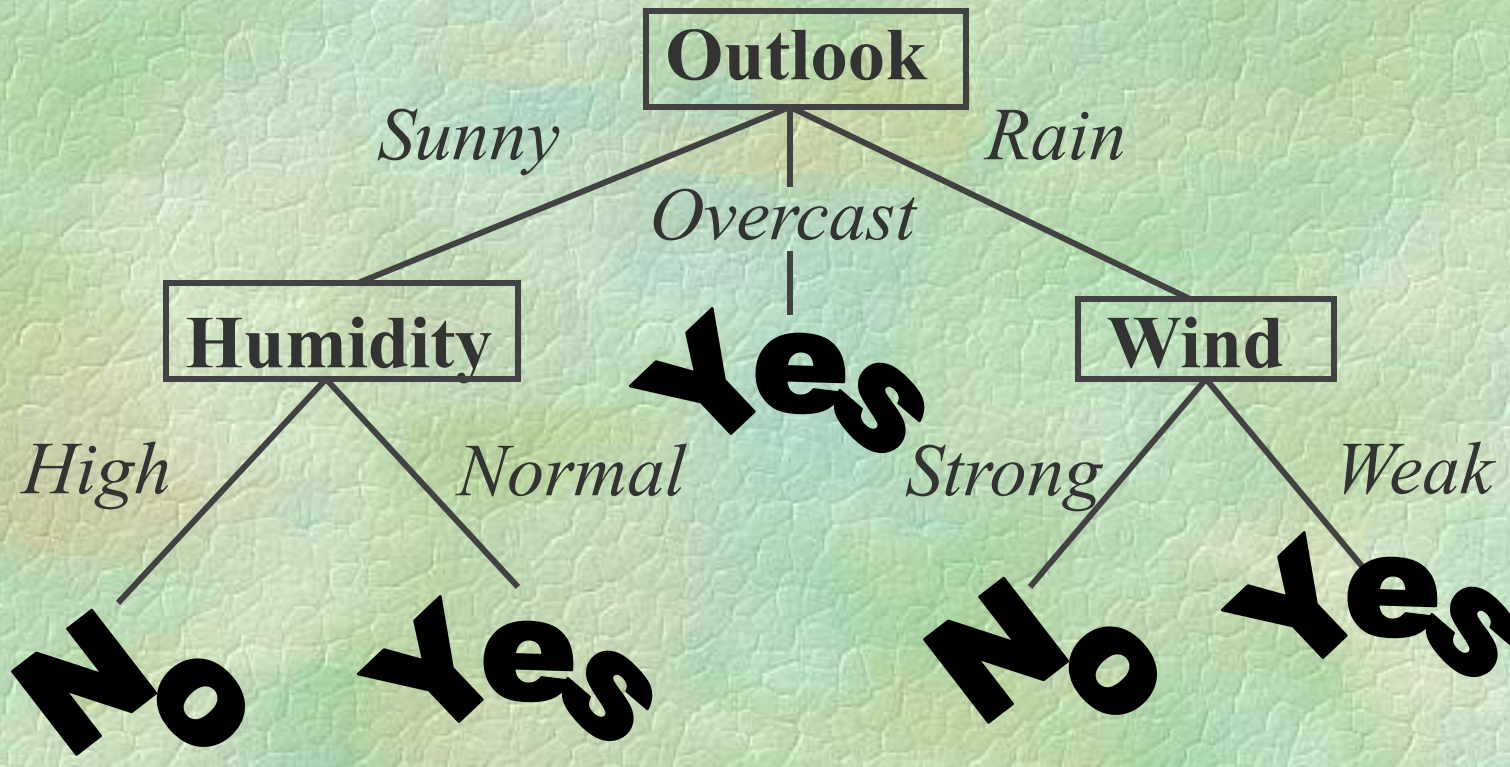
Machine Learning: Lecture 3

Decision Tree Learning

(Based on Chapter 3 of Mitchell T.., Machine Learning, 1997)

Decision Tree Representation



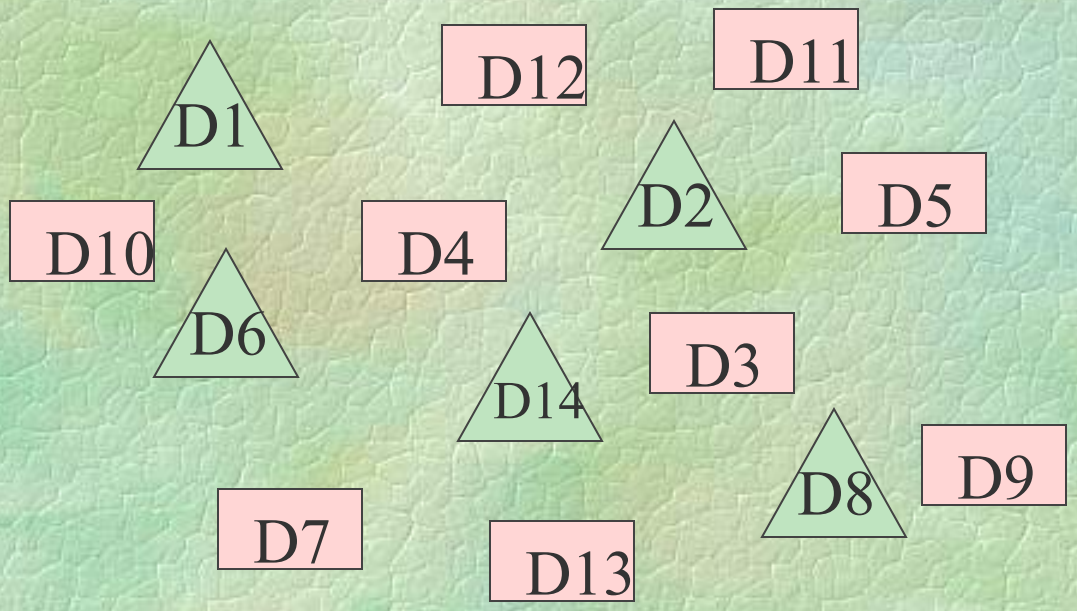
A Decision Tree for the concept *PlayTennis*

Appropriate Problems for Decision Tree Learning

- Instances are represented by discrete attribute-value pairs (though the basic algorithm was extended to real-valued attributes as well)
- The target function has discrete output values (can have more than two possible output values --> classes)
- Disjunctive hypothesis descriptions may be required
- The training data may contain errors
- The training data may contain missing attribute values

ID3: The Basic Decision Tree Learning Algorithm

- Database, See [Mitchell, p. 59]

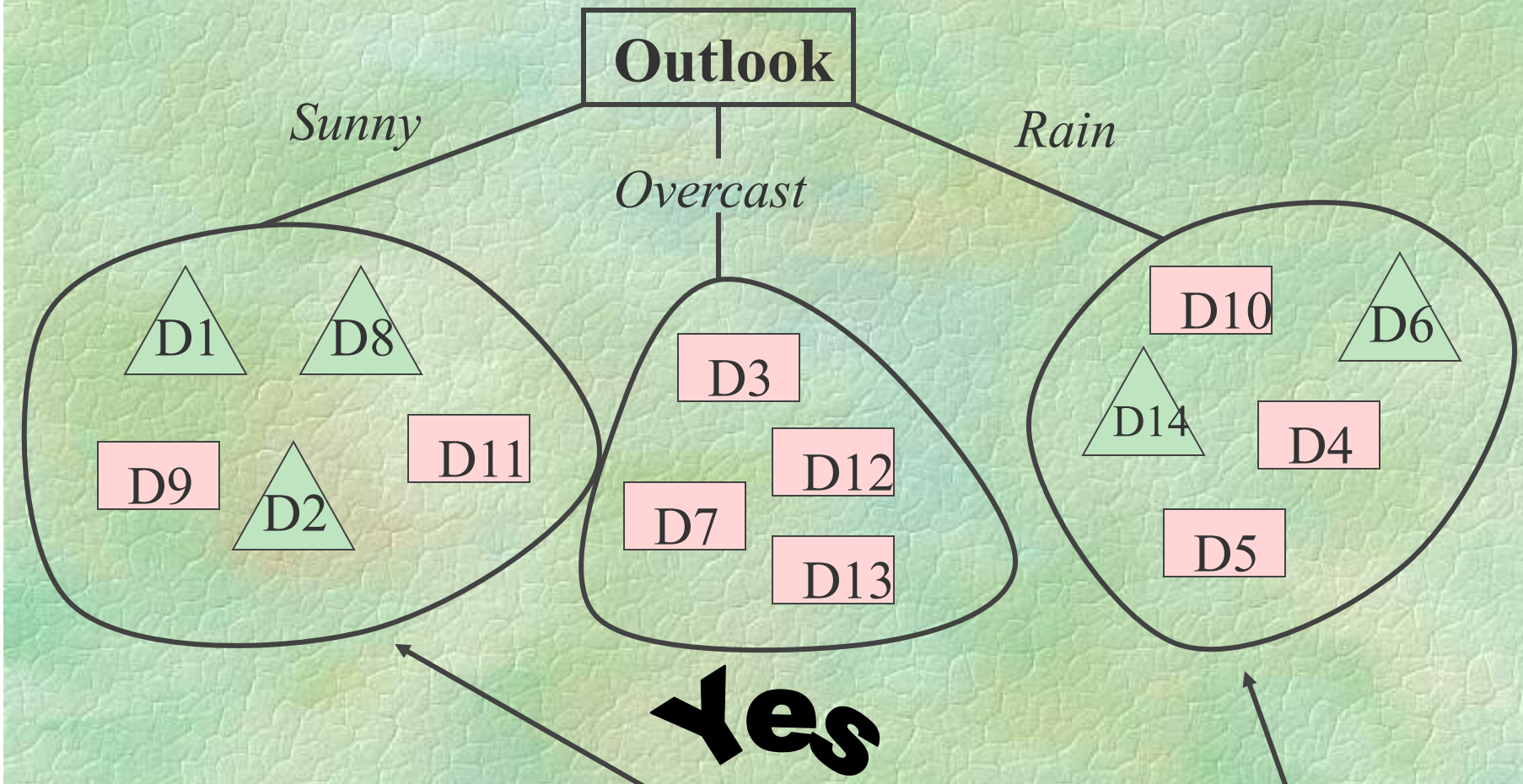


What is the "best" attribute?

Answer: Outlook

["best" = with highest information gain]

ID3 (Cont'd)



What are the "best" attributes? **Humidity** and **Wind**

What Attribute to choose to "best" split a node?

- Choose the attribute that minimize the **Disorder (or Entropy)** in the subtree rooted at a given node.
- **Disorder** and **Information** are related as follows: the more disorderly a set, the more information is required to correctly guess an element of that set.
- **Information:** What is the best strategy for guessing a number from a finite set of possible numbers? i.e., how many questions do you need to ask in order to know the answer (we are looking for the minimal number of questions). Answer Log_2(S), where S is the set of numbers and |S|, its cardinality.

Q1: is it smaller than 5?
Q2: is it smaller than 2?

E.g.: 0 1 2 3 4 5 6 7 8 9 10

Q2 Q1

What Attribute to choose to "best" split a node? (Cont'd)

- Log_2 |S| can also be thought of as the information value of being told x (the number to be guessed) instead of having to guess it.
- Let U be a subset of S. What is the informatin value of being told x after finding out whether or not x∈ U? **Ans: Log_2|S|-[P(x ∈ U) Log_2|U|+ P(s ∉U) Log_2|S-U|**
- Let S = P ∪N (positive and negative data). The information value of being told x after finding out whether x ∈ U or x ∈ N is **I({P,N})=Log_2(|S|)-|P|/|S| Log_2|P| -|N|/|S| Log_2|N|**

What Attribute to choose to "best" split a node? (Cont'd)

- We want to use this measure to choose an attribute that minimizes the disorder in the partitions it creates. Let $\{S_i \mid 1 \le i \le n\}$ be a partition of S resulting from a particular attribute. The disorder associated with this partition is:

$$\mathbf{V(\{S_i \mid 1 \le i \le n\}) = \Sigma |S_i|/|S| . I(\{P(S_i), N(S_i)\})}$$

Set of positive examples in S_i

Set of negative examples in S_i

Hypothesis Space Search in Decision Tree Learning

- **Hypothesis Space**: Set of possible decision trees (i.e., complete space of finie discrete-valued functions).
- **Search Method**: Simple-to-Complex *Hill-Climbing* Search (only a single current hypothesis is maintained (≠ from candidate-elimination method)). **No Backtracking!!!**
- **Evaluation Function**: Information Gain Measure
- **Batch Learning:** ID3 uses all training examples at each step to make statistically-based decisions (≠ from candidate-elimination method which makes decisions incrementally). ==> the search is less sensitive to errors in individual training examples.

Inductive Bias in Decision Tree Learning

- **ID3's Inductive Bias:** *Shorter* trees are preferred over longer trees. Trees that place *high information gain attributes close to the root* are preferred over those that do not.
- **Note:** this type of bias is different from the type of bias used by Candidate-Elimination: the inductive bias of ID3 follows from its search strategy (***preference*** or ***search*** bias) whereas the inductive bias of the Candidate-Elimination algorithm follows from the definition of its hypothesis space (***restriction*** or ***language*** bias).

Why Prefer Short Hypotheses?

- **Occam's razor:**
  Prefer the simplest hypothesis that fits the data
  [William of Occam (Philosopher), circa 1320]
- Scientists seem to do that: E.g., Physicist seem to prefer simple explanations for the motion of planets, over more complex ones
- **Argument:** Since there are fewer short hypotheses than long ones, it is less likely that one will find a short hypothesis that coincidentally fits the training data.
- **Problem with this argument:** it can be made about many other constraints. Why is the "short description" constraint more relevant than others?
- **Nevertheless:** Occam's razor was shown experimentally to be a successful strategy!

Issues in Decision Tree Learning: I. Avoiding Overfitting the Data

- **Definition:** Given a hypothesis space H, a hypothesis $h \in H$ is said to ***overfit*** the training data if there exists some alternative hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances. (See curves in [Mitchell, p.67])
- There are two approaches for overfitting avoidance in Decision Trees:
  - Stop growing the tree before it perfectly fits the data
  - Allow the tree to overfit the data, and then ***post-prune*** it.

Issues in Decision Tree Learning: II. Other Issues

- Incorporating Continuous-Valued Attributes
- Alternative Measures for Selecting Attributes
- Handling Training Examples with Missing Attribute Values
- Handling Attributes with Differing Costs