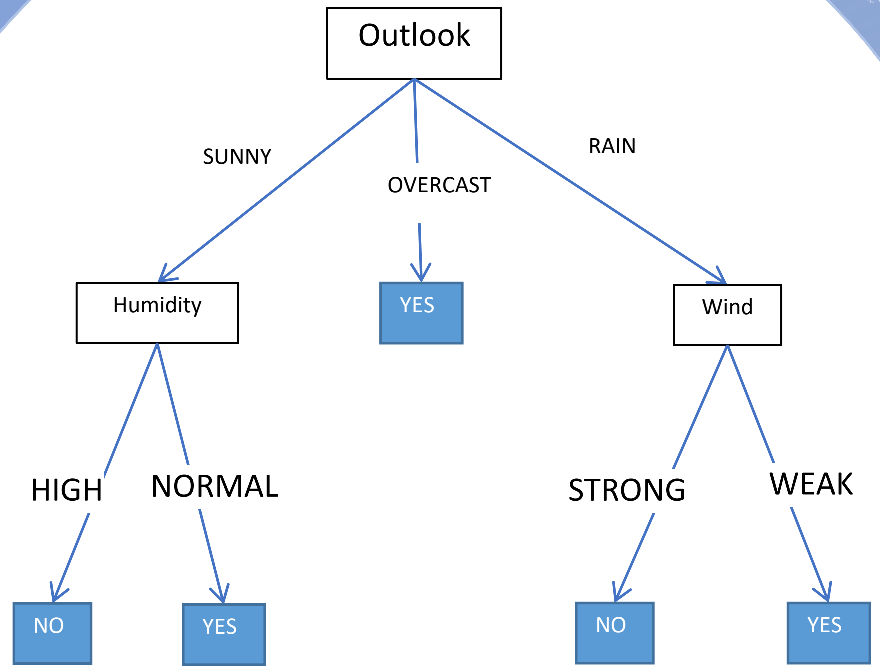
Decision Trees

Part 2: Entropy

The Tennis Problem

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



If(Outlook = sunny AND Humidity = high)	then NO
If(Outlook = sunny AND Humidity = normal)	then YES
If(Outlook = overcast)	then YES
If(Outlook = rain AND Wind = strong)	then NO
If(Outlook = rain AND Wind = weak)	then YES

Decision Tree Learning Algorithm (sometimes called "ID3")

```
function buildTree( subsample, depth )

    //BASE CASE:
    if (depth == 0) OR (all examples have same label) then
        return most common label in the subsample
    end if

    //RECURSIVE CASE:
    for each feature do
        Try splitting the data (i.e. build a decision stump)
        Calculate the cost for this stump
    end for
    Pick feature with minimum cost

    Find left/right subsamples
    Add left branch ← buildTree( leftSubSample, depth − 1 )
    Add right branch ← buildTree( rightSubSample, depth − 1 )

    return tree

end function
```

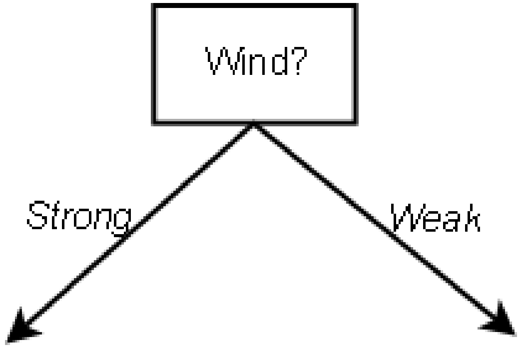
The Tennis Problem

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Note: 9 examples say “YES”, while 5 say “NO”.

Partitioning the data…

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



	Outlook	Temp	Humid	Wind	Play?
2	Sunny	Hot	High	Strong	No
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
14	Rain	Mild	High	Strong	No

3 examples say yes, 3 say no.

	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
13	Overcast	Hot	Normal	Weak	Yes

6 examples say yes, 2 examples say no.

Thinking in Probabilities...

Before the split : 9 'yes', 5 'no', ... $p('yes') = \frac{9}{14} \approx 0.64$

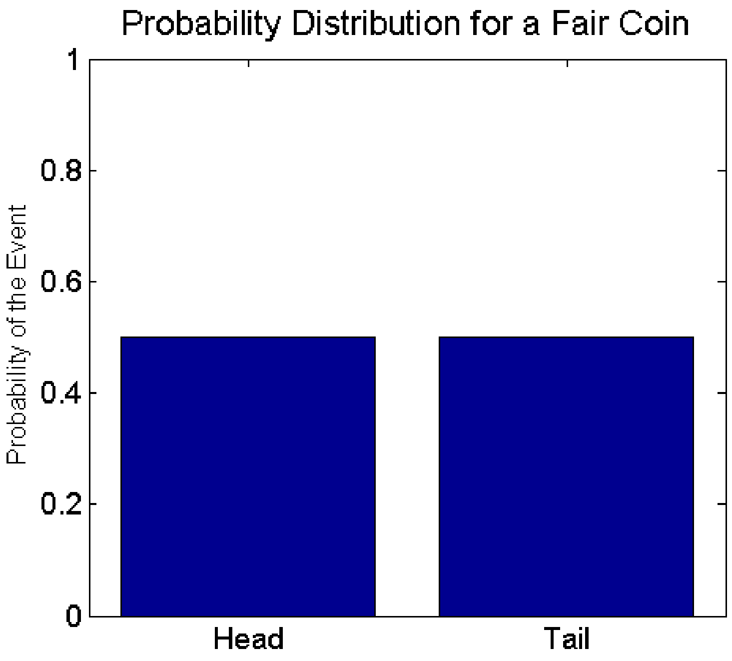
On the left branch : 3 'yes', 3 'no', ... $p('yes') = \frac{3}{6} = 0.5$

On the right branch : 6 'yes', 2 'no', ... $p('yes') = \frac{6}{8} = 0.75$

Remember... $p('no') = 1 - p('yes')$

The “Information” in a feature

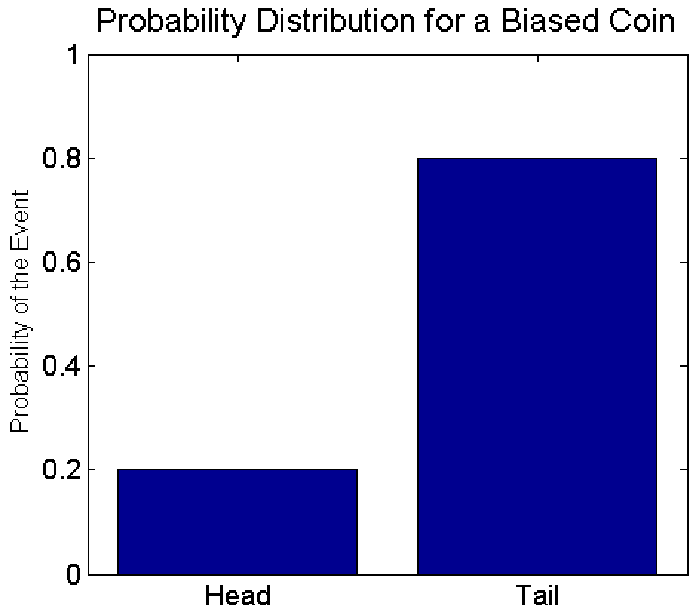
More uncertainty = less information



$$H(X) = 1$$

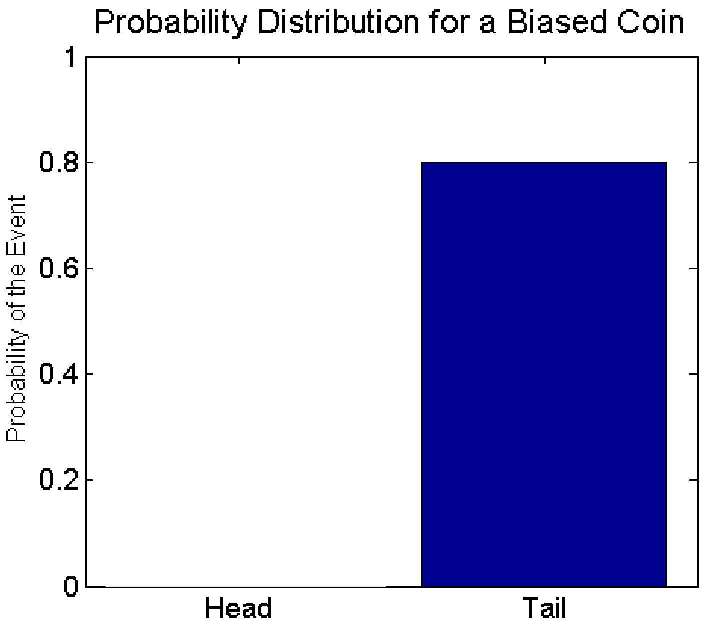
The “Information” in a feature

Less uncertainty = more information



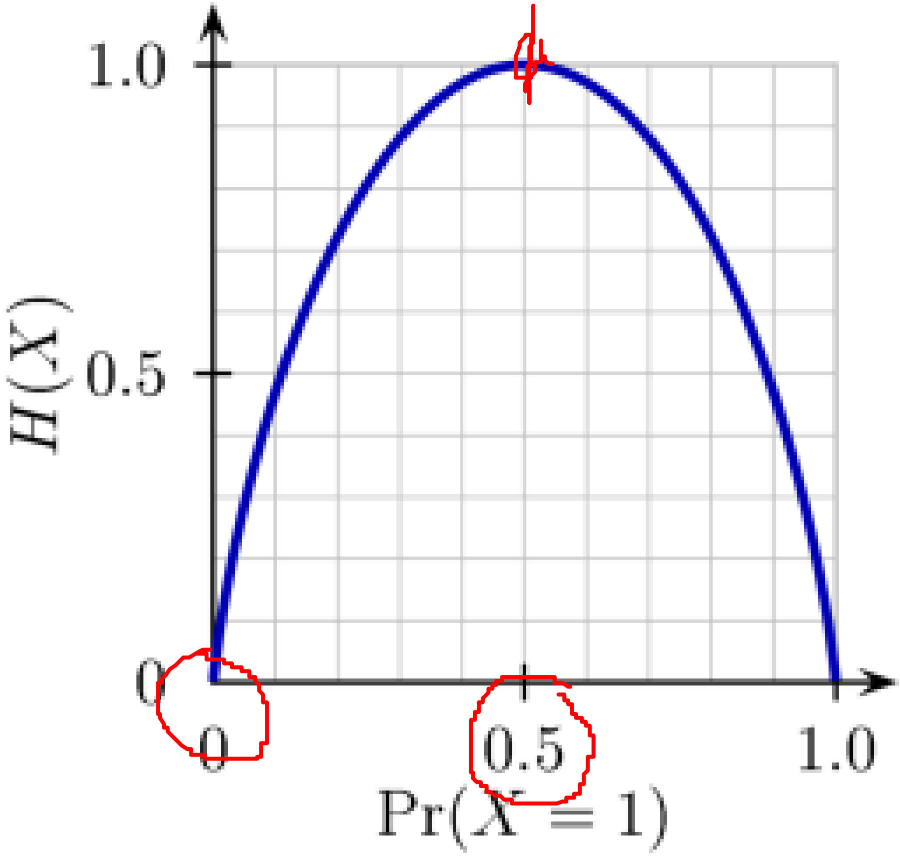
$$H(X) = 0.72193$$

The “Information” in a feature



$$H(X) = 0$$

Entropy



$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

$$\begin{aligned} H(X) &= -\Big(p(head)\log p(head) + p(tail)\log p(tail)\Big) \\ &= -\Big(0.5\log 0.5 + 0.5\log 0.5\Big) \\ &= -\Big((-0.5) + (-0.5)\Big) = 1 \end{aligned}$$

Calculating Entropy

The variable of interest is “T” (for tennis), taking on 'yes' or 'no' values. Before the split : 9 'yes', 5 'no', ...
$p('yes') = \frac{9}{14} \approx 0.64$

In the whole dataset, the entropy is:

$$\begin{aligned} H(T) &= -\sum_i p(x_i) \log p(x_i) \\ &= -\left\{ \frac{5}{14} log \frac{5}{14} + \frac{9}{14} log \frac{9}{14} \right\} = 0.94029 \end{aligned}$$

$H(T)$ is the entropy **before** we split.

See worked example in the supporting material.

Information Gain, also known as "Mutual Information"

$H(T)$ is the entropy before we split.

$H(T|W = strong)$ is the entropy of the data on the left branch.
$H(T|W = weak)$ is the entropy of the data on the right branch.

$H(T|W)$ is the weighted average of the two.

Choose the feature with maximum value of $H(T) - H(T|W)$.

See worked example in the supporting material.

Why don't we just measure the number of errors?!

x_1	x_2	y
1	1	1
1	0	0
1	1	1
1	0	1
0	0	0
0	0	0
0	0	0
0	0	1

Errors = 0.25 … for both!

Mutual information

$I(X1,Y) = 0.1887$

$I(X2;Y) = 0.3113$

Decision Tree Learning Algorithm (sometimes called "ID3")

function BUILDTREE(subsample, depth)

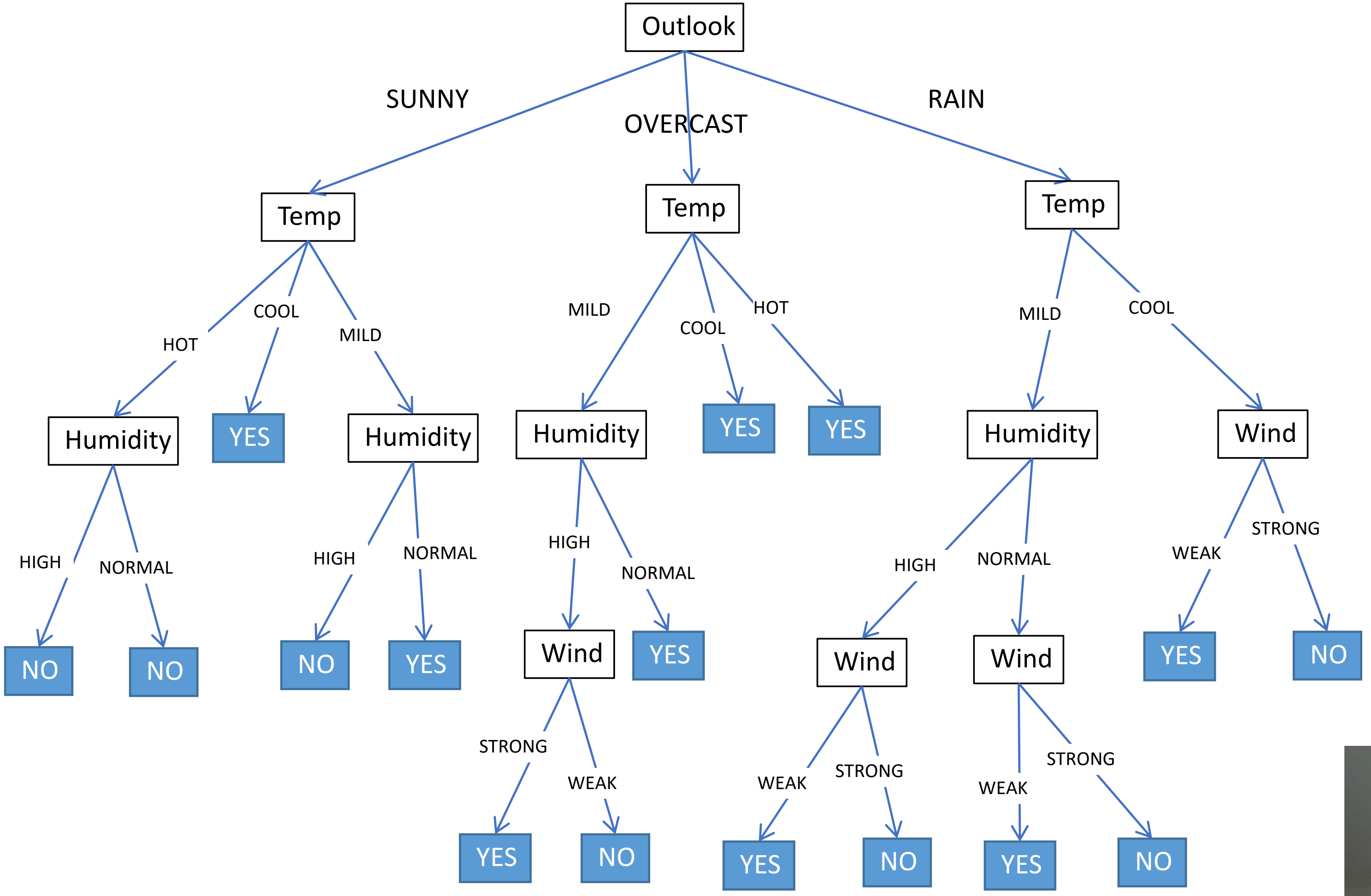
  //**BASE CASE:**
  if ($depth$ == 0) OR (all examples have same label) **then**
    return most common label in the subsample
  end if

  //**RECURSIVE CASE:**
  for each feature **do**
    Try splitting the data (i.e. build a decision stump)
    Calculate **gain** for this stump
  end for
  Pick feature with ~~minimum~~ ~~cost~~
  maximum information gain

  Find left/right subsamples
  Add left branch ← BUILDTREE($leftSubSample$, $depth - 1$)
  Add right branch ← BUILDTREE($rightSubSample$, $depth - 1$)

  return $tree$

end function


Outlook
SUNNY
OVERCAST
RAIN
Temp
Temp
Temp
HOT
COOL
MILD
MILD
COOL
HOT
MILD
COOL
Humidity
YES
Humidity
Humidity
YES
YES
Humidity
Wind
HIGH
NORMAL
HIGH
NORMAL
HIGH
NORMAL
HIGH
NORMAL
WEAK
STRONG
NO
NO
NO
YES
Wind
YES
Wind
Wind
YES
NO
STRONG
WEAK
WEAK
STRONG
WEAK
STRONG
YES
NO
YES
NO
YES
NO