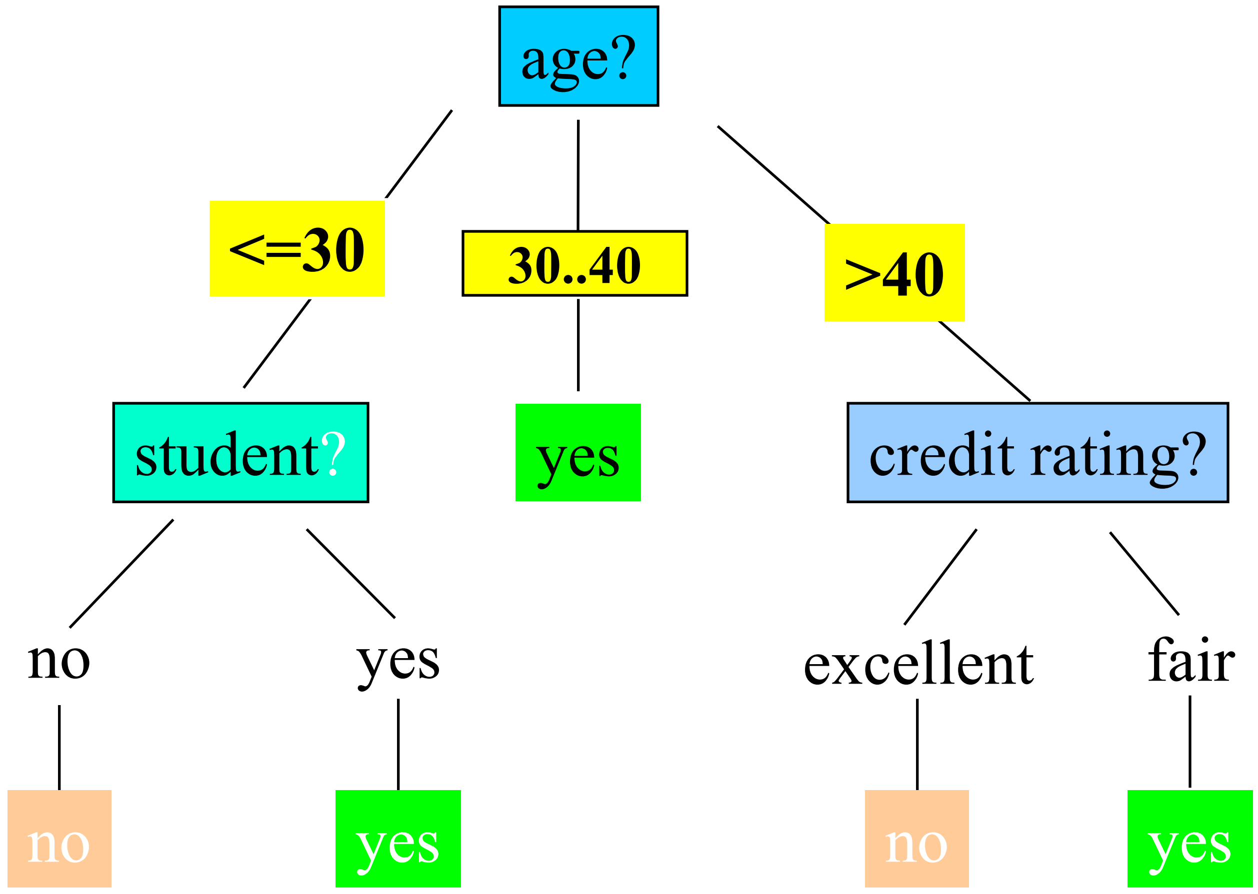
Classification

Decision Tree Algorithm

Building a decision tree: an example training dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Output: A Decision Tree for *"buys_computer"*



DECISION TREES

Decision Tree Representation

The target function can be Boolean or discrete valued

Each node corresponds to an attribute

Each branch corresponds to an attribute value

Each leaf node assigns a class

DECISION TREES

Basic Decision Tree Learning Algorithm

Most algorithms for growing decision trees are variants of a basic algorithm

An example of this core algorithm is the ID3 algorithm

It employs a top-down, greedy search through the space of possible decision trees

DECISION TREES

Basic Decision Tree Learning Algorithm

First of all we *select* the best attribute to be tested at the root of the tree

For making this selection each attribute is evaluated using a statistical test to determine how well it alone classifies the training examples

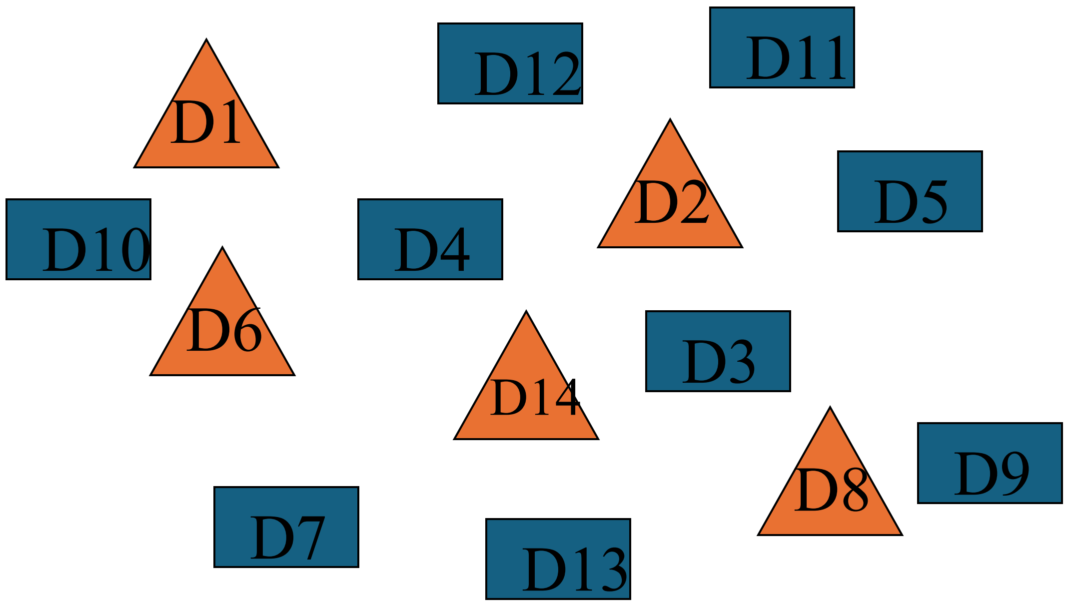
DECISION TREES

Basic Decision Tree Learning Algorithm

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

DECISION TREES

Basic Decision Tree Learning Algorithm



We have

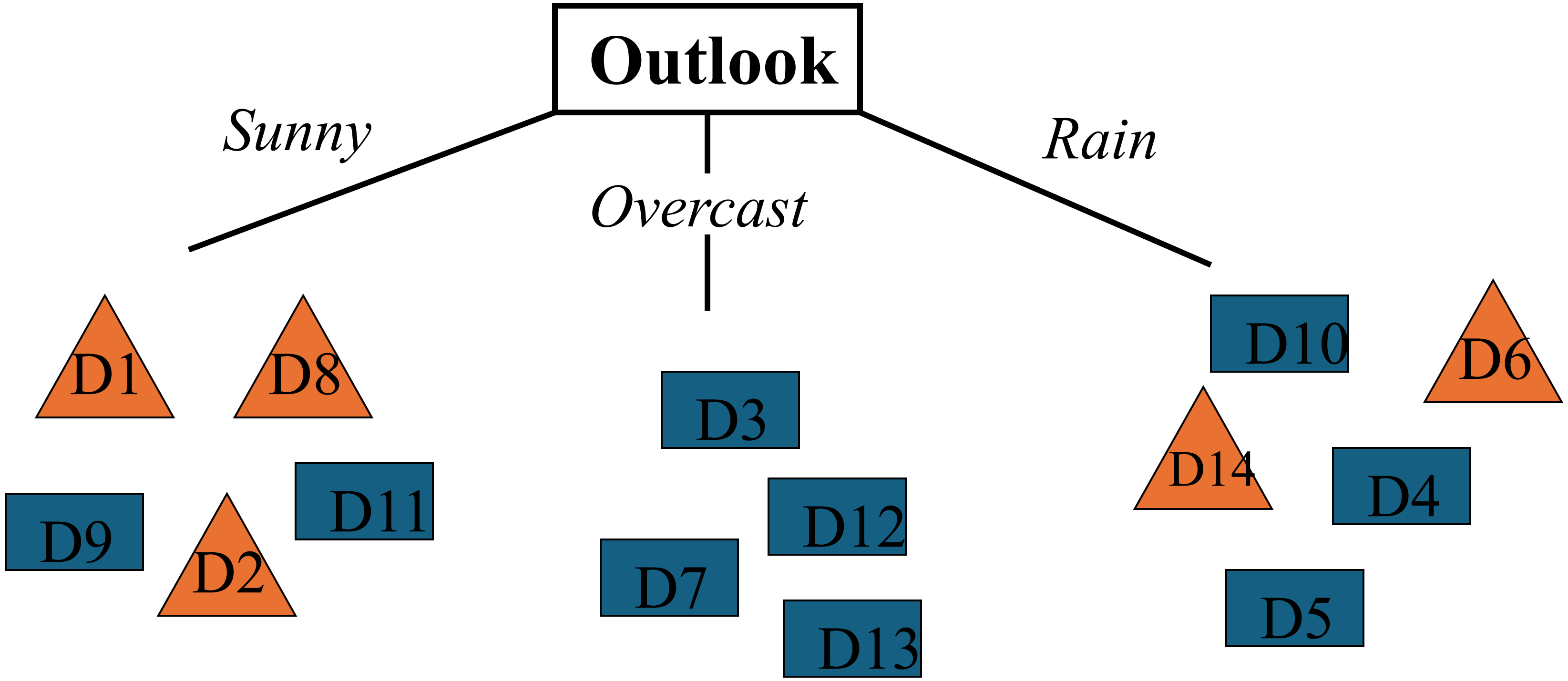
- 14 observations

- 4 attributes
- **Outlook**
- **Temperature**
- **Humidity**
- **Wind**

- 2 classes (Yes, No)

DECISION TREES

Basic Decision Tree Learning Algorithm

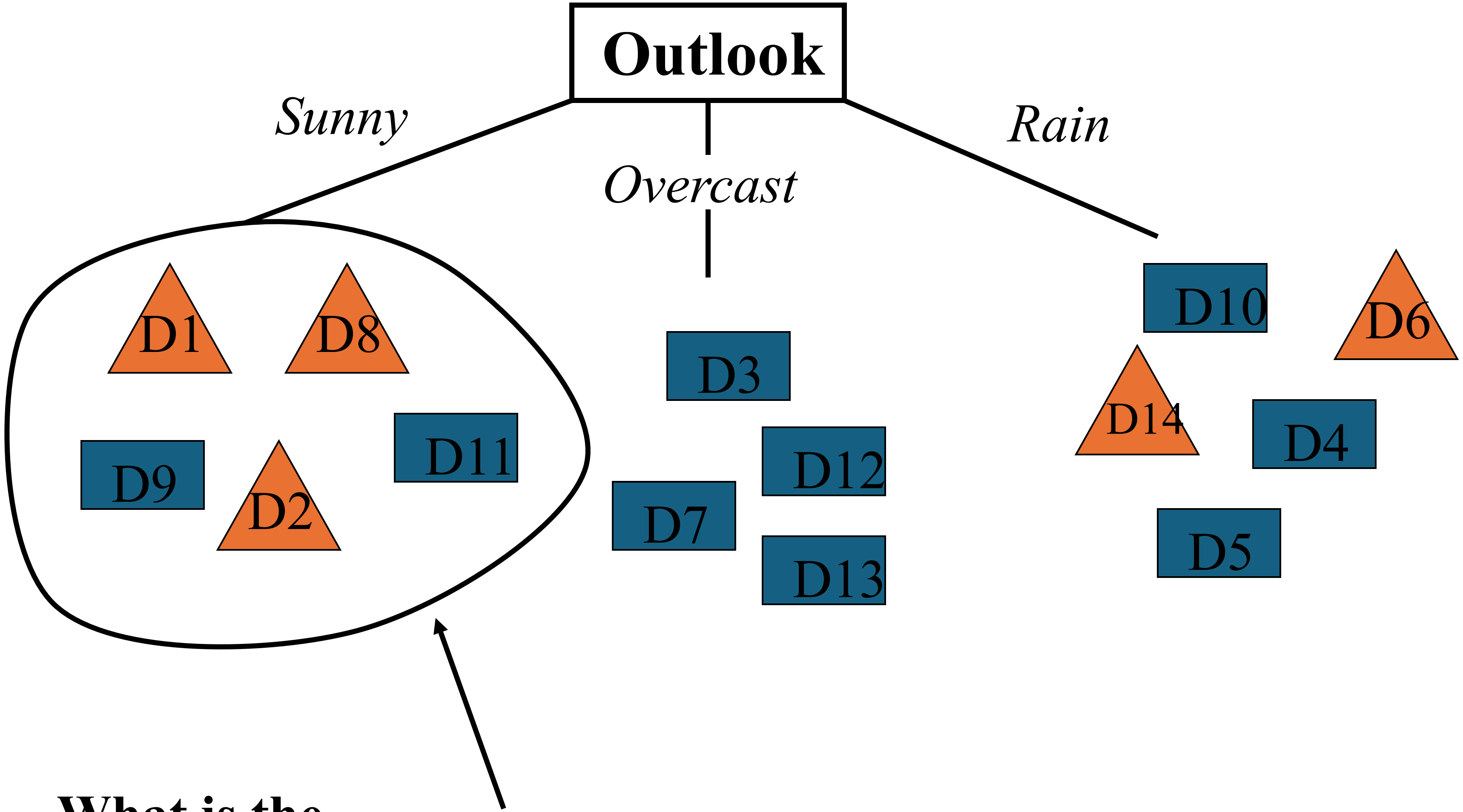


DECISION TREES

Basic Decision Tree Learning Algorithm

The selection process is then repeated using the training examples associated with each descendant node to select the best attribute to test at that point in the tree

DECISION TREES



What is the
“best” attribute to test at this point? The possible choices are
Temperature, Wind & Humidity

DECISION TREES

Basic Decision Tree Learning Algorithm

This forms a greedy search for an acceptable decision tree, in which the algorithm never backtracks to reconsider earlier choices

DECISION TREES

Which Attribute is the Best Classifier?

The central choice in the ID3 algorithm is selecting which attribute to test at each node in the tree

We would like to select the attribute which is most useful for classifying examples

For this we need a good quantitative measure

For this purpose a statistical property, called *information gain* is used

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

In order to define information gain precisely, we begin by defining entropy

Entropy characterizes the impurity of an arbitrary collection of examples

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

This formula is called Entropy H

$$H(X) = -\sum_{j=1}^{m} p_j \log_2 p_j$$

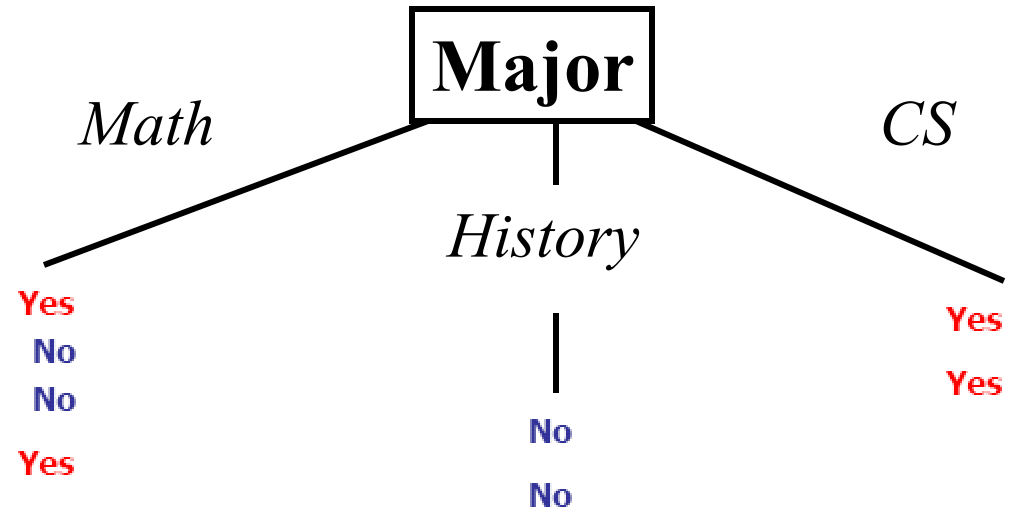
High Entropy means that the examples have more equal probability of occurrence (and therefore not easily predictable)

Low Entropy means easy predictability

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

Suppose we are trying to predict output Y (Like Film Gladiator) & we have input X (College Major = v)



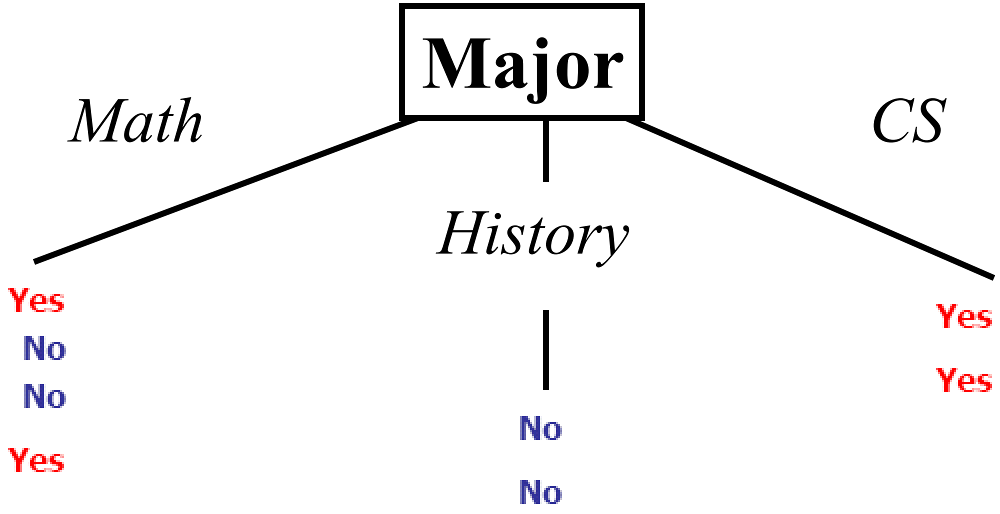
X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

Conditional Entropy $\quad H(Y \mid X = v)$

The Entropy of Y among only those records in which X = v



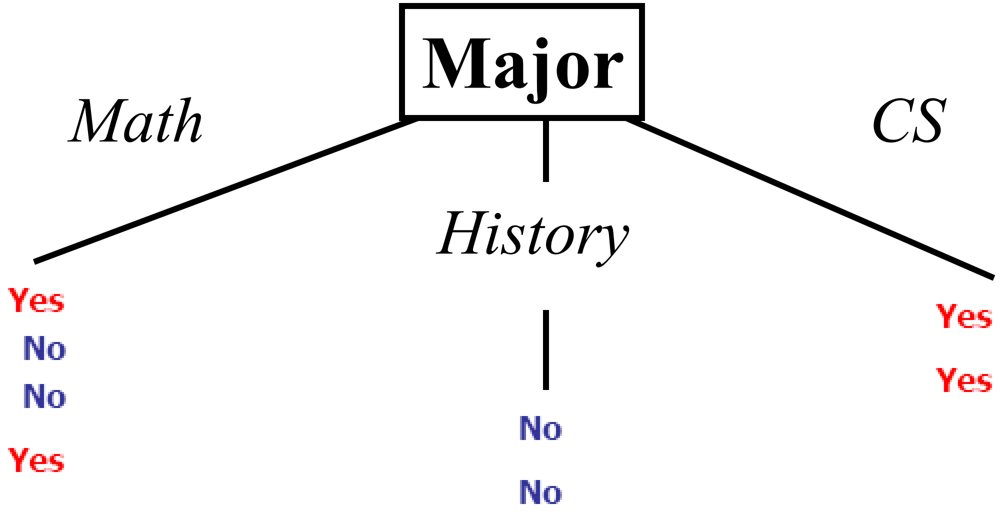
X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

Conditional Entropy of Y

$$H(Y \mid X = \text{Math}) = 1.0$$
$$H(Y \mid X = \text{History}) = 0$$
$$H(Y \mid X = \text{CS}) = 0$$



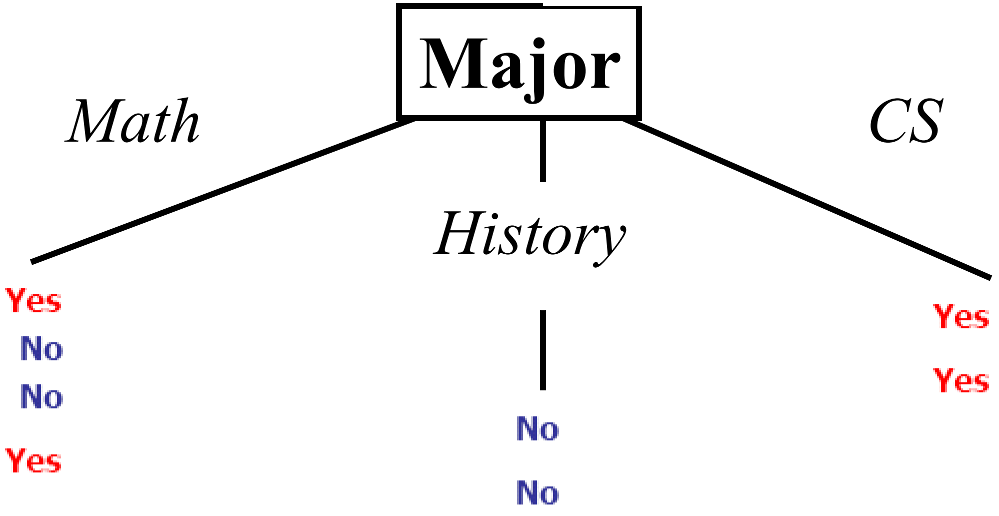
X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

DECISION TREES

Which Attribute is the Best Classifier?: Definition of Entropy

Average Conditional Entropy of Y

$$H(Y \mid X) = \Sigma_j \, Prob(X=v_j) \, H(Y \mid X = v_j)$$



v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$$

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

Information Gain is the expected reduction in entropy caused by partitioning the examples according to an attribute's value

$$\text{Info Gain } (Y \mid X) = H(Y) - H(Y \mid X) = 1.0 - 0.5 = 0.5$$

In general, we write Gain (S, A)
Where S is the collection of examples & A is an attribute

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Let's investigate the attribute *Wind*

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

The collection of examples has 9 positive values and 5 negative ones

$$Entropy(S) = 0.940$$

Eight (6 positive and 2 negative ones) of these examples have the attribute value *Wind = Weak*

Six (3 positive and 3 negative ones) of these examples have the attribute value *Wind = Strong*

DECISION TREES

Which Attribute is the Best Classifier?: Information Gain

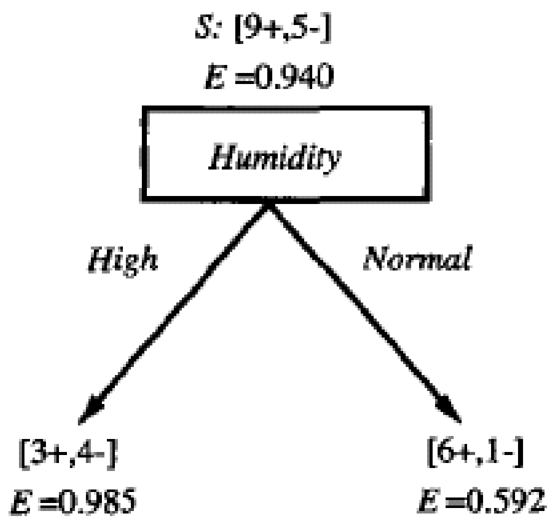
The information gain obtained by separating the examples according to the attribute *Wind* is calculated as:

$$
\begin{aligned}
Gain(S, Wind) &= Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S) - (8/14) Entropy(S_{Weak}) \\
&\quad - (6/14) Entropy(S_{Strong}) \\
&= 0.940 - (8/14)0.811 - (6/14)1.00 \\
&= 0.048
\end{aligned}
$$

DECISION TREES

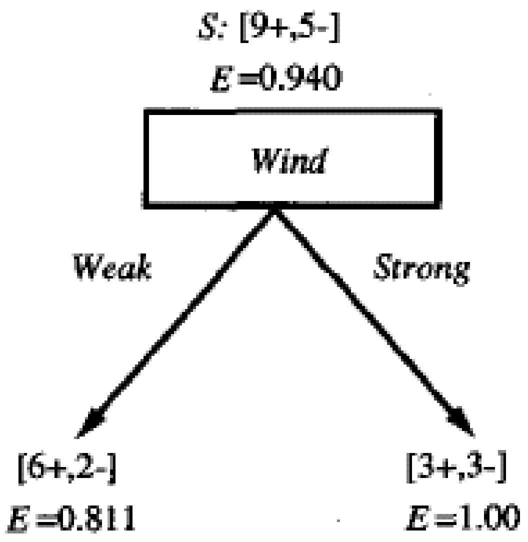
Which Attribute is the Best Classifier?: Information Gain

We calculate the Info Gain for each attribute and select the attribute having the highest Info Gain



Gain (S, Humidity)

$= .940 - (7/14).985 - (7/14).592$
$= .151$



Gain (S, Wind)

$= .940 - (8/14).811 - (6/14)1.0$
$= .048$

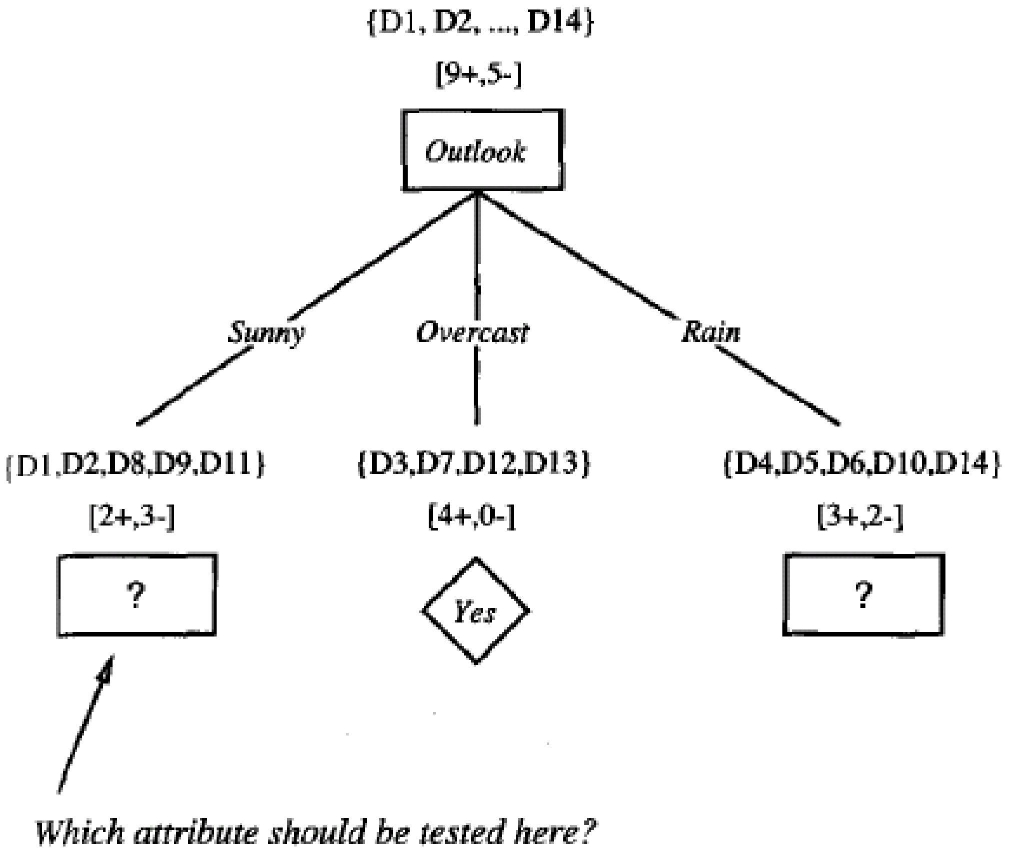
Example

Which attribute should be selected as the first test?

$$Gain(S, Outlook) = 0.246$$
$$Gain(S, Humidity) = 0.151$$
$$Gain(S, Wind) = 0.048$$
$$Gain(S, Temperature) = 0.029$$

“Outlook” provides the most information

DECISION TREES



$S_{sunny} = \{D1, D2, D8, D9, D11\}$

$Gain(S_{sunny}, Humidity) = .970 - (3/5)\,0.0 - (2/5)\,0.0 = .970$

$Gain(S_{sunny}, Temperature) = .970 - (2/5)\,0.0 - (2/5)\,1.0 - (1/5)\,0.0 = .570$

$Gain(S_{sunny}, Wind) = .970 - (2/5)\,1.0 - (3/5)\,.918 = .019$

DECISION TREES

Example

The process of selecting a new attribute is now repeated for each (non-terminal) descendant node, this time using only training examples associated with that node

Attributes that have been incorporated higher in the tree are excluded, so that any given attribute can appear at most once along any path through the tree

DECISION TREES

Example

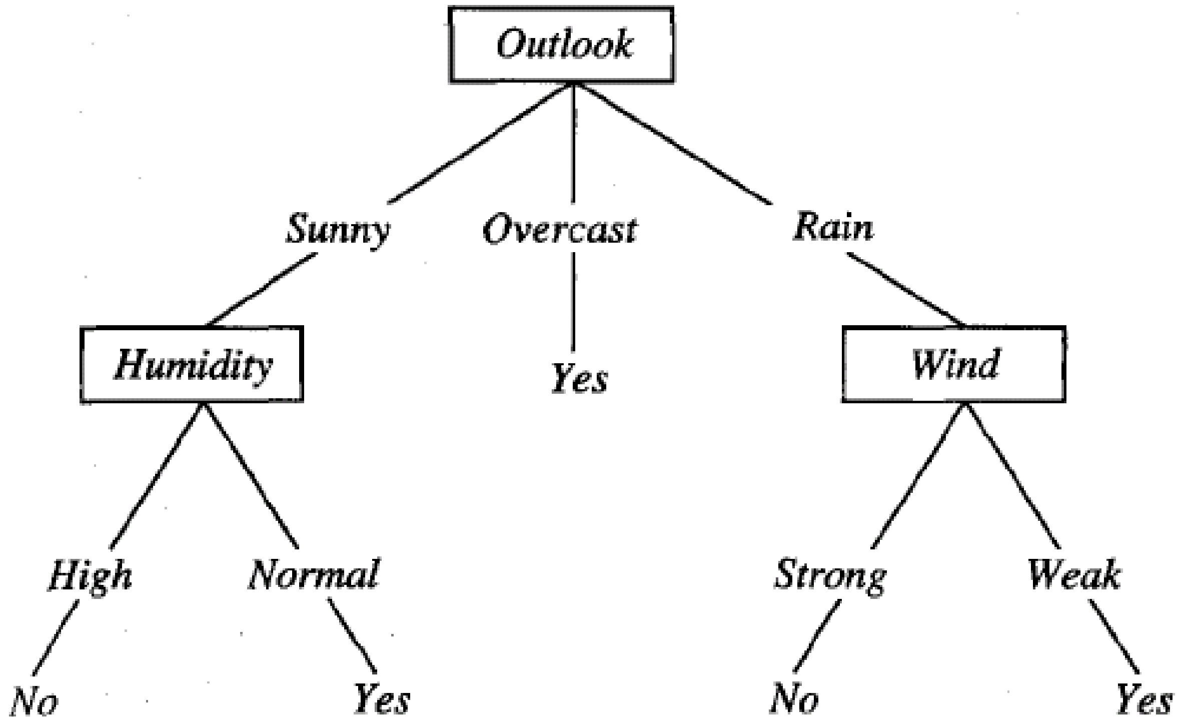
This process continues for each new leaf node until either:

1. **Every attribute has already been included along this path through the tree**

2. **The training examples associated with a leaf node have zero entropy**

DECISION TREES

Example



DECISION TREES

From Decision Trees to Rules

Next Step: Make rules from the decision tree

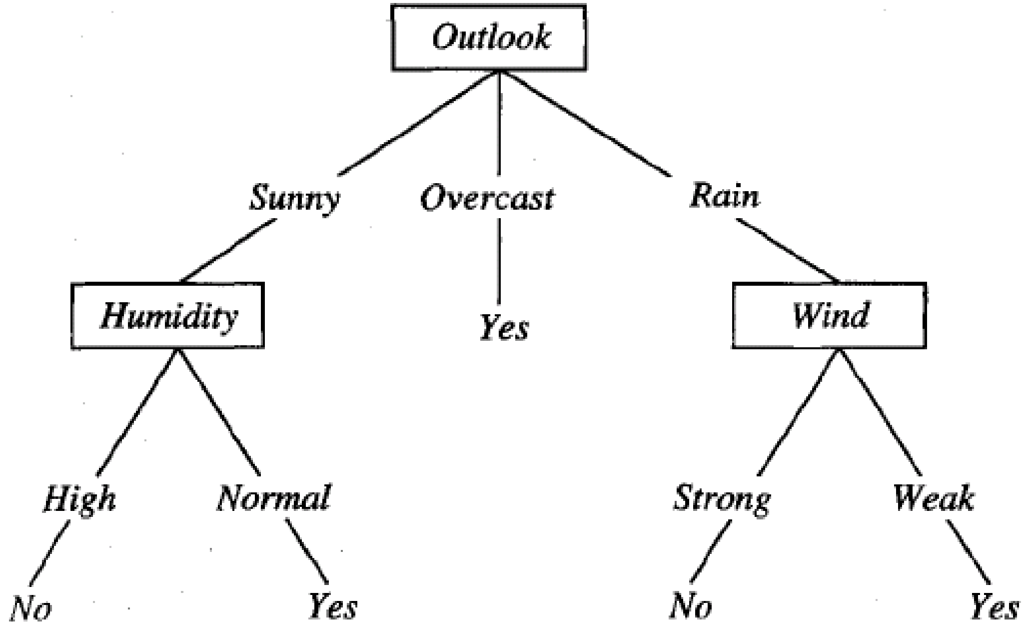
After making the identification tree, we trace each path from the root node to leaf node, recording the test outcomes as *antecedents* and the leaf node classification as the *consequent*

For our example we have:

If the Outlook is Sunny and the Humidity is High then No
If the Outlook is Sunny and the Humidity is Normal then Yes
...

DECISION TREES

Decision Tree Representation



$$(Outlook = Sunny \wedge Humidity = Normal)$$
$$\vee \quad (Outlook = Overcast)$$
$$\vee \quad (Outlook = Rain \wedge Wind = Weak)$$