ID3, C4.5 Decision Trees

Presented by

Elizabeth Leon

Decision trees

- The hypothesis space is a *complete space*
  - The space is the set of all decision trees defined over the given set of attributes
  - Any (finite) discrete value function can be represented by some decision tree.

ID3 Learning Algorithm

- Performs a simple to complex, hill climbing search through the hypothesis space
- Performs no backtracking in its search
- Maintains only a single current hypothesis
- Uses all training instances at each step of the search.

ID3 Learning Algorithm

- Nominal inputs
- Uses *information gain* as evaluation function
- Preference for short trees
- Preference for trees with high information gain attributes near the root.

ID3 Algorithm

1. Determine the attribute with the highest information gain on the training set.
2. Use this attribute as the root, create a branch for each of the values the attribute can have.
3. For each branch, repeat the process with subset of the training set that is classified by that branch.

   Until the entropy in the subset is 0 (all samples are classified in the same class)

   The classes are the leaves

Example: Playing tennis

Outlook	Temp.	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes

Outlook	Temp.	Humidity	Wind	play
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Data set “playing tennis”

Example: Playing tennis Entropy

- Two class problem: *yes* and *no*
- 14 samples: 9 classified *yes*,

  5 classified *no*

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

$$E(S_{yes}) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) = 0.41$$

$$E(S_{no}) = -\left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = 0.53$$

$$E(S) = 0.41 + 0.53 = 0.94$$

Example: Playing tennis Information Gain

$$\text{Gain(S, A)} = \text{E(S)} - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

where $S_V = \{ s \in S \mid A(s) = V\}$

$$\text{Gain(S, Wind)} = \text{E(S)} - \sum_{v \in (weak, strong)} \frac{|S_v|}{|S|} E(S_v)$$

$$= \text{E(S)} - \left[\left(\frac{8}{14} \right) E(S_{weak}) + \left(\frac{6}{14} \right) E(S_{strong}) \right]$$

$$= 0.94 - \left[\left(\frac{8}{14} \right) 0.811 + \left(\frac{6}{14} \right) 1.0 \right]$$

$$= 0.048$$

$$E(S_{weak}) = -\left(\frac{5}{8}\right)\log_2\left(\frac{5}{8}\right) - \left(\frac{3}{8}\right)\log_2\left(\frac{3}{8}\right)$$

$$= 0.811$$

$$E(S_{strong}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right)$$

$$= 1.0$$

Example: Playing tennis Information Gain

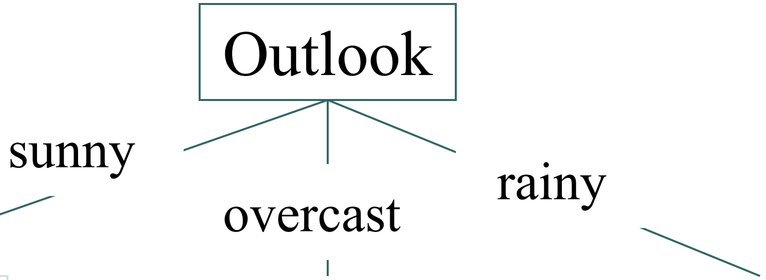
$$\text{Gain}(S, \text{Outlook}) = 0.246$$

$$\text{Gain}(S, \text{Humidity}) = 0.151$$

$$\text{Gain}(S, \text{Wind}) = 0.048$$

$$\text{Gain}(S, \text{Temperature}) = 0.029$$

Decision Tree For Playing Tennis



Outlook	Temp	Hum	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Outlook	Temp	Hum	Wind	Play
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

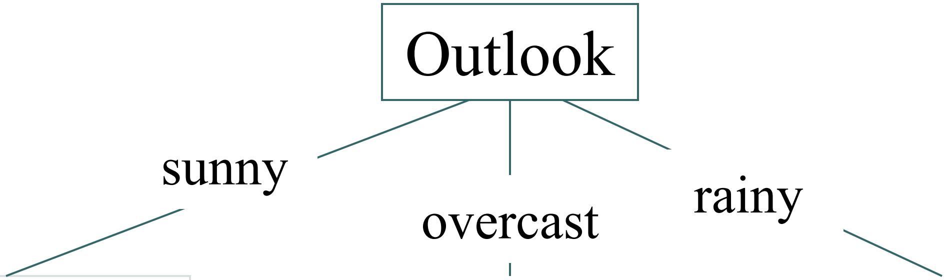
Outlook	Temp	Hum	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

$$Gain\ (S_{sunny}\ ,\ Humidity) = .970 - (3/5)\ 0.0 - (2/5)\ 0.0 = .970$$

$$Gain\ (S_{sunny}\ ,\ Temperature) = .970 - (2/5)\ 0.0 - (2/5)\ 1.0 - (1/5)\ 0.0 = .570$$

$$Gain\ (S_{sunny}\ ,\ Wind) = .970 - (2/5)\ 1.0 - (3/5)\ .918 = .019$$

Decision Tree For Playing Tennis



Outlook

sunny

overcast

rainy

Outlook	Temp	Hum	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Outlook	Temp	Hum	Wind	Play
Overcast	Hot	High	Weak	Yes
Overcast	Cool	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes

Outlook	Temp	Hum	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

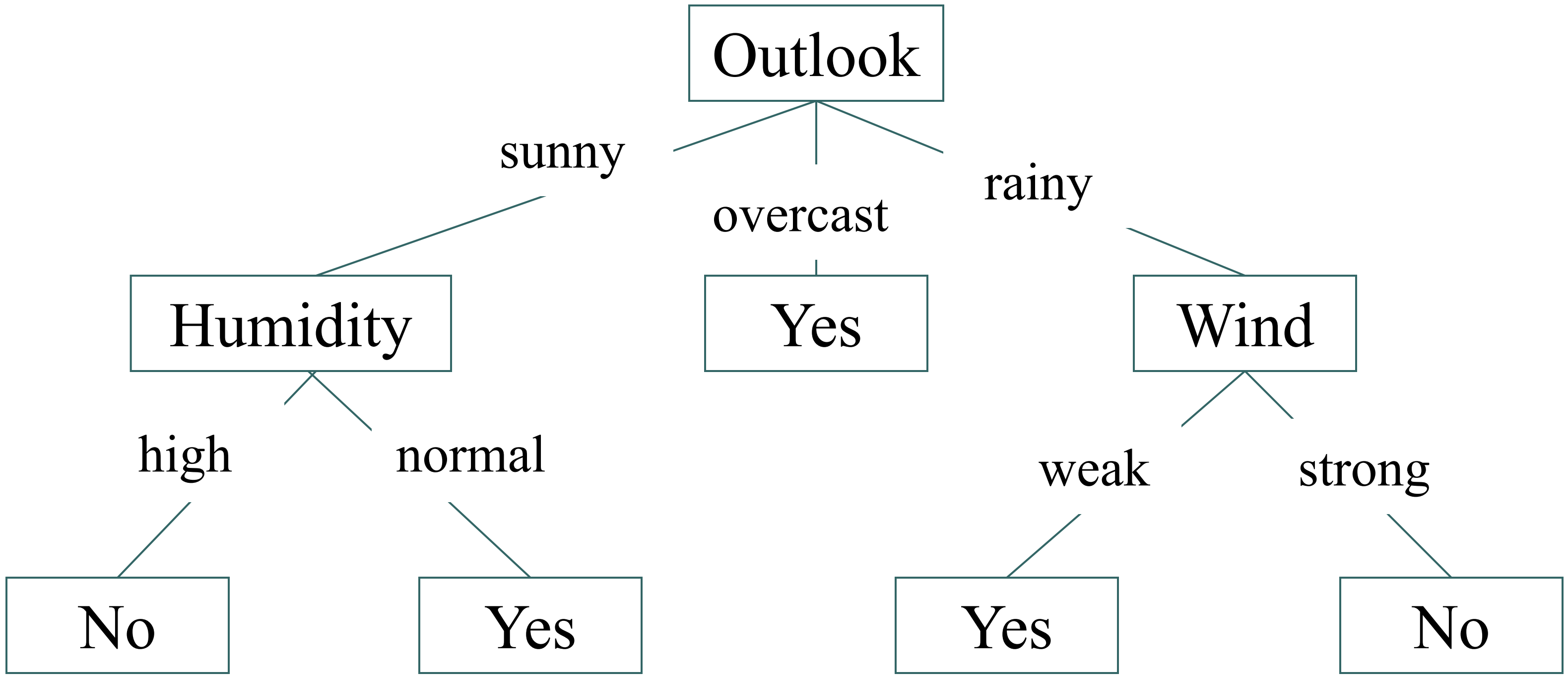
Entropy is 0. All the samples belong to the same class “Yes”

$$Gain\ (S_{rainy}\ , Humidity) = .019$$

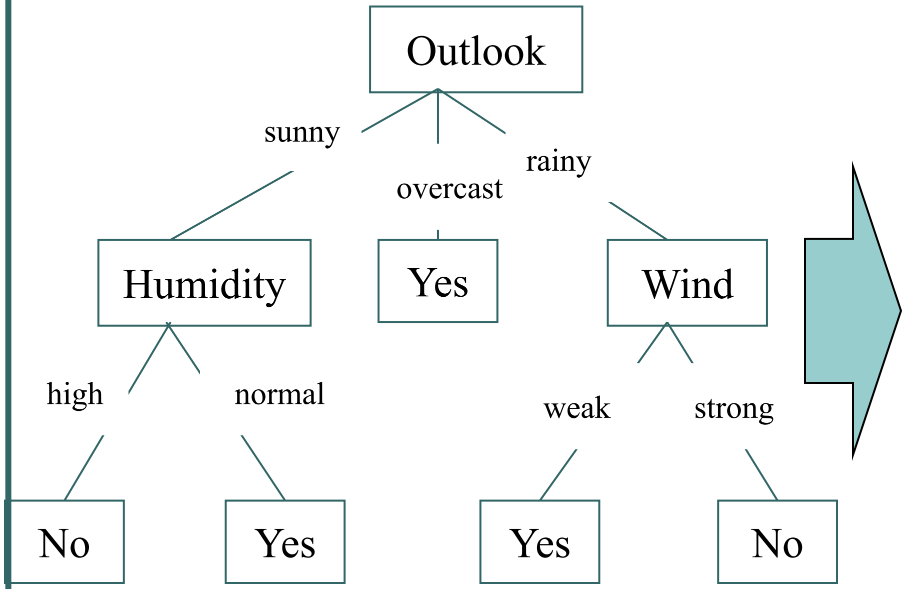
$$Gain\ (S_{rainy}\ , Temperature) = .019$$

$$Gain\ (S_{rainy}\ , Wind) = .970$$

Decision Tree For Playing Tennis



Rules



IF (Outlook = Sunny) & (Humidity = High)
THEN PlayTennis = No

IF (Outlook = Sunny) & (Humidity = Normal)
THEN PlayTennis = Yes

IF (Outlook = overcast)
THEN PlayTennis = Yes

IF (Outlook = Rainy) & (Wind = weak) THEN
PlayTennis = Yes

IF (Outlook = Rainy) & (Wind = Strong) THEN
PlayTennis = No

C4.5 Algorithm

- Extension of ID3 algorithm
- Extends from categorical to numeric attributes
- Uses *Gain ratio* measure

$$GainRatio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \qquad SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

- Deals with missing attribute values

C4.5 Algorithm

- Uses postpruning approach (pessimistic pruning)
  - Compare predicted error before and after merging leaves with parent node
- Provision of pruning based on the rules derived from the learned tree (C4.5Rules)

C4.5 Algorithm
Extension to numeric attributes

- Three types of tests
  - Discrete attribute: the “standard” test
  - Numeric attribute: binary test with outcomes Y<=Z and Y>Z (Y: attribute , Z : threshold)
  - The values are allocated to a group with one outcome and branch for each group
- Threshold Z
  - Training samples are sorted $\{V_1,V_2,\ldots, V_m\}$
  - m-1 possible splits: $\{V_i,V_{i+1}\}$. All of which should be examined to obtain an optimal split
  - Usual choose the midpoint. C4.5 chooses the smaller value V_i for every interval $\{V_i,V_{i+1}\}$

Example: Playing tennis

Outlook	Humidity	Wind	Play
Sunny	85	Weak	No
Sunny	90	Strong	No
Overcast	78	Weak	Yes
Rainy	96	Weak	Yes
Rainy	80	Weak	Yes
Rainy	70	Strong	No
Overcast	65	Strong	Yes

Outlook	Humidity	Wind	play
Sunny	95	Weak	No
Sunny	70	Weak	Yes
Rainy	80	Weak	Yes
Sunny	70	Strong	Yes
Overcast	90	Strong	Yes
Overcast	75	Weak	Yes
Rainy	80	Strong	No

Data set “playing tennis” with numerical attributes

Example: Playing tennis

- The set of values for humidity= {65,70,75,78,80,85,90,95,96}
- The set of potential threshold values= {65,70,75,78,80,85,90,95}

✔<=65 or >65

✔<=70 or >70

✔<=75 or >75

✔<=78 or >78

✔<=80 or >80

✔<=85 or >85

✔<=90 or >90

✔<=95 or >95

✔<=96 or >96

Example: Playing tennis Measure

- **humidity <= 65 or humidity > 65**

$$\mathrm{Gain(S, Temp)} = \mathrm{E(S)} - \sum_{v \in (<=65, >65)} \frac{|S_v|}{|S|} E(S_v)$$

$$= \mathrm{E(S)} - \left[\left(\frac{1}{14}\right)(-1/1\ \log_2(1/1) - 0/1\ \log_2\ (0/1)) + \left(\frac{13}{14}\right)(-8/13\ \log_2(8/13) - 5/13\ \log_2\ (5/13)) \right]$$

$$= 0.94 - 0.892577$$

$$= 0.047423$$

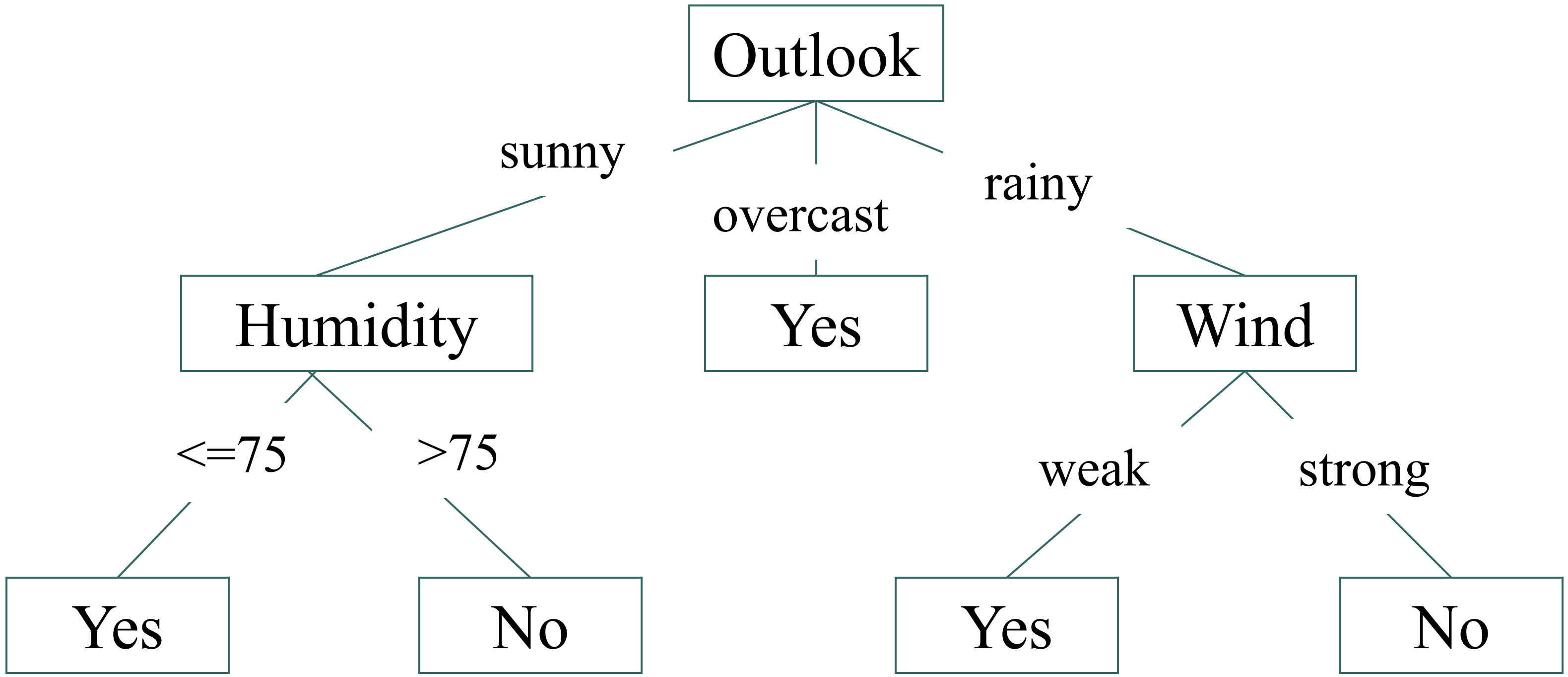
$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}$$

where: $$SplitInformation(S, A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$\mathrm{GainRatio(S, Humidity)} = 0.047423 / (-1/14 \log_2(1/14) - 13/14 \log_2(13/14))$$

$$= 0.127745$$

Decision tree using c4.5 Playing tennis



C4.5 Algorithm
Unknown attributes values

- It doesn't compute surrogate splits
  - Instead, C4.5 follows all B possible answers to the descendent nodes and ultimately B leaf nodes.
  - The classification is based on the labels of the B leaf nodes, weighted by the decision probabilities at the node

Example:
Unknown attribute values

Outlook	Humidity	Wind	Play
Sunny	85	Weak	No
Sunny	90	Strong	No
?	78	Weak	Yes
Rainy	96	Weak	Yes
Rainy	80	Weak	Yes
Rainy	70	Strong	No
Overcast	65	Strong	Yes

Outlook	Humidity	Wind	play
Sunny	95	Weak	No
Sunny	70	Weak	Yes
Rainy	80	Weak	Yes
Sunny	70	Strong	Yes
Overcast	90	Strong	Yes
Overcast	75	Weak	Yes
Rainy	80	Strong	No

Unknown attribute values (Cont.)

$$\mathrm{E(S) = - 8/13\ log(8/13) - 5/13\ log(5/13)) = 0.961}$$

$$\begin{aligned} Gain(S, Outlook) &= 5/13(-2/5\log(2/5) - 3/5\log(3/5)) \\ &+ 3/13(-3/3\log(3/3) - 0/3\log(0/3)) \\ &+ 5/13(-3/5\log(3/5) - 2/5\log(2/5)) \\ &= 0.747 \end{aligned}$$

The information gained is corrected with the factor F= 13/14

$$Gain(S, Outlook) = 13/14(0.961 - 0.747) = 0.199$$

$$\begin{aligned} \mathrm{SplitInfo(S, Temp)} &= \mathrm{-(\ 5/13\log(5/13) + 3/13\log(3/13)} \\ &\mathrm{+ 5/13\log(5/13) + 1/13\log(1/13))} \\ &= 1.876 \end{aligned}$$

Unknown attribute values (Cont.)

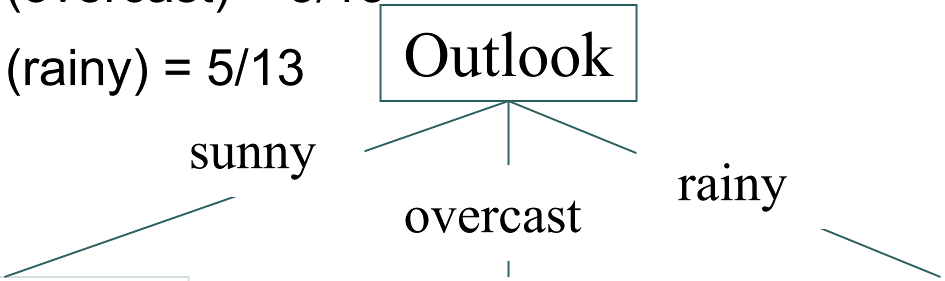
C4.5 associates a weigh with each sample (with missing values) in each subset (representing the probability that the case belongs to each subset).

- W_{new} after spliting is equal to the old parameter W_{old} before splitting multiplied by the probability that the samples belongs to each subset.

$$W_{new} = W_{old} P(S_i)$$

Unknown attribute values (Cont.)

- Splitting S on outlook:
  - Wnew(sunny) = 5/13
  - Wnew (overcast) = 3/13
  - Wnew (rainy) = 5/13



Outlook	Hum	Wind	Play	w
Sunny	70	Strong	Yes	1
Sunny	90	Weak	No	1
Sunny	85	Weak	No	1
Sunny	95	Weak	No	1
Sunny	70	Strong	Yes	1
Sunny	**90**	**Strong**	**Yes**	**5/13**

Outlook	Hum	Wind	Play	w
Overcast	**90**	**Strong**	**Yes**	**3/13**
Overcast	78	Weak	Yes	1
Overcast	65	Strong	Yes	1
Overcast	75	Weak	Yes	1

Outlook	Hum	Wind	Play	w
Rainy	80	Strong	No	1
Rainy	70	Strong	No	1
Rainy	80	Weak	Yes	1
Rainy	80	Weak	Yes	1
Rainy	96	Weak	Yes	1
Rainy	**90**	**Strong**	**Yes**	**5/13**

Unknown attribute values (Cont.)

- Similarly number of elements in each partition

  $|S_i|$ = # with known value + sum of ***weights*** for ***unknown*** values.

  - $|S_{sunny}| = 5+(5/13)$
  - $|S_{overcast}| = 3+(3/13)$
  - $|S_{rainy}| = 5+(5/13)$

Unknown attribute values (Cont.)

- The decision tree leaves are defined with two new parameters: ($|S_i|$ / E)
  - $|S_i|$ is the sum of the proportion of samples that reach the leaf and E is the number of samples that belong to other class

Unknown attribute values
Final rules

IF (Outlook = Sunny) & (Humidity <=70) THEN PlayTennis = Yes (2/0)

IF (Outlook = Sunny) & (Humidity >70) THEN PlayTennis = No (3.4/0.4)

IF (Outlook = overcast) THEN PlayTennis = Yes (3.2/0)

IF (Outlook = Rainy) & (Wind = weak) THEN PlayTennis = Yes (2.4/0)

IF (Outlook = Rainy) & (Wind = Strong) THEN PlayTennis = No (3/0)

C5.0 (compared vs c4.5)

- Better generation of rules set. Much faster and much less memory
- Trees with similar accuracy. Faster and smaller
- Boosting (technique for generating and combining multiple classifiers)

http://www.cse.unsw.edu.au/~quinlan/

References

- Duda, R., Pattern Classification.Wiley, 2000.
- Mitchell, T., Machine Learning, McGraw Hill,New York, 1997.
- Kantardzic, M.,Data Mining: concepts, models, methods and algorithms.

References

- http://www.cse.unsw.edu.au/~quinlan/
- http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html
- http://pilat.org/projects.html