Decision Tree

By Zhang Hongxin

State Key Lab of CAD&CG, ZJU

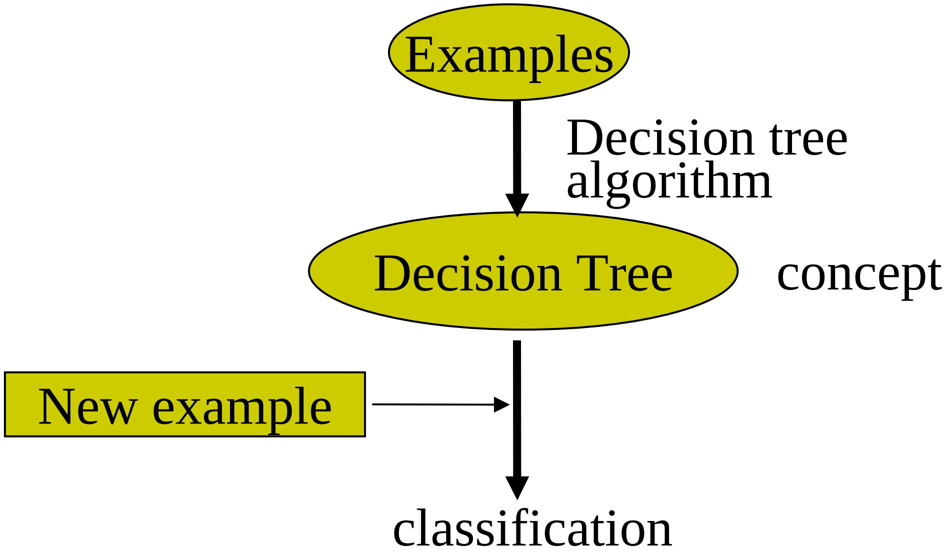
2005-06-16

Review

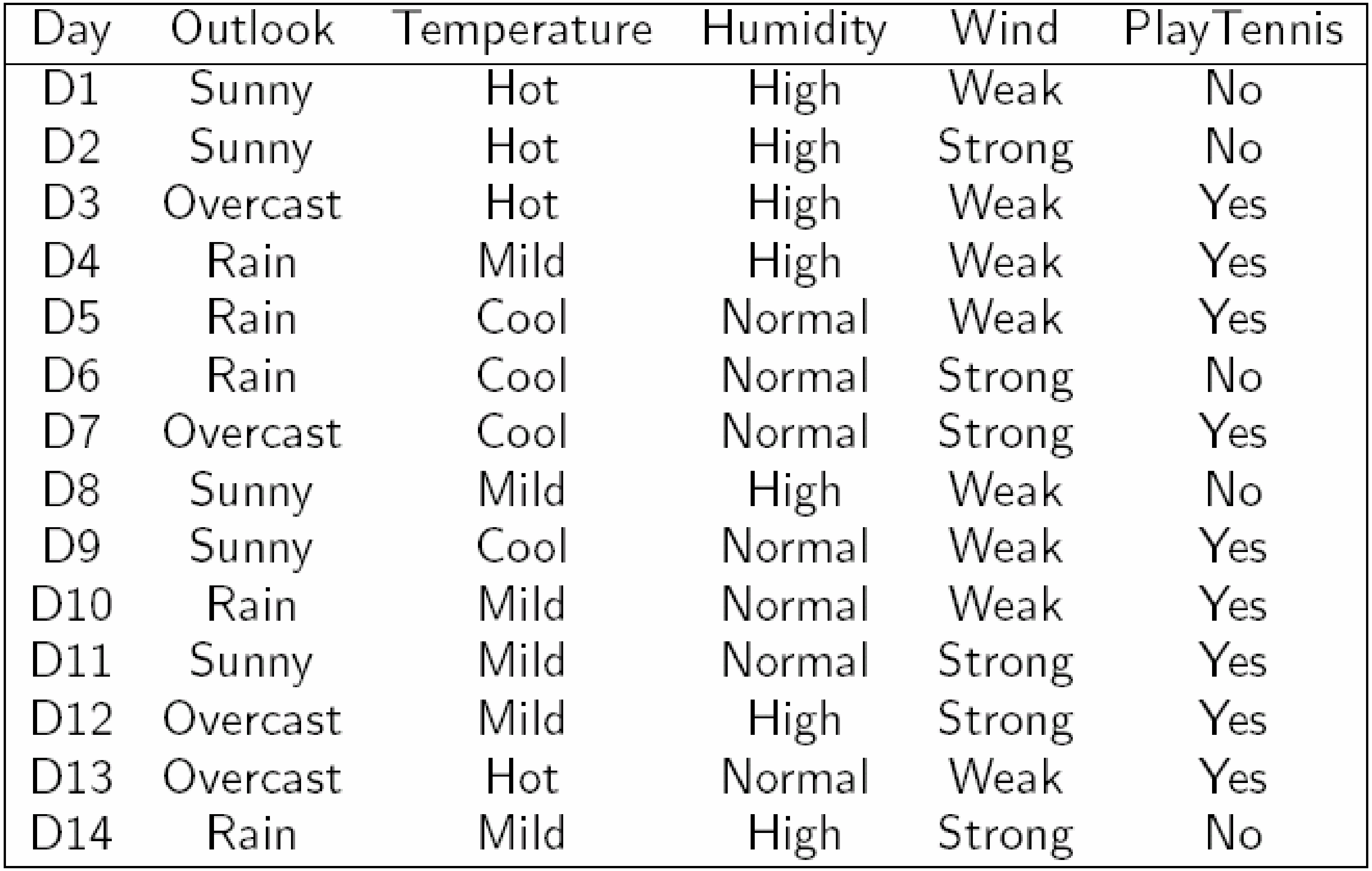
- Concept learning
  - Induce Boolean function from a sample of positive/negative training examples.
  - Concept learning can be cast as searching through predefined hypotheses space
- Searching Algorithm:
  - FIND-S
  - LIST-THEN-ELIMINATE
  - CANDIDATE-ELIMINATION

Decision Tree

1. Decision tree learning is a method for approximating discrete-valued target functions (Classifier), in which the learned function is represented by a decision tree.
2. Decision tree algorithm induces concepts from examples.
3. Decision tree algorithm is a general-to-specific searching strategy

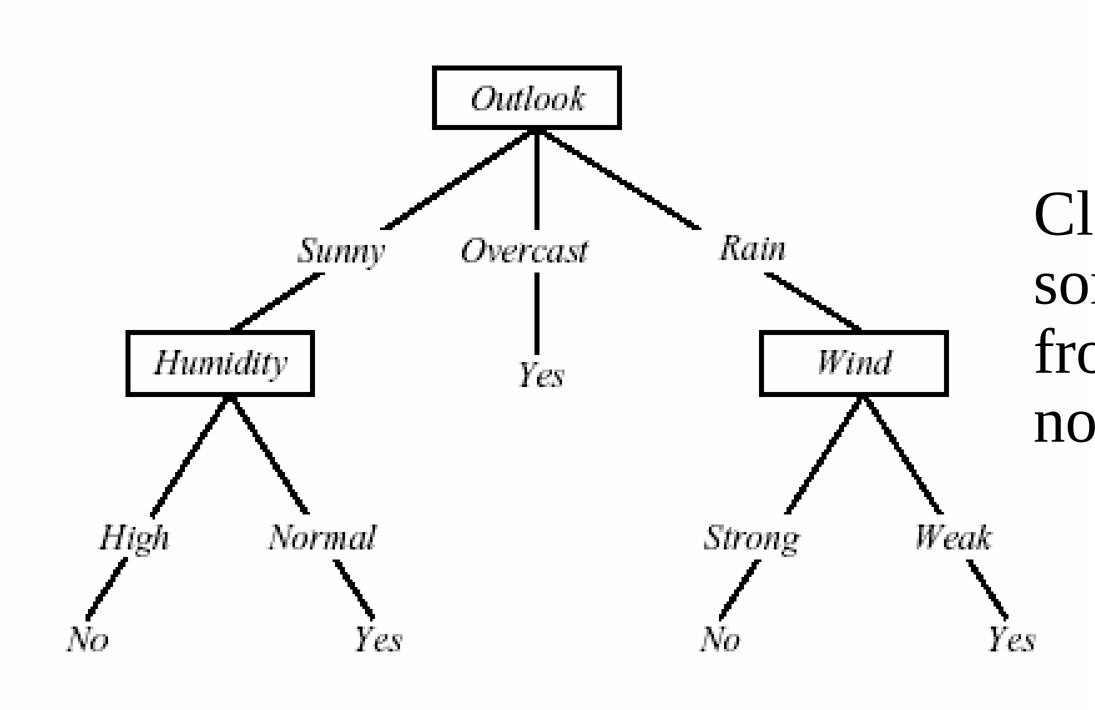


A Demo Task – *Play Tennis*



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 Tree Representation



Classify instances by sorting them down the tree from the root to some leaf node

- Each branch corresponds to attribute value
- Each leaf node assigns a classification

Decision Tree Representation

- Each path from the tree root to a leaf corresponds to a conjunction of attribute tests

  (Overlook = Sunny) ^ (Humidity = Normal)

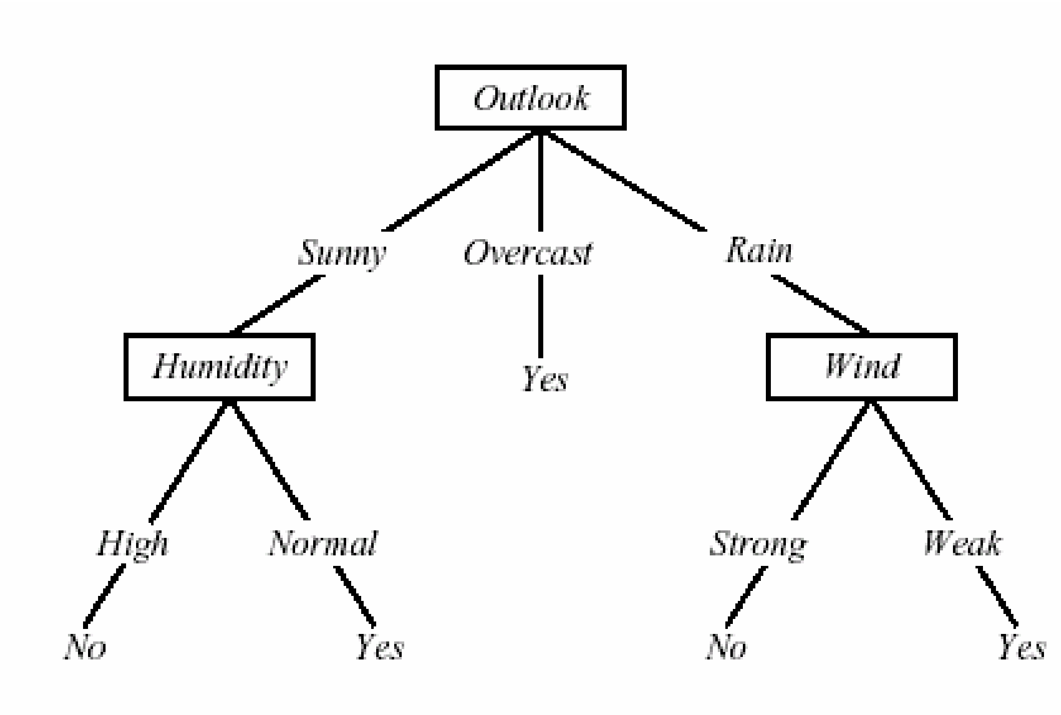
The tree itself corresponds to a disjunction of these conjunctions

(Overlook = Sunny ^ Humidity = Normal)
V (Outlook = Overcast)
V (Outlook = Rain ^Wind = Weak)

Top-Down Induction of Decision Trees

Main loop:

1. *A*   *the “best” decision attribute for next node*
2. *Assign A as decision attribute for node*
3. *For each value of A, create new descendant of node*
4. *Sort training examples to leaf nodes*
5. *If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes*

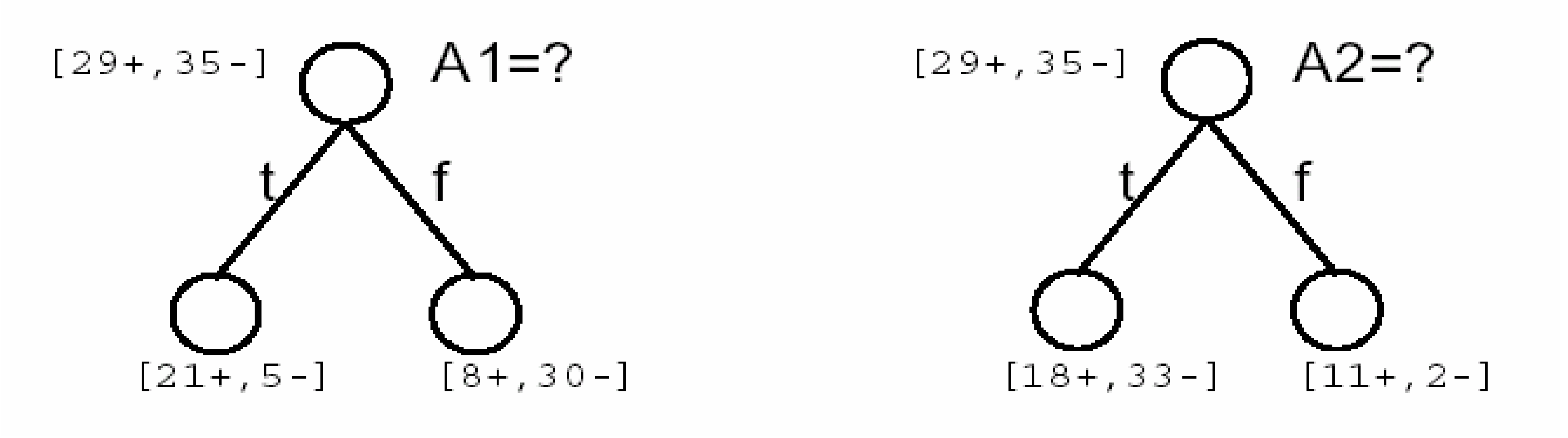


{ Outlook = Sunny, Temperature = Hot, Humidity = High, Wind = Strong }

Tests attributes along the tree

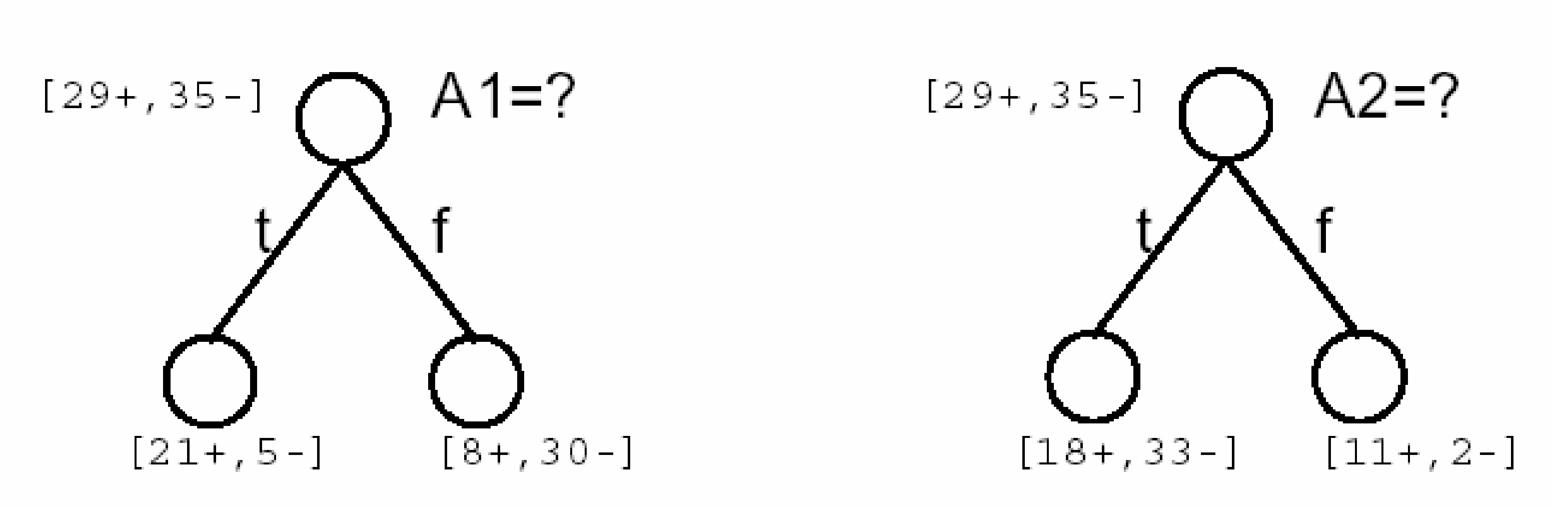
- typically, equality test (e.g., “Wind=Strong”)
- other tests (such as inequality) are possible

Which Attribute is Best?



- Occam’s razor: (year 1320)
  - Prefer the simplest hypothesis that fits the data.
- Why?
  - It’s a philosophical problem.
    - Philosophers and others have debated this question for centuries, and the debate remains unresolved to this day.

Simple is beauty



- Shorter trees are preferred over lager Trees
- Idea: want attributes that classifies examples well. The best attribute is selected.
- How well an attribute alone classifies the training data?
  - information theory

Information theory

- A branch of mathematics founded by Claude Shannon in the 1940s.
- What is it?
  - A method for quantifying the flow of information across tasks of varying complexity
- What is information?
  - The amount our uncertainty is reduced given new knowledge

Information Measurement

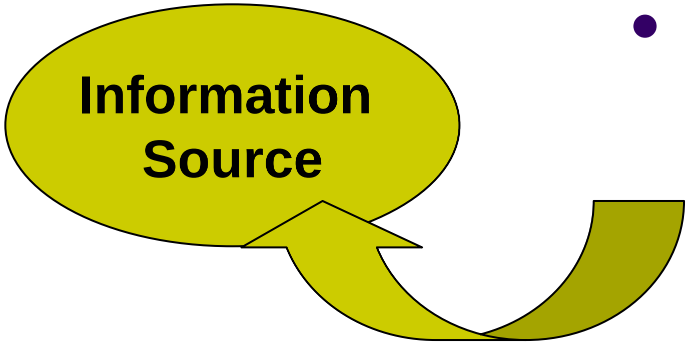
- Information Measurement
  - The amount of information about an event is closely related to its probability of occurrence.
- Units of information: *bits*

Messages containing knowledge of a low probability of occurrence convey relatively large amount of information.

$$P\downarrow I\uparrow$$

Messages containing knowledge of high probability of occurrence convey relatively little information.

$$P\uparrow I\downarrow$$



- Source alphabet of *n* symbols

  $\{S_1, S_2, S_3, \ldots, S_n\}$

Let the probability of producing be

$$P(S_i) = P_i \qquad \text{for} \qquad P_i \geq 0, \sum_i P_i = 1$$

Question

A. If a receiver receives the symbol S_i in a message, how much information is received?

B. If a receiver receives in a M - symbols message, how much information is received on average?

Question A

- The information of a single symbol S_i in a n symbols message
  - Case I: $n = 1$
  - Answer: S_1 is transmitted for sure. Therefore, no information. $I(1) = 0$
  - Case II: $n > 1$
  - Answer: Consider a symbol S_i ,then S_j the received information is $I(S_i S_j) = I(S_i) + I(S_j)$

So the amount of information or information content in the k^{th}symbols is $I(S_i) = -\log_2 P_i$

Question B

- The information is received on average Message
  - S_1 will occur, on average, $P_i N$ times for $N \to \infty$
  - Therefore, total information of the M-symbol message is

$$I_t = -\sum_{i=1}^{M} N P_i \log_2 P_i$$

  - The average information per symbol is $I_i / N = E$ and

$$E_t = -\sum_{i=1}^{M} P_i \log_2 P_i$$

Entropy

Entropy in Classification

- A collection S, containing positive and negative examples, the entropy to this boolean classification is

$$E(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\Theta} \log_2 p_{\Theta}$$

- Generally

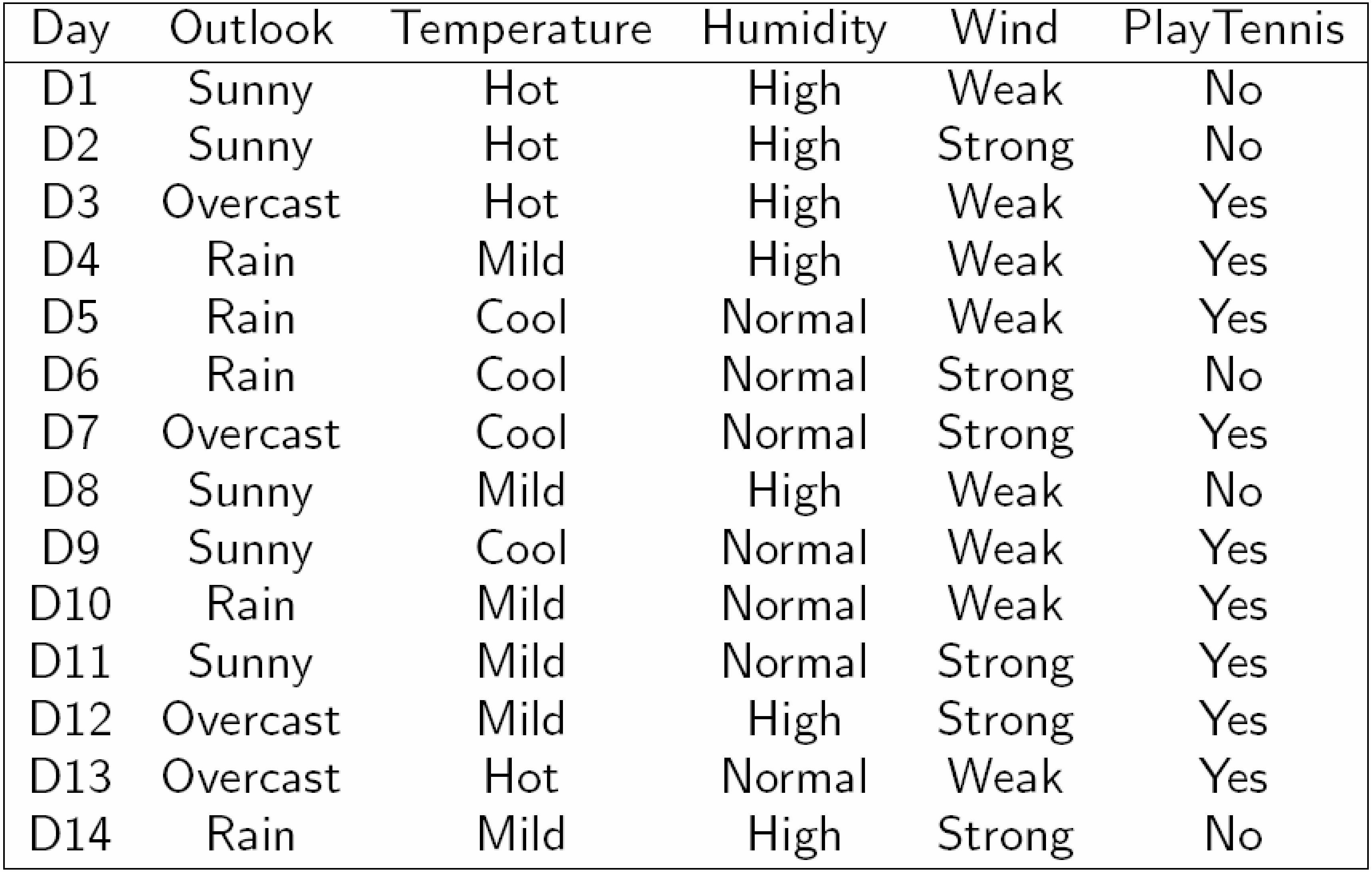
$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Information Gain

- What is the uncertainty removed by splitting on the value of A?
- The information gain of S relative to attribute A is the expected reduction in entropy caused by knowing the value of A
  - S_v: the set of examples in S where attribute A has value v

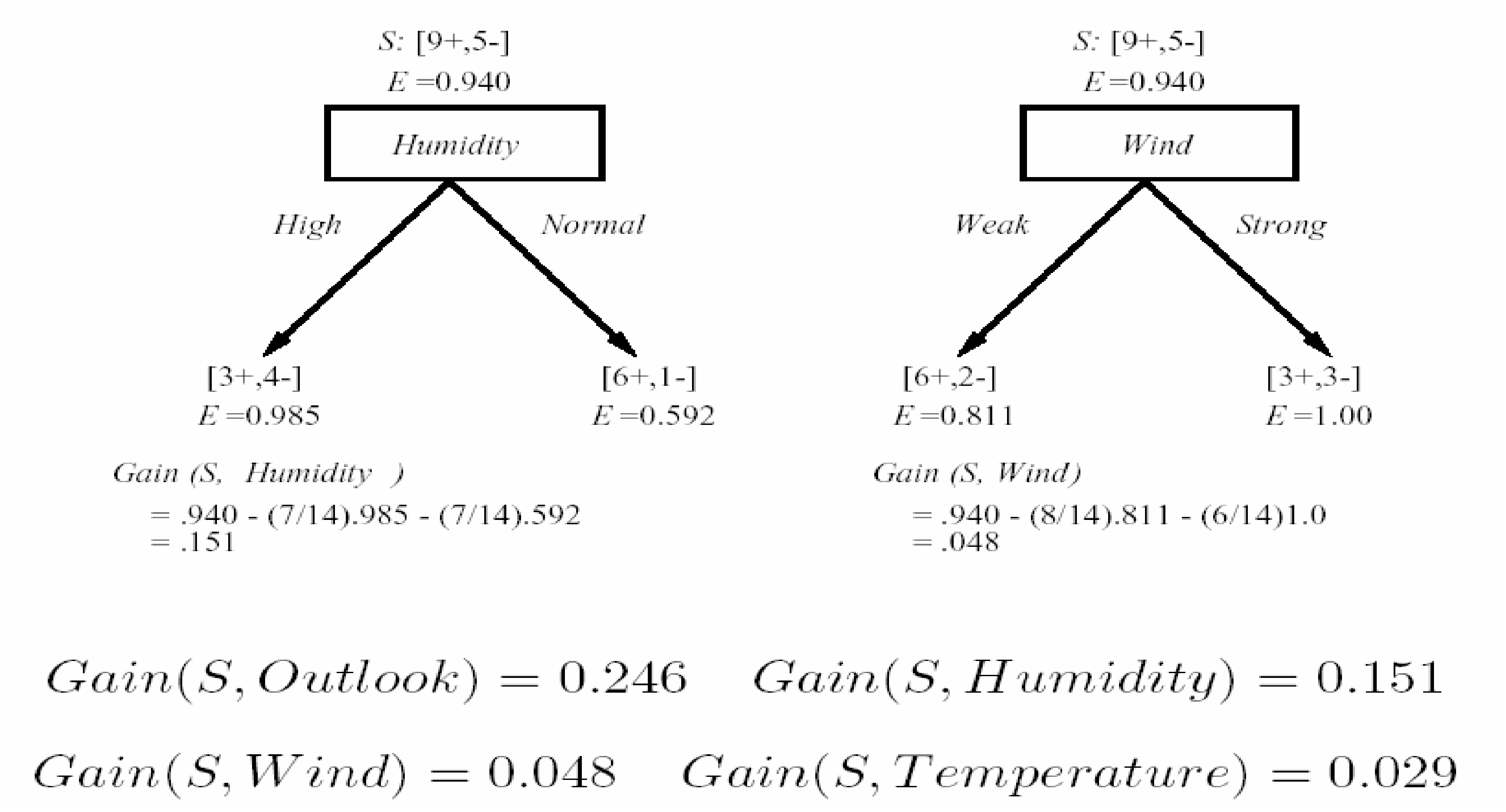
$$G(S,A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

PlayTennis



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

Which attribute is the best classifier?



Gain (S, Humidity)

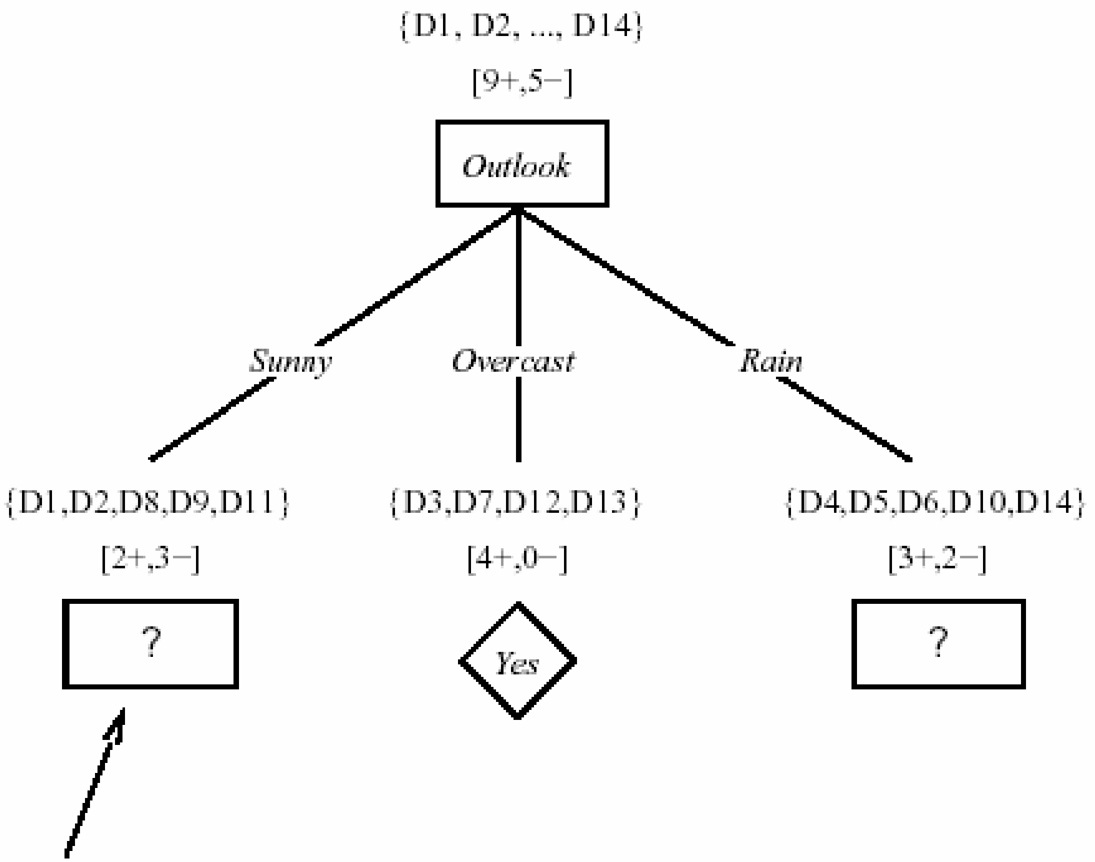
= .940 - (7/14).985 - (7/14).592
= .151

Gain (S, Wind)

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

$$Gain(S, Outlook) = 0.246 \quad Gain(S, Humidity) = 0.151$$

$$Gain(S, Wind) = 0.048 \quad Gain(S, Temperature) = 0.029$$



{D1, D2, ..., D14}

[9+,5−]

Outlook

Sunny *Overcast* *Rain*

{D1,D2,D8,D9,D11} [2+,3−] ?

{D3,D7,D12,D13} [4+,0−] *Yes*

{D4,D5,D6,D10,D14} [3+,2−] ?

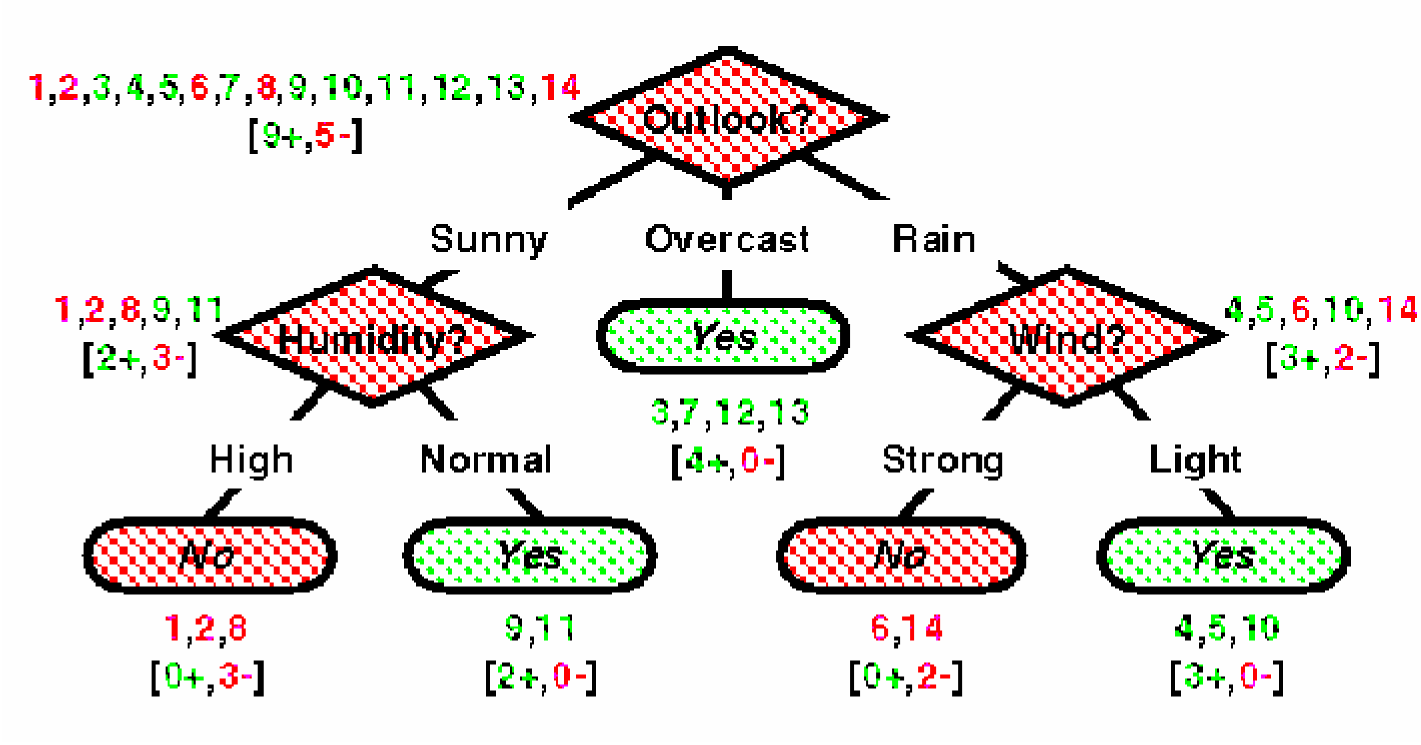
Which attribute should be tested here?

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

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

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

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



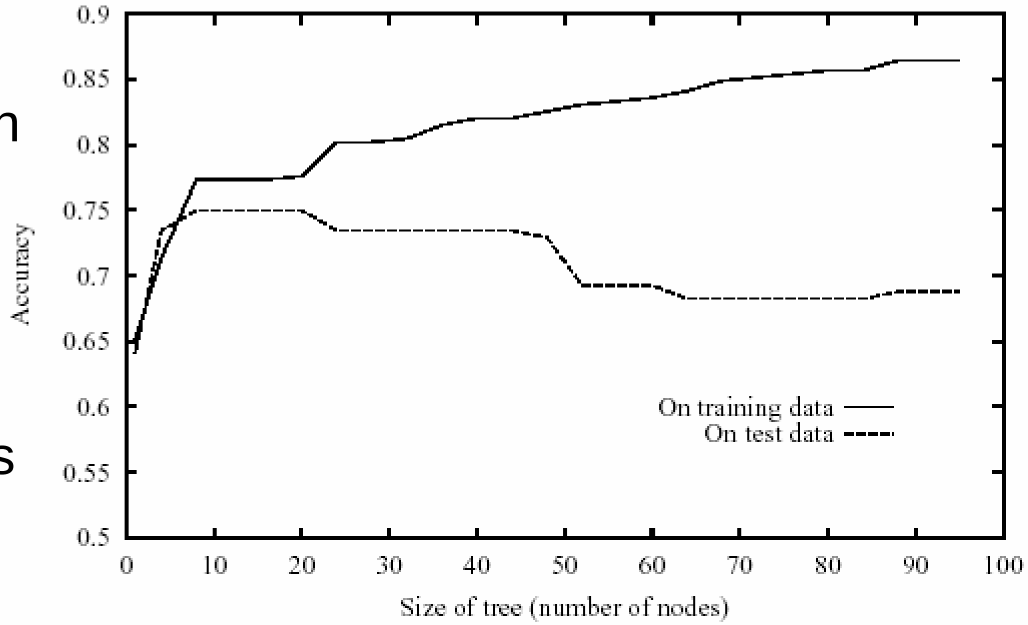
```
A1 = overcast: + (4.0)
A1 = sunny:
| A3 = high: - (3.0)
| A3 = normal: + (2.0)
A1 = rain:
| A4 = weak: + (3.0)
| A4 = strong: - (2.0)
```

See/C 5.0

Issues in Decision Tree

- Over-fitting
  - Hypothesis $h \in H$ **overfits** the training data if there is an alternative hypothesis $h' \in H$ such that

1. h has smaller error than h' over the training examples, but
2. h' has a smaller error than h over the entire distribution of instances

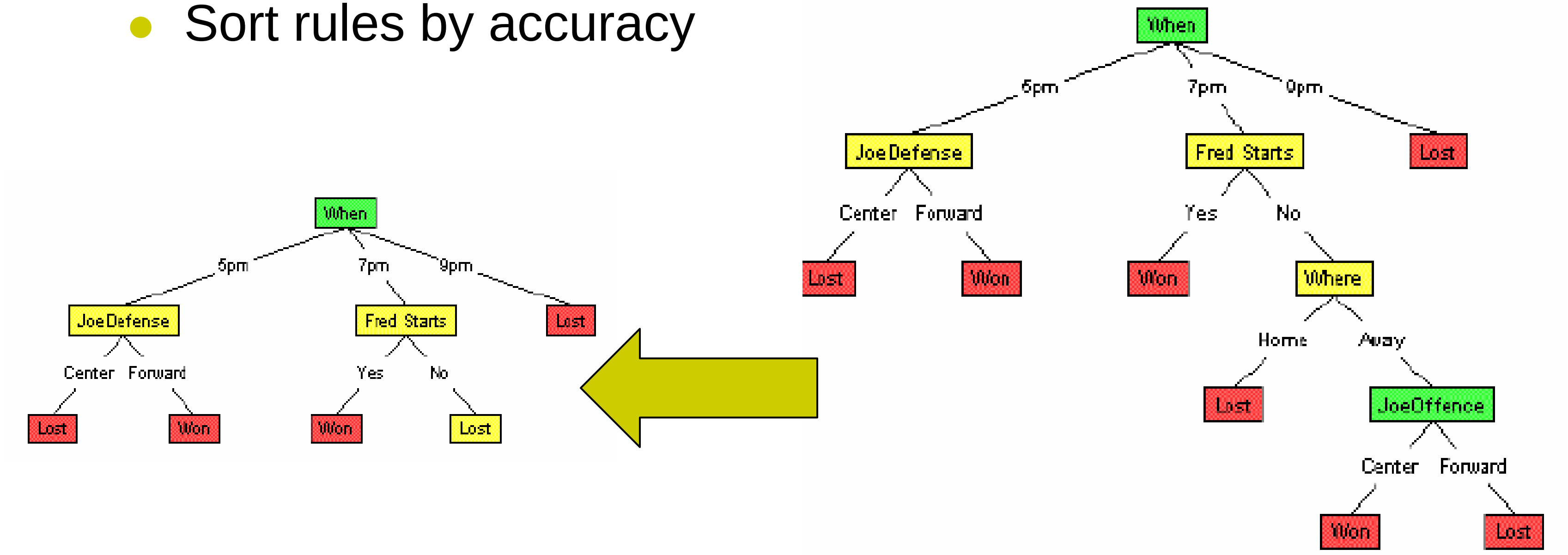


- Solution
  - Stop growing the tree earlier
    - Not successful in practice
  - Post-prune the tree
    - Reduced Error Pruning
    - Rule Post Pruning

- Implementation
  - Partition the available (training) data into two sets
    - Training set: used to form the learned hypothesis
    - Validation set : used to estimate the accuracy of this hypothesis over subsequent data

Tree Pruning

- Reduced Error Pruning
  - Nodes are removed if the resulting pruned tree performs no worse than the original over the validation set.
- Rule Post Pruning
  - Convert tree to set of rules.
  - Prune each rules by improving its estimated accuracy
  - Sort rules by accuracy



More considerations

- Continuous-Valued Attributes
  - Dynamically defining new discrete-valued attributes that partition the continuous attribute value into a discrete set of intervals.
- Alternative Measures for Selecting Attributes
  - Based on some measure other than information gain.
- Training Data with Missing Attribute Values
  - Assign a probability to the unknown attribute value.
- Handling Attributes with Differing Costs
  - Replacing the information gain measure by other measures

$$\frac{Gain^2(S,A)}{Cost(A)} \quad \text{or} \quad \frac{2^{Gain(S,A)}-1}{(Cost(A)+1)^w}$$

Acknowledgment

- Thanks to Mr. Wang Rui for writing the initial version of these slides.