Decision Tree

Nazerfard, Ehsan
nazerfard@aut.ac.ir

Decision Boundary

- What kind of Decision Boundary for this data?

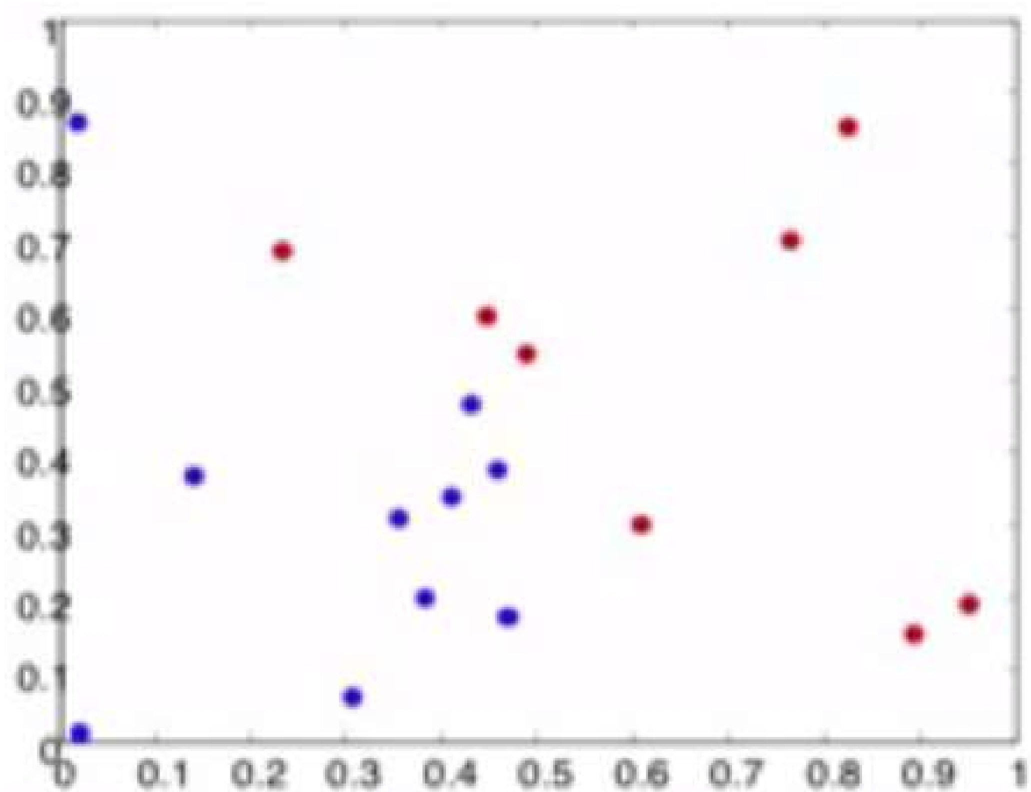
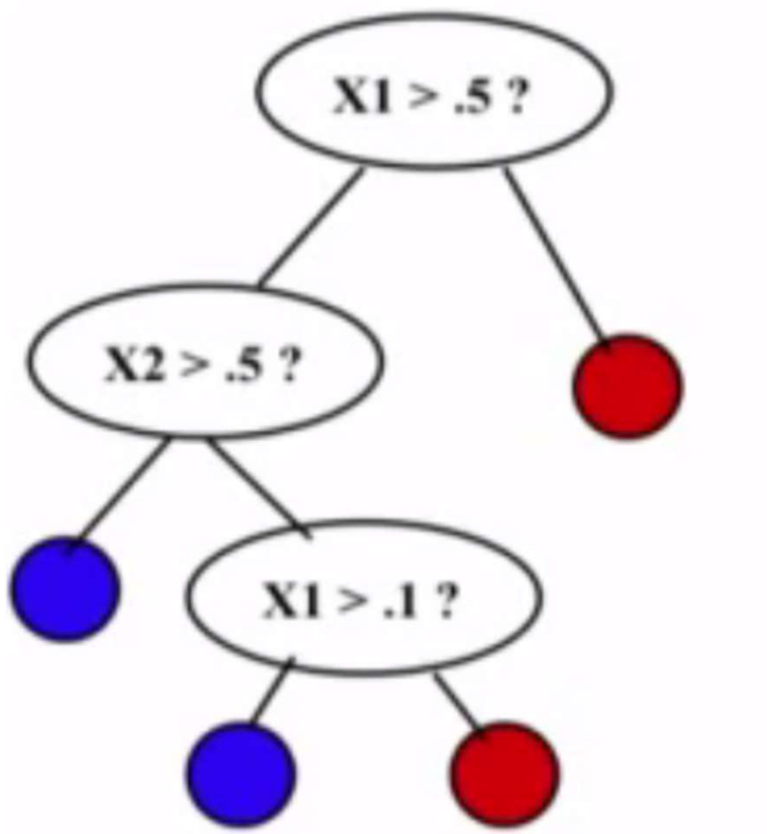
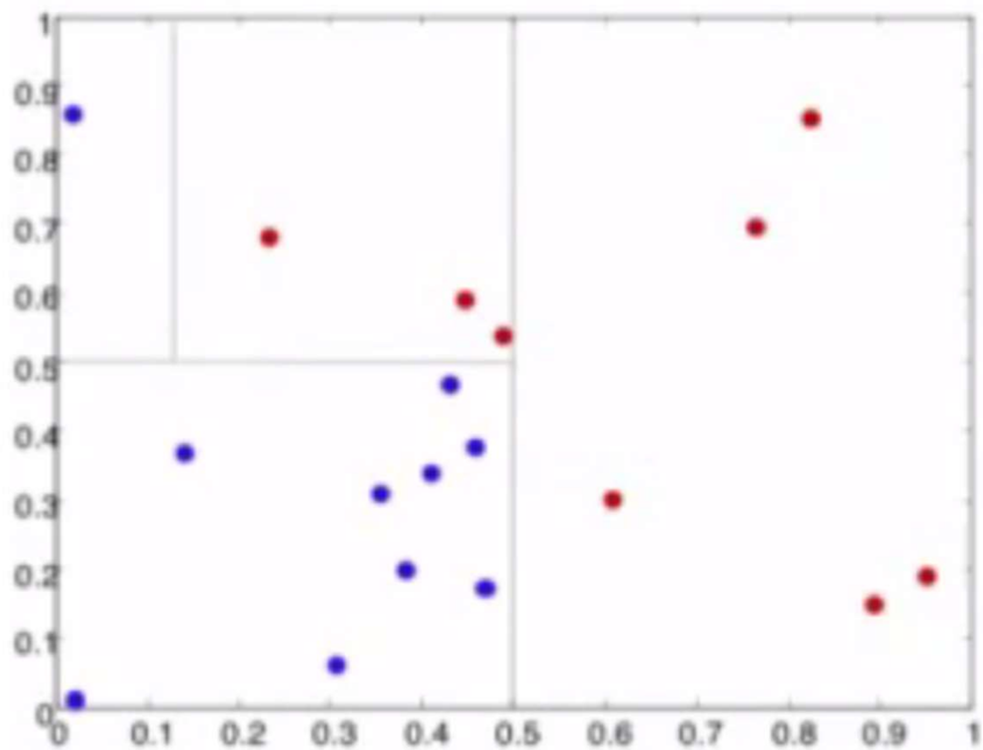


Figure © Alex Ihler, Lecture on Decision Trees, UCI

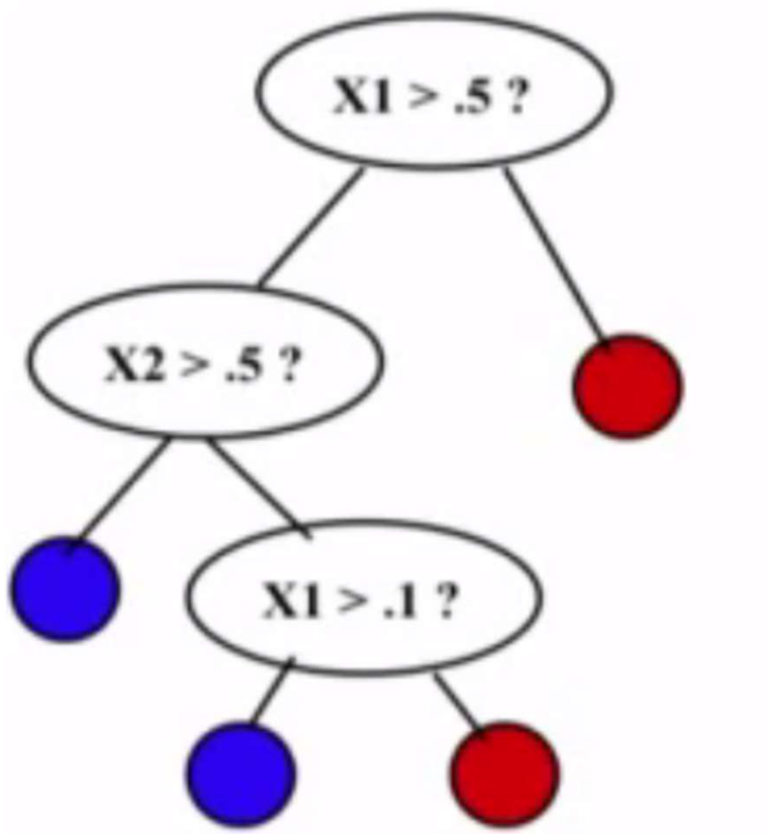
Decision Tree Classifier

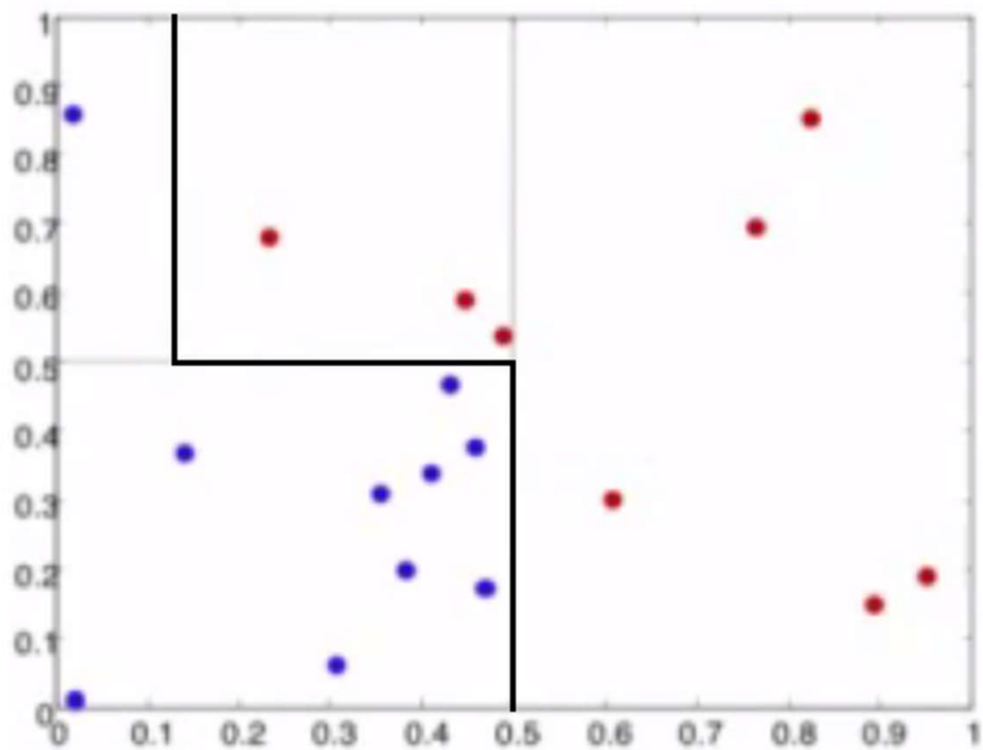




Figures © Alex Ihler, Lecture on Decision Trees, UCI

Decision Tree Classifier (cont.)





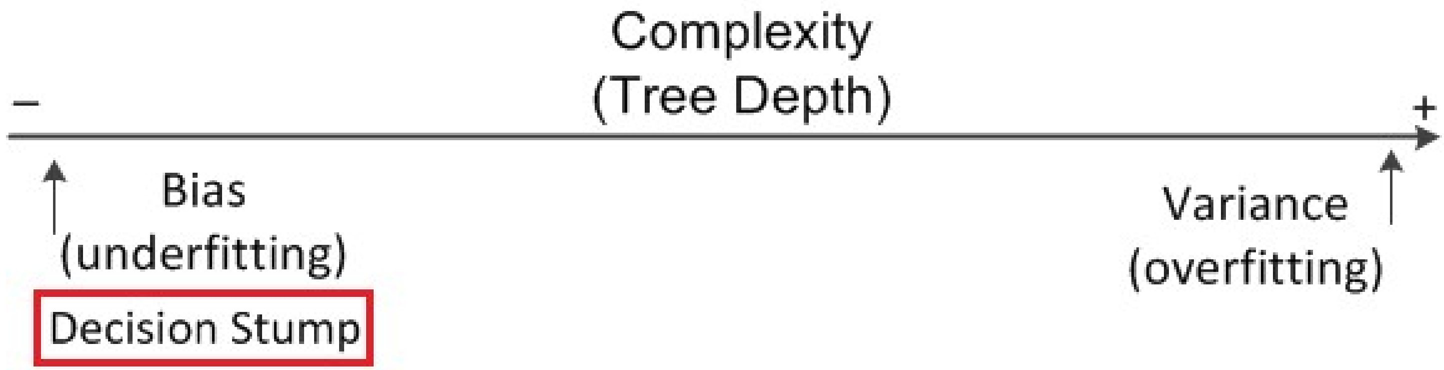
What are the challenges?

Decision Tree Properties

- How interpretable is DT?

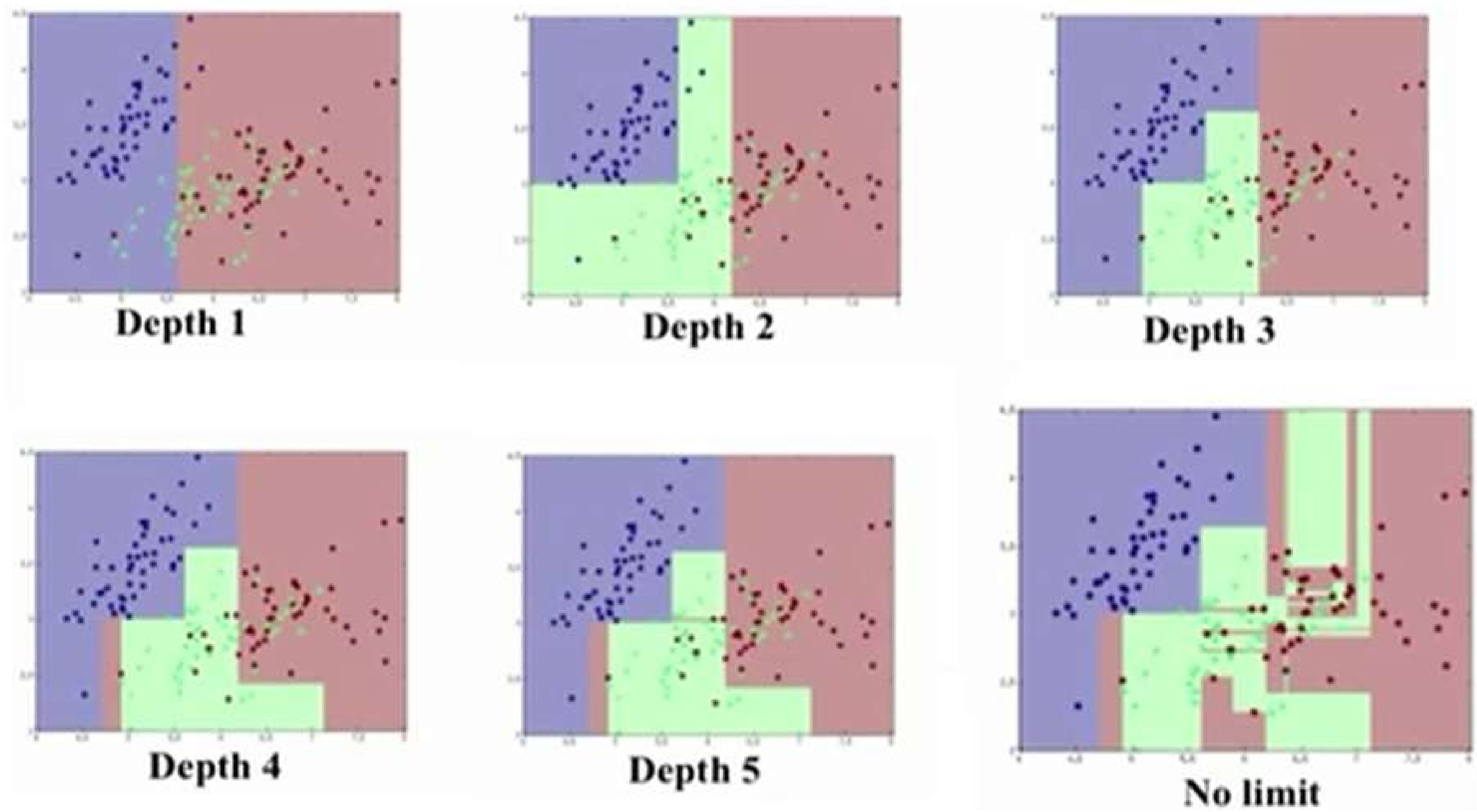
- Non-Parametric

- Bias-Variance tradeoff



Controlling Complexity

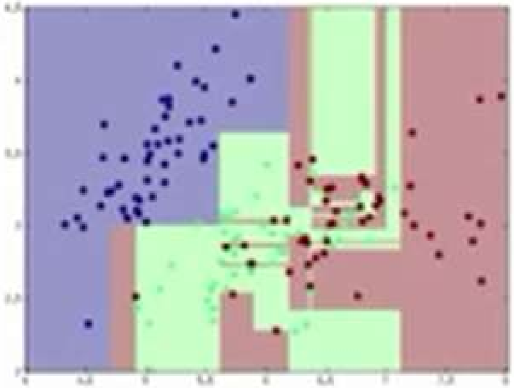
- Maximum depth cutoff



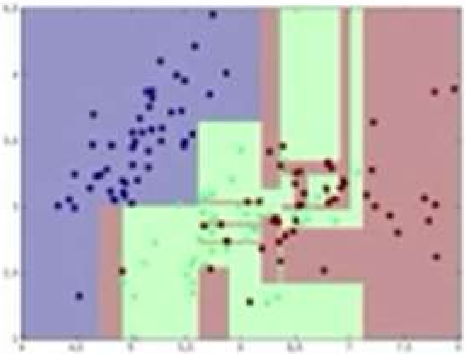
Figures © Alex Ihler, Lecture on Decision Trees, UCI

Controlling Complexity (cont.)

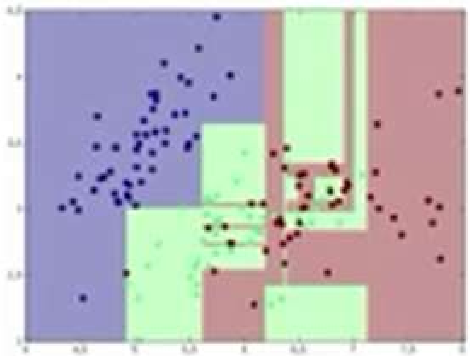
- Minimum # parent data



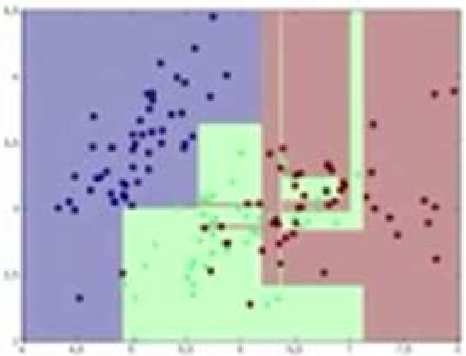
minParent 1



minParent 3



minParent 5



minParent 10

The Akinator

- What would be the most important question to ask first?

Figure © en.akinator.com

Entropy

- Entropy → Information Gain
- Entropy is a measure of randomness

- Example:
  - Communicating fair coin tosses:

    H T T H T H H H T T H T T …
  - Communicating my daily lottery results

    0 0 0 0 0 0 …

  - The latter, takes less work to communicate. Why?

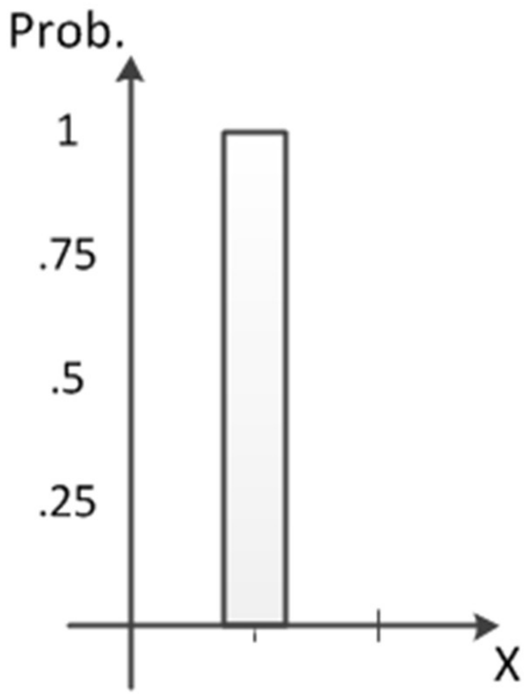
Entropy (cont.)

- Formal Definition
  - $H(X) = \sum_i p(x_i) \log_2^{\frac{1}{p(x_i)}}$
  - Shannon got the letter H from Boltzmann's H-theorem

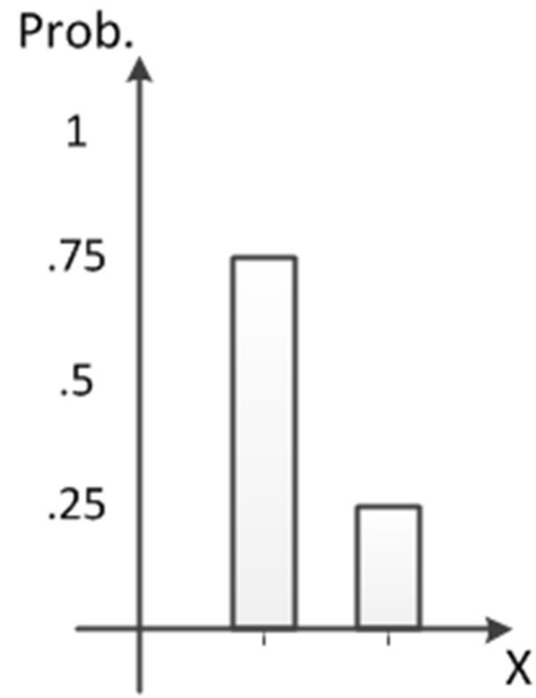
- $E(X)$ measures the randomness of random variable X.

$$E(X) = -\sum_i p(x_i) \log_2^{p(x_i)}$$

Entropy (cont.)



$$E(X) = -(\log_2 1 + 0 \times \log_2 0) = 0$$



$$E(X) = -(0.75 \log_2 0.75 + 0.25 \log_2 0.25) = 0.81$$

Entropy (cont.)

- $E(S)$ measures the impurity of set S.

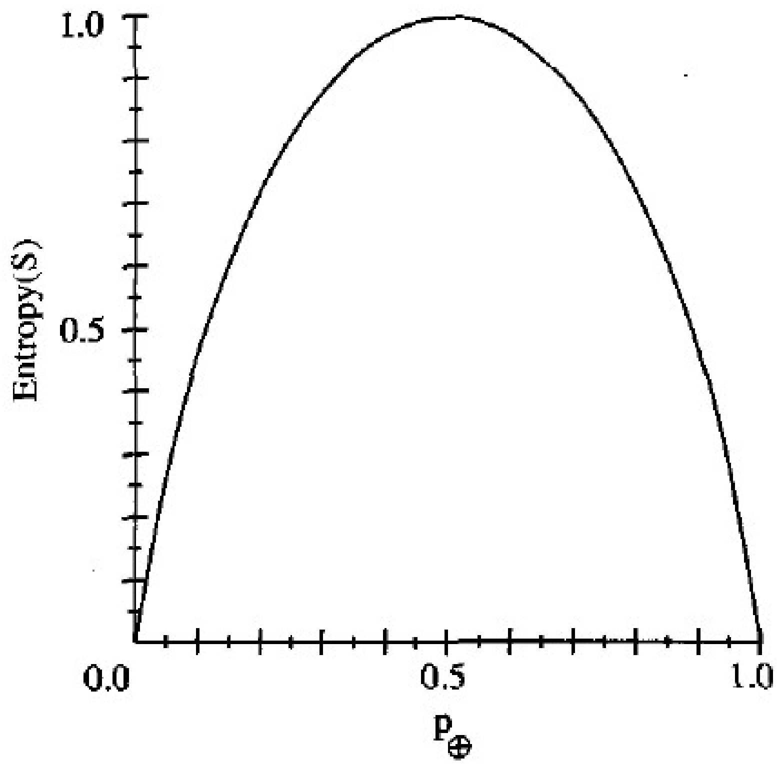
- Given a collection of S, containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is:

$$E(S) = -(p_{\oplus} log_2{}^{p_{\oplus}} + p_{\ominus} log_2{}^{p_{\ominus}})$$

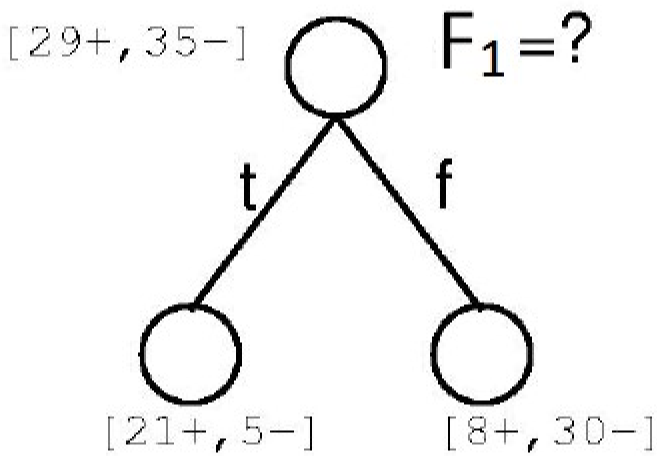
  - $p_{\oplus}$ refers to the proportion of positive examples.

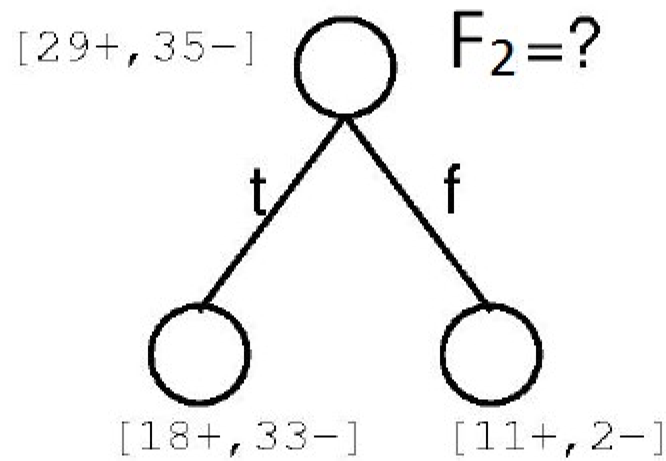
Entropy Function

- The entropy function relative to a boolean classification, as $p_{\oplus}$ varies between 0 and 1.



Which Feature is more Effective*?

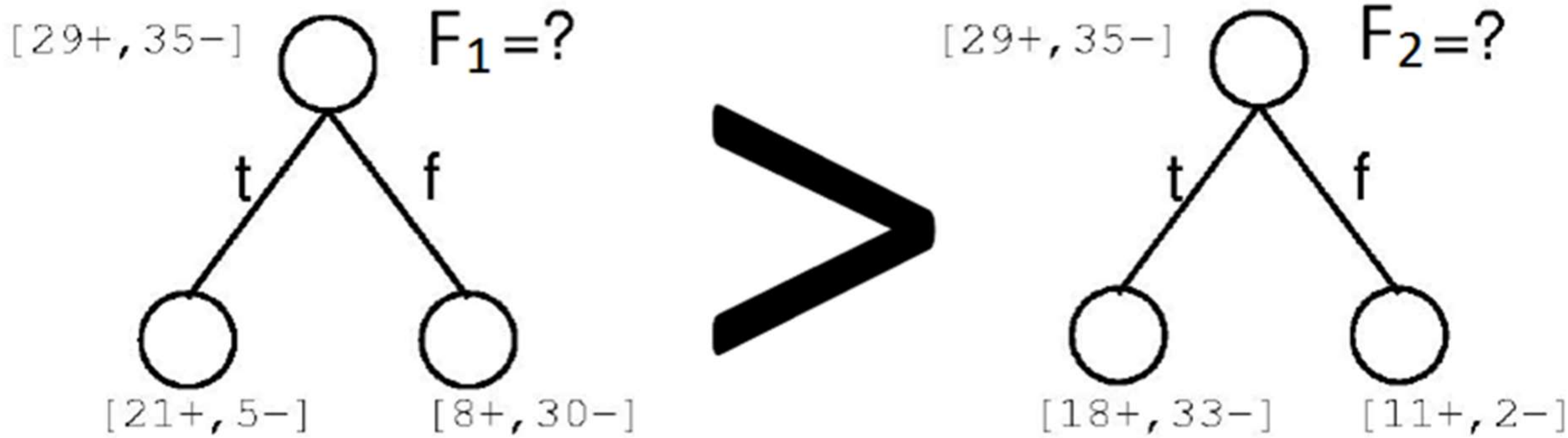




*effective here indicates dichotomiser

Example © Larry Holder, Lecture on Decision Trees, WSU

How Come?



IG([29+, 35-], F_1) = 0.26

IG([29+, 35-], F_2) = 0.12

IG (Information Gain)

Example © Larry Holder, Lecture on Decision Trees, WSU

Information Gain

- Information Gain measures the expected reduction in entropy.
- Formally, the information gain, $Gain(S,F)$ of a feature F, relative to a collection of examples S, is defined as:

$$Gain(S,F) = Entropy(S) - \sum_{v \in Values(F)} \frac{|S_v|}{|S|} Entropy(S_v)$$

- where $Values(F)$ is the set of all possible values for feature F, and S_v is the subset of S for which feature F has value v.

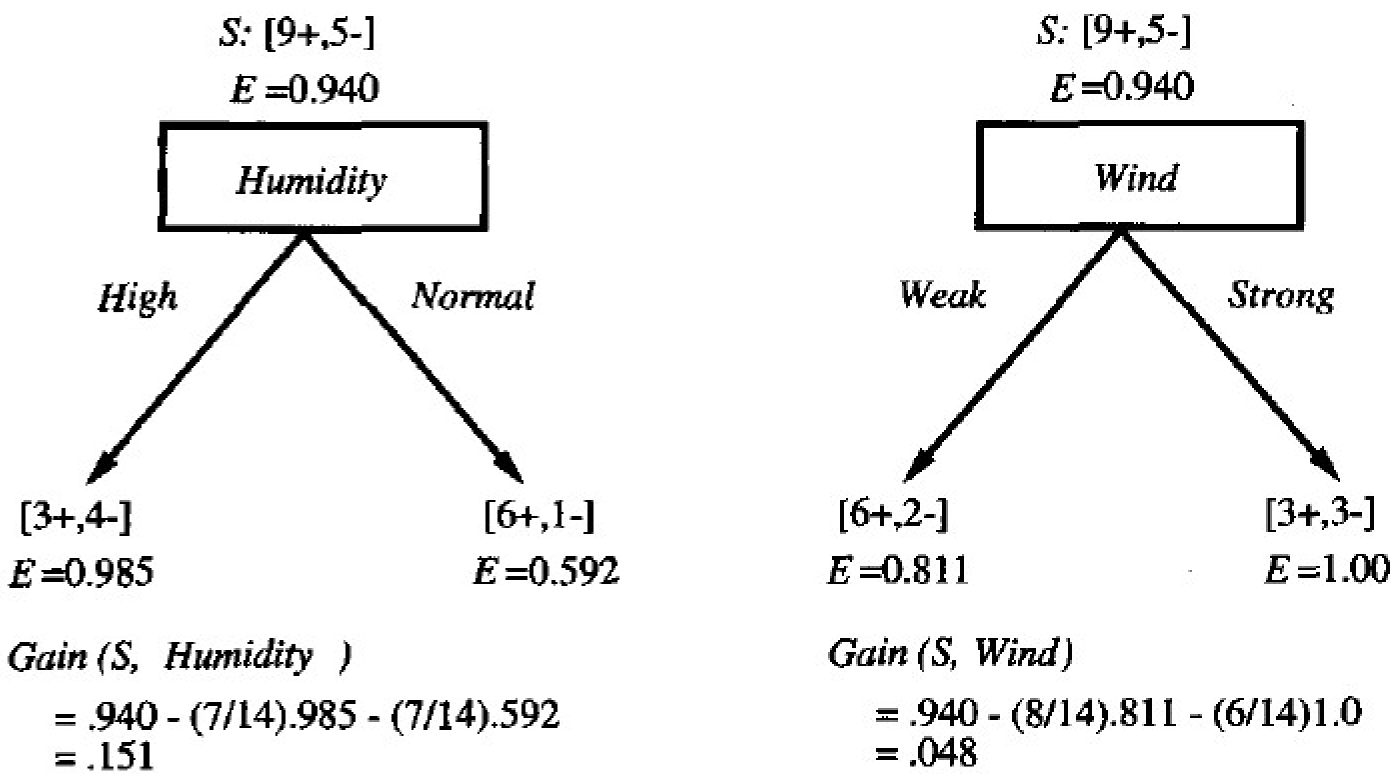
The PlayTennis/Weather Data

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

The Wind Feature

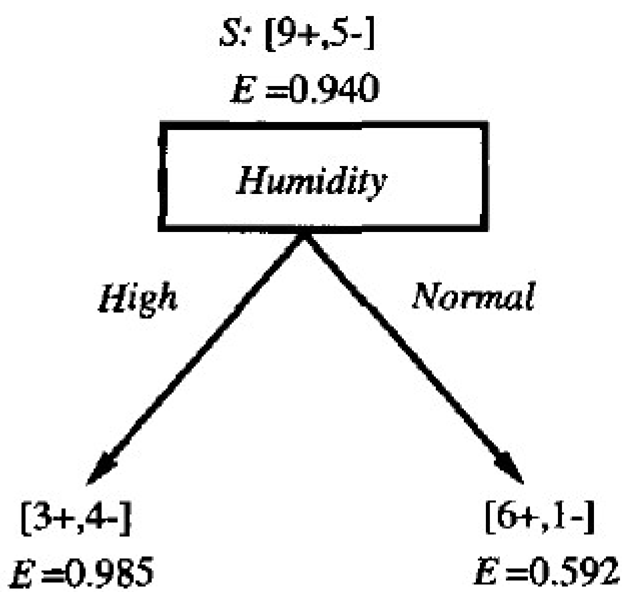
$$
\begin{aligned}
Values(Wind) &= Weak, Strong \\
S &= [9+, 5-] \\
S_{Weak} &\leftarrow [6+, 2-] \\
S_{Strong} &\leftarrow [3+, 3-] \\
Gain(S, Wind) &= Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&= Entropy(S) - (8/14) Entropy(S_{Weak}) \\
&\quad - (6/14) Entropy(S_{Strong}) \\
&= 0.940 - (8/14)0.811 - (6/14)1.00 \\
&= 0.048
\end{aligned}
$$

Which Feature is more Effective?



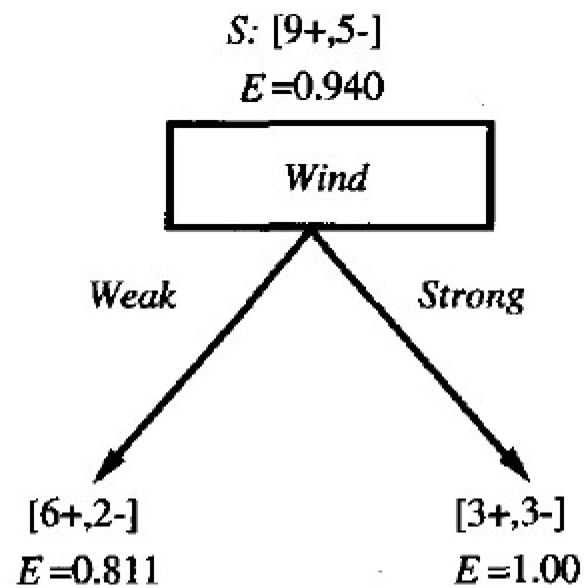
$$Gain(S, Humidity) = 0.151 \qquad Gain(S, Wind) = 0.048$$

Which Feature is the most Effective?



Gain (S, Humidity)

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



Gain (S, Wind)

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

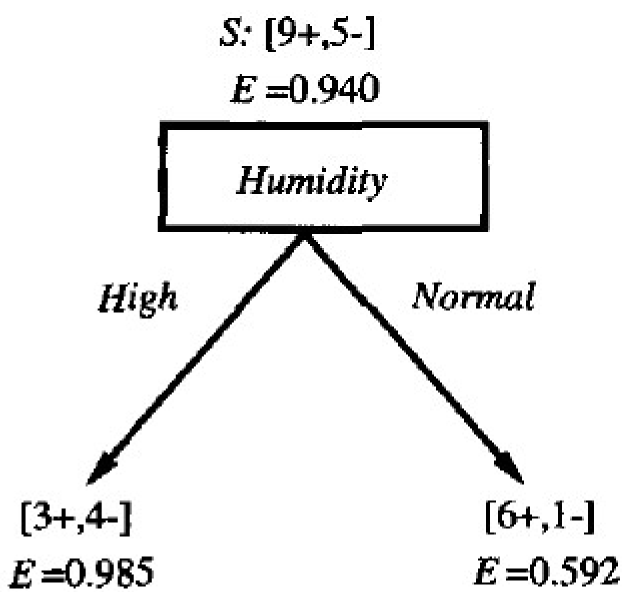
$$Gain(S, Humidity) = 0.151$$

$$Gain(S, Outlook) = 0.246$$

$$Gain(S, Wind) = 0.048$$

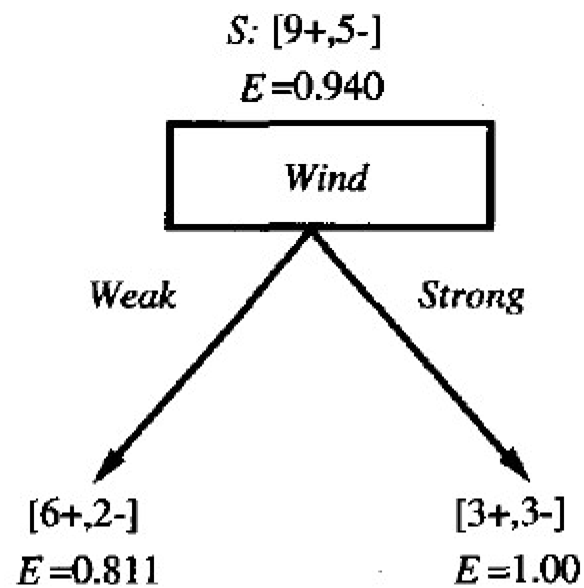
$$Gain(S, Temp) = 0.029$$

The Outlook Feature is the Best



Gain (S, Humidity)

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



Gain (S, Wind)

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

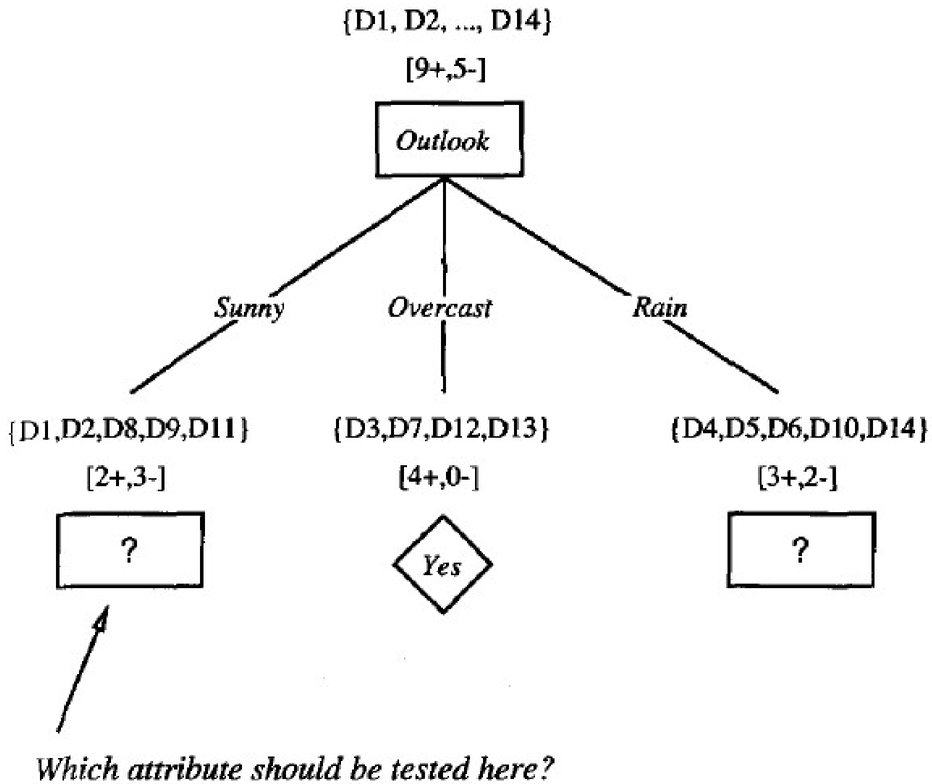
$$Gain(S, Humidity) = 0.151$$

$$\boxed{Gain(S, Outlook) = 0.246}$$

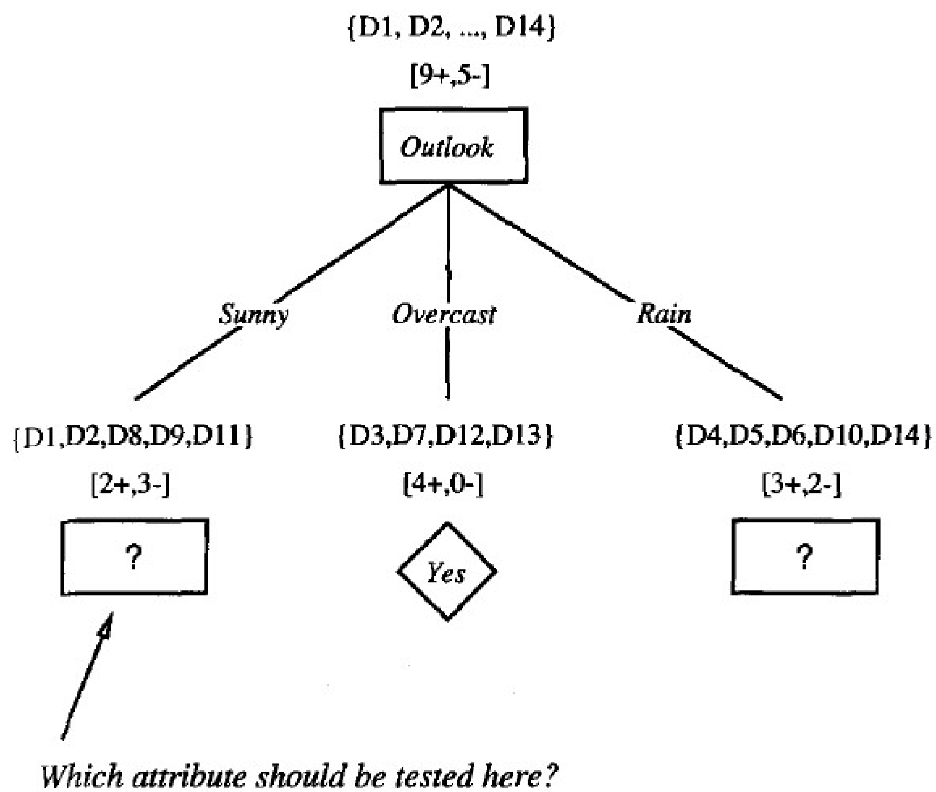
$$Gain(S, Wind) = 0.048$$

$$Gain(S, Temp) = 0.029$$

What about the Next Feature



What about the Next Feature



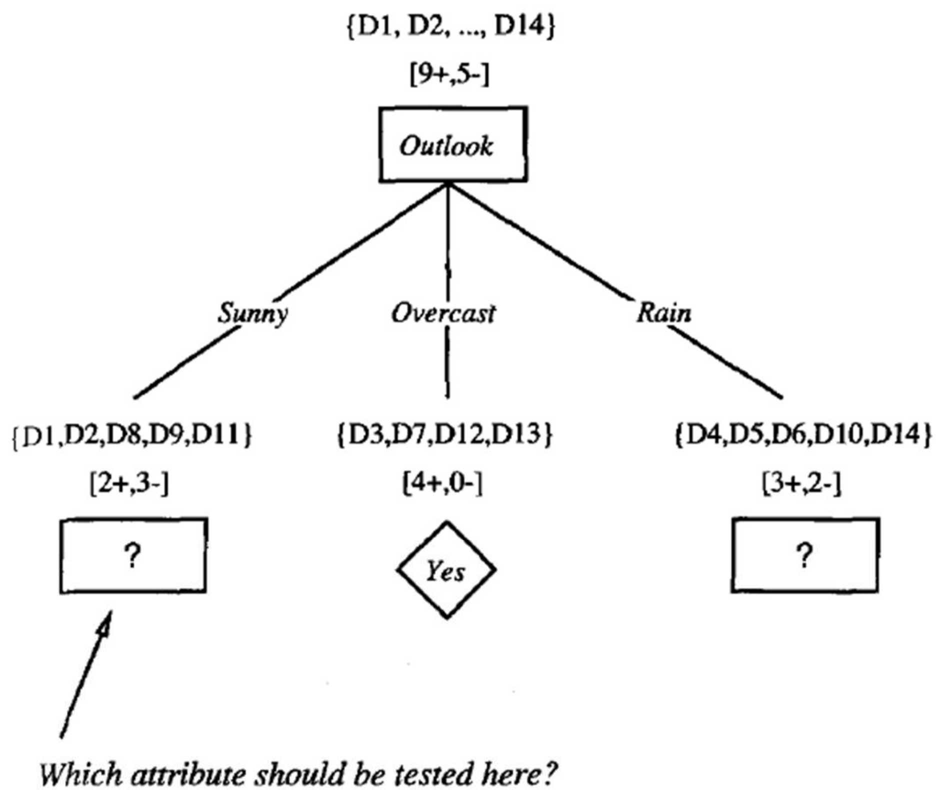
$S_{sunny} = \{D1, D2, D8, D9, D11\}$

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

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

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

The Humidity Feature



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

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

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

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

Hypothesis Space Search

- ID3 algorithm

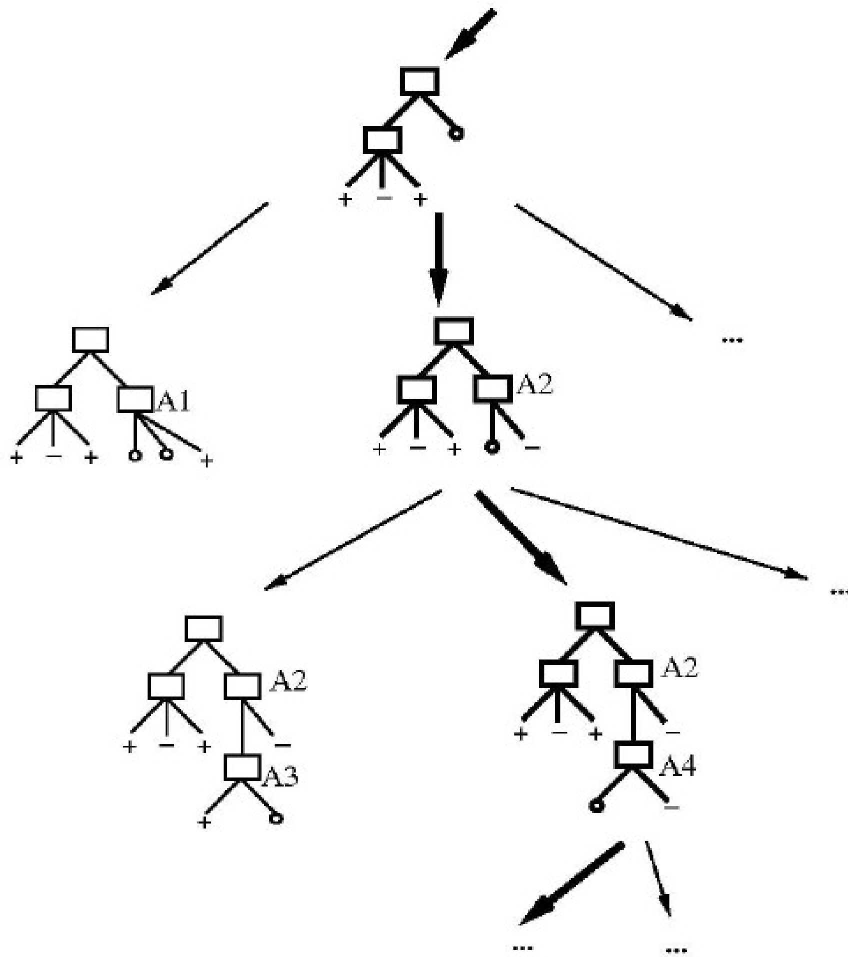


Figure © Tom Mitchell, Machine Learning, 1997

The Final Decision Tree

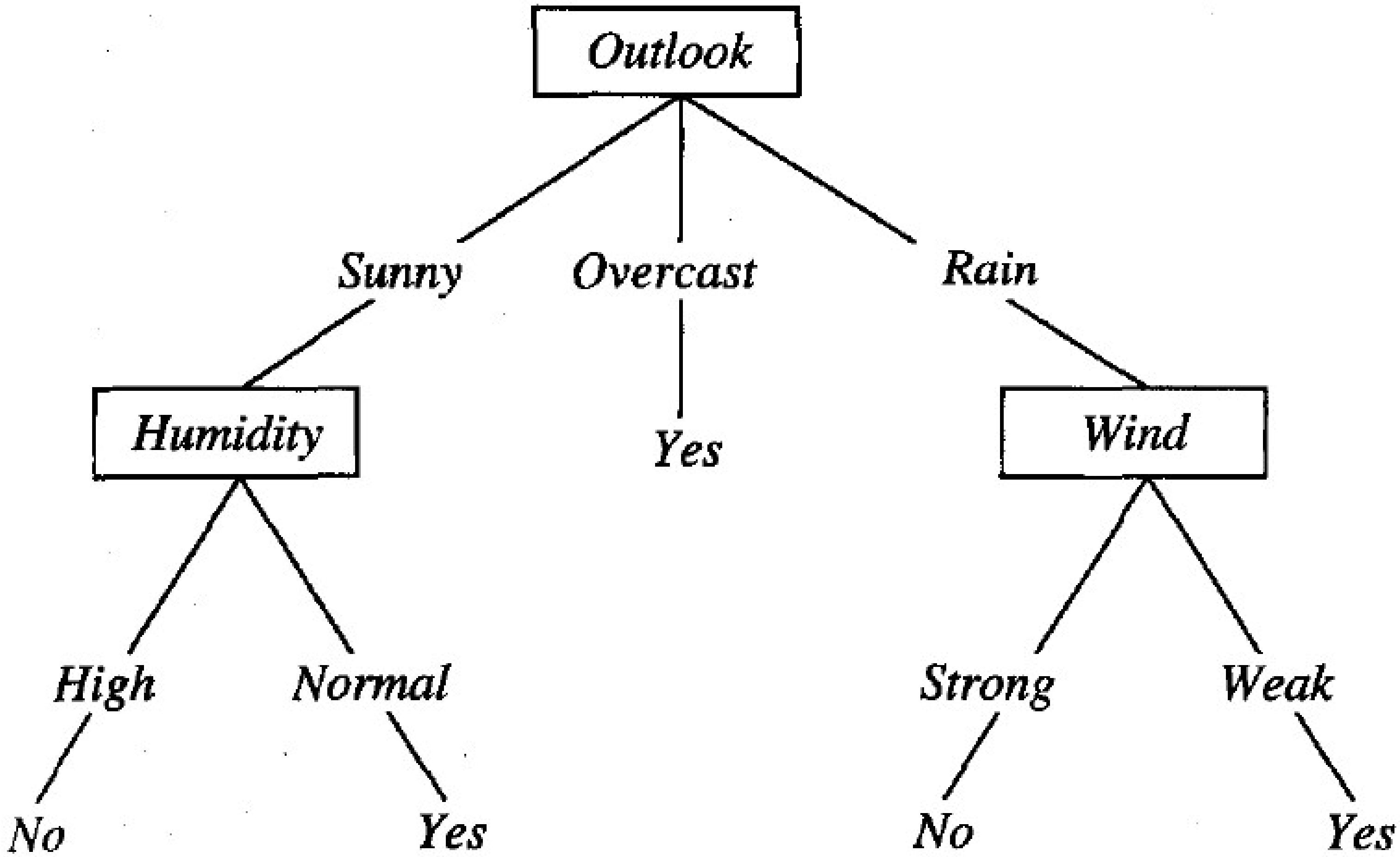


Figure © Tom Mitchell, Machine Learning, 1997

The Weather Data

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

Observations

- The Weather example by Weka

- Observations?
  - Where has the Temperature feature gone?
  - Dealing with random noise in the training data:

$$D_{15} = (< Outlook = Sunny, Temperature = Hot, Humidity = Normal, Wind = Strong >, Play = No)$$

  - Regarding the Day feature

C4.5 (J48 in Weka)

- C4.5 vs. ID3

	ID3	C4.5
Pruning	×	√
Features	nominal	nominal + numeric
missing values	×	√
Gain	Basic Info Gain	Gain Ratio

The C4.5 Algorithm

- C4.5 vs. ID3

	ID3	C4.5
Pruning	×	√
Features	nominal	nominal + numeric
missing values	×	√
Gain	Basic Info Gain	Gain Ratio

- MDLPC

Temp:	40	48	60	72	80	90
Play:	−	−	+	+	+	−

The C4.5 Algorithm

- C4.5 vs. ID3

	ID3	C4.5
Pruning	×	√
Features	nominal	nominal + numeric
missing values	×	√
Gain	Basic Info Gain	Gain Ratio

$$GainRatio(S,F) = \frac{Gain(S,F)}{SplitEntropy(S,F)}$$

$$SplitEntropy(S,F) = -\sum_{i=1}^{|Values\ F|} \frac{|S_i|}{|S|} log^{\frac{|S_i|}{|S|}}$$

Further Reading

- The CART algorithm
  - Classification and Regression Tree
- Gini index

- Random Forest (later on in this course)

Source

- Tom Mitchell, Machine Learning, McGraw-Hill, 1997