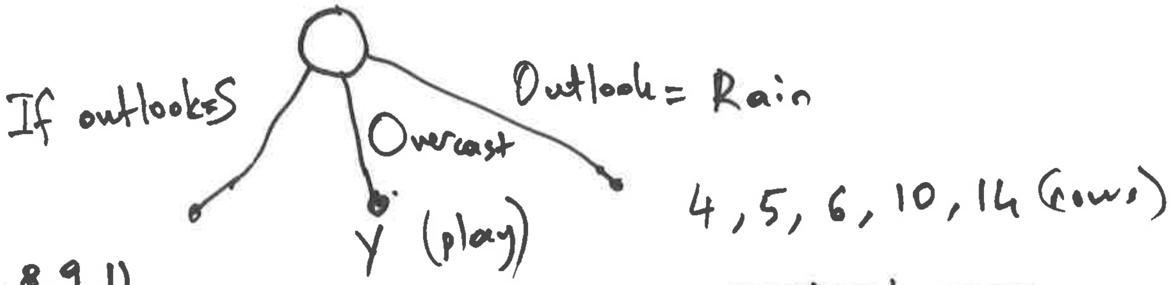
2 d) Let X= Wind , $G(\text{Wind}) = 0.048$

, split the tree according to Outlook.

$$\text{Outlook} = \arg\max_{X} G(X)$$

Outlook is the attribute with the maximum information gain.

3)



rows: 1, 2, 8, 9, 11)

	H	W	T	
1	high	Weak	hot	N
2	H	H	S	N
8				N
9				Y
11				Y

	H	W	T	Play
4	M	H	W	Y
5	C	N	W	Y
6				N
10				Y
14				N

$$H(S\overset{small}{T}1) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$

$$H(ST2) = H(ST1)$$

⋮

Iteratively continue.

Will I play tennis today?

"9+" & "5−" Bernoulli distribution.
$P_+ = \frac{9}{14}$, $P_- = \frac{5}{14}$

Entropy : $H(\text{Table}) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$

	O	T	H	W	Play?
✓ 1	S	H	H	W	-
✓ 2	S	H	H	S	-
3	O	H	H	W	+
4	R	M	H	W	+
5	R	C	N	W	+
6	R	C	N	S	-
7	O	C	N	S	+
✓ 8	S	M	H	W	-
✓ 9	S	C	N	W	+
10	R	M	N	W	+
✓ 11	S	M	N	S	+
12	O	M	H	S	+
13	O	H	N	W	+
14	R	M	H	S	-

Outlook: S(unny),
O(vercast),
R(ainy)

Temperature: H(ot),
M(edium),
C(ool)

Humidity: H(igh),
N(ormal),
L(ow)

Wind: S(trong),
W(eak)

$H(\text{Table}) = 0.94$
$H(\text{Play?}) = 0.94.$

Example|| Given the table

A1	A2	Result	Pr
0	0	0	$\Rightarrow \frac{1}{2}$
0	1	0	
1	0	1	$\Rightarrow \frac{1}{2}$
1	1	1	

$$H(R) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$H(R) = 1.$$

Gain of A1:

A1 "0" → "0"   Pr. 0, 1

A1 "1" → "1"   Pr. 1, 0

Entropy

$$H(A1) = \frac{1}{2}\,0 + \frac{1}{2}\,0 = 0$$

Entropy

A1	Yes	No	Pr.
0	0	2	0, 1
1	2	0	1 0

$$H(\text{"0"}) = -0\log 0 - 1\log 1 = 0$$

$$H(\text{"1"}) = -1\log 1 - 0\log 0 = 0$$

$G(A1) = 1 = 1 - 0.$ (very high) (1*)

Gain of A2:

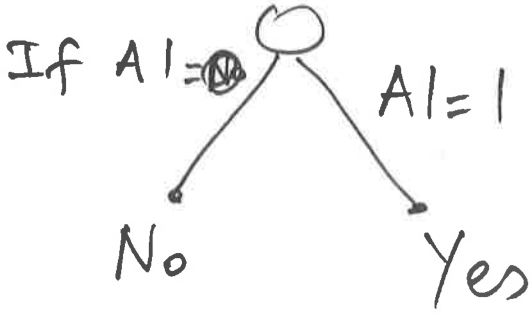
$$H(A_2) = \frac{1}{2}\underbrace{\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)}_{H(\text{"0"})} + \frac{1}{2}\underbrace{\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)}_{H(\text{"1"})}$$

A2	Yes	No	Pr.
0	1	1	1/2, 1/2
1	1	1	1/2, 1/2

$$H(A_2) = 1$$

$G(A2) = 0 = 1 - 1 = 0$ (low) (2*)

From (1*) & (2*) split according to A1.

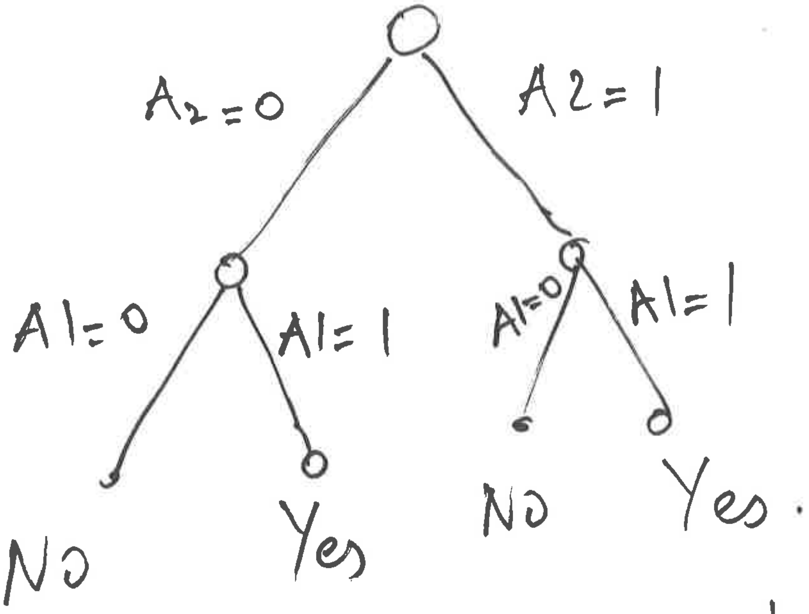


Simple tree.

At the leaves of the tree there is no uncertainity. That is why we try to reduce entropy.

Inefficient tree:

Split wrt A2



Tree 2

Computational complexity of the tree is determined by the # of comparisons (questions) that you have to make (evaluate).

∴ Tree 2 is inefficient.