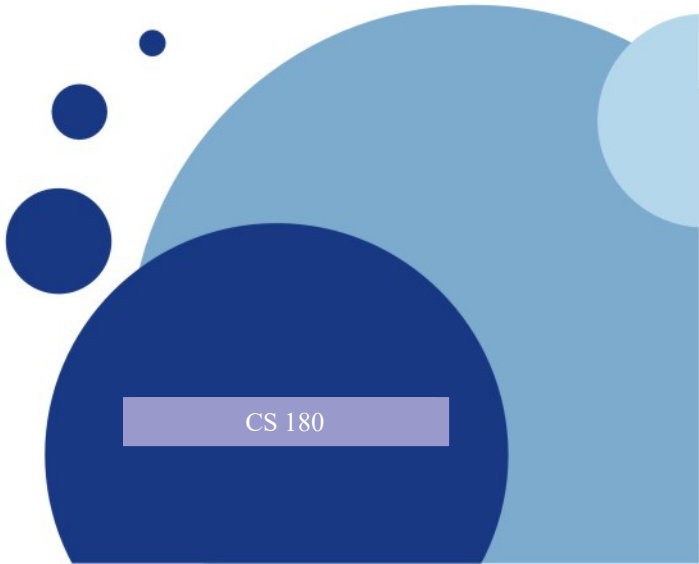
AI

Decision Trees



CS 180

AI

Today: Decision Trees

- Decision tree representation
- ID3 learning algorithm
- Entropy, information gain

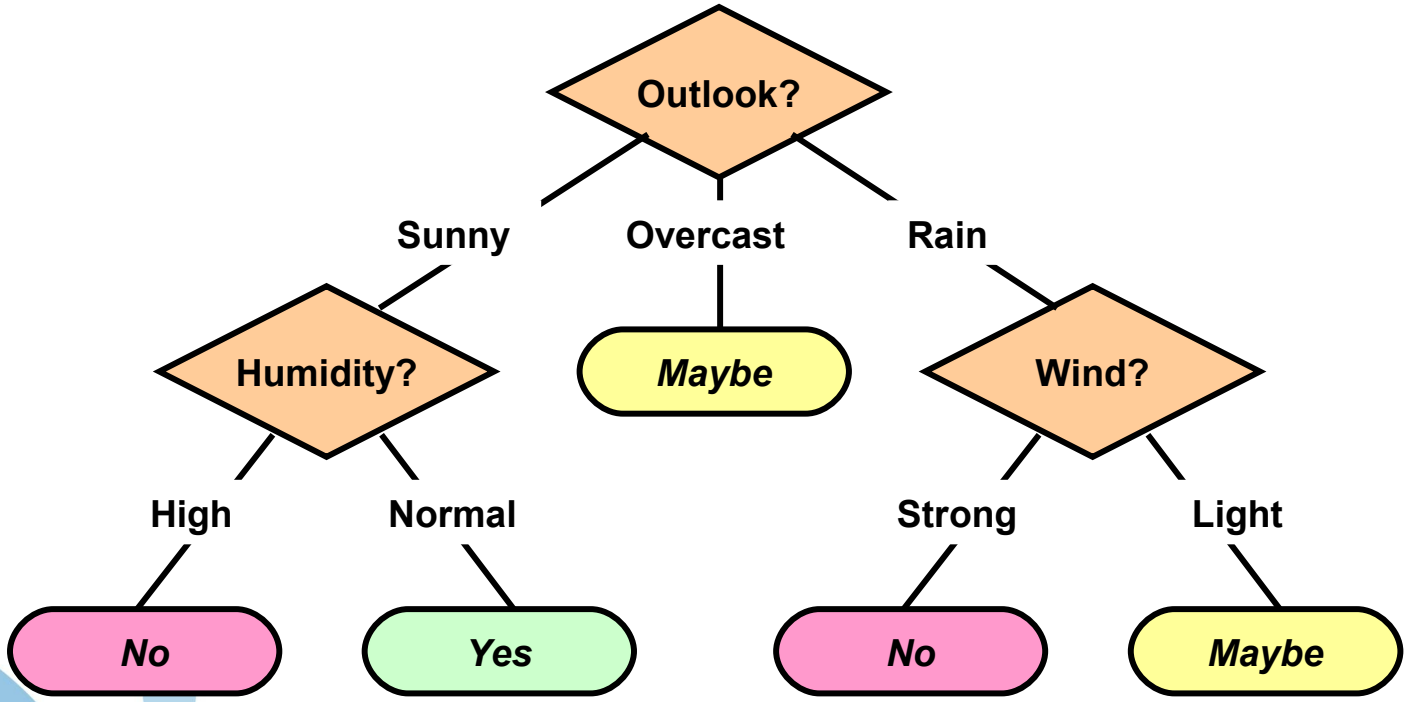
AI

Tree Classifiers

- The terminology Tree is graphic
- A decision tree grows from the root downward
- The idea is to send the examples down the tree, using the concept of information entropy

AI

Decision Tree for PlayTennis



AI

Decision Trees

Decision tree representation:

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification

AI

When to Consider Using Decision Trees

- Instances describable by attribute-value pairs
- Target function is discrete valued
- Disjunctive hypothesis may be required
- Possibly noisy training data

- Examples:
  - Equipment or medical diagnosis
  - Credit-risk analysis

AI

Decision Tree Learning: Top-Down Induction (*ID3*)

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

AI

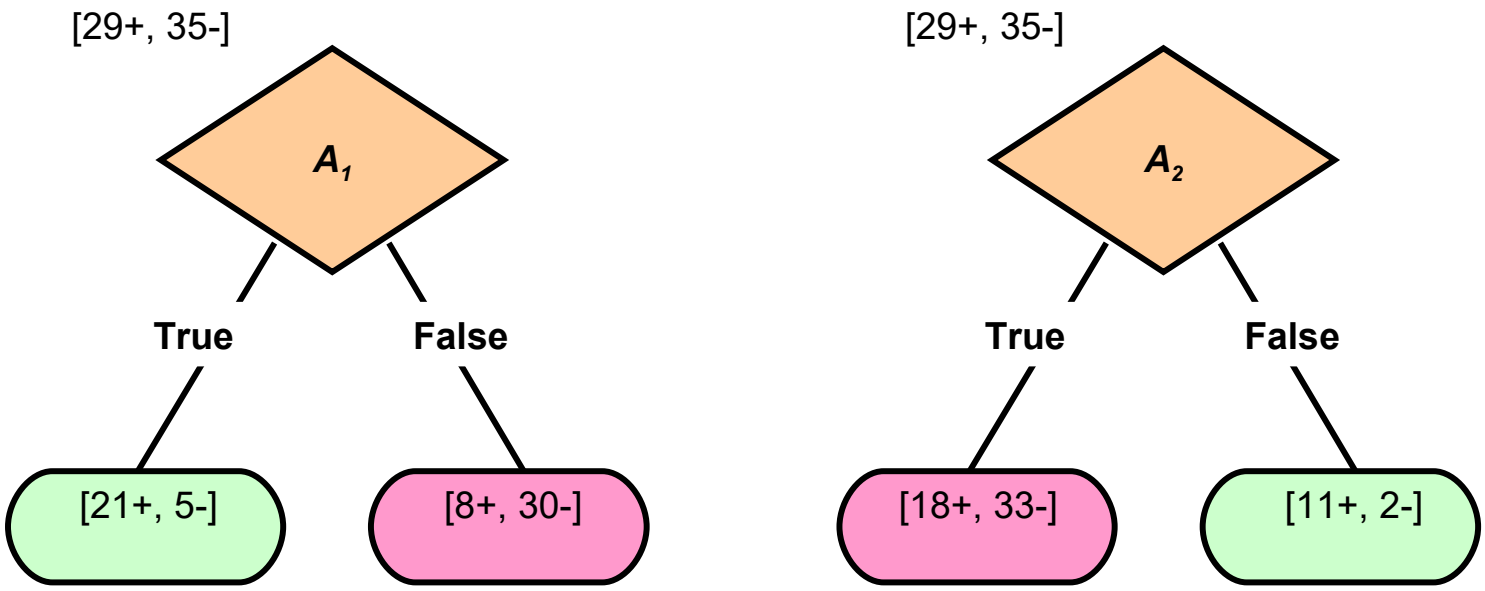
Decision Tree Learning: Top-Down Induction (*ID3*)

Main Loop:

1. A ← the "best" decision attribute for next *node*
2. Assign A as decision attribute for *node*
3. For each value of A, create new descendant of *node*
4. Sort training examples to leaf nodes
5. If training examples perfectly classified, Then STOP, else iterate over new leaf nodes

AI

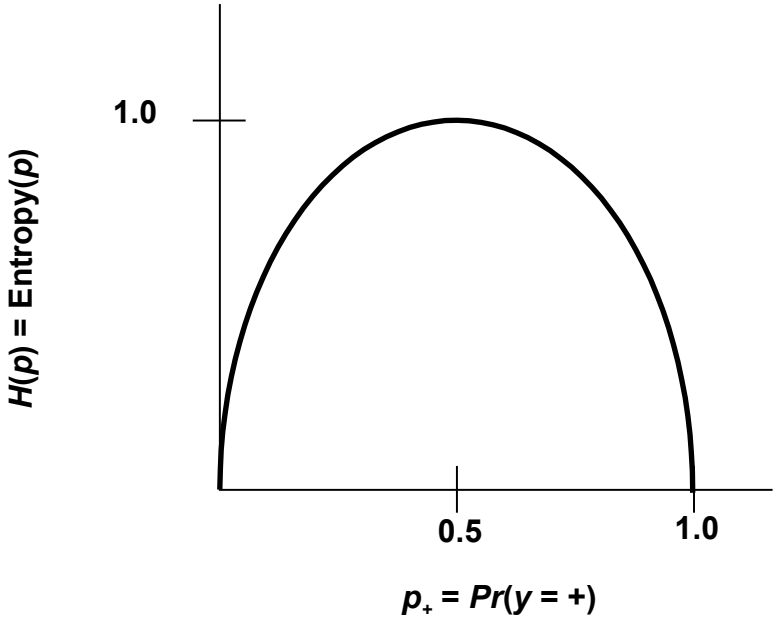
But which attribute is best?



AI

Entropy: A measure of homogeneity

- S is a sample of training examples
- p_+ is the proportion of positive examples
- p_- is the proportion of negative examples
- Entropy measures the impurity of S
- Entropy(S) = $-p_+ \log_2 p_+ - p_- \log_2 p_-$



AI

Entropy: A measure of homogeneity

- In general, for c classes:

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

AI

Entropy Example

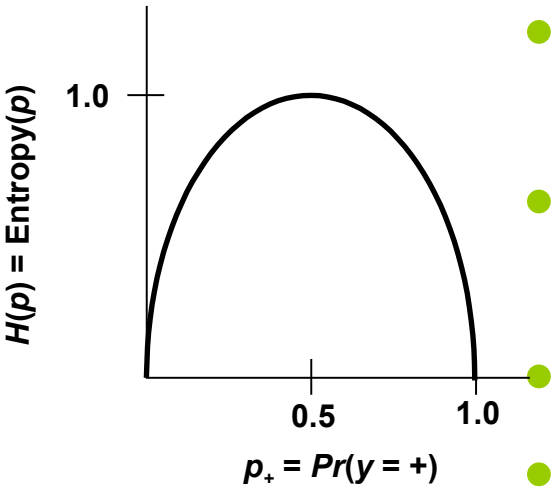
- Suppose S is a collection of 14 examples of some boolean concept
- There are 9 positive and 5 negative examples ([9+, 5-])
- Then the entropy of S relative to this boolean classification is

$$\text{Entropy}([9+,5-]) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$$

$$= 0.940$$

AI

Entropy



- Entropy is 0 if all members belong to the same class
- Example: if all members are positive, then p- is 0
- Entropy(S) = $-1\log_2(1) - 0\log_2(0) = 0$
- The entropy is 1 if the collection contains an equal number of positive and negative examples
- If the collection contains unequal numbers of positive and negative examples, the entropy is between 0 and 1

AI

Information Gain

- Gain(S,A): expected reduction in entropy due to sorting S on attribute A

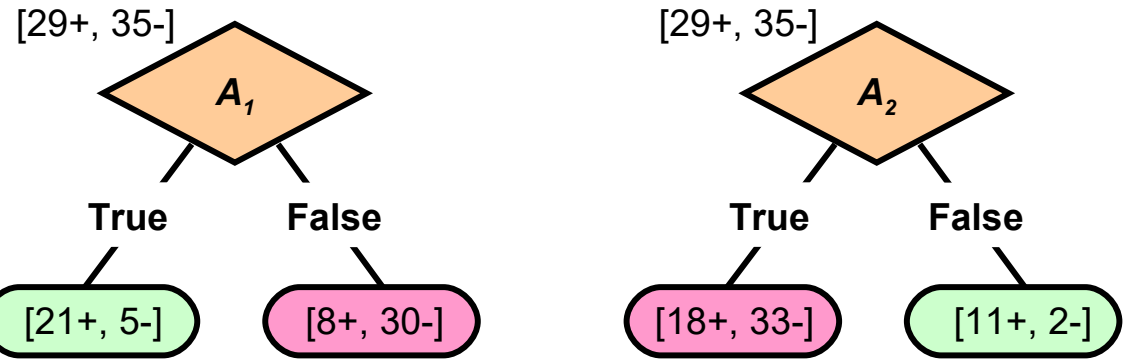
$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

  - where S_v is $\{x \in S: x.A = v\}$, the set of examples in S where attribute *A* has value *v*

AI

Information Gain

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

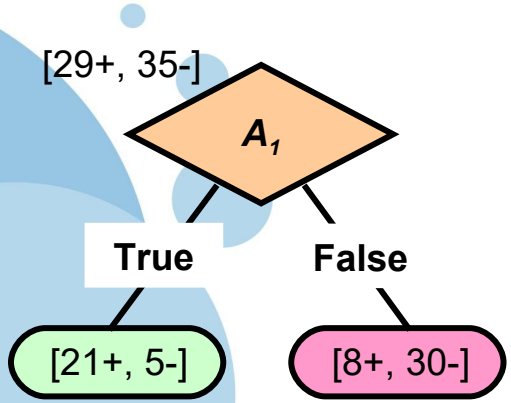


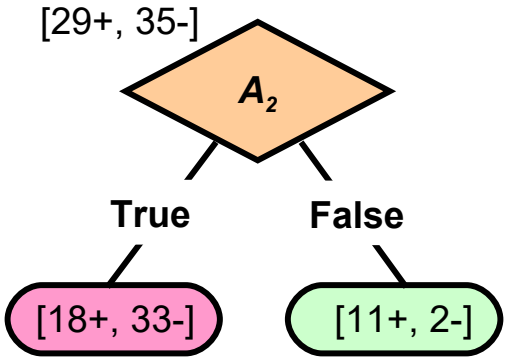
AI

Information Gain

- Entropy([21+,5-]) = 0.71
- Entropy([8+,30-]) = 0.74
- Gain(S,A1)=Entropy(S)
  -26/64*Entropy([21+,5-])
  -38/64*Entropy([8+,30-])
- =0.27

- Entropy([18+,33-]) = 0.94
- Entropy([11+,2-]) = 0.62
- Gain(S,A2)=Entropy(S)
  -51/64*Entropy([18+,33-])
  -13/64*Entropy([11+,2-])
- =0.12





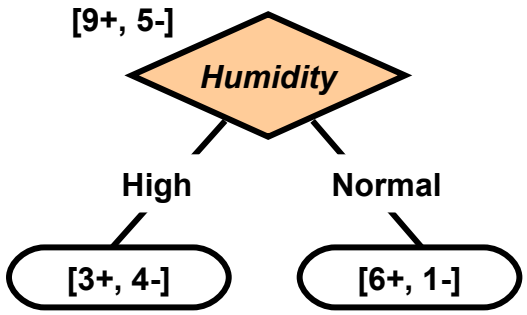
AI

An Illustrative Example: PlayTennis

Day	Outlook	Temperature	Humidity	Wind	*PlayTennis*?
1	Sunny	Hot	High	Light	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Light	Yes
4	Rain	Mild	High	Light	Yes
5	Rain	Cool	Normal	Light	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Light	No
9	Sunny	Cool	Normal	Light	Yes
10	Rain	Mild	Normal	Light	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Light	Yes
14	Rain	Mild	High	Strong	No

AI

Constructing A Decision Tree for *PlayTennis* using *ID3* [1]

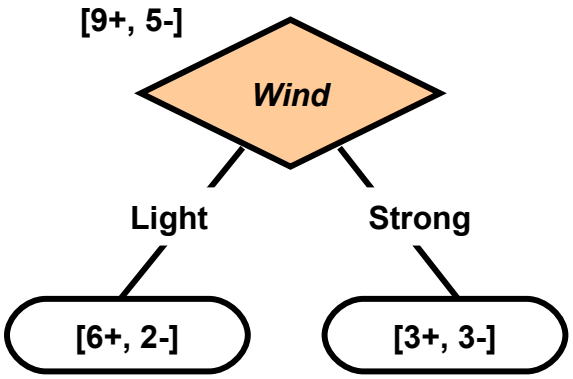


- Prior (unconditioned) distribution: 9+, 5-
  - *H*(S) = -(9/14) lg (9/14) - (5/14) lg (5/14) = 0.94
  - *H*(S, *Humidity* = High) = -(3/7) lg (3/7) - (4/7) lg (4/7) = 0.985
  - *H*(S, *Humidity* = Normal) = -(6/7) lg (6/7) - (1/7) lg (1/7) = 0.592
  - *Gain*(S, *Humidity*) = 0.94 - ((7/14) * 0.985 + (7/14) * 0.592) = 0.151

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

AI

Constructing A Decision Tree for *PlayTennis* using *ID3* [2]

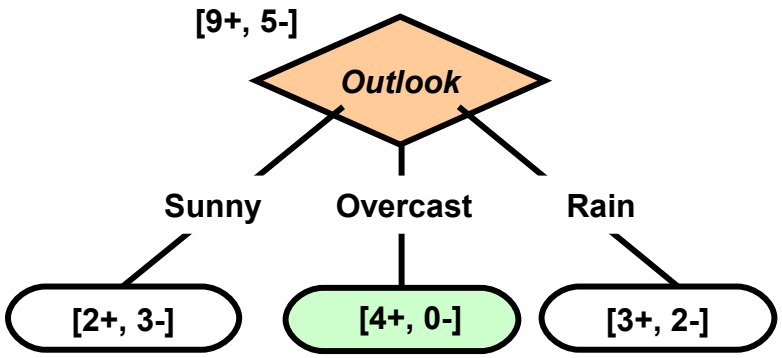


- *Gain* (S, *Wind*) = 0.94 - ((8/14) * 0.811 + (6/14) * 1.0) = 0.048

AI

Constructing A Decision Tree for *PlayTennis* using *ID3 [3]*

- Selecting the root attribute
  - *Gain*(*D*, *Humidity*) = 0.151
  - *Gain*(*D*, *Wind*) = 0.048
  - *Gain*(*D*, *Temperature*) = 0.029
  - *Gain*(*D*, *Outlook*) = 0.246



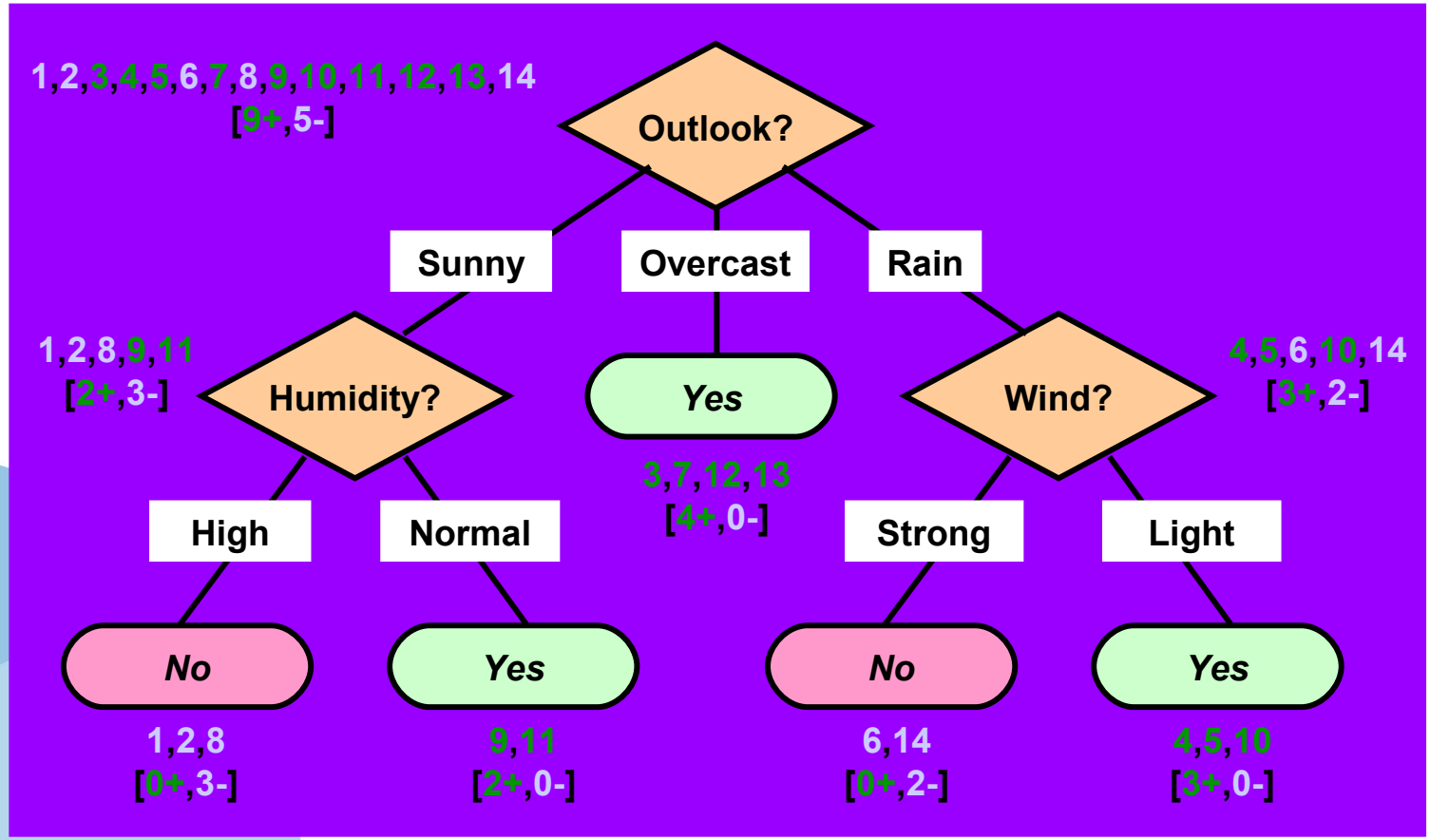
AI

Constructing A Decision Tree for *PlayTennis* using *ID3* [4]

- Selecting the next attribute (root of subtree)
  - Convention: lg (0/a) = 0
  - *Gain*(S_{Sunny}, *Humidity*) = 0.97 - ((3/5) * 0 - (2/5) * 0) = <u>0.97</u>
  - *Gain*(S_{Sunny}, *Wind*) = 0.97 - ((2/5) * 1 - (3/5) * 0.92) = 0.02
  - *Gain*(S_{Sunny}, *Temperature*) = 0.57

AI

Constructing A Decision Tree for *PlayTennis* using *ID3* [5]



AI

Hypothesis Space Search ID3

- Hypothesis space is complete!
  - Target function surely in there…
- Outputs a single hypothesis
- No backtracking on selected attributes (greedy search)
  - Local minima (suboptimal splits)
- Statistically-based search choices
  - Robust to noisy data
- Inductive bias (search bias)
  - Prefer shorter trees over longer ones
  - Place high information gain attributes close to the root

AI

Inductive Bias in ID3

- H is the power set of instances X
- Preference for short trees, and for those with high information gain attributes near the root
- Bias is a *preference* for some hypotheses (ID3), rather than a *restriction* of the hypothesis space H (Candidate-elimination)
- Occam's razor: prefer the shortest (simplest) hypothesis that fits the data

AI

Occam's Razor

Why prefer short hypotheses?

Argument in favor:

- Fewer short hypotheses than long hypotheses
- A short hypothesis that fits the data is unlikely to be a coincidence
- A long hypothesis that fits the data might be a coincidence

Argument opposed:

- There are many ways to define small sets of hypotheses
- What is so special about small sets based on *size* of hypothesis?

AI

Summary

- Decision trees
- ID3
- Entropy, Information gain

- Announcements:
  - No class on Friday, March 17, 2017 (deadline for MP1). Use this time to finish your MP1.
  - Quiz on decision trees on Wednesday, March 22, 2017. Bring calculator.