# Decision Trees

Rodrigo Fernandes de Mello

Invited Professor at Télécom ParisTech

Associate Professor at Universidade de São Paulo, ICMC, Brazil

http://www.icmc.usp.br/~mello

mello@icmc.usp.br

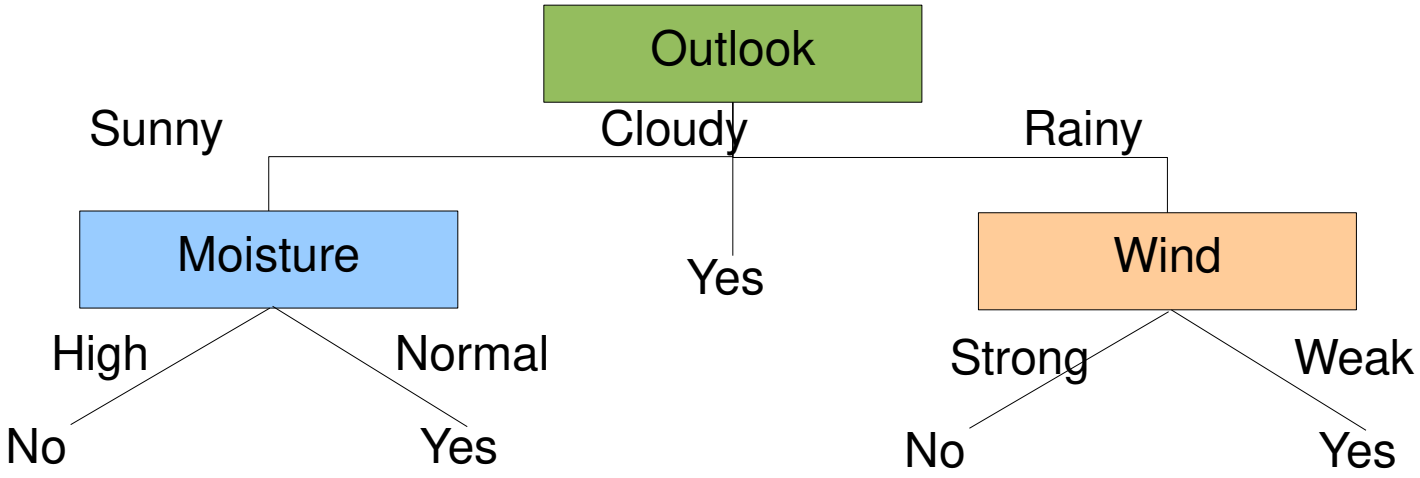

# Decision Trees

- They build up trees to organize the attributes as internal nodes and answers as leaves
  - The importance of each attribute is related to its tree height
  - Intended to be used on discrete data
    - If not, one needs to discretize it

- Common applications:
  - Health diagnosis systems
  - Bank credit analysis

# Decision Trees

- For example, consider the problem of “Play some Sport”
  - Let’s classify if a day is good to play



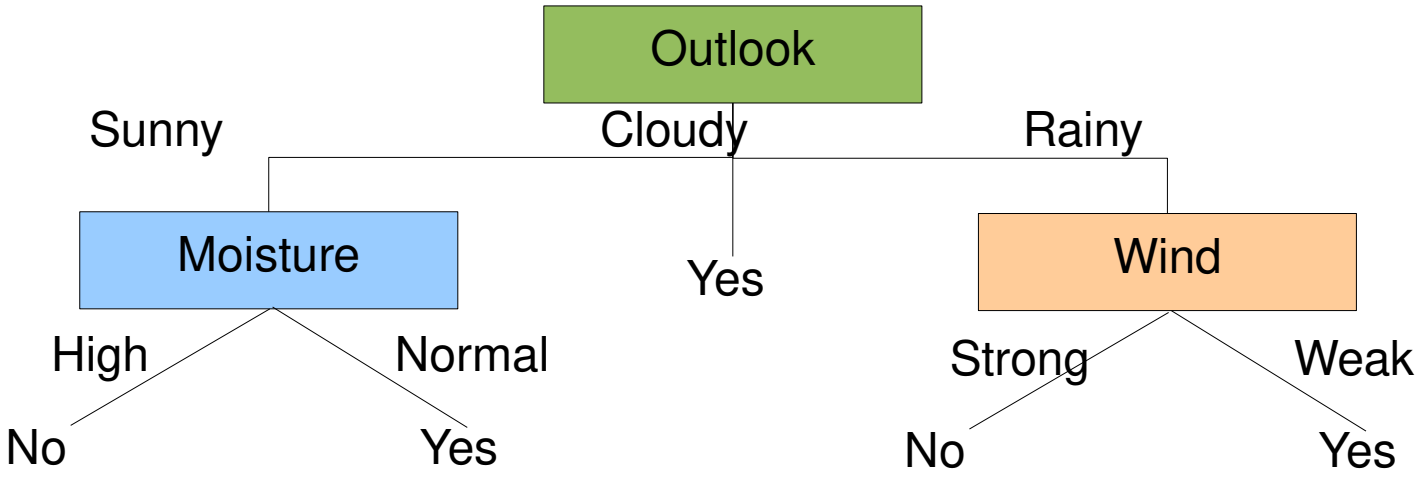

  - For example:
    - Having the query instance:

      <Outlook=Sunny, Temperature=Hot, Moisture=High>

    - Output:
      - No

# Decision Trees

- After building up the tree, one can organize the rules employed in a next classification stage:




<Outlook=Sunny AND Moisture=Normal>

OR <Outlook=Cloudy>

OR <Outlook=Rainy AND Wind=Weak>

# Decision Trees

- Most well-known algorithms:
  - ID3 (Quinlan, 1986)
  - C4.5 (Quinlan, 1993)
    - J48

- The ID3 algorithm
  - It builds up a tree based on a top-down approach
    - It attempts to position the most discriminative attribute as close as possible to the tree root
  - Every attribute must be tested in order to define their relevances
    - For each attribute, branches are created according to its possible values

# Decision Trees

- ID3 employs the Information Gain approach to decide on the most important attributes
  - That depends on the Shannon's Entropy

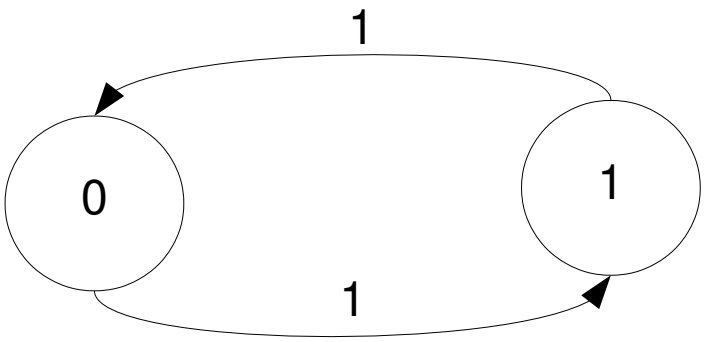
# Shannon's Entropy

- Entropy:
  - A Thermodynamic property to determine the amount of useful energy in a given system
  - Gibbs stated that the best interpretation for the Entropy in the Statistical Mechanics is as an Uncertainty Measure

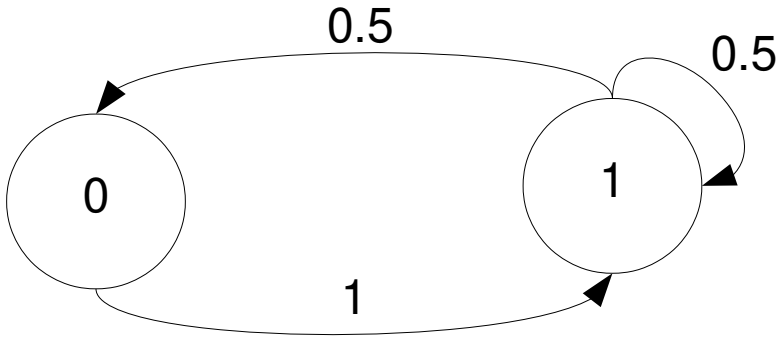
- A little about the History:
  - Entropy starts with Lazare Carnot (1803)
  - Rudolf Clausius (1850s-1860s) brings new interpretations in physics
  - Claude Shannon (1948) designs the concept of Entropy in the context of **Information Theory**

# Shannon's Entropy

- Let the system start as follows:




- And change to:



# Shannon's Entropy

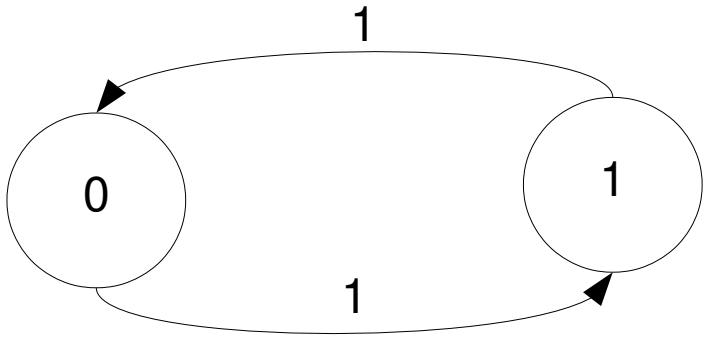
- This is the Equation proposed by Shannon:

$$E = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

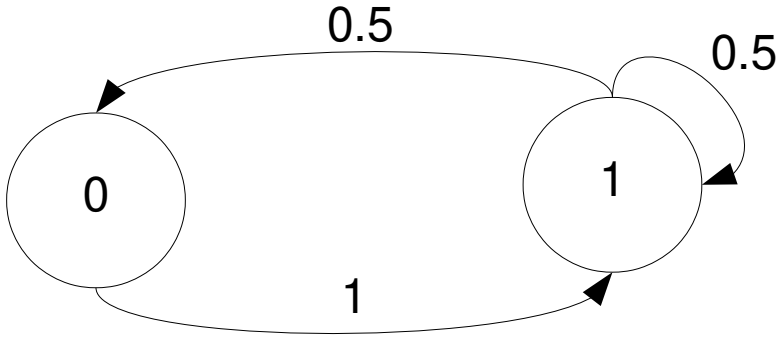
- It measures the total energy of a system:
  - It considers the system is at a given state i and transitions to j
  - The function $\log_2$ supports to find the number of bits of Information

# Shannon's Entropy

- Thus:




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




**After modifying the behavior, the system agregated a greater level of uncertainty or energy (Ex: Win the Lottery)**

$$E = -(1\log_2(1) + 0.5\log_2(0.5) + 0.5\log_2(0.5)) = 1$$

# ID3 using the Entropy

- Let a collection of instances S, including positive and negative examples
  - i.e., two distinct and discrete classes/labels

- Let the probability of pertaining to each class of S
  - The Entropy is given by:

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

# ID3 using the Entropy

- For illustration purposes, consider S has 14 examples:
  - 9 positives
  - 5 negatives

- Then, the Entropy of such a set is:

$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

  - Observe a binary-class scenario has maximum Entropy equals to 1

# ID3 using the Entropy

- Other scenarios:
  - Having [7+, 7-]

$$E(S) = -\frac{7}{14}\log_2\frac{7}{14} - \frac{7}{14}\log_2\frac{7}{14} = 0.99\ldots \approx 1$$

  - Having [0+, 14-] or [14+, 0-]

$$E(S) = -\frac{14}{14}\log_2\frac{14}{14} = 0$$

  - Entropy measures the level of uncertainty about some event/ source

# ID3 using the Entropy

- We may generalize Entropy to more classes:

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

- Several other systems can be studied using Entropy:
  - For instance:
    - Time Series, after some quantization process

- Why the Log function is used?
  - It allows to measure information in bits
  - A great example is about codifying some message
    - Codify: TORONTO

# ID3 uses Information Gain

- After defining Entropy, we can define Information Gain
  - It measures how effective an attribute is to classify some dataset
  - A different point of view:
    - It measures the Entropy reduction when partitioning the dataset using a given attribute

$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

    - The second term measures the Entropy after partitioning the dataset with attribute A
  - Therefore:
    - IG measures the Entropy reduction, i.e. the uncertainty decrease, after selecting attribute A to compose the tree

# ID3 uses Information Gain

- To illustrate, take S and the attribute Wind (Weak or Strong)
  - S contains 14 instances [9+, 5-]
  - Now consider that:
    - 6 of the positive examples and 2 of the negative ones have Wind=Weak (8 in total)
    - There are 3 instances with Wind=Strong for both the positive and the negative classes (6 in total)
- Then, the Information Gain while selecting the attribute Wind to take the root of our decision tree is given by:

$$S = [9+, 5-]$$
$$S_{\text{weak}} \leftarrow [6+, 2-]$$
$$S_{\text{strong}} \leftarrow [3+, 3-]$$
$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

# ID3 uses Information Gain

- So that:

$$S = [9+, 5-]$$
$$S_{\text{weak}} \leftarrow [6+, 2-]$$
$$S_{\text{strong}} \leftarrow [3+, 3-]$$

$$\mathbf{GI}(S, A) = E(S) - \sum_{v \in \mathbf{Valores}(A)} \frac{S_v}{S} E(S_v)$$

$$\mathbf{GI}(S, A) = 0.94 - \frac{8}{14} E(S_{\text{weak}}) - \frac{6}{14} E(S_{\text{strong}})$$

# ID3 uses Information Gain

- Therefore:

$$S = [9+, 5-]$$
$$S_{\text{weak}} \leftarrow [6+, 2-]$$
$$S_{\text{strong}} \leftarrow [3+, 3-]$$

$$E(S_{\text{weak}}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.811$$

$$E(S_{\text{strong}}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.00$$

$$\mathbf{GI}(S, A) = 0.94 - \frac{8}{14}0.811 - \frac{6}{14}1.00 = 0.048$$

- This is the Information Gain measure used by ID3 along the iterations to build up a decision tree
  - Observe this scenario has only reduced the uncertainty a little
  - Thus, is this attribute good enough to take the root? You will see it is not!

# Running ID3

- Consider the concept of play some sport

| Outlook | Temperature | Moisture | Wind | Play Tennis |
|---|---|---|---|---|
| Sunny | Warm | High | Weak | No |
| Sunny | Warm | High | Strong | No |
| Cloudy | Warm | High | Weak | Yes |
| Rainy | Pleasant | High | Weak | Yes |
| Rainy | Cold | Normal | Weak | Yes |
| Rainy | Cold | Normal | Strong | No |
| Cloudy | Cold | Normal | Strong | Yes |
| Sunny | Pleasant | High | Weak | No |
| Sunny | Cold | Normal | Weak | Yes |
| Rainy | Pleasant | Normal | Weak | Yes |
| Sunny | Pleasant | Normal | Strong | Yes |
| Cloudy | Pleasant | High | Strong | Yes |
| Cloudy | Warm | Normal | Weak | Yes |
| Rainy | Pleasant | High | Strong | No |

# Running ID3

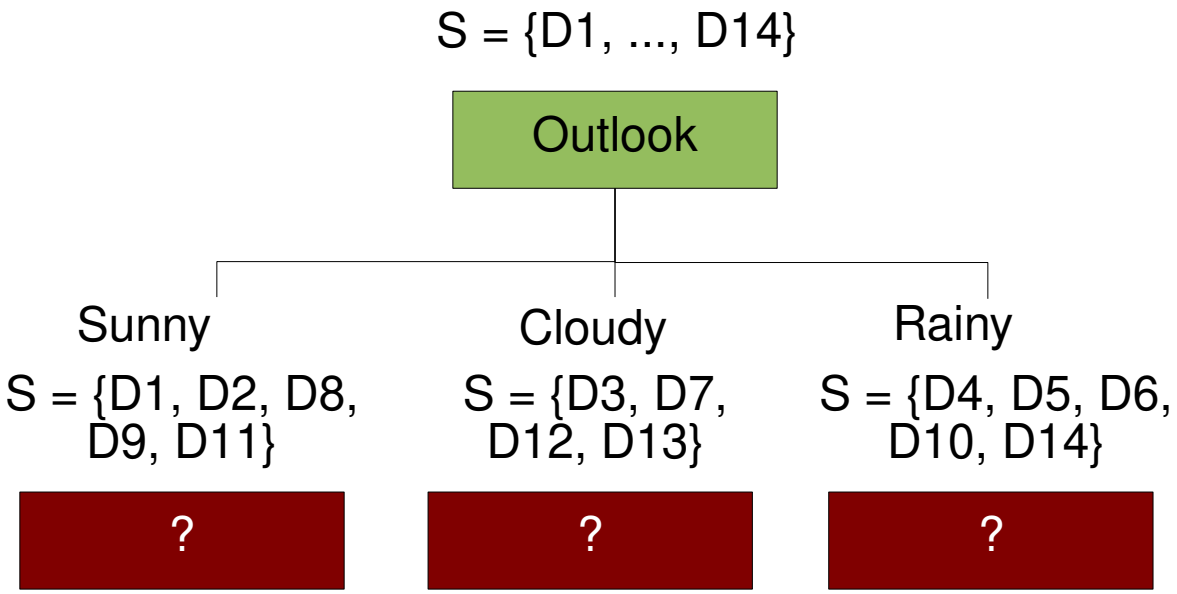
- First step:
  - Compute the Information Gain for each of its attributes:

$$\mathbf{GI}(S, \text{Outlook}) = 0.246$$
$$\mathbf{GI}(S, \text{Moisture}) = 0.151$$
$$\mathbf{GI}(S, \text{Wind}) = 0.048$$
$$\mathbf{GI}(S, \text{Temperature}) = 0.029$$

  - The attribute with greater IG is selected to be the tree root
    - This is the one that reduces the most the uncertainty!
    - The children nodes are created according to the possible values for the root attribute

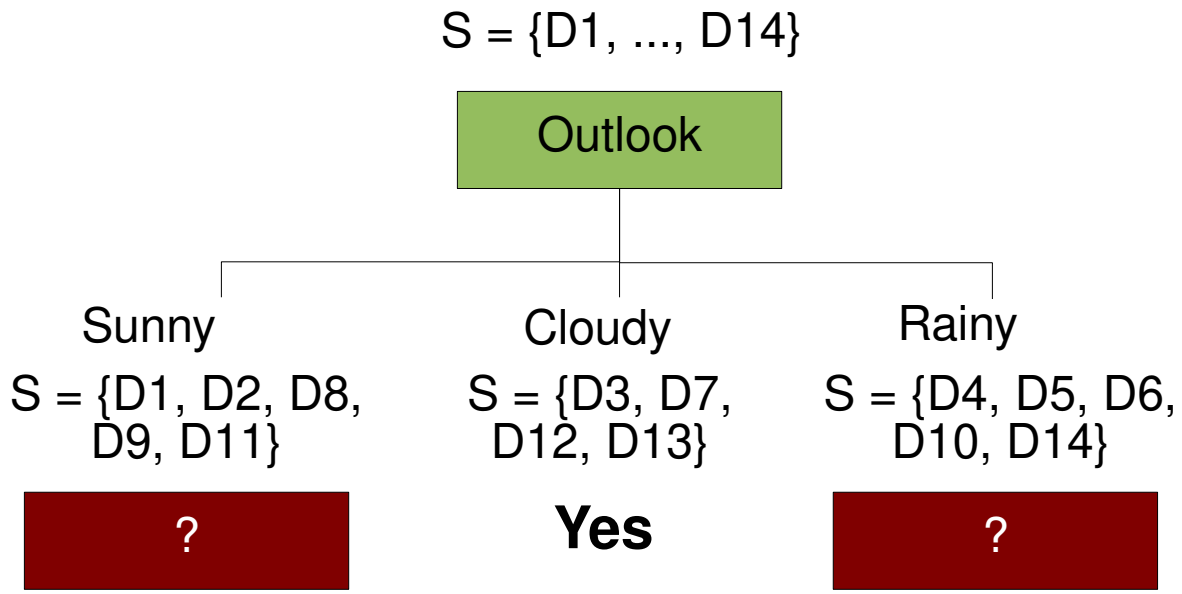
# Running ID3

- Having the tree root:
  - We must proceed in the same way to each of its branches
  - We only consider the examples filtered by each branch
    - If there is divergence



# Running ID3

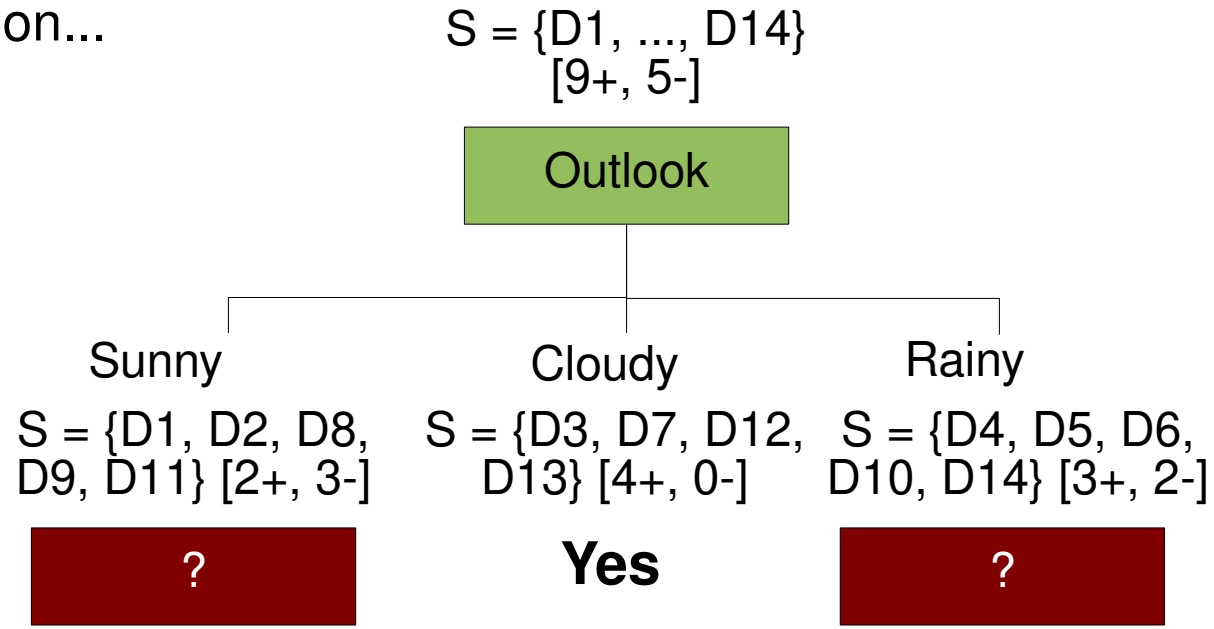
- Observe there is no divergence for one of the branches, so its Entropy is zero and a leaf label is defined:




- The attributes already incorporated throughout the path to reach a node are not considered in the Information Gain
  - Observe the Outlook will never be assessed again for both branches

# Running ID3

- Carrying on...




- Computing the Information Gain to the Sunny branch:
  - Entropy is E(S = Sunny)

$$E(S = \text{Sunny}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.97$$

# Running ID3

- Computing the Information Gain to the Sunny branch:

$$\mathbf{GI}(S, \text{ Moisture }) = 0.97 - \frac{3}{5}0.0 - \frac{2}{5}0.0 = 0.97$$

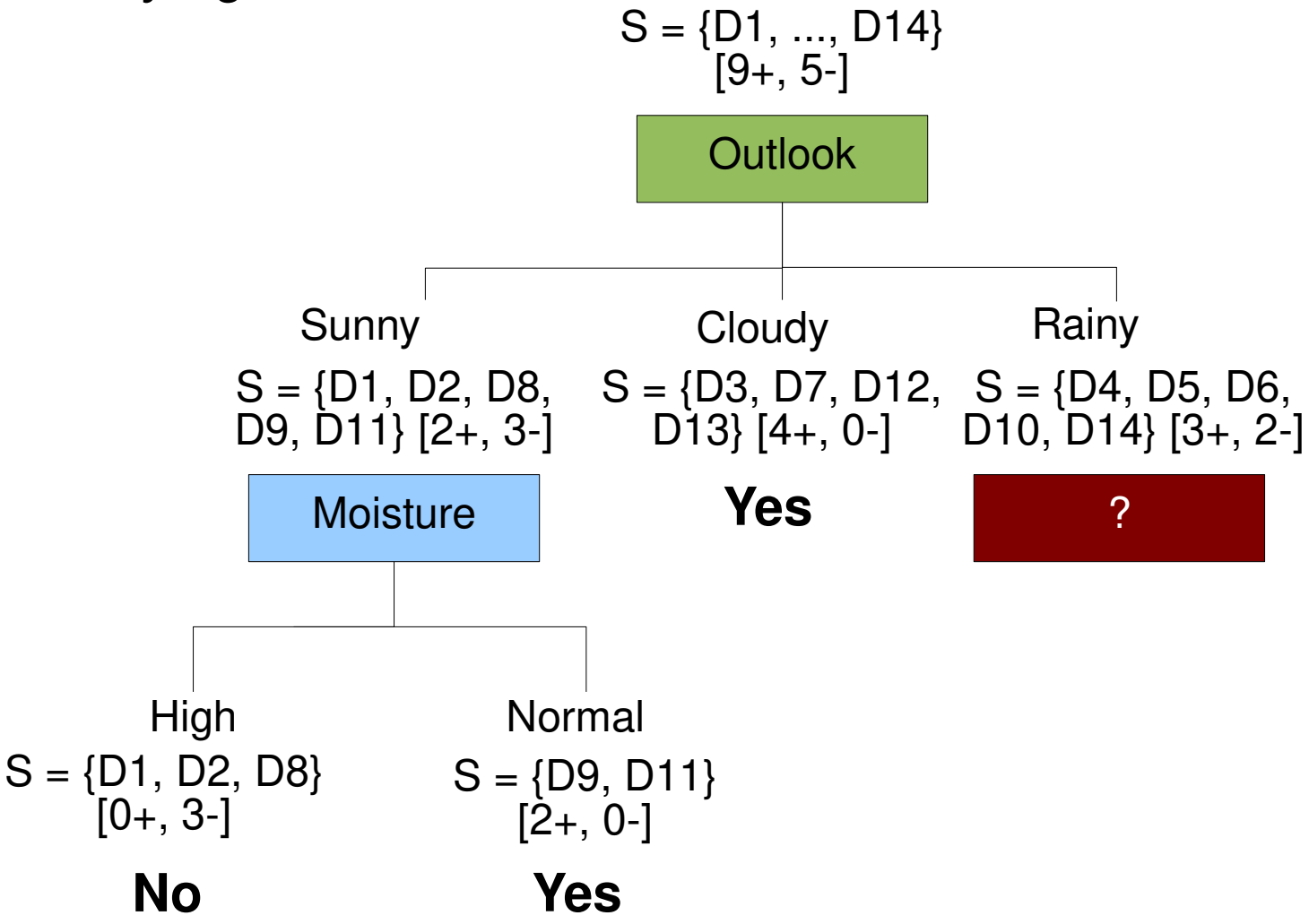
$$\mathbf{GI}(S, \text{ Temperature }) = 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1.0 = 0.57$$

$$\mathbf{GI}(S, \text{ Wind }) = 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = 0.019$$

- Observe the Moisture is the best choice!

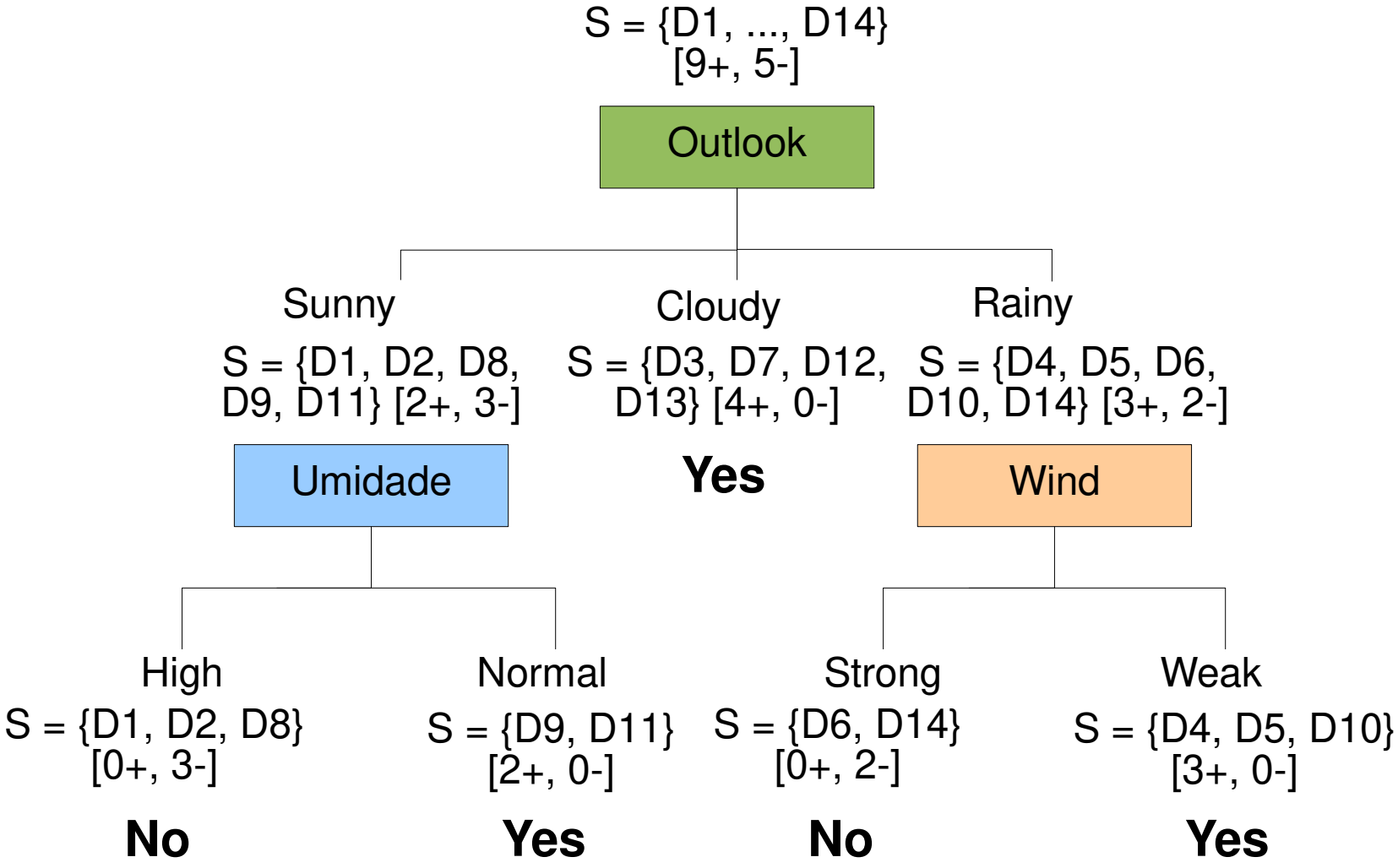
# Running ID3

- Carrying on...



# Running ID3

- And then...



# Running ID3

- ID3 carries on until one of the following conditions is satisfied:
  - (1) Either all attributes were included in the path from the root to the leaves
  - (2) Or the training examples at a given branch have a single class/label

- Observe ID3 starts with a null tree
  - Then it builds the tree up from the scratch
  - And It builds up a single tree

# Discussion

- C4.5 is an extension of ID3
  - It proceeds in the same way but it has a backtracking process afterwards in attempt to reduce the number of tree nodes
  - It is a copyrighted solution
  - J48 is based on the C4.5 documentation and is open

- The Information Gain is a way of reducing the tree height
  - But why?
    - William of Occam, 1320

# Suggestion

- Implement ID3
  - Discrete version
  - Continuous version

- Test it with the UCI datasets
  - http://archive.ics.uci.edu/ml/

# Complementary references


- Tom Mitchell, Machine Learning, 1994
- Hudson (2006) Signal Processing Using Mutual Information, IEEE Signal Processing Magazine
- Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.