

Machine Learning, Spring 2020

Decision Trees

Reading Assignment: Chapter 6 & 7

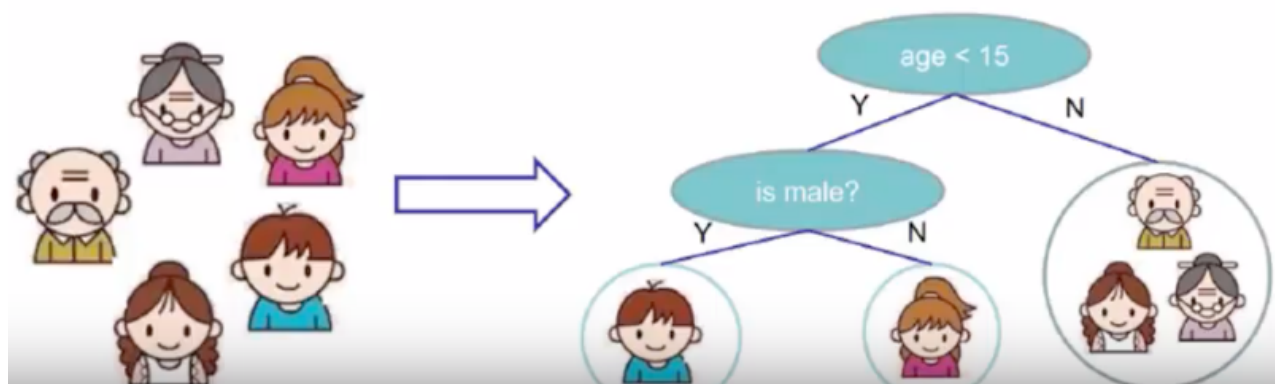
Python tutorial: <http://learnpython.org/>

TensorFlow tutorial: <https://www.tensorflow.org/tutorials/>

PyTorch tutorial: <https://pytorch.org/tutorials/>

Decision Trees Model

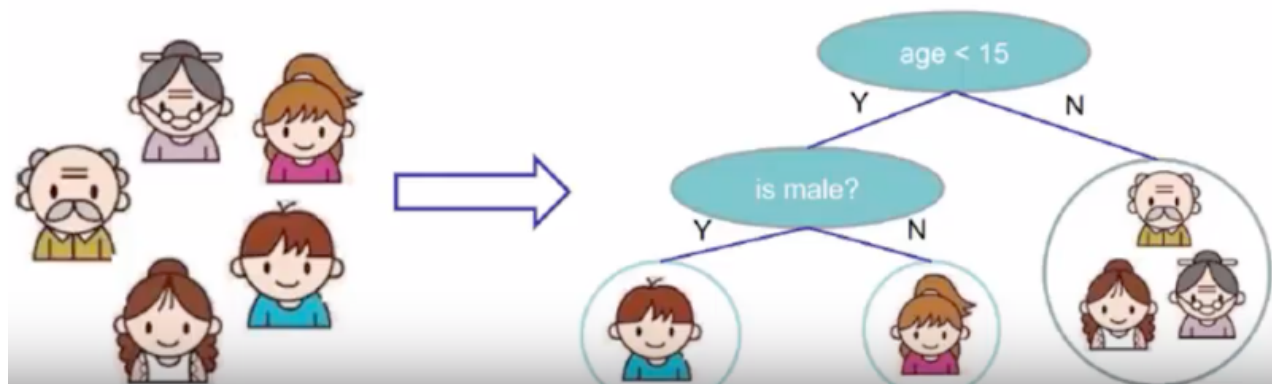
Decision tree: A decision tree decision is an inductive learning task that uses particular facts to make more generalized conclusions. Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.



Who will play the game?

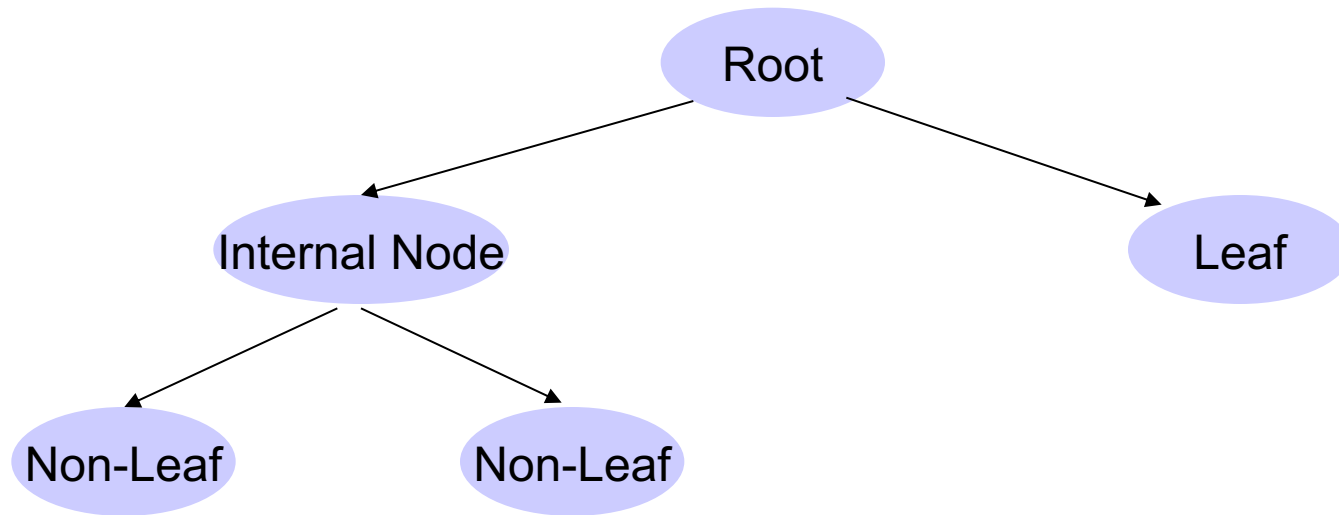
Decision for Computer Game

- Age > 15?
- Is male?



Decision Tree

- Root node
- Internal node
- Leaf node



Tree Training

- Training data
- Construct the tree
- Testing with a trained tree

Tree Training Procedure

- Choose one attribute for the root node
- Decision made based on the node
- Choose second attribute for the internal node
- Decision made based on the second node
- Iterate the steps above

Which attribute is best?

- Which criteria should be used to evaluate the quality of a attribute?
- Answer: Entropy

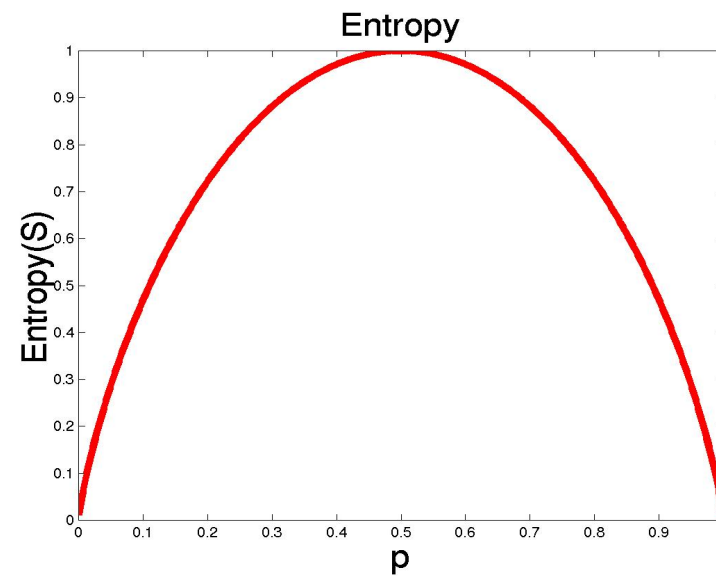
Entropy

- Entropy(S)= Expected number of bits needed to encode class (+ or -) of randomly drawn members of S (under the optimal, shortest length-code)

$$H(X) = - \sum p_i * \log p_i, i=1,2, \dots, n$$

- Example: [1,1,1,1,1,1,2,2] and [1,2,3,4,5,6,7,8], which entropy is smaller?
- Binary case: encode (+ or -) of random member of S:
$$-p_+ \log_2 p_+ - p_- \log_2 p_-$$

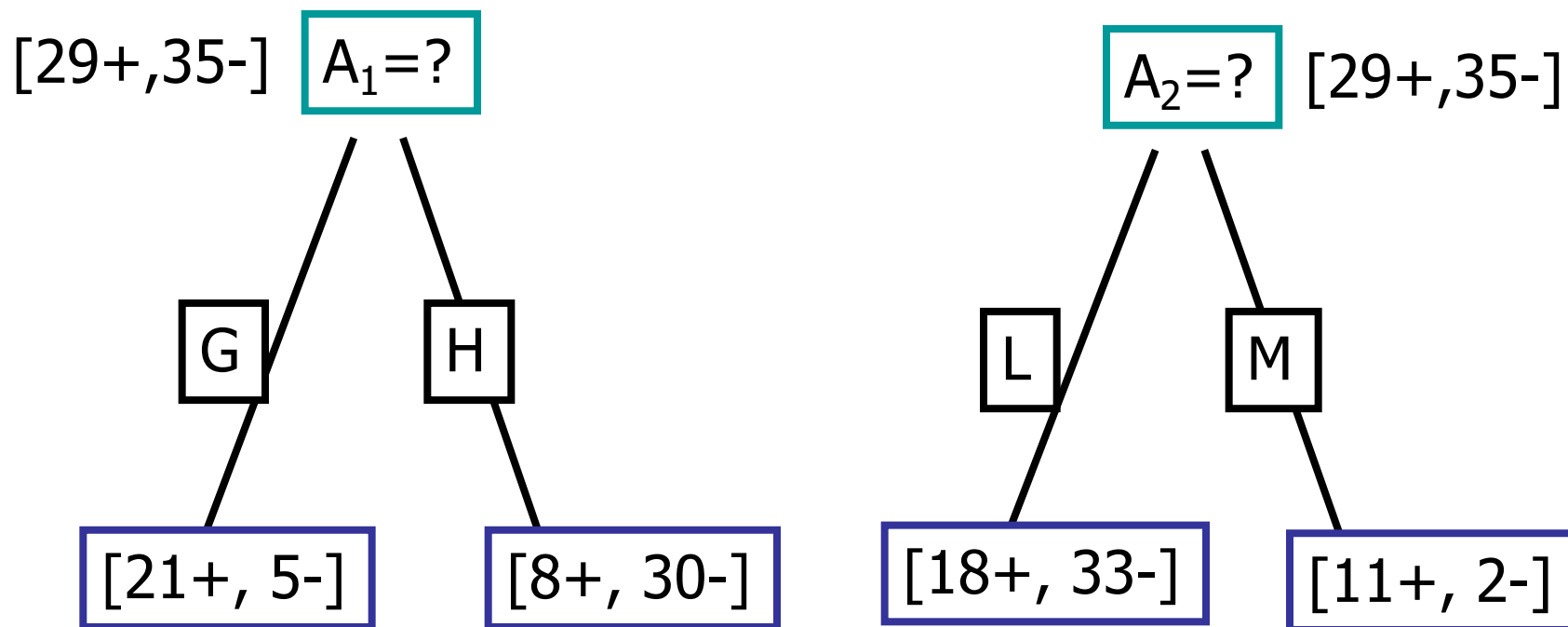
Entropy



Entropy

- A statistical property called information gain, measures how well a given attribute separates the training examples
- Information gain uses the notion of entropy, commonly used in information theory
- Information gain = expected reduction of entropy

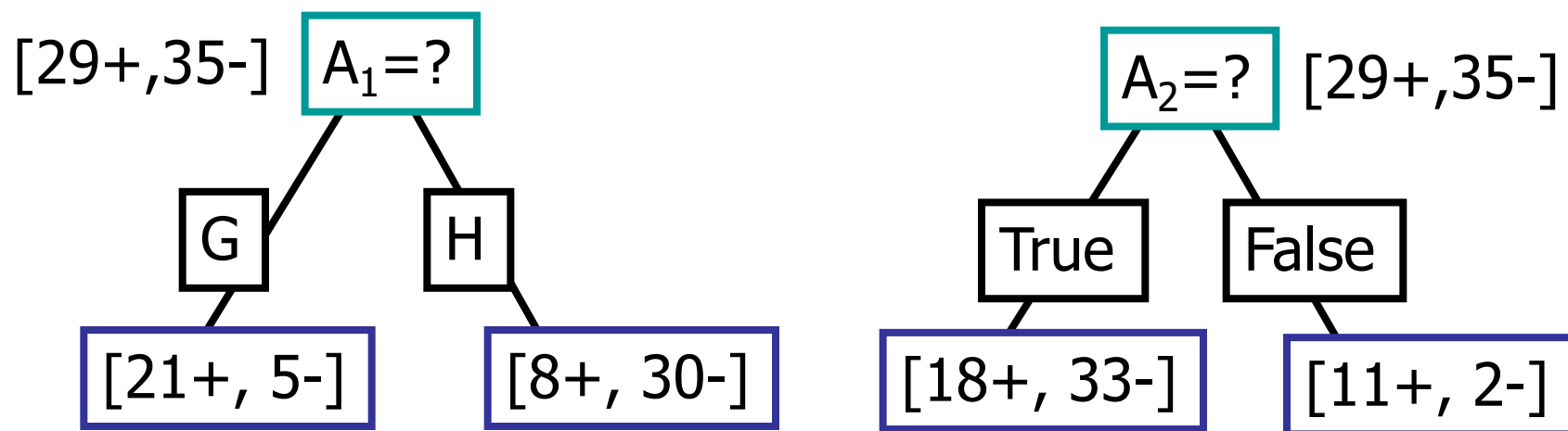
Example of Entropy Calculation



- S is a sample of training examples
- p_+ is the proportion of positive examples
- p_- is the proportion of negative examples

Before Classification

$$\text{Entropy}([29+, 35-]) = -29/64 \log_2 29/64 - 35/64 \log_2 35/64 \\ = 0.99$$



Information Gain

$$\text{Entropy}([21+, 5-]) = 0.71$$

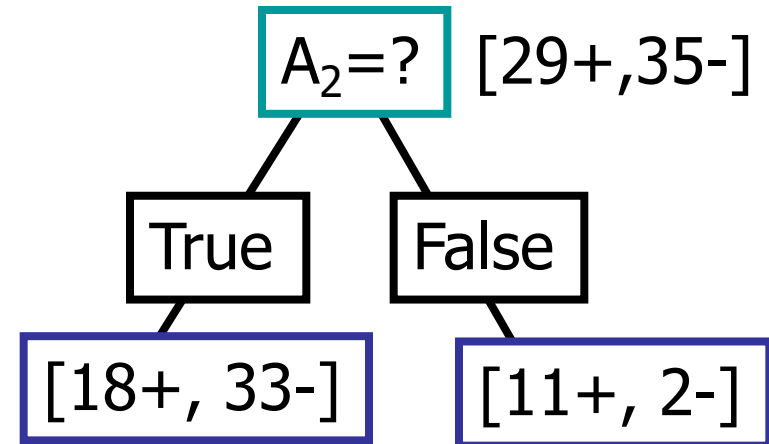
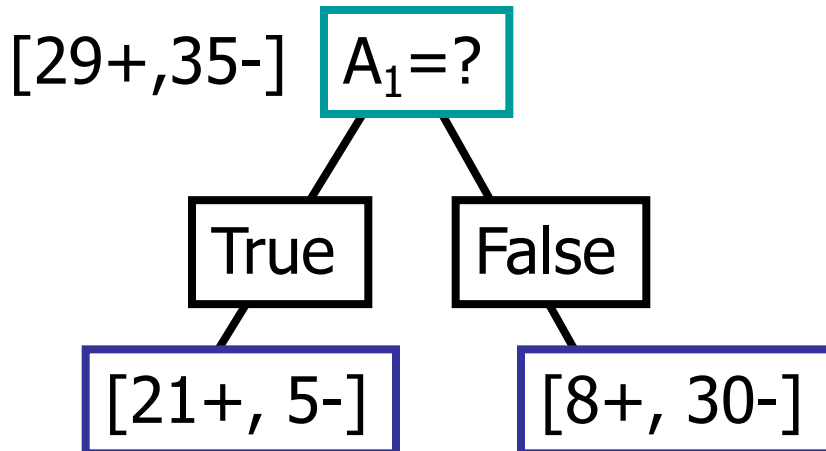
$$\text{Entropy}([8+, 30-]) = 0.74$$

$$\begin{aligned} \text{Gain}(S, A_1) &= \text{Entropy}(S) \\ &\quad - 26/64 * \text{Entropy}([21+, 5-]) \\ &\quad - 38/64 * \text{Entropy}([8+, 30-]) \\ &= 0.27 \end{aligned}$$

$$\text{Entropy}([18+, 33-]) = 0.94$$

$$\text{Entropy}([11+, 2-]) = 0.62$$

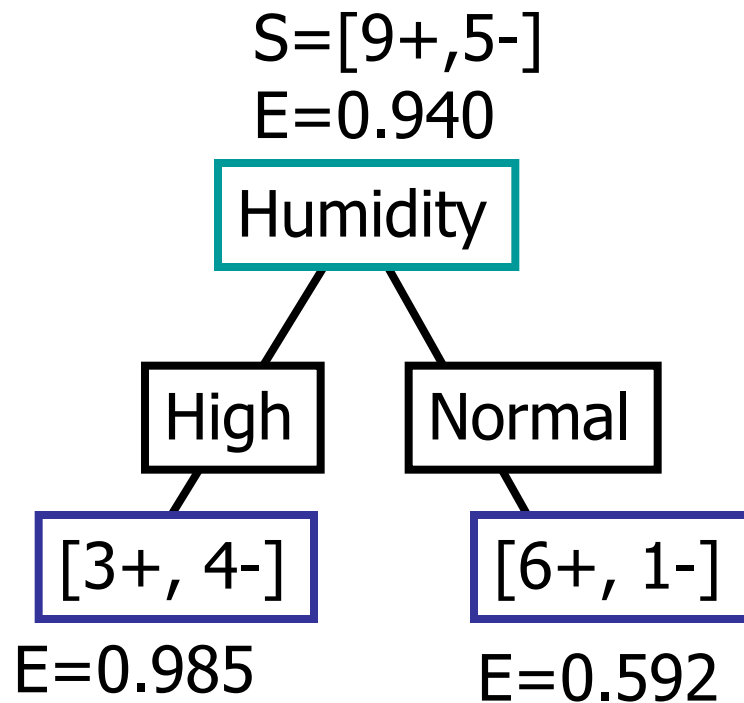
$$\begin{aligned} \text{Gain}(S, A_2) &= \text{Entropy}(S) \\ &\quad - 51/64 * \text{Entropy}([18+, 33-]) \\ &\quad - 13/64 * \text{Entropy}([11+, 2-]) \\ &= 0.12 \end{aligned}$$



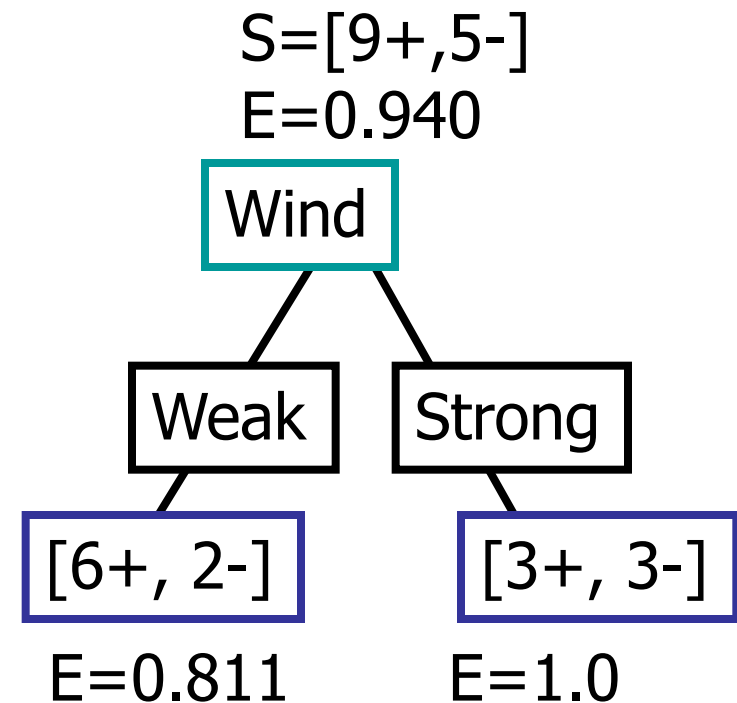
Training A Tree

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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

Selecting the Next Attribute



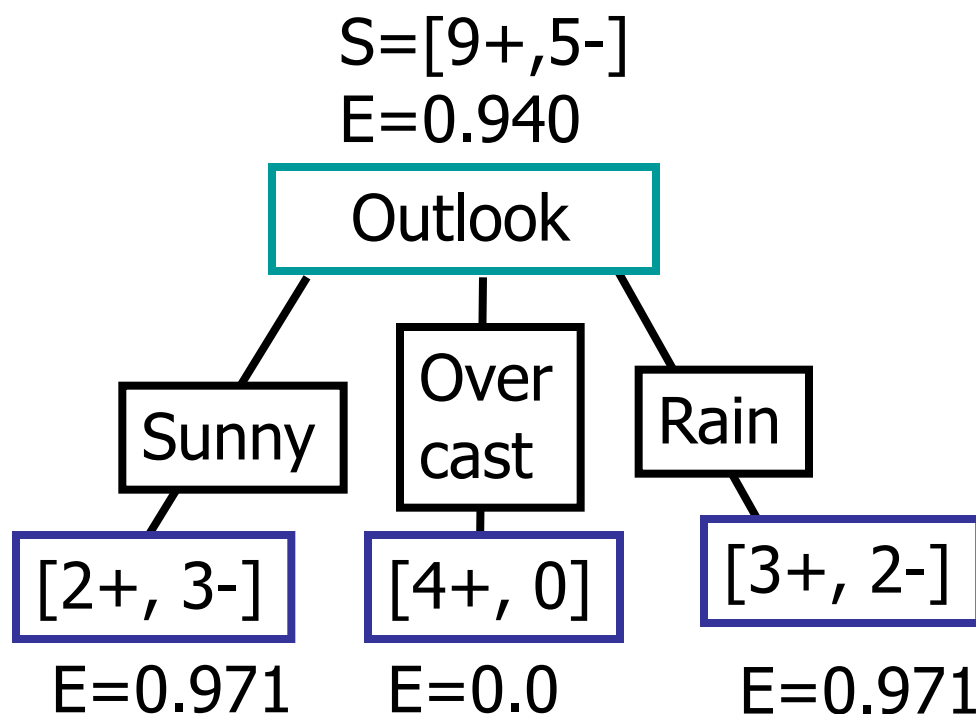
$$\begin{aligned}
 \text{Gain}(S, \text{Humidity}) &= 0.940 - (7/14) * 0.985 \\
 &\quad - (7/14) * 1.0 \\
 &= 0.151
 \end{aligned}$$



$$\begin{aligned}
 \text{Gain}(S, \text{Wind}) &= 0.940 - (8/14) * 0.811 \\
 &\quad - (6/14) * 1.0 \\
 &= 0.048
 \end{aligned}$$

Humidity provides greater info. gain than Wind, w.r.t target classification.

Selecting the Next Attribute



$$\begin{aligned}
 &\text{Gain}(S, \text{Outlook}) \\
 &= 0.940 - (5/14) * 0.971 \\
 &\quad - (4/14) * 0.0 - (5/14) * 0.0971 \\
 &= 0.247
 \end{aligned}$$

Selecting the Next Attribute

The information gain values for the 4 attributes are:

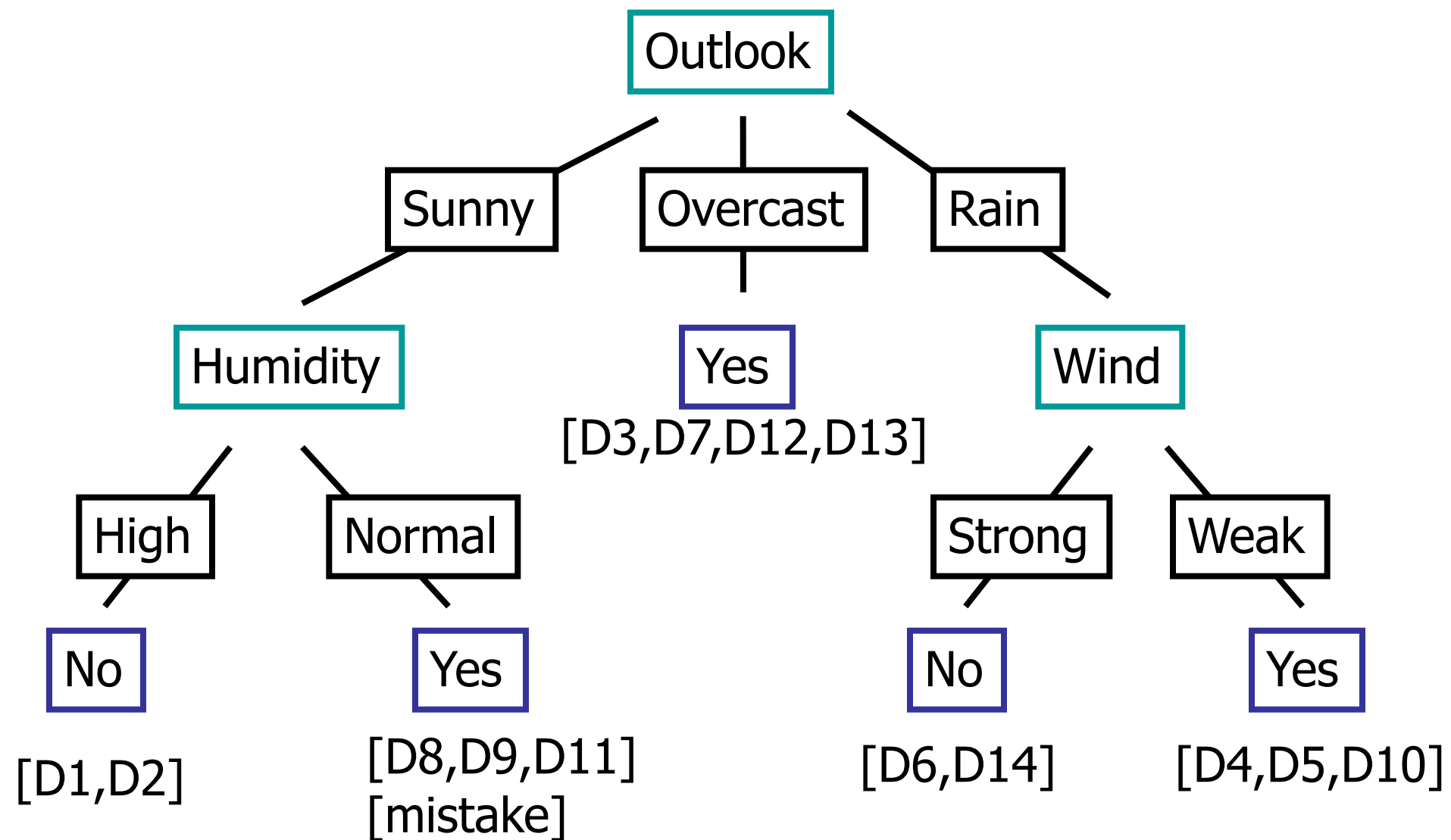
- $\text{Gain}(S, \text{Outlook}) = 0.247$
- $\text{Gain}(S, \text{Humidity}) = 0.151$
- $\text{Gain}(S, \text{Wind}) = 0.048$
- $\text{Gain}(S, \text{Temperature}) = 0.029$

where S denotes the collection of training

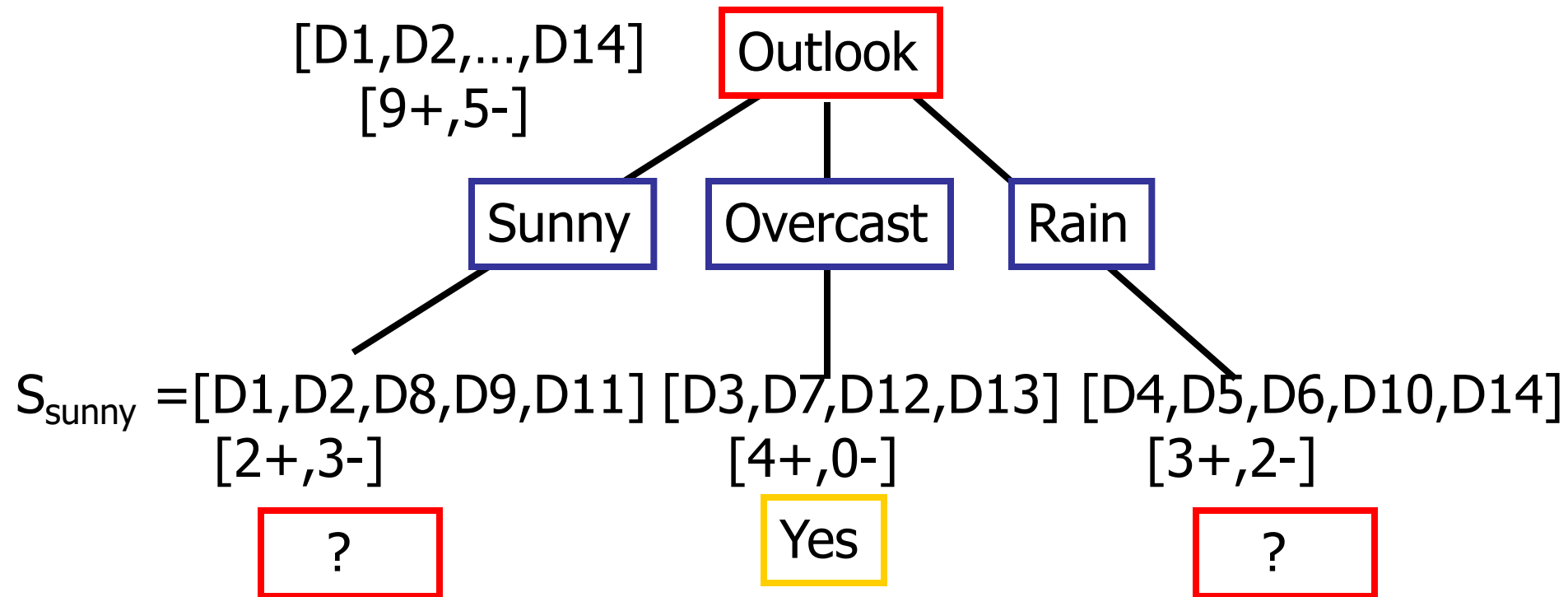
ID3 Algorithm for Decision Tree

- ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset.

ID3 Algorithm

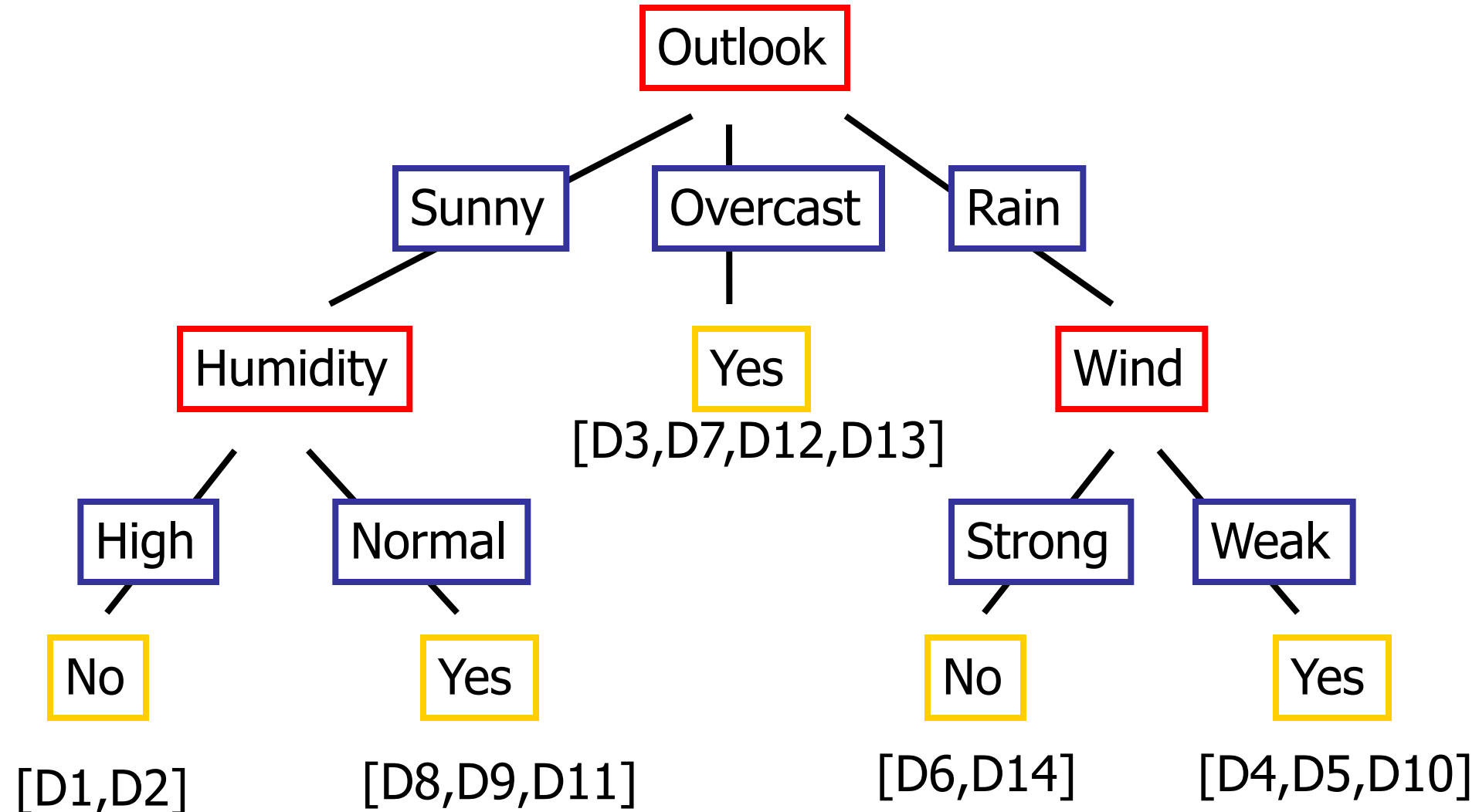


ID3 Algorithm



$$\begin{aligned} \text{Gain}(S_{\text{sunny}}, \text{Humidity}) &= 0.970 - (3/5)0.0 - 2/5(0.0) = 0.970 \\ \text{Gain}(S_{\text{sunny}}, \text{Temp.}) &= 0.970 - (2/5)0.0 - 2/5(1.0) - (1/5)0.0 = 0.570 \\ \text{Gain}(S_{\text{sunny}}, \text{Wind}) &= 0.970 - (2/5)1.0 - 3/5(0.918) = 0.019 \end{aligned}$$

ID3 Algorithm



ID3 Algorithm

- What is the challenging issue for ID3 algorithm?
- Sparse distributed data (each class only contains few points)

C4.5 Algorithm

- C4.5 is an extension of Quinlan's earlier ID3 algorithm.
- The splitting criterion is the normalized information gain (difference in entropy).

GINI Coefficient

- The Gini index or Gini coefficient is a statistical measure of distribution

$$Gini(p) = \sum_{k=1}^K p_k (1 - p_k) = 1 - \sum_{k=1}^K p_k^2$$

Overfitting of Decision Tree

Avoid Overfitting

- Pre-prune: stop growing when split not statistically significant
- Post-prune: stop growing when split not statistically significant

Pre-prune

- Limits the depth of the tree
- Limits the number of samples of a internal node
- Limits the information gain

Pre-prune

