Decision Trees

Md. Mohsin Uddin

East West University

mmuddin@ewubd.edu

May 28, 2019

- Slides borrowed from Vibhav Gogate, University of Texas at Dallas and Tom.M.Mitchell

Learning Decision Trees

Decision trees provide a very popular and efficient hypothesis space.

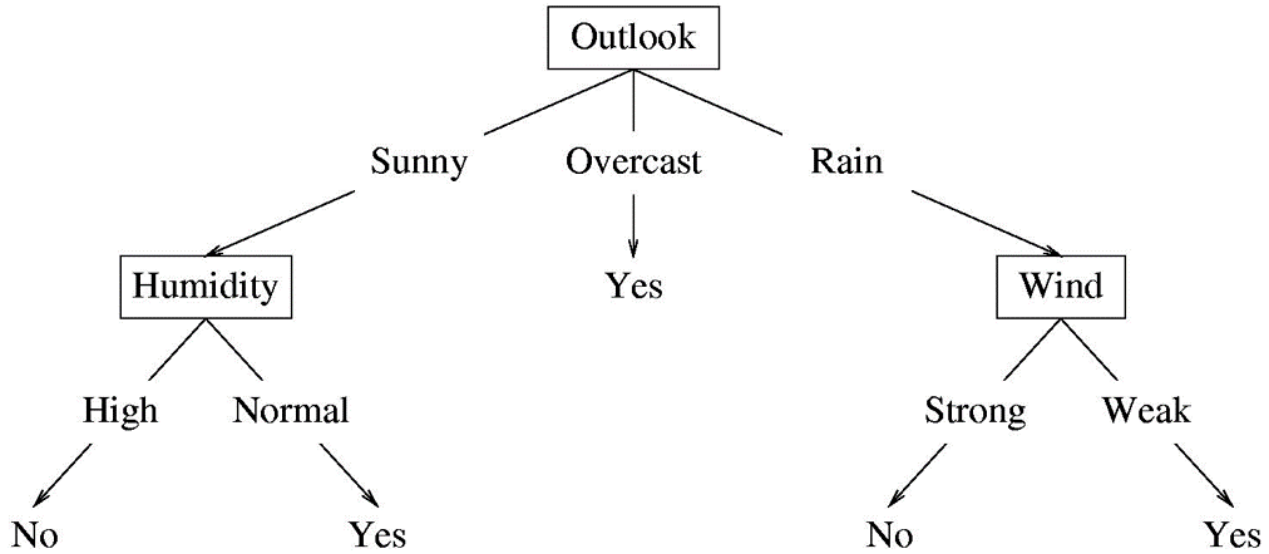
- Variable Size. Any boolean function can be represented
- Deterministic
- Discrete and Continuous Parameters.

Learning algorithms for decision trees can be described as

- Constructive Search. The tree is built by adding nodes.
- Batch (although online algorithms do exist).

Decision Tree Hypothesis Space

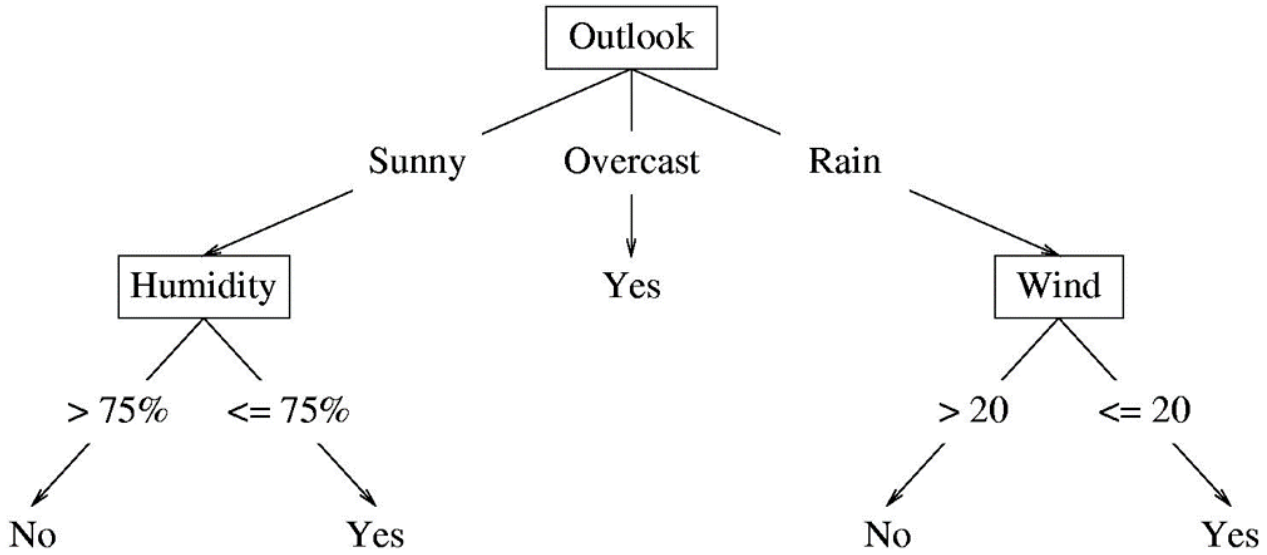
- **Internal nodes** test the value of particular features x_j and branch according to the results of the test.
- **Leaf nodes** specify the class $h(\mathbf{x})$.



Suppose the features are **Outlook** (x_1), **Temperature** (x_2), **Humidity** (x_3), and **Wind** (x_4). Then the feature vector $\mathbf{x} = (Sunny, Hot, High, Strong)$ will be classified as **No**. The **Temperature** feature is irrelevant.

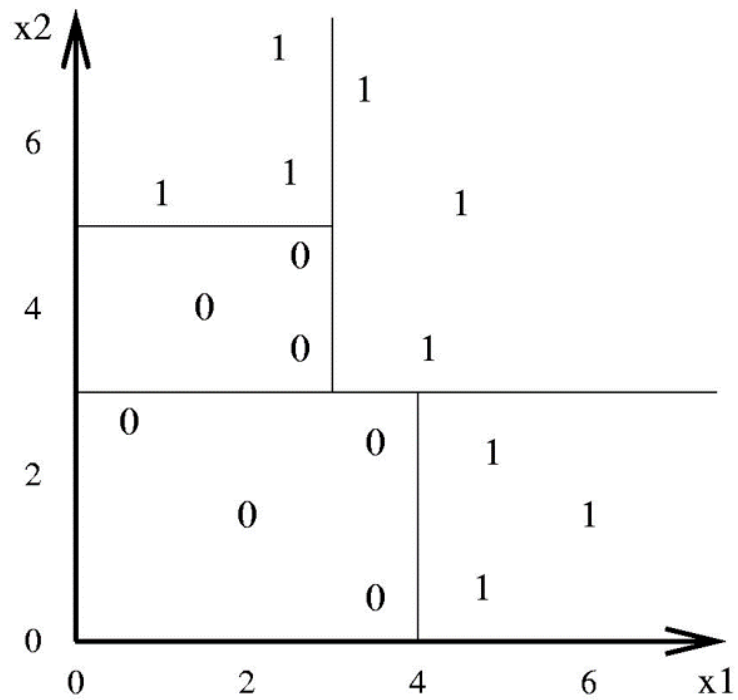
Decision Tree Hypothesis Space

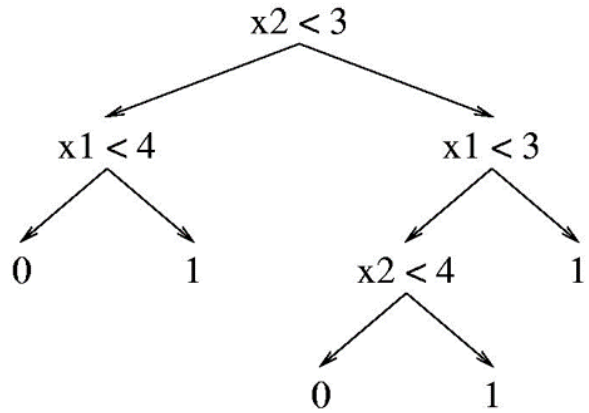
If the features are continuous, internal nodes may test the value of a feature against a threshold.



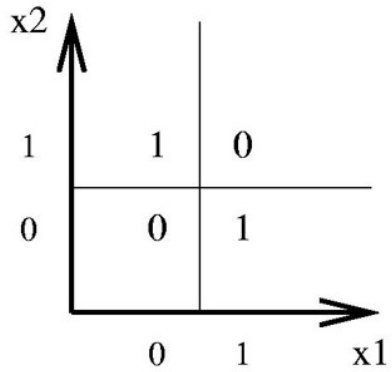
Decision Tree Decision Boundaries

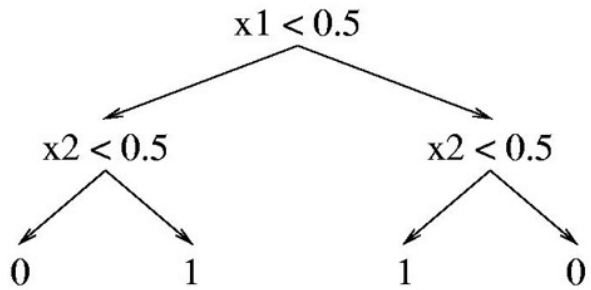
Decision trees divide the feature space into axis-parallel rectangles, and label each rectangle with one of the K classes.





Decision Trees Can Represent Any Boolean Function





The tree will in the worst case require exponentially many nodes, however.

Decision Trees Provide Variable-Size Hypothesis Space

As the number of nodes (or depth) of tree increases, the hypothesis space grows

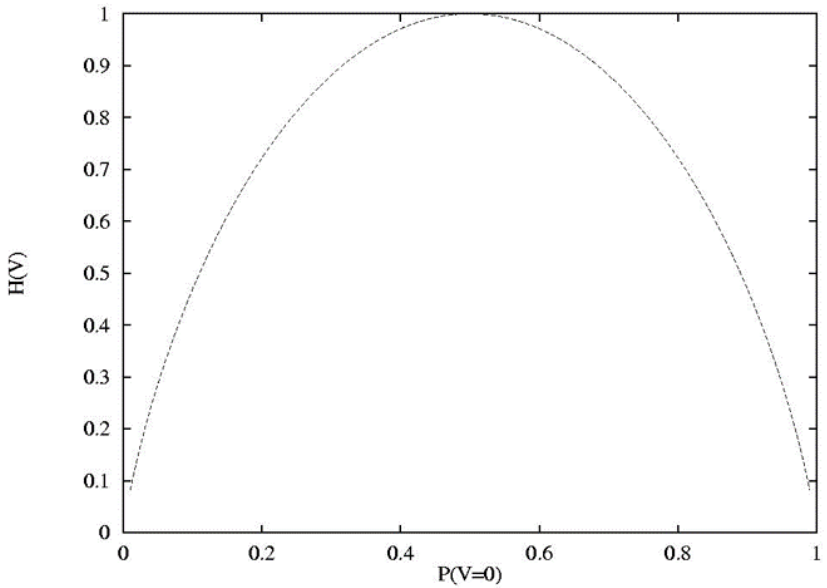
- **depth 1** ("decision stump") can represent any boolean function of one feature.
- **depth 2** Any boolean function of two features; some boolean functions involving three features (e.g., $(x_1 \wedge x_2) \vee (\neg x_1 \wedge \neg x_3)$
- **etc.**

Entropy

The *entropy* of V, denoted $H(V)$ is defined as follows:

$$H(V) = \sum_{v=0}^{1} -P(H = v) \lg P(H = v).$$

This is the average surprise of describing the result of one "trial" of V (one coin toss).



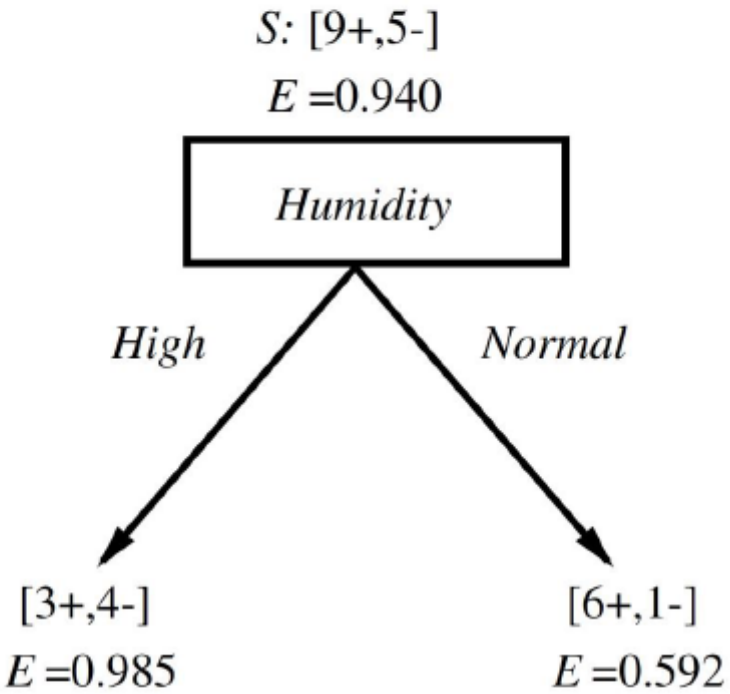
Entropy can be viewed as a measure of uncertainty.

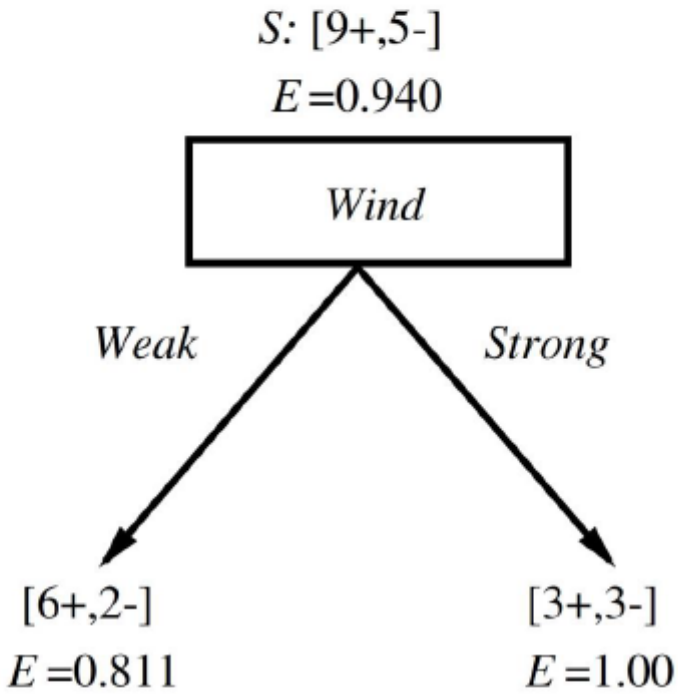
Information Gain: Using Entropy to make decisions

- $Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} P(v) Entropy(S_v)$
- Entropy is impurity in data
- Entropy(S): current impurity
- The second term measures the expected impurity after partitioning the data with respect to the A, i.e. new impurity
- Gain=Reduction in impurity
  - We want to be as pure as possible, i.e. maximize reduction in impurity
- So we want to maximize Gain!

When do I 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





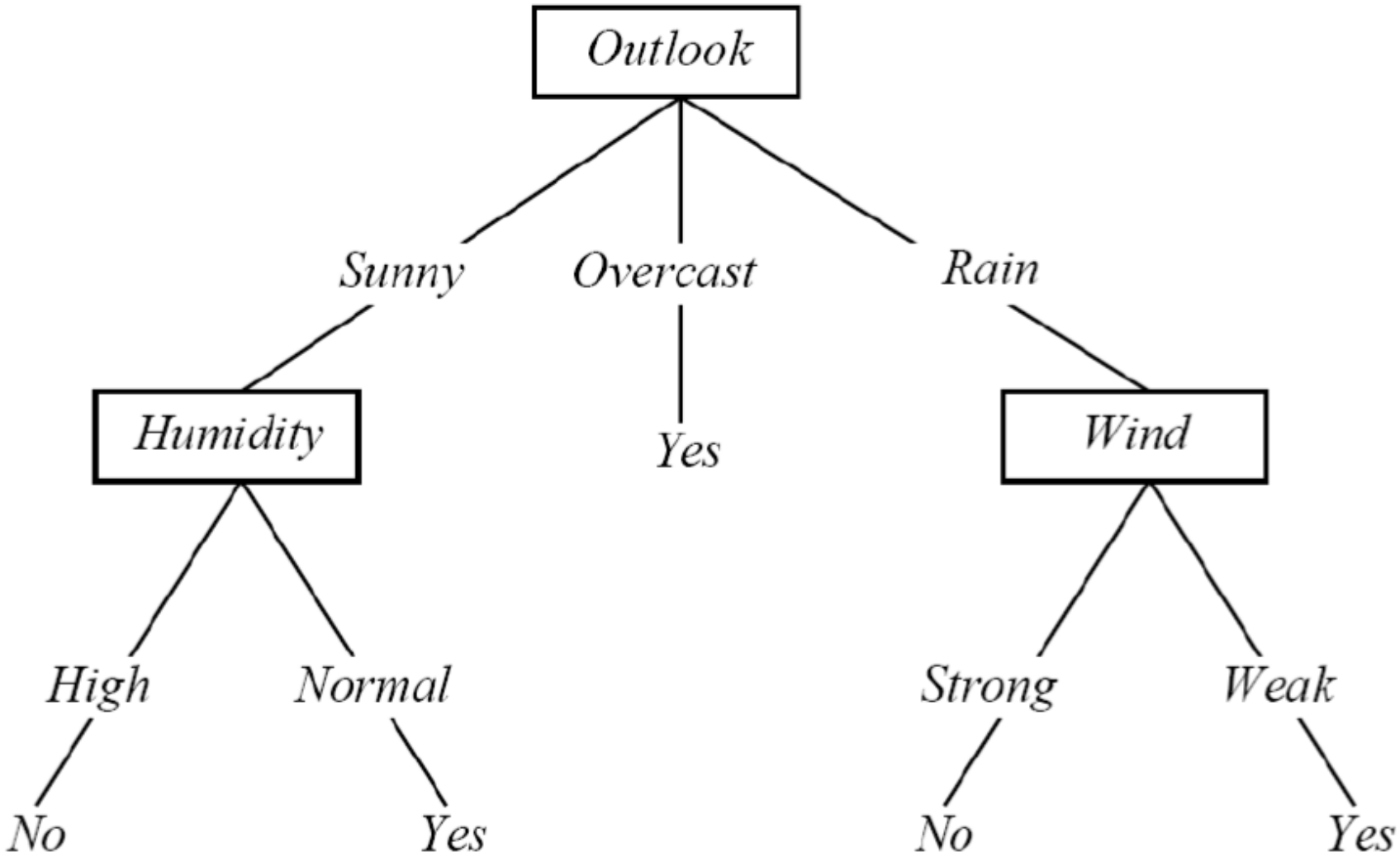
Gain (S, Humidity)

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

Gain (S, Wind)

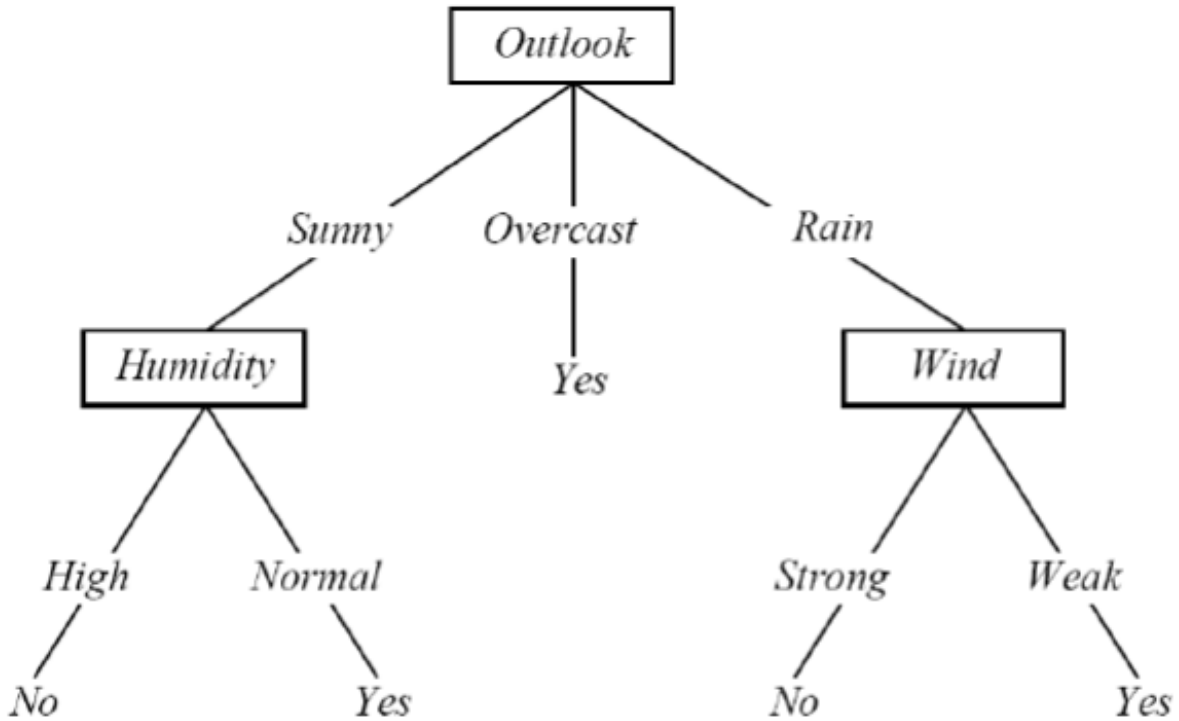
= .940 - (8/14).811 - (6/14)1.0
= .048

Decision Tree



Is the decision tree correct?

- Let's check whether the split on Wind attribute is correct.
- We need to show that Wind attribute has the highest information gain.



To select node following Rain path:

Wind attribute – 5 records match

Note: calculate the entropy only on examples that got "routed" in our branch of the tree (Outlook=Rain)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1					No
D2					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

Calculation

- $S = \{D4, D5, D6, D10, D14\}$
- Entropy:

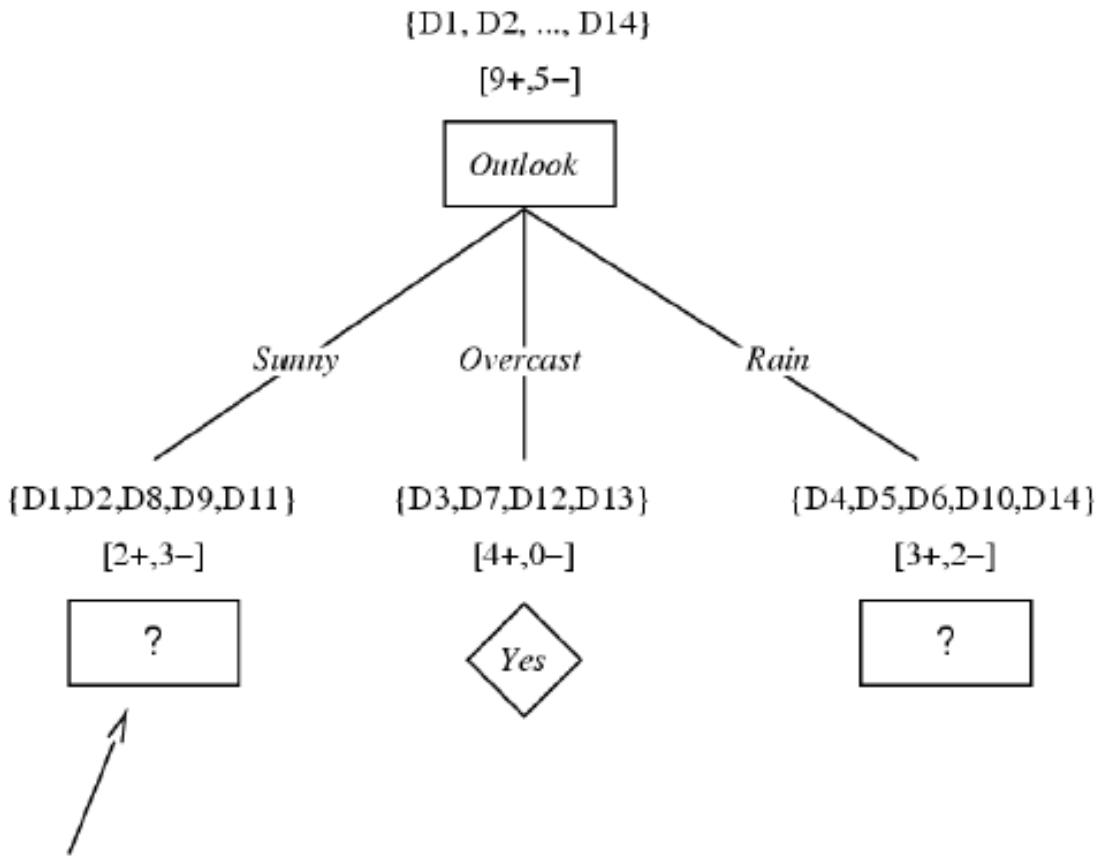
  $H(S) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971$
- Information Gain

  $IG(S, Temp) = H(S) - H(S|Temp) = 0.01997$

  $IG(S, Humidity) = H(S) - H(S|Humidity) = 0.01997$

  $IG(S, Wind) = H(S) - H(S|Wind) = 0.971$

To select node following Sunny path:



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

References

 Tom Mitchell, Machine learning. McGraw-Hill, 1997