# CS 5710-Machine Learning

Lecture 03: Decision Tree

# DT: binary classification

Training data
- 3 attributes
  - Outlook
  - Humidity
  - Wind
- Output
  - Play or not play
- Summary
  - 9 yes / 5 no

Training examples: **9 yes / 5 no**

| Day | Outlook | Humidity | Wind | Play |
|---|---|---|---|---|
| D1 | Sunny | High | Weak | No |
| D2 | Sunny | High | Strong | No |
| D3 | Overcast | High | Weak | Yes |
| D4 | Rain | High | Weak | Yes |
| D5 | Rain | Normal | Weak | Yes |
| D6 | Rain | Normal | Strong | No |
| D7 | Overcast | Normal | Strong | Yes |
| D8 | Sunny | High | Weak | No |
| D9 | Sunny | Normal | Weak | Yes |
| D10 | Rain | Normal | Weak | Yes |
| D11 | Sunny | Normal | Strong | Yes |
| D12 | Overcast | High | Strong | Yes |
| D13 | Overcast | Normal | Weak | Yes |
| D14 | Rain | High | Strong | No |

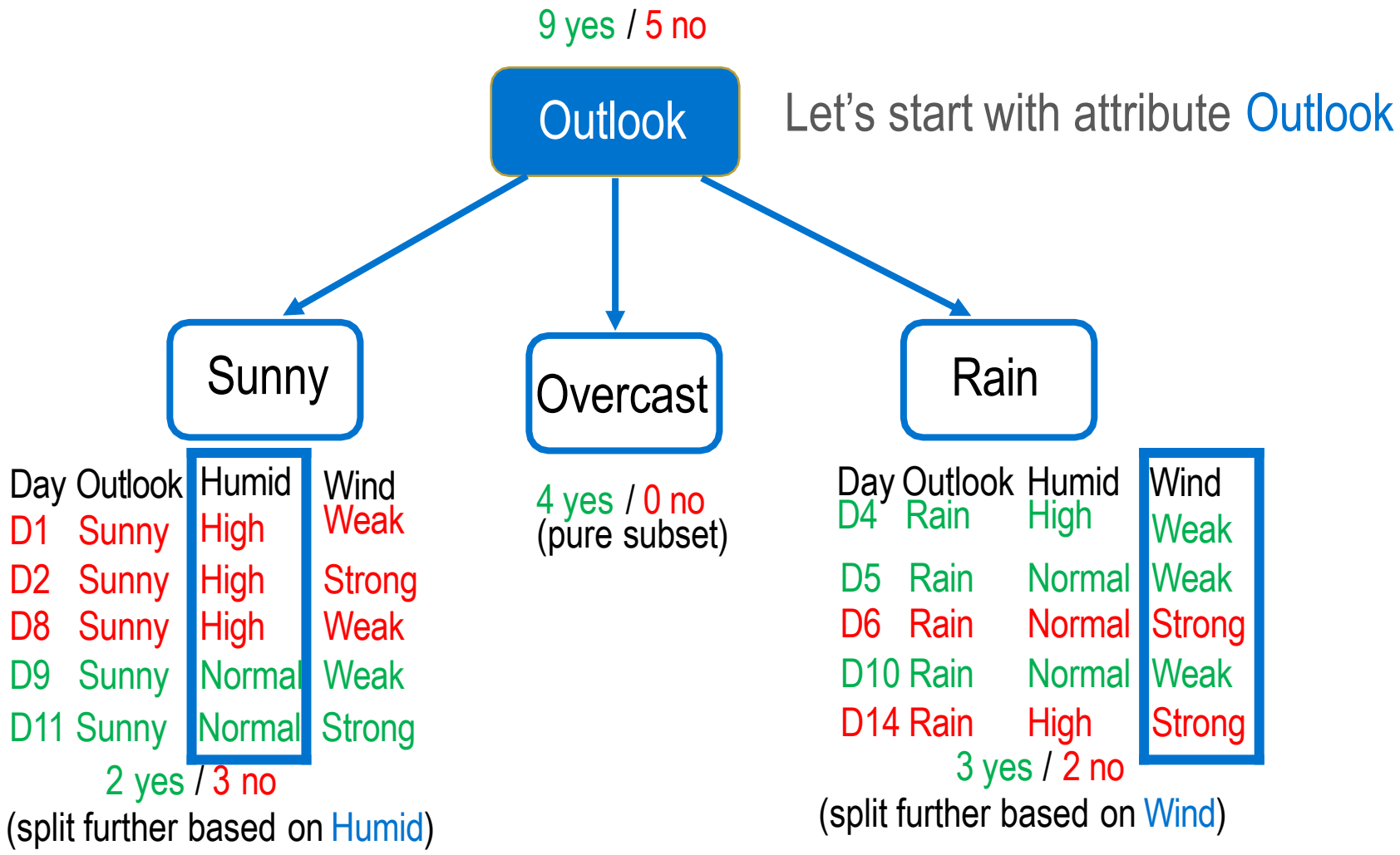
Test data D15 attributes: Rain High Weak

Question: play or not play? (Guess?)

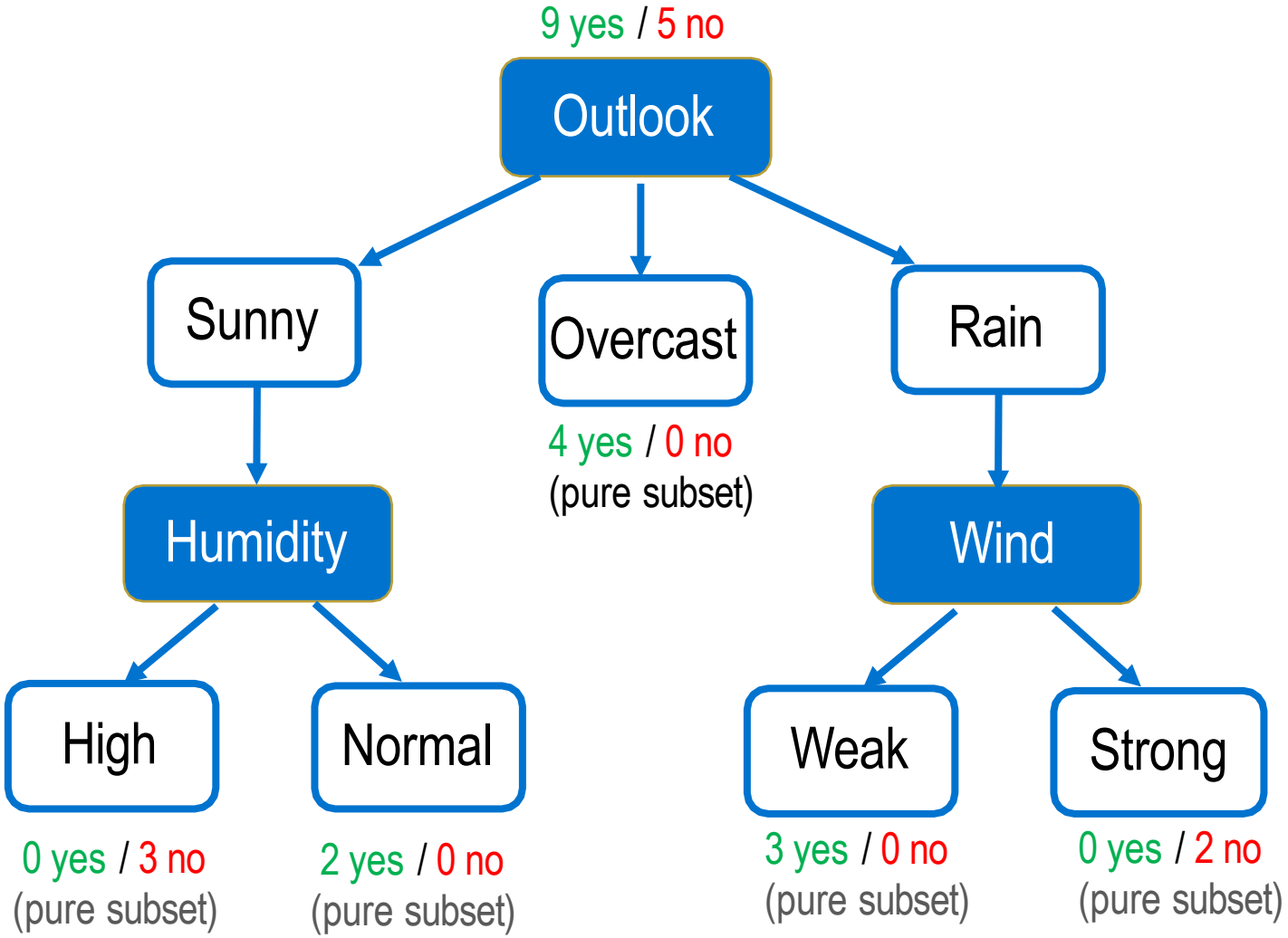
# Divide & conquer

- Split training data into subsets (divide)
  - Pick an attribute
  - Check whether subsets are pure (all yes or no)
  - If yes: stop
  - If no: split again
- Check which subset new test data falls into (conquer)

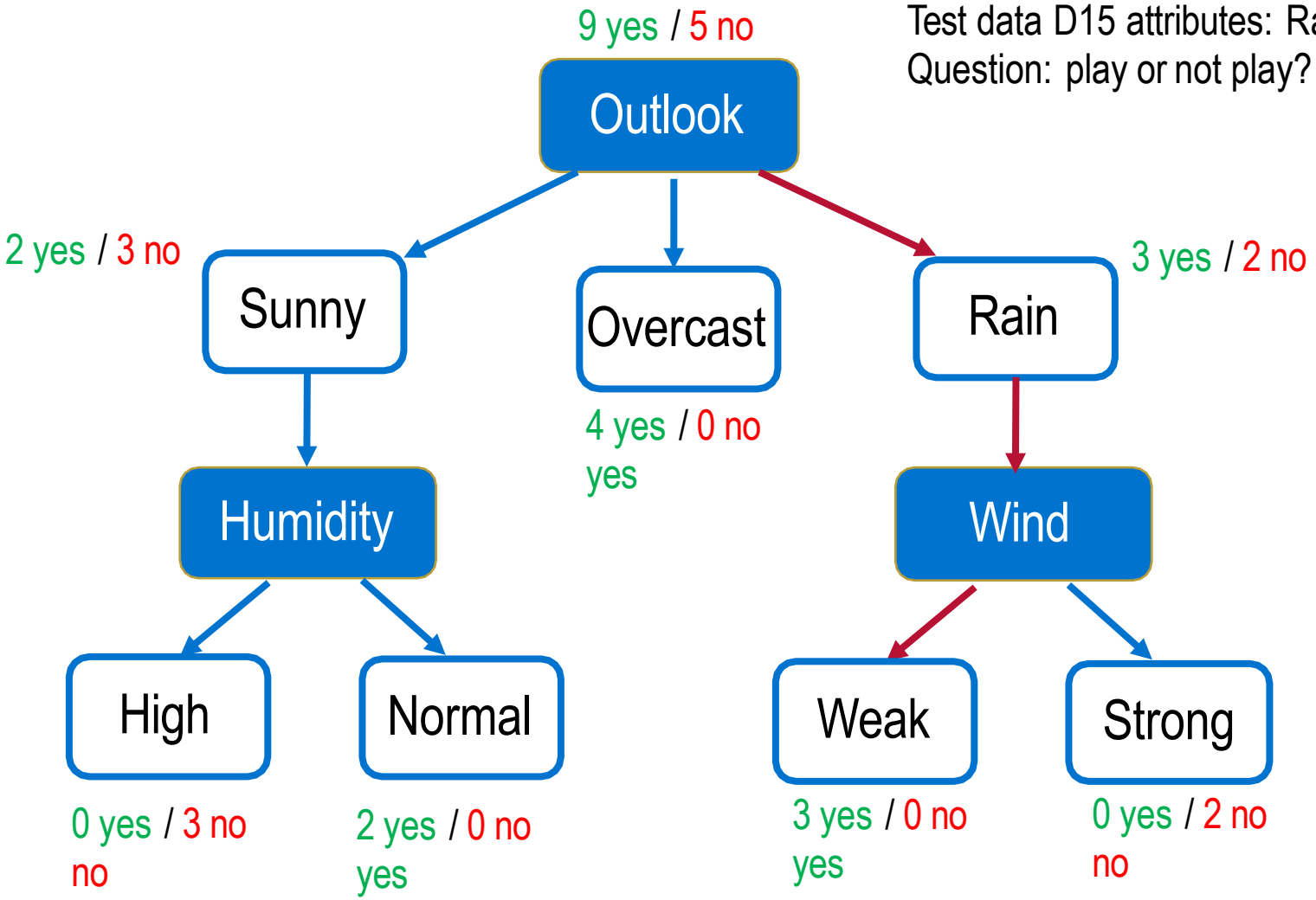
# Split training data (1)



# Split training data (2)



# Split training data (3)



# How to learn decision trees

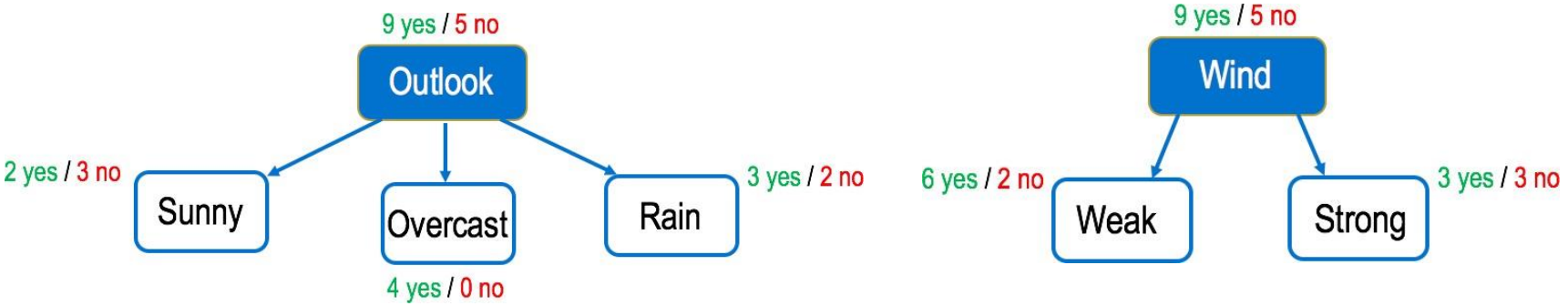
- Build an empty decision tree → split → recurse (choosing a good attribute for splitting is important)
- Some examples: ID3, C4.5, CART

# ID3(Iterative Dichotomiser 3)algorithm

ID3 (node, {training data}) # Generate a DT

1. Pick an attribute (A) with the maximum information gain for the considered training data
2. For each value of A, create new child node
3. Split training data to child nodes
4. Check subset for each child node
   - If subset is pure: stop
   - Else: ID3 (child node, {subset data})

# How to select an attribute?




Start from the original training data set.
Consider attributes Outlook and Wind. Which one is better?

We hope that uncertainty can be reduced after the split

- 4 yes / 0 no: 100 % certain about yes
- 3 yes / 3 no: 50% certain about yes

How to measure reduced uncertainty by using an attribute?

# Entropy (2)

Consider a training set S with binary labels

$$H(S) = - p_{(+)} \log_2 p_{(+)} - p_{(-)} \log_2 p_{(-)}$$

- $p_{(+)}$ : (# of yes ) / (# of total training examples)
- $p_{(.)}$ : (# of no) / (# of total training examples)
- Interpretation: assume a training example $x \in S$.
  How many bits are needed to tell if x has a label yes or no.

Entropy example

- 4 yes / 0 no: $H(S) = -1 \log_2 1 - 0 = 0$ bits
- 3 yes / 3 no: $H(S) = -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{2}\log_2 \frac{1}{2} = 1$ bits

# Information gain (1)

## Expected drop in entropy after split

$$\text{Gain}(S, A) = H(S) - \sum_{V \in \text{Values}(A)} \frac{|S_V|}{|S|} H(S_V)$$

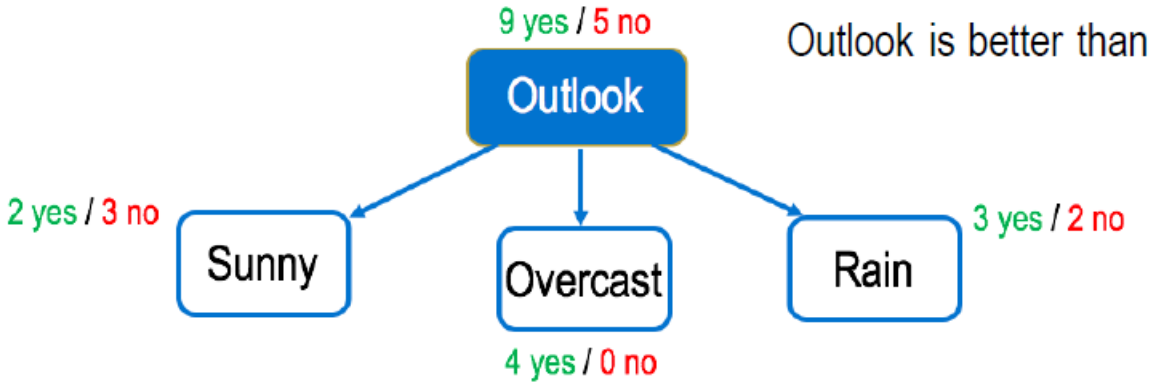
uncertainty before split

uncertainty after split

- A: attribute
- S: set of training examples
- V: possible values of attribute A
- $S_V$: set of training examples with the value of attribute A = V
- Subsets with more examples have a larger effect

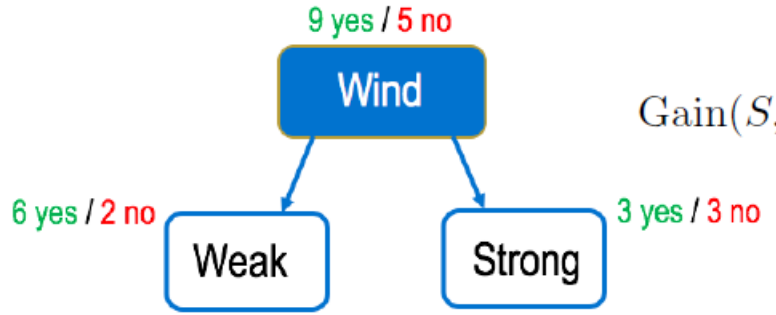
Maximizing Gain(S, A) is equivalent to minimizing uncertainty after split

# Information gain (2)




Outlook is better than Wind based on information gain

$$\text{Gain}(S, \text{Outlook}) = H(S) - \frac{5}{14}H(S_{\text{sunny}}) - \frac{4}{14}H(S_{\text{overcast}}) - \frac{5}{14}H(S_{\text{rain}})$$
$$= 0.94 - \frac{5}{14} * 0.97 - \frac{4}{14} * 0 - \frac{5}{14} * 0.97 = 0.25$$

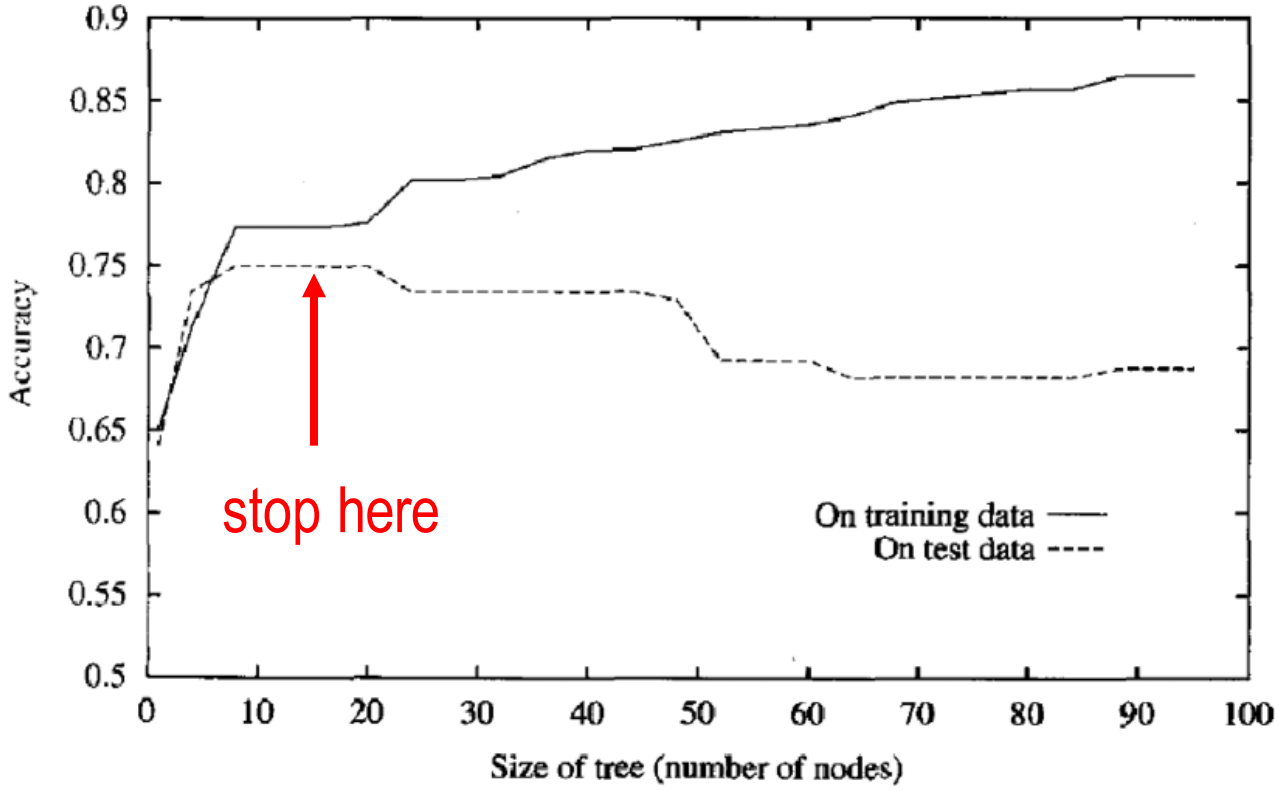



$$\text{Gain}(S, \text{Wind}) = H(S) - \frac{8}{14}H(S_{\text{weak}}) - \frac{6}{14}H(S_{\text{strong}})$$
$$= 0.94 - \frac{8}{14} * 0.81 - \frac{6}{14} * 1 = 0.049$$

# Properties on ID3

- Non-robust
  - A small change in the training data can result in a big change in the tree, and thus a big change in final predictions
- Overfitting
  - If there is no label noise, training set error is always 0!
  - Need strategies for a simpler tree, e.g., fixed depth, pruning
- Harder to use for continuous data

# Overfitting




It may be unnecessary to grow a complete tree

# Tree pruning

- Pre-pruning (forward / online-pruning)
  - Use a termination condition to decide when to stop splitting, e.g., fixed depth
  - Strict conditions result in oversimplified trees, whereas lose conditions result in little simplification
- Post-pruning (backward pruning)
  - First generate DT then remove non-significant branches
  - Trim nodes in a bottom-up fashion (use validation set)
  - After trimming, replace subtree with a leaf node
  - Class label of leaf node is determined by majority

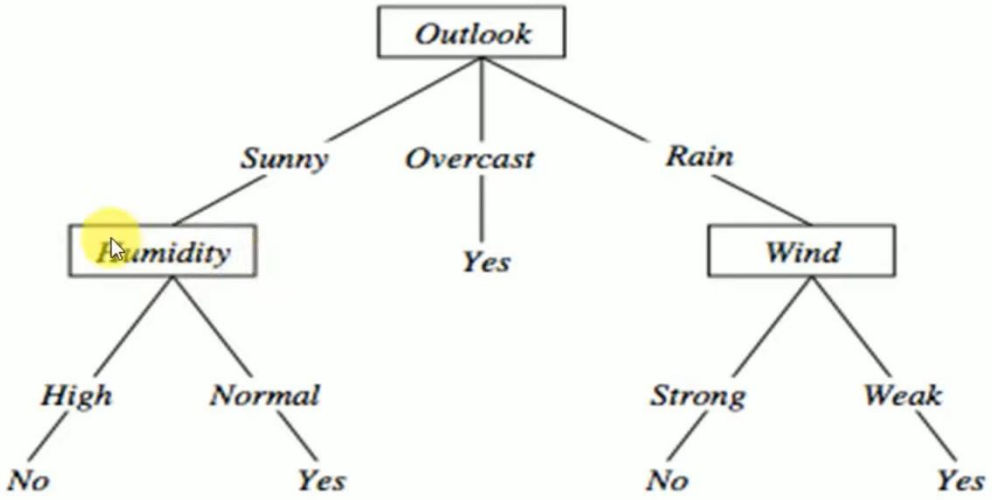
# Avoiding overfitting in Decision Tree

## Reduced-error pruning

1. Each node is a candidate for pruning
2. *Pruning* consists in removing a subtree rooted in a node: the node becomes a leaf and is assigned the most common classification
3. Nodes are removed only if the resulting tree performs better on the **validation set.**
4. Nodes are pruned iteratively: at each iteration the node whose removal most increases accuracy on the validation set is pruned.

# Avoiding overfitting in Decision Tree

**Reduced-error pruning**



# Avoiding overfitting in Decision Tree

**Rule post-pruning**

1. Create the decision tree from the training set
2. Convert the tree into an equivalent set of rules
   - Each path corresponds to a rule
   - Each node along a path corresponds to a pre-condition
   - Each leaf classification to the post-condition

# Avoiding overfitting in Decision Tree

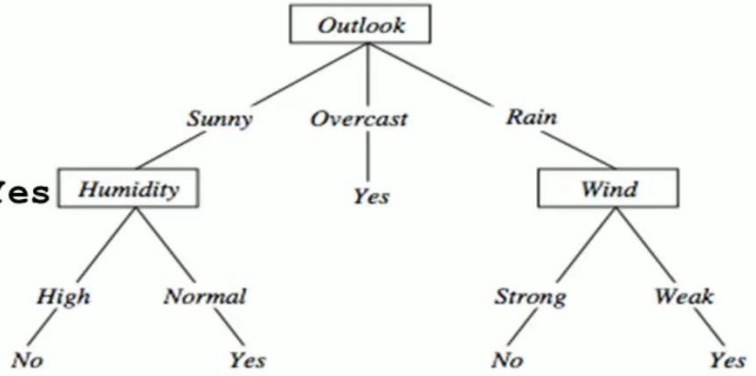
**Rule post-pruning**

1. Create the decision tree from the training set
2. Convert the tree into an equivalent set of rules
   - Each path corresponds to a rule
   - Each node along a path corresponds to a pre-condition
   - Each leaf classification to the post-condition
3. Prune (generalize) each rule by removing those preconditions whose removal improves accuracy over validation set
4. Sort the rules in estimated order of accuracy, and consider them in sequence when classifying new instances

# Avoiding overfitting in Decision Tree

## Rule post-pruning

1. Outlook=sunny ^ humidity=high -> No
2. Outlook=sunny ^ humidity=normal -> Yes
3. Outlook=overcast -> Yes
4. Outlook=rain ^ wind=strong -> No
5. Outlook=rain ^ wind=weak -> Yes




Compare first rule to:

**Outlook=sunny->No**

**Humidity=high->No**

Calculate accuracy of 3 rules based on validation set and pick best version.

# Avoiding overfitting in Decision Tree

## Rule post-pruning

### Why converting to rules?

1. Each distinct path produces a different rule: a condition removal may be based on a local (contextual) criterion. Node pruning is global and affects all the rules
2. In rule form, tests are not ordered and there is no book-keeping involved when conditions (nodes) are removed
3. Converting to rules improves readability for humans

# Continuous attributes (1)

Example: predict whether customers will buy a product

- Attributes: gender, income, and age (the last two are continuous)
- Continuous attributes
  - Infinite number of possible split values

# Continuous attributes (2)

## Threshold split on attribute A

- One branch: A < t (e.g., age < 40)
- The other branch: A >= t (e.g., age >= 40)

## How to find t?

- Find several candidates t
  - Sort values of A in ascending order $\{a_1, a_2, \dots, a_m\}$
  - Candidates: $a_i + \frac{a_{i+1} - a_i}{2}, i = 1, 2, \dots, m-1$
- Choose the best t (how?)

# Continuous attributes (3)

Choose the best t with the maximum information gain (or equivalently, with the minimum uncertainty after split)

$$t^* = \arg\min_{t \in T} \underbrace{\frac{|S_{A<t}|}{|S|} H(S_{A<t}) + \frac{|S_{A \geq t}|}{|S|} H(S_{A \geq t})}_{\text{uncertainty after split}}$$

- $T$: set of candidates t
- $S_{A<t}$: set of training examples with value of attribute A < t
- $S_{A \geq t}$: set of training examples with value of attribute A >= t

# Multiclass classification

- Prediction: use most frequent class label in the subset
- Entropy $H(S) = -\sum_{c} p_{(c)} \log_2 p_{(c)}$

$p_{(c)}$ : (# of class c ) / (# of total training examples)

| Day | Outlook | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 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 |

## Attribute: Outlook

$$Values\,(Outlook) = Sunny, Overcast, Rain$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{Sunny} \leftarrow [2+, 3-] \qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.971$$

$$S_{Overcast} \leftarrow [4+, 0-] \qquad Entropy(S_{Overcast}) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$$

$$S_{Rain} \leftarrow [3+, 2-] \qquad Entropy(S_{Rain}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.971$$

$$Gain\,(S, Outlook) = Entropy(S) - \sum_{v \in \{Sunny, Overcast, Rain\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Outlook) = Entropy(S) - \frac{5}{14}Entropy(S_{Sunny}) - \frac{4}{14}Entropy(S_{Overcast}) - \frac{5}{14}Entropy(S_{Rain})$$

$$Gain(S, Outlook) = 0.94 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971 = 0.2464$$

| Day | Outlook | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 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 |

## Attribute: Temp

$$Values\ (Temp) = Hot, Mild, Cool$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$$

$$S_{Hot} \leftarrow [2+, 2-] \qquad Entropy(S_{Hot}) = -\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4} = 1.0$$

$$S_{Mild} \leftarrow [4+, 2-] \qquad Entropy(S_{Mild}) = -\frac{4}{6} log_2 \frac{4}{6} - \frac{2}{6} log_2 \frac{2}{6} = 0.9183$$

$$S_{Cool} \leftarrow [3+, 1-] \qquad Entropy(S_{Cool}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.8113$$

$$Gain\ (S, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp)$$

$$= Entropy(S) - \frac{4}{14} Entropy(S_{Hot}) - \frac{6}{14} Entropy(S_{Mild})$$

$$- \frac{4}{14} Entropy(S_{Cool})$$

$$Gain(S, Temp) = 0.94 - \frac{4}{14} 1.0 - \frac{6}{14} 0.9183 - \frac{4}{14} 0.8113 = 0.0289$$

| Day | Outlook | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 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 |

## Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94$$

$$S_{High} \leftarrow [3+, 4-] \qquad Entropy(S_{High}) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.9852$$

$$S_{Normal} \leftarrow [6+, 1-] \qquad Entropy(S_{Normal}) = -\frac{6}{7}log_2\frac{6}{7} - \frac{1}{7}log_2\frac{1}{7} = 0.5916$$

$$Gain\,(S, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Humidity)$$

$$= Entropy(S) - \frac{7}{14}Entropy(S_{High}) - \frac{7}{14}Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14}0.9852 - \frac{7}{14}\,0.5916 = 0.1516$$

| Day | Outlook | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|---|
| 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 |

## Attribute: Wind

$$Values\,(Wind) = Strong, Weak$$

$$S = [9+, 5-] \qquad Entropy(S) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$$

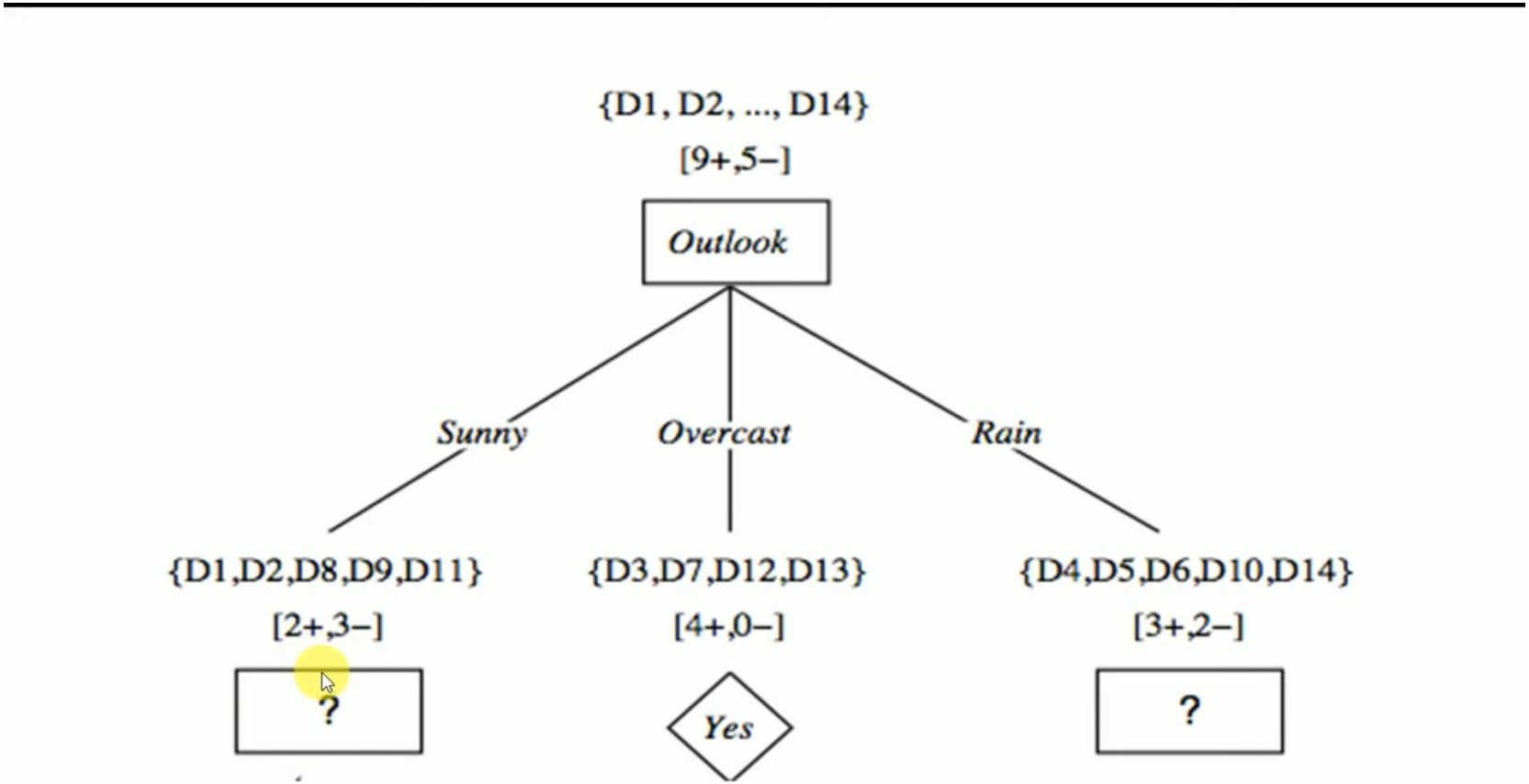
$$S_{Strong} \leftarrow [3+, 3-] \qquad Entropy(S_{Strong}) = 1.0$$

$$S_{Weak} \leftarrow [6+, 2-] \qquad Entropy(S_{Weak}) = -\frac{6}{8} log_2 \frac{6}{8} - \frac{2}{8} log_2 \frac{2}{8} = 0.8113$$

$$Gain\,(S, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{Strong}) - \frac{8}{14} Entropy(S_{Weak})$$

$$Gain(S, Wind) = 0.94 - \frac{6}{14}\ 1.0 - \frac{8}{14}\ 0.8113 = 0.0478$$


{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−]
?

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D1 | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| D11 | Mild | Normal | Strong | Yes |

## Attribute: Temp

$$Values\,(Temp) = Hot, Mild, Cool$$

$$S_{Sunny} = [2+, 3-] \qquad Entropy(S_{Sunny}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$$

$$S_{Hot} \leftarrow [0+, 2-] \qquad Entropy(S_{Hot}) = 0.0$$

$$S_{Mild} \leftarrow [1+, 1-] \qquad Entropy(S_{Mild}) = 1.0$$

$$S_{Cool} \leftarrow [1+, 0-] \qquad Entropy(S_{Cool}) = 0.0$$

$$Gain\,(S_{Sunny}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Temp)$$

$$= Entropy(S) - \frac{2}{5}Entropy(S_{Hot}) - \frac{2}{5}Entropy(S_{Mild})$$

$$- \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S_{sunny}, Temp) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = 0.570$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| Dl | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| Dl1 | Mild | Normal | Strong | Yes |

## Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S_{Sunny} = [2+, 3-] \qquad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$$

$$S_{high} \leftarrow [0+, 3-] \qquad Entropy(S_{High}) = 0.0$$

$$S_{Normal} \leftarrow [2+, 0-] \qquad Entropy(S_{Normal}) = 0.0$$

$$Gain\,(S_{Sunny}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Humidity) = Entropy(S) - \frac{3}{5}Entropy(S_{High}) - \frac{2}{5}Entropy(S_{Normal})$$

$$Gain(S_{sunny}, Humidity) = 0.97 - \frac{3}{5}\,0.0 - \frac{2}{5}0.0 = 0.97$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| Dl | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| Dl1 | Mild | Normal | Strong | Yes |

## Attribute: Wind

$Values\ (Wind) = Strong, Weak$

$S_{Sunny} = [2+, 3-]$ $\quad Entropy(S) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.97$

$S_{Strong} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Strong}) = 1.0$

$S_{Weak} \leftarrow [1+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.9183$

$$Gain\ (S_{Sunny}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Sunny}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

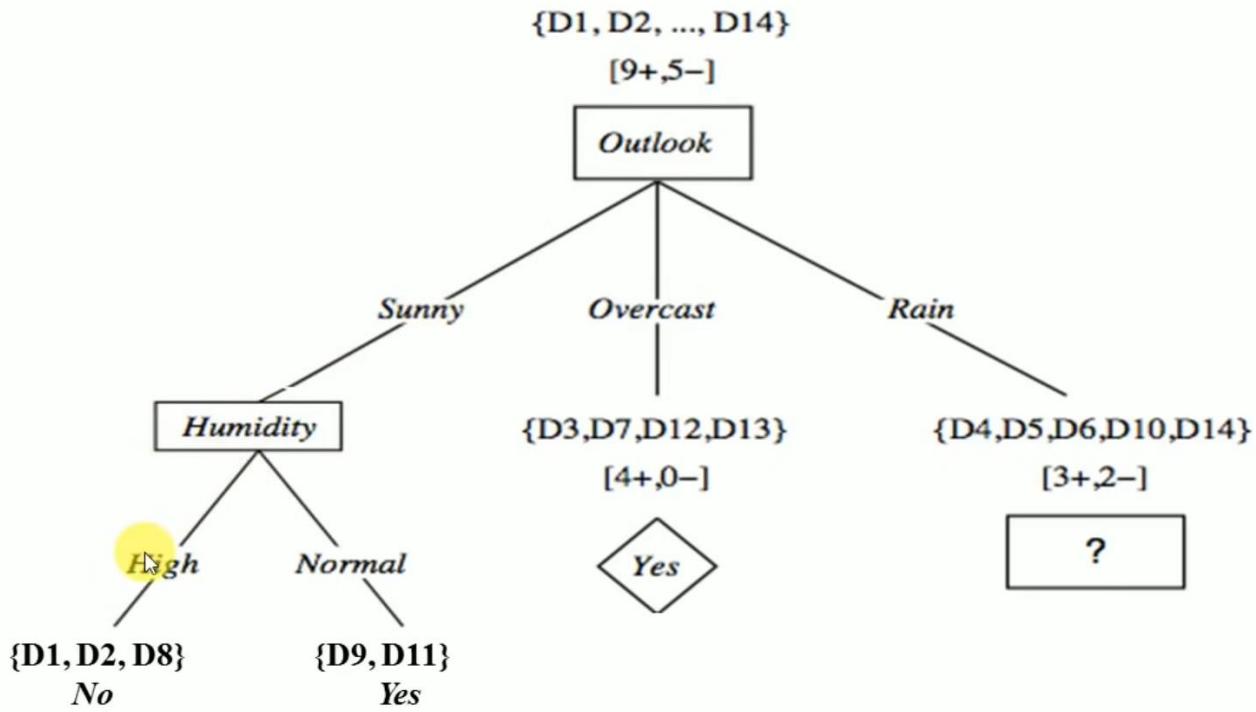
$$Gain(S_{sunny}, Wind) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}\,0.918 = 0.0192$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D1 | Hot | High | Weak | No |
| D2 | Hot | High | Strong | No |
| D8 | Mild | High | Weak | No |
| D9 | Cool | Normal | Weak | Yes |
| D11 | Mild | Normal | Strong | Yes |

$$Gain(S_{sunny}, Temp) = 0.570$$

$$Gain(S_{sunny}, Humidity) = 0.97$$

$$Gain(S_{sunny}, Wind) = 0.0192$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
{D4,D5,D6,D10,D14}
[3+,2−]
High
Normal
Yes
?
{D1, D2, D8}
No
{D9, D11}
Yes

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D4 | Mild | High | Weak | Yes |
| D5 | Cool | Normal | Weak | Yes |
| D6 | Cool | Normal | Strong | No |
| D10 | Mild | Normal | Weak | Yes |
| D14 | Mild | High | Strong | No |

## Attribute: Temp

$$Values\,(Temp) = Hot, Mild, Cool$$

$S_{Rain} = [3+, 2-]$ $\quad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$

$S_{Hot} \leftarrow [0+, 0-]$ $\quad Entropy(S_{Hot}) = 0.0$

$S_{Mild} \leftarrow [2+, 1-]$ $\quad Entropy(S_{Mild}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$

$S_{Cool} \leftarrow [1+, 1-]$ $\quad Entropy(S_{Cool}) = 1.0$

$$Gain\,(S_{Rain}, Temp) = Entropy(S) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Temp)$$

$$= Entropy(S) - \frac{0}{5}Entropy(S_{Hot}) - \frac{3}{5}Entropy(S_{Mild})$$

$$- \frac{2}{5}Entropy(S_{Cool})$$

$$Gain(S_{Rain}, Temp) = 0.97 - \frac{0}{5}0.0 - \frac{3}{5}\,0.918 - \frac{2}{5}\,1.0 = 0.0192$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D4 | Mild | High | Weak | Yes |
| D5 | Cool | Normal | Weak | Yes |
| D6 | Cool | Normal | Strong | No |
| Dl0 | Mild | Normal | Weak | Yes |
| Dl4 | Mild | High | Strong | No |

## Attribute: Humidity

$$Values\,(Humidity) = High, Normal$$

$$S_{Rain} = [3+, 2-] \qquad Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

$$S_{High} \leftarrow [1+, 1-] \qquad Entropy(S_{High}) = 1.0$$

$$S_{Normal} \leftarrow [2+, 1-] \qquad Entropy(S_{Normal}) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.9183$$

$$Gain\,(S_{Rain}, Humidity) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Humidity) = Entropy(S) - \frac{2}{5}Entropy(S_{High}) - \frac{3}{5}Entropy(S_{Normal})$$

$$Gain(S_{Rain}, Humidity) = 0.97 - \frac{2}{5}\,1.0 - \frac{3}{5}0.918 = 0.0192$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D4 | Mild | High | Weak | Yes |
| D5 | Cool | Normal | Weak | Yes |
| D6 | Cool | Normal | Strong | No |
| Dl0 | Mild | Normal | Weak | Yes |
| Dl4 | Mild | High | Strong | No |

## Attribute: Wind

$$Values\ (wind) = Strong, Weak$$

$S_{Rain} = [3+, 2-]$

$S_{Strong} \leftarrow [0+, 2-]$

$S_{Weak} \leftarrow [3+, 0-]$

$$Entropy(S_{Sunny}) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

$$Entropy(S_{Strong}) = 0.0$$

$$Entropy(S_{weak}) = 0.0$$

$$Gain\ (S_{Rain}, Wind) = Entropy(S) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S_{Rain}, Wind) = Entropy(S) - \frac{2}{5}Entropy(S_{Strong}) - \frac{3}{5}Entropy(S_{Weak})$$

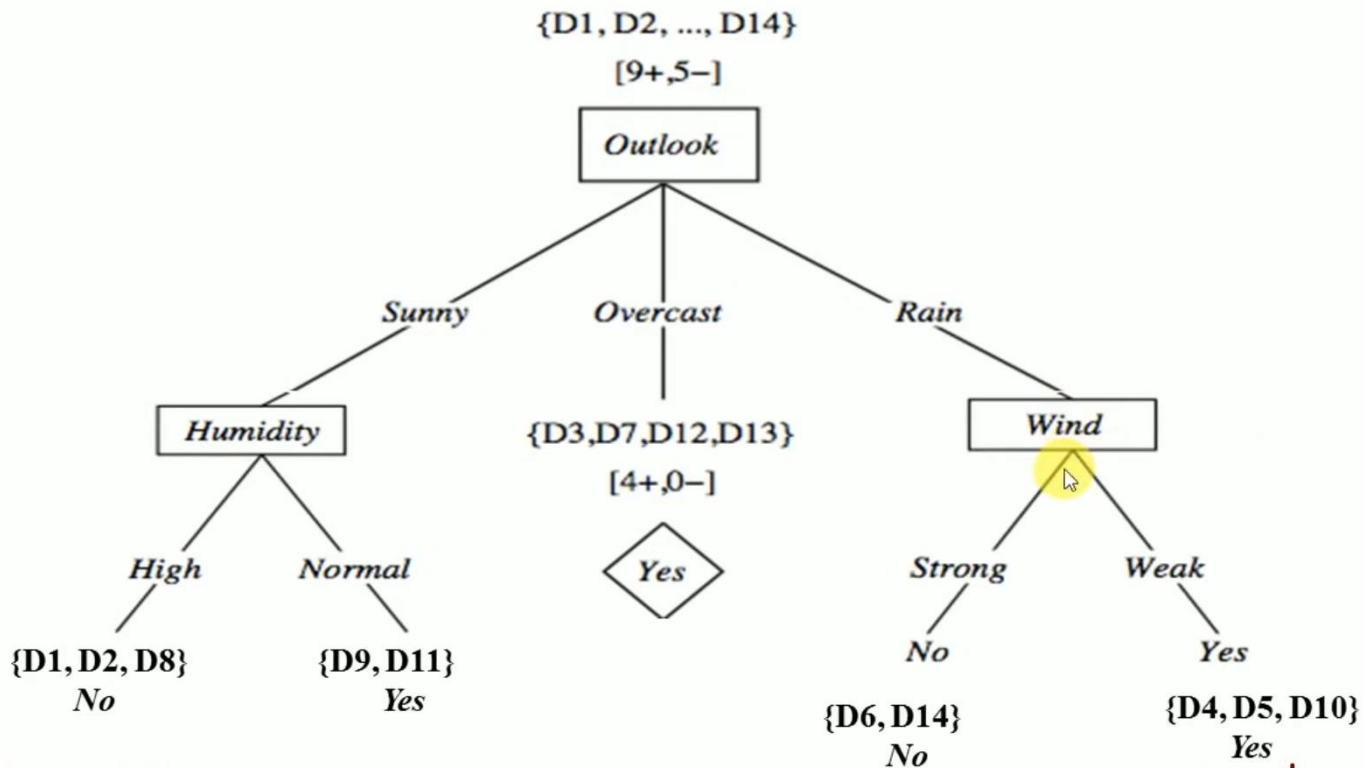
$$Gain(S_{Rain}, Wind) = 0.97 - \frac{2}{5}\ 0.0 - \frac{3}{5}0.0 = 0.97$$

| Day | Temp | Humidity | Wind | Play Tennis |
|---|---|---|---|---|
| D4 | Mild | High | Weak | Yes |
| D5 | Cool | Normal | Weak | Yes |
| D6 | Cool | Normal | Strong | No |
| Dl0 | Mild | Normal | Weak | Yes |
| Dl4 | Mild | High | Strong | No |

$$Gain(S_{Rain}, Temp) = 0.0192$$

$$Gain(S_{Rain}, Humidity) = 0.0192$$

$$Gain(S_{Rain}, Wind) = 0.97$$


{D1, D2, ..., D14}
[9+,5−]
Outlook
Sunny
Overcast
Rain
Humidity
{D3,D7,D12,D13}
[4+,0−]
Wind
High
Normal
Yes
Strong
Weak
{D1, D2, D8}
No
{D9, D11}
Yes
No
Yes
{D6, D14}
No
{D4, D5, D10}
Yes

| Instance | a1 | a2 | a3 | Classification |
|---|---|---|---|---|
| 1 | True | Hot | High | No |
| 2 | True | Hot | High | No |
| 3 | False | Hot | High | Yes |
| 4 | False | Cool | Normal | Yes |
| 5 | False | Cool | Normal | Yes |
| 6 | True | Cool | High | No |
| 7 | True | Hot | High | No |
| 8 | True | Hot | Normal | Yes |
| 9 | False | Cool | Normal | Yes |
| 10 | False | Cool | High | Yes |

**Attribute: a1**

$$Values(a1) = True,\ False$$

$$S = [6+,\ 4-] \qquad Entropy(S) = -\frac{6}{10}log_2\frac{6}{10} - \frac{4}{10}log_2\frac{4}{10} = 0.9709$$

$$S_{True} = [1+, 4-] \qquad Entropy(S_{True}) = -\frac{1}{5}log_2\frac{1}{5} - \frac{4}{5}log_2\frac{4}{5} = 0.7219$$

$$S_{Flase} \leftarrow [5+,\ 0-] \qquad Entropy(S_{False}) = 0.0$$

$$Gain\ (S, a1) = Entropy(S) - \sum_{v \in \{True, False\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a1) = Entropy(S) - \frac{5}{10}Entropy(S_{True}) - \frac{5}{10}Entropy(S_{False})$$

$$Gain(S, a1) = 0.9709 - \frac{5}{10} * 0.7219 - \frac{5}{10} * 1 = 0.6099$$

| Instance | a1 | a2 | a3 | Classification |
|---|---|---|---|---|
| 1 | True | Hot | High | No |
| 2 | True | Hot | High | No |
| 3 | False | Hot | High | Yes |
| 4 | False | Cool | Normal | Yes |
| 5 | False | Cool | Normal | Yes |
| 6 | True | Cool | High | No |
| 7 | True | Hot | High | No |
| 8 | True | Hot | Normal | Yes |
| 9 | False | Cool | Normal | Yes |
| 10 | False | Cool | High | Yes |

**Attribute: a2**

$$Values(a2) = Hot,\ Cool$$

$$S = [6+,\ 4-] \qquad Entropy(S) = -\frac{6}{10}log_2\frac{6}{10} - \frac{4}{10}log_2\frac{4}{10} = 0.9709$$

$$S_{Hot} = [2+, 3-] \qquad Entropy(S_{Hot}) = -\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5} = 0.9709$$

$$S_{Cool} \leftarrow [4+,\ 1-] \qquad Entropy(S_{Cool}) = -\frac{4}{5}log_2\frac{4}{5} - \frac{1}{5}log_2\frac{1}{5} = 0.7219$$

$$Gain\ (S, a2) = Entropy(S) - \sum_{v \in \{Hot, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a2) = Entropy(S) - \frac{5}{10}Entropy(S_{Hot}) - \frac{5}{10}Entropy(S_{Cool})$$

$$Gain(S, a2) = 0.9709 - \frac{5}{10} * 0.9709 - \frac{5}{10} * 0.7219 = 0.1245$$

| Instance | a1 | a2 | a3 | Classification |
|---|---|---|---|---|
| 1 | True | Hot | High | No |
| 2 | True | Hot | High | No |
| 3 | False | Hot | High | Yes |
| 4 | False | Cool | Normal | Yes |
| 5 | False | Cool | Normal | Yes |
| 6 | True | Cool | High | No |
| 7 | True | Hot | High | No |
| 8 | True | Hot | Normal | Yes |
| 9 | False | Cool | Normal | Yes |
| 10 | False | Cool | High | Yes |

**Attribute: a3**

$$Values\,(a3) = High,\ Normal$$

$$S = [6+,\ 4-] \qquad Entropy(S) = -\frac{6}{10}log_2\frac{6}{10} - \frac{4}{10}log_2\frac{4}{10} = 0.9709$$

$$S_{High} = [2+, 4-] \qquad Entropy(S_{High}) = -\frac{2}{6}log_2\frac{2}{6} - \frac{4}{6}log_2\frac{4}{6} = 0.9183$$

$$S_{Normal} \leftarrow [4+,\ 0-] \qquad Entropy(S_{Normal}) = 0.0$$

$$Gain\,(S, a3) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a3) = Entropy(S) - \frac{6}{10}Entropy(S_{High}) - \frac{4}{10}Entropy(S_{Normal})$$

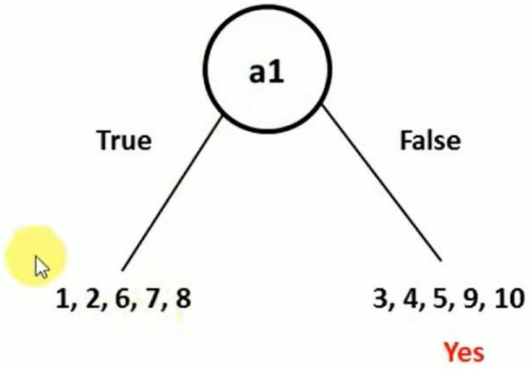
$$Gain(S, a3) = 0.9709 - \frac{6}{10} * 0.9183 - \frac{4}{10} * 0.0 = 0.4199$$

| Instance | a1 | a2 | a3 | Classification |
|---|---|---|---|---|
| 1 | True | Hot | High | No |
| 2 | True | Hot | High | No |
| 3 | False | Hot | High | Yes |
| 4 | False | Cool | Normal | Yes |
| 5 | False | Cool | Normal | Yes |
| 6 | True | Cool | High | No |
| 7 | True | Hot | High | No |
| 8 | True | Hot | Normal | Yes |
| 9 | False | Cool | Normal | Yes |
| 10 | False | Cool | High | Yes |

$$Gain(S, a1) = 0.6099 \; - \; Maximum\ Gain$$

$$Gain(S, a2) = 0.1245$$

$$Gain(S, a3) = 0.4199$$



| Instance | a2 | a3 | Classification |
|---|---|---|---|
| 1 | Hot | High | No |
| 2 | Hot | High | No |
| 6 | Cool | High | No |
| 7 | Hot | High | No |
| 8 | Hot | Normal | Yes |

## Attribute: a2

$$Values\,(a2) = Hot,\ Cool$$

$$S_{a1} = [1+,\ 4-] \qquad Entropy(S_{a1}) = -\frac{1}{5}log_2\frac{1}{5} - \frac{4}{5}log_2\frac{4}{5} = 0.7219$$

$$S_{Hot} = [1+, 3-] \qquad Entropy(S_{Hot}) = -\frac{1}{4}log_2\frac{1}{4} - \frac{3}{4}log_2\frac{3}{4} = 0.8112$$

$$S_{Cool} \leftarrow [0+,\ 1-] \qquad Entropy(S_{Cool}) = 0.0$$

$$Gain\,(S, a2) = Entropy(S) - \sum_{v \in \{Hot, Cool\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, a2) = Entropy(S) - \frac{4}{5}Entropy(S_{Hot}) - \frac{1}{5}Entropy(S_{Cool})$$

$$Gain(S, a2) = 0.9709 - \frac{4}{5} * 0.8112 - \frac{1}{5} * 0.0 = 0.3219$$

| Instance | a2 | a3 | Classification |
|---|---|---|---|
| 1 | Hot | High | No |
| 2 | Hot | High | No |
| 6 | Cool | High | No |
| 7 | Hot | High | No |
| 8 | Hot | Normal | Yes |

## Attribute: a3

$$Values\,(a3) = High,\ Normal$$

$$S_{a1} = [1+,\ 4-] \qquad Entropy(S_{a1}) = -\frac{1}{5}log_2\frac{1}{5} - \frac{4}{5}log_2\frac{4}{5} = 0.7219$$

$$S_{High} = [0+, 4-] \qquad Entropy(S_{High}) = 0.0$$

$$S_{Normal} \leftarrow [1+,\ 0-] \qquad Entropy(S_{Normal}) = 0.0$$

$$Gain\,(S, a3) = Entropy(S) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

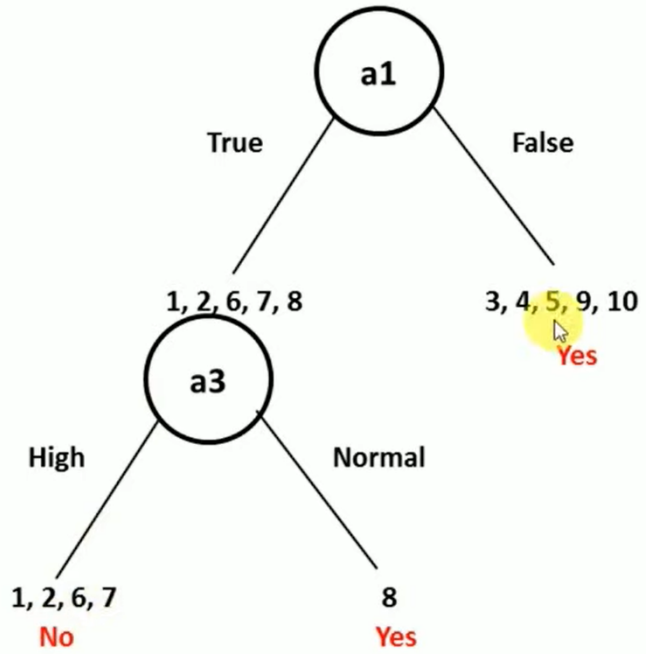
$$Gain(S, a3) = Entropy(S) - \frac{4}{5}Entropy(S_{High}) - \frac{1}{5}Entropy(S_{Normal})$$

$$Gain(S, a3) = 0.9709 - \frac{4}{5} * 0.0 - \frac{1}{5} * 0.0 = 0.7219$$

| Instance | a2 | a3 | Classification |
|---|---|---|---|
| 1 | Hot | High | No |
| 2 | Hot | High | No |
| 6 | Cool | High | No |
| 7 | Hot | High | No |
| 8 | Hot | Normal | Yes |

$$Gain(S_{a1}, a2) = 0.3219$$

$$Gain(S_{a1}, a3) = 0.7219 - \textit{Maximum Gain}$$



# Dealing with continuous-valued attributes

**Example. Temperature in the PlayTennis example**

- Sort the examples according to Temperature

| Temperature | 40 | 48 | 60 | 72 | 80 | 90 |
|---|---|---|---|---|---|---|
| PlayTennis | No | No | Yes | Yes | Yes | No |

- Determine candidate thresholds by averaging consecutive values where there is a change in classification: **(48+60)/2=54** and **(80+90)/2=85**
- Evaluate information gain of candidate thresholds (attributes) $\text{Temperature}_{>54}$ and $\text{Temperature}_{>85}$. Then Select the threshold based on the information gain.

# Dealing with continuous-valued attributes

| Temperature | 40 | 48 | 60 | 72 | 80 | 90 |
|---|---|---|---|---|---|---|
| PlayTennis | No | No | Yes | Yes | Yes | No |

**Information gain of $Temperature_{>54}$**

$$Values\ (Temp_{>54}) = <54, >54$$

$$S = [3+, 3-] \qquad Entropy(S) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = 1.0$$

$$S_{<54} \leftarrow [0+, 2-] \qquad Entropy(S_{<54}) = 0.0$$

$$S_{>54} \leftarrow [3+, 1-] \qquad Entropy(S_{>54}) = -\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4} = 0.8113$$

$$Gain\ (S, Temp_{>54}) = Entropy(S) - \sum_{v \in \{<54, >54\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp_{>54}) = Entropy(S) - \frac{2}{6} Entropy(S_{<54}) - \frac{4}{6} Entropy(S_{>54})$$

$$Gain(S, Temp_{>54}) = 1.0 - \frac{2}{6}\ 0.0 - \frac{4}{6}\ 0.8113 = 0.4591$$

# Dealing with continuous-valued attributes

| | | | | | | |
|---|---|---|---|---|---|---|
| **Temperature** | **40** | **48** | **60** | **72** | **80** | **90** |
| **PlayTennis** | **No** | **No** | **Yes** | **Yes** | **Yes** | **No** |

**Information gain of $Temperature_{>85}$**

$$Values\ (Temp_{>85}) = < 85, > 85$$

$$S = [3+, 3-] \qquad Entropy(S) = -\frac{3}{6} log_2 \frac{3}{6} - \frac{3}{6} log_2 \frac{3}{6} = 1.0$$

$$S_{<85} \leftarrow [3+, 2-] \qquad Entropy(S_{<85}) = -\frac{3}{5} log_2 \frac{3}{5} - \frac{2}{5} log_2 \frac{2}{5} = 0.971$$

$$S_{>85} \leftarrow [0+, 1-] \qquad Entropy(S_{>85}) = 0.0$$

$$Gain\ (S, Temp_{>85}) = Entropy(S) - \sum_{v \in \{<85, >85\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Temp_{>85}) = Entropy(S) - \frac{5}{6} Entropy(S_{<85}) - \frac{1}{6} Entropy(S_{>85})$$

$$Gain(S, Temp_{>85}) = 1.0 - \frac{5}{6}\ 0.971 - \frac{1}{6}\ 0.0 = 0.1908$$

# Dealing with continuous-valued attributes

| Temperature | 40 | 48 | 60 | 72 | 80 | 90 |
|---|---|---|---|---|---|---|
| PlayTennis | No | No | Yes | Yes | Yes | No |

$$Gain(S, Temp_{>54}) = 0.4591$$

$$Gain(S, Temp_{>85}) = 0.1908$$

# Summary of DT

- One of the most popular ML tools
  - Easy to understand (white-box), implement, and use
  - Computationally cheap (for heuristic algorithms)
  - Fast at testing time: O(depth)
- Important to select a good attribute
  - e.g., use information gain
- Overfitting
  - Need to use strategies to find simple trees

# Questions?