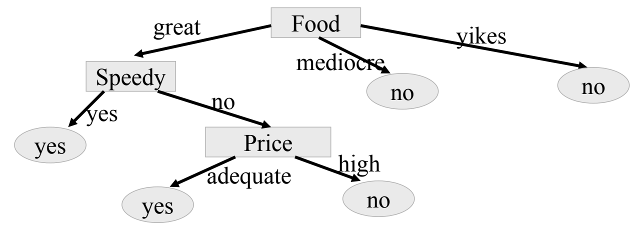
Decision Tree Learning

CS472/CS473 – Fall 2005

Decision Tree Example: BigTip



	Food (3)	Chat (2)	Speedy (2)	Price (2)	Bar (2)	BigTip
1	great	yes	yes	adequate	no	yes
2	great	no	yes	adequate	no	yes
3	mediocre	yes	no	high	no	no
4	great	yes	yes	adequate	yes	yes

Top-Down Induction of DT (simplified)

Training Data: $D = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)\}$

TDIDF(D,c_{def})

- IF(all examples in D have same class c)
  - Return leaf with class c (or class c_{def}, if D is empty)
- ELSE IF(no attributes left to test)
  - Return leaf with class c of majority in D
- ELSE
  - Pick A as the "best" decision attribute for next node
  - FOR each value v_i of A create a new descendent of node
    - $D_i = \{(\vec{x}, y) \in D : \text{attribute A of } \vec{x} \text{ has value } v_i\}$
    - Subtree t_i for v_i is TDIDT(D_i,c_{def})
  - RETURN tree with A as root and t_i as subtrees

Example: Text Classification

- **Task: Learn rule that classifies Reuters Business News**
  - Class +: "Corporate Acquisitions"
  - Class -: Other articles
  - 2000 training instances
- **Representation:**
  - Boolean attributes, indicating presence of a keyword in article
  - 9947 such keywords (more accurately, word "stems")

LAROCHE STARTS BID FOR NECO SHARES +

Investor David F. La Roche of North Kingstown, R.I., said he is offering to purchase 170,000 common shares of NECO Enterprises Inc at 26 dlrs each. He said the successful completion of the offer, plus shares he already owns, would give him 50.5 pct of NECO's 962,016 common shares. La Roche said he may buy more, and possible all NECO shares. He said the offer and withdrawal rights will expire at 1630 EST/2130 gmt, March 30, 1987.

SALANT CORP 1ST QTR FEB 28 NET -

Oper shr profit seven cts vs loss 12 cts. Oper net profit 216,000 vs loss 401,000. Sales 21.4 mln vs 24.9 mln. NOTE: Current year net excludes 142,000 dlr tax credit. Company operating in Chapter 11 bankruptcy.

Example: TDIDT

TDIDF(D,c_{def})

- IF(all examples in D have same class c)
  - Return leaf with class c (or c_{def}, if D=Ø)
- ELSE IF(no attributes left to test)
  - Return leaf with class c of majority in D
- ELSE
  - A ← "best" decision attribute for node
  - FOR each value v_i of A create a new descendent of node
    - $D_i = \{(\vec{x}, y) \in D : \text{attrib. A of } \vec{x} \text{ has val. } v_i\}$
    - Subtree t_i for v_i is TDIDT(D_i,c_{def})
  - RETURN tree with A as root and t_i as subtrees

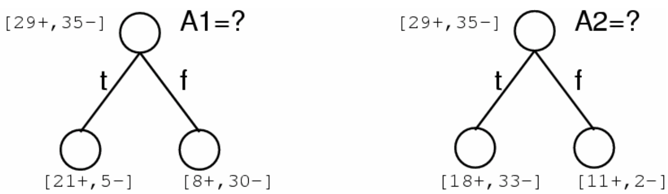
Training Data D:

F S P	BigTip
$\vec{x}_1 = (\text{g, y, a})$	$f(\vec{x}_1) = 1$
$\vec{x}_2 = (\text{g, n, h})$	$f(\vec{x}_2) = 0$
$\vec{x}_3 = (\text{g, y, h})$	$f(\vec{x}_3) = 1$
$\vec{x}_4 = (\text{g, n, a})$	$f(\vec{x}_4) = 1$
$\vec{x}_5 = (\text{m, y, a})$	$f(\vec{x}_5) = 0$
$\vec{x}_6 = (\text{y, y, a})$	$f(\vec{x}_6) = 0$
$\vec{x}_7 = (\text{g, y, a})$	$f(\vec{x}_7) = 1$
$\vec{x}_8 = (\text{g, y, h})$	$f(\vec{x}_8) = 1$
$\vec{x}_9 = (\text{m, y, a})$	$f(\vec{x}_9) = 0$
$\vec{x}_{10} = (\text{g, y, a})$	$f(\vec{x}_{10}) = 1$

Picking the Best Attribute to Split

- **Ockham's Razor:**
  - All other things being equal, choose the simplest explanation
- **Decision Tree Induction:**
  - Find the smallest tree that classifies the training data correctly
- **Problem**
  - Finding the smallest tree is computationally hard
- **Approach**
  - Use heuristic search (greedy search)

Which Attribute is "Best"?



- **Heuristics**
  - Pick split that decreases training error the most
  - Pick split that maximizes information (Information Gain)
  - Other statistical tests

Which Attribute is "Best"?

- **Information Gain**
  - Idea: Measure how much information an attribute conveys
  - Entropy: Number of bits to transmit one label (~disorder)
    (n= fract. neg examples in D / p= fract. pos examples in D)

$$Entropy_y(D) \equiv -p \log_2 p - n \log_2 n$$

  - Information Gain: Reduction in entropy, if attribute value known

$$Gain(D, A) = Entropy_y(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy_y(D_v)$$

Decision Tree for "Corporate Acq."

- vs = 1: -
- vs = 0:
- | export = 1:

…
- | export = 0:
- | | rate = 1:
- | | | stake = 1: +
- | | | stake = 0:
- | | | | debenture = 1: +
- | | | | debenture = 0:
- | | | | | takeover = 1: +
- | | | | | takeover = 0:
- | | | | | | file = 0: -
- | | | | | | file = 1:
- | | | | | | | share = 1: +
- | | | | | | | share = 0: -

… and many more

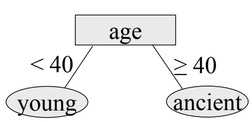
Total size of tree:
- 299 nodes

Note: word stems expanded for improved readability.

How Expressive are Decision Trees?

- **What functions h: X → Y can a decision tree represent?**
  - Assume that X is finite (only finite number of instances)
    → Decision trees can represent any function over a finite instance space X.
  - What if X is not finite (e.g. integer-valued attributes)?
  - What if X is not discrete (e.g. real-valued attributes)?
  - What if the data contains noise?
    - In the most extreme case, examples can have the same attribute values, but different labels.

TDIDT Extensions



- **Numerical (continuous) attributes**
  - Use > and < in attribute tests
- **Finite attributes with many values**
  - Example:
    - Target concept is "brakes defect"
    - Instances: all cars in the US
    - Attributes: Manufacturer (3 values), VIN (100.000.000 values)
  - Which attribute will Information Gain select? →GainRatio
- **Numerical (continuous) target attribute (regression)**
  - E.g. pick attribute test so that target values become more similar
  - E.g. predict mean value of examples in each leaf
- **Early stopping and Pruning**