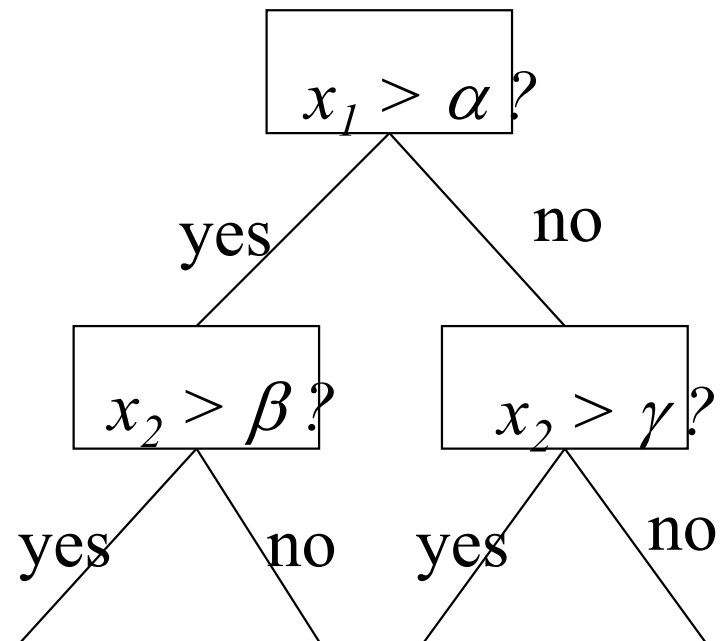


Learning from Observation

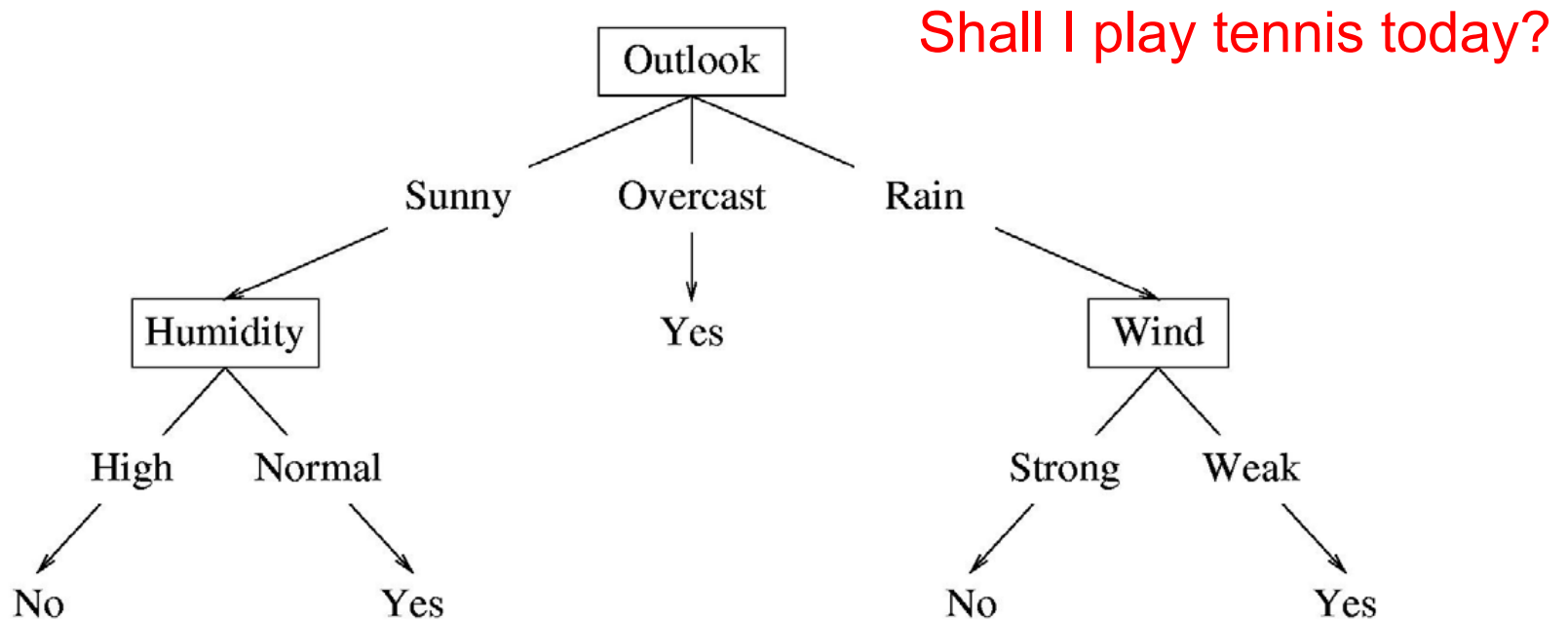
Method: Decision trees

- “Divide and conquer”:
Split data into smaller
and smaller subsets.
- Splits usually on a
single variable



Decision Tree Hypothesis Space

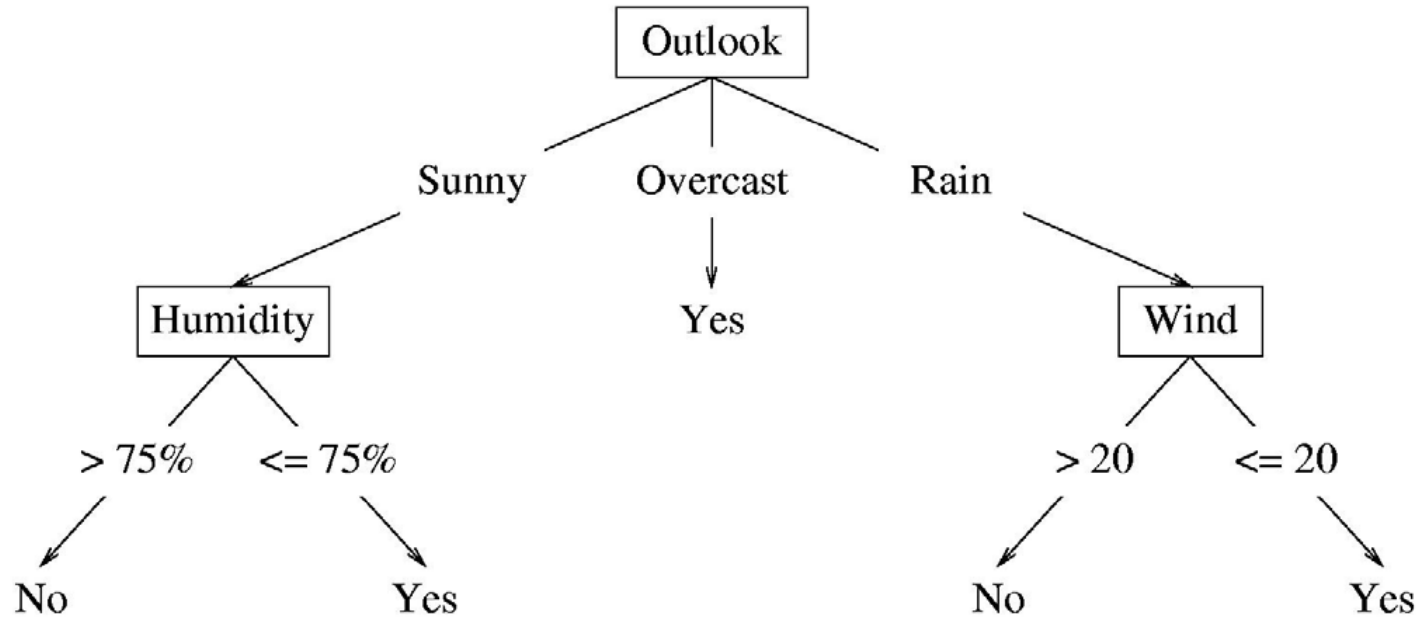
- **Internal nodes** test the value of particular features x_j and branch according to the results of the test.
- **Leaf nodes** specify the class $h(\mathbf{x})$.



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

Decision Tree Hypothesis Space

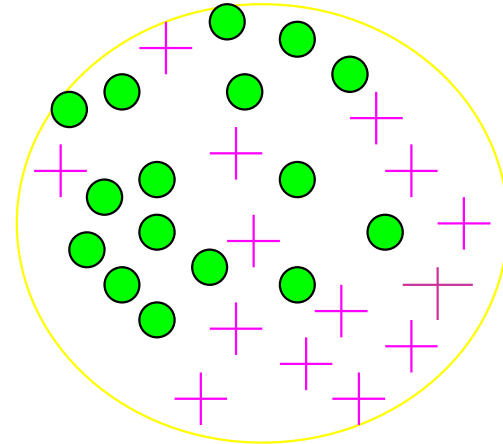
If the features are continuous, internal nodes may test the value of a feature against a threshold.



Entropy: a common way to measure impurity

- Entropy =
$$\sum_i -p_i \log_2 p_i$$

p_i is the probability of class i



16/30 are green circles; 14/30 are pink crosses

$\log_2(16/30) = -.9$; $\log_2(14/30) = -1.1$

Entropy = $-(16/30)(-.9) - (14/30)(-1.1) = .99$

- Entropy comes from information theory. The higher the entropy the more the information content.

What does that mean for learning from examples?

Entropy

The entropy is maximal when all possibilities are equally likely.

The goal of the decision tree is to decrease the entropy in each node.

Entropy is zero in a pure "yes" node (or pure "no" node).

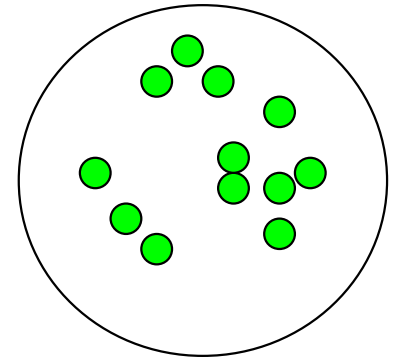
Entropy

- What is the entropy of a group in which all examples belong to the same class?

– entropy = $-1 \log_2 1 = 0$

not a good training set for learning

Minimum impurity

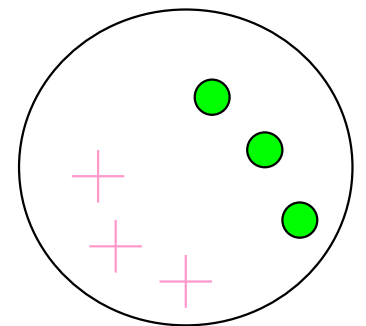


- What is the entropy of a group with 50% in either class?

– entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

good training set for learning

Maximum impurity



Information Gain

- We want to determine **which attribute** in a given set of training feature vectors is **most useful** for discriminating between the classes to be learned.
- **Information gain** tells us how important a given attribute of the feature vectors is.
- We will use it to decide the ordering of attributes in the nodes of a decision tree.

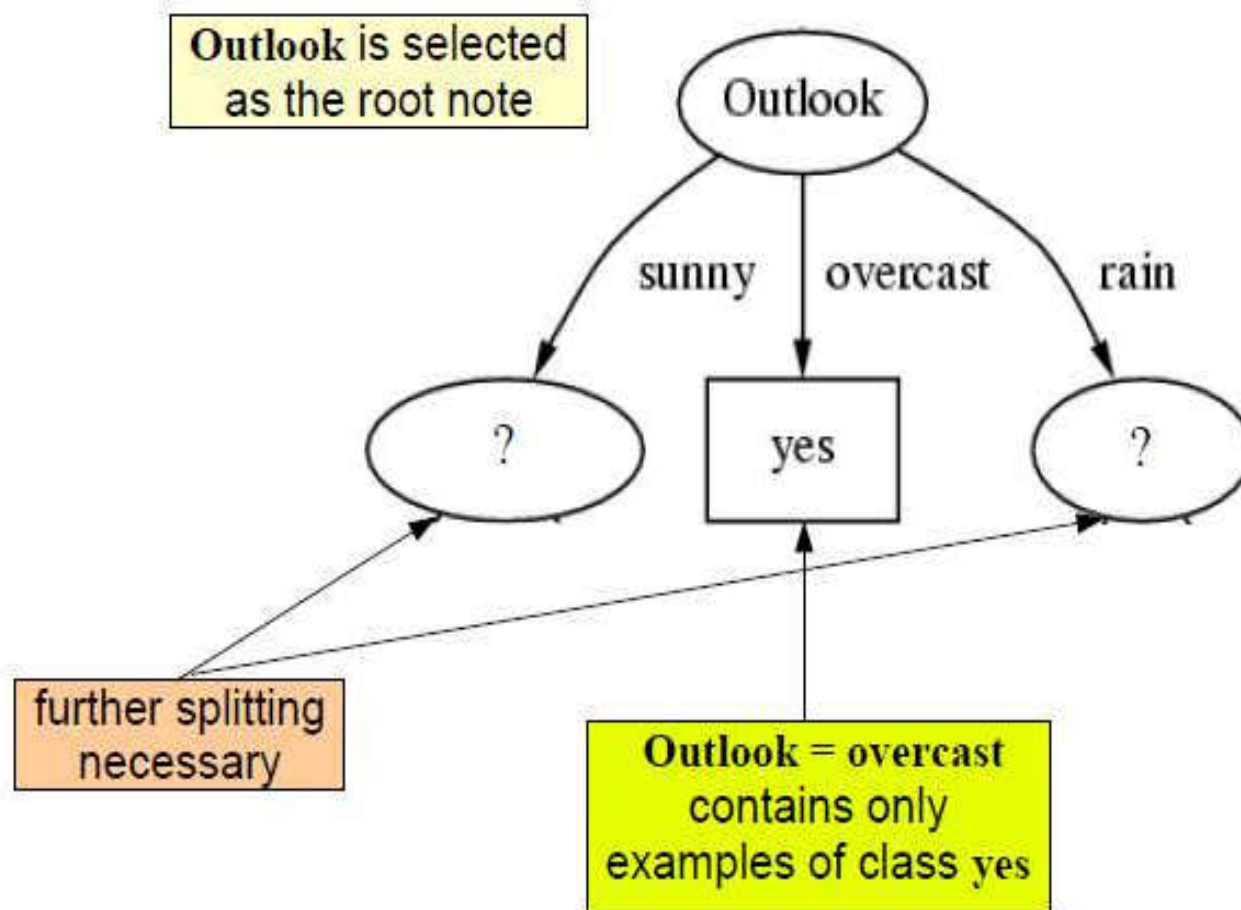
Decision tree learning algorithm

- Create pure nodes whenever possible
- If pure nodes are not possible, choose the split that leads to the highest gain.

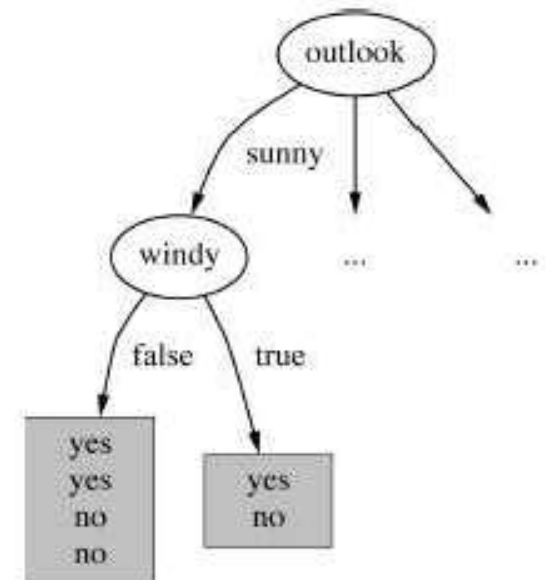
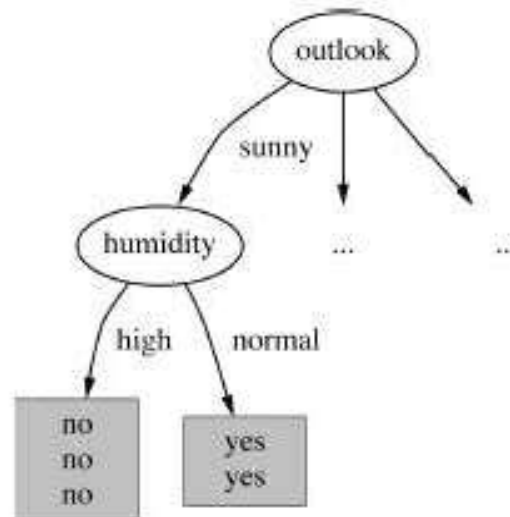
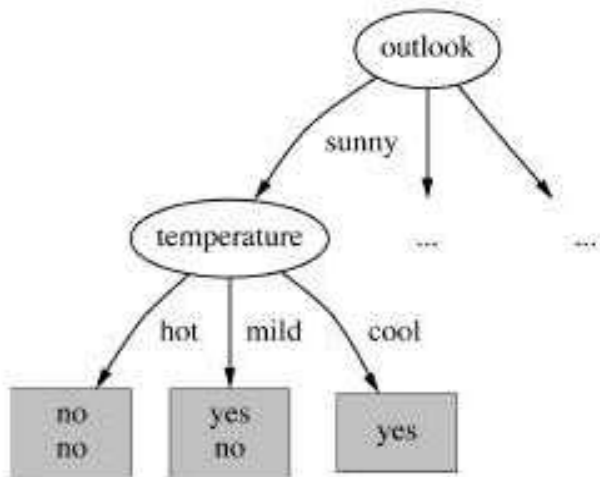
Playing Tennis Example

Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Example (Cont...)



Example (Cont...)



$Gain(Temperature)$

$= 0.571$ bits

$Gain(Humidity)$

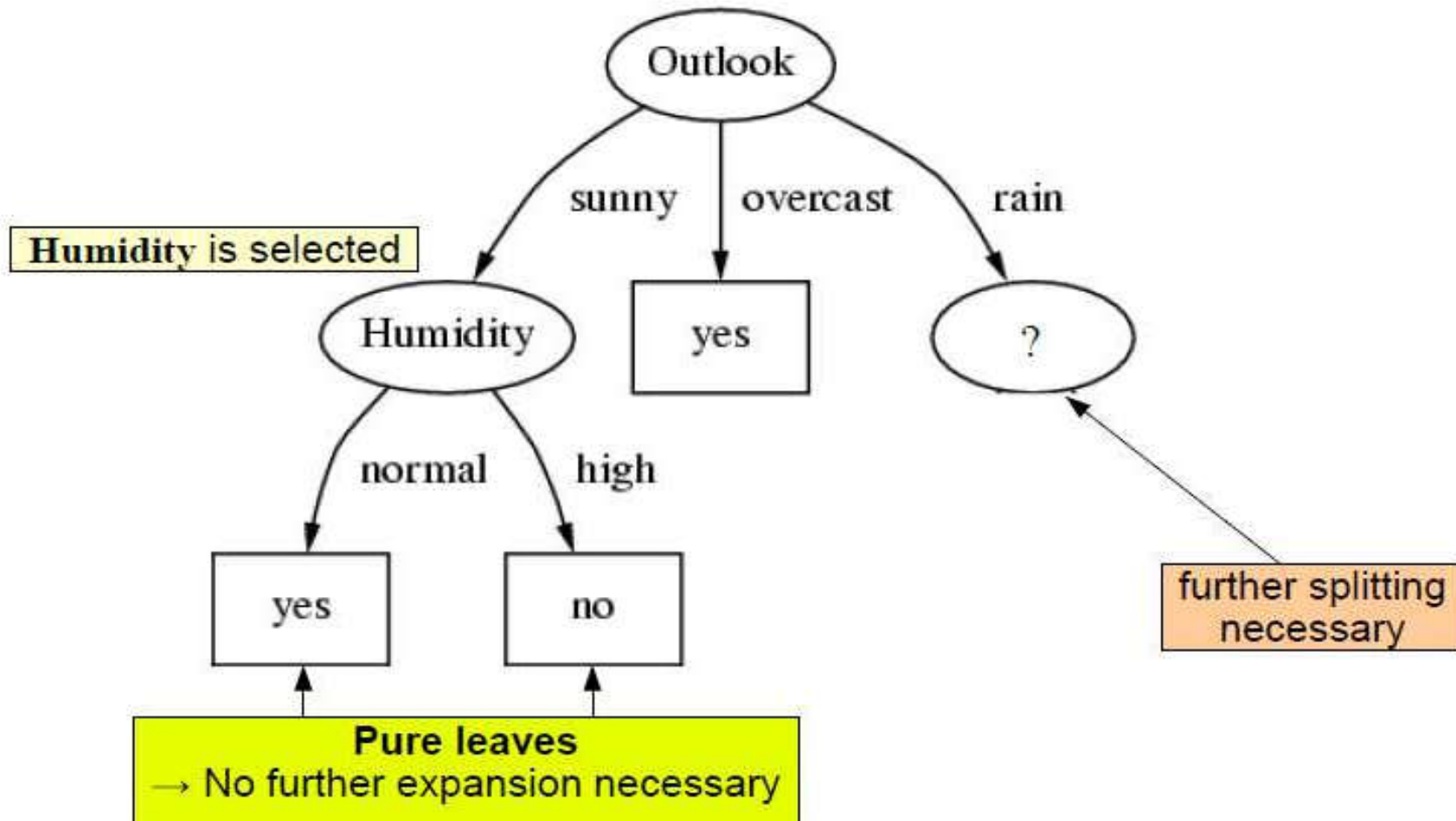
$= 0.971$ bits

$Gain(Windy)$

$= 0.020$ bits

Humidity is selected

Example (Cont...)



Final Decision Tree

