Decision Trees

Dr. Mustafa Shiple

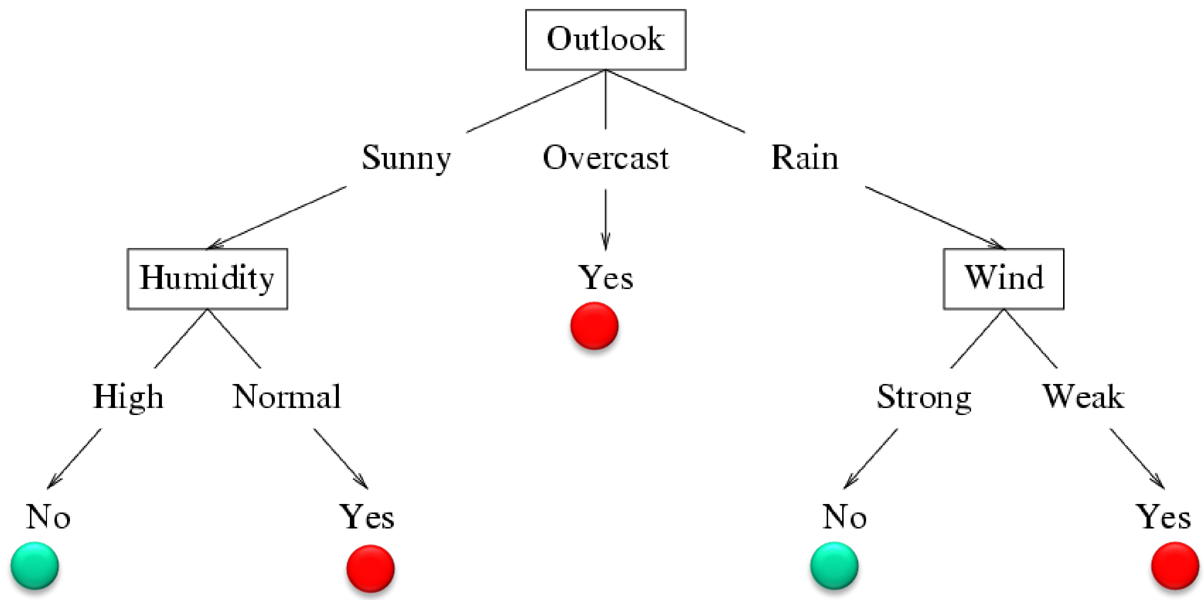
Training Data Example: Goal is to Predict When This Player Will Play Tennis?

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

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 $\mathbf{x} = (Sunny, Hot, High, Strong)$ will be classified as **No**. The **Temperature** feature is irrelevant.

Learning Algorithm for Decision Trees

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$$

$$\mathbf{x} = (x_1, \ldots, x_d)$$

$$x_j, y \in \{0,1\}$$

$\textsc{GrowTree}(S)$

if $(y = 0 \text{ for all } \langle \mathbf{x}, y \rangle \in S)$ **return** new leaf(0)

else if $(y = 1 \text{ for all } \langle \mathbf{x}, y \rangle \in S)$ **return** new leaf(1)

else

choose best attribute x_j

$S_0 = \text{all } \langle \mathbf{x}, y \rangle \in S \text{ with } x_j = 0;$

$S_1 = \text{all } \langle \mathbf{x}, y \rangle \in S \text{ with } x_j = 1;$

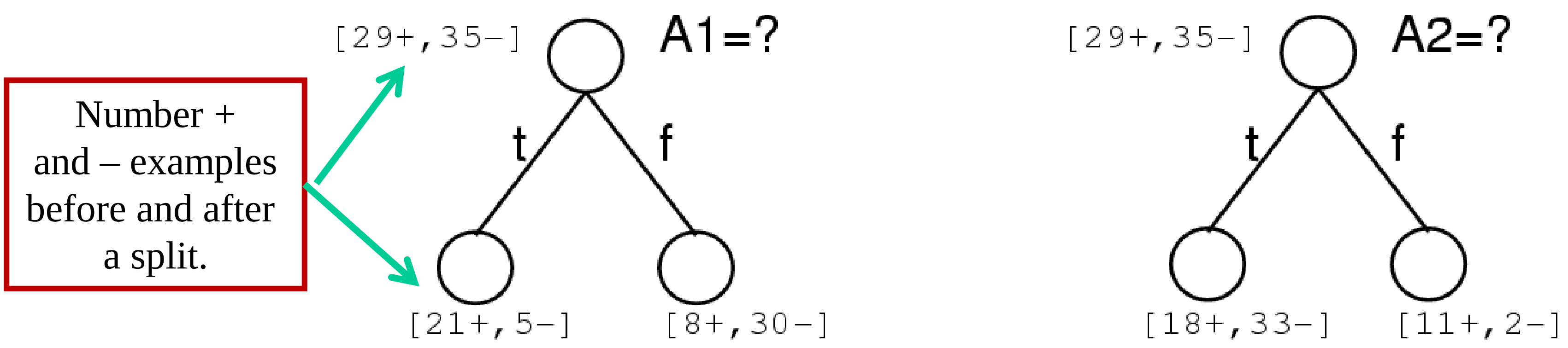
return new node$(x_j, \textsc{GrowTree}(S_0), \textsc{GrowTree}(S_1))$

What happens if features are not binary? What about regression?

Choosing the Best Attribute

A1 and A2 are “attributes” (i.e. features or inputs).

Which attribute is best?



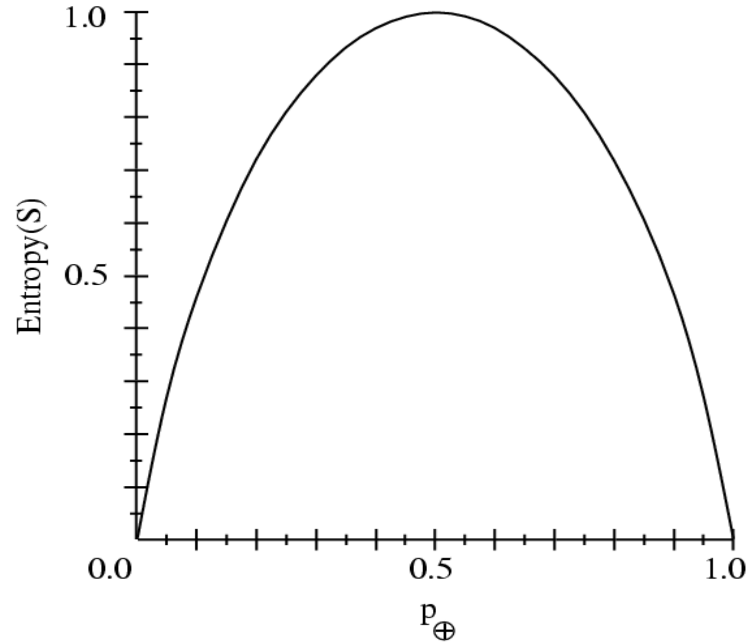
- Many different frameworks for choosing BEST have been proposed!
- We will look at Entropy Gain.

Entropy

- $p_{\oplus}$ is the proportion of positive examples in S
- $p_{\ominus}$ is the proportion of negative examples in S
- Entropy measures the impurity of S

$$Entropy(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

Entropy



- S is a sample of training examples

Entropy is like a measure of impurity…

Entropy

High Entropy – High level of Uncertainty

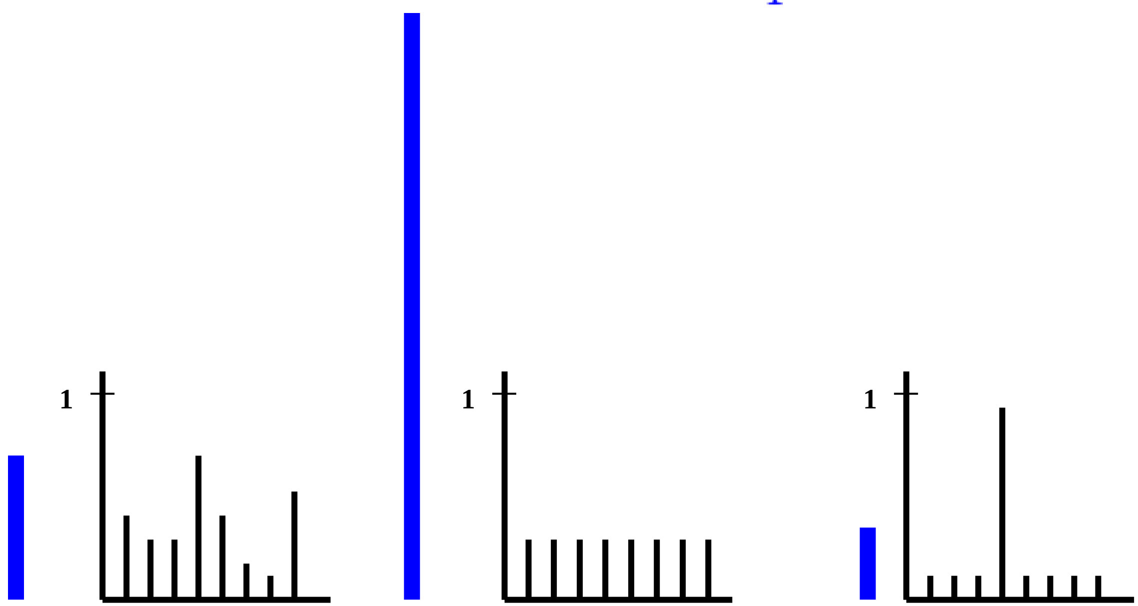
Low Entropy – No Uncertainty.

(Convince yourself that the max value would be $\log(k)$)

(Also note that the base of the log only introduces a constant factor; therefore, we'll think about base 2)

$$Entropy(S[p_1, p_2, \ldots, p_k]) = -\sum_{1}^{k} p_i \log(p_i)$$

Test yourself again: assign high, medium, low to each of these distributions.
For the middle distribution, try to guess the value of the entropy.



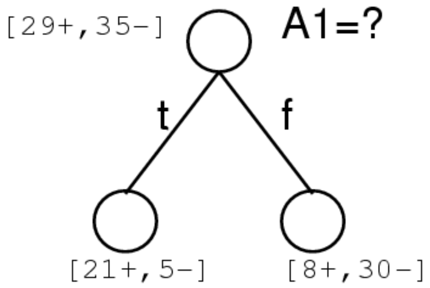
Information Gain

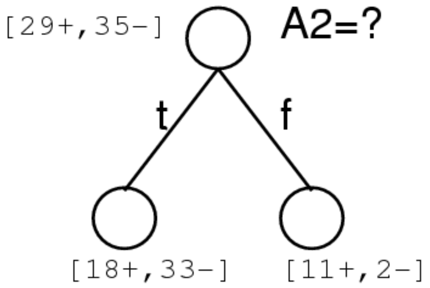
$Gain(S, A)$ = expected reduction in entropy due to sorting on A

High Entropy – High level of Uncertainty

Low Entropy – No Uncertainty.

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





Information Gain

- Let's assume each element of S consists of a set of features
- Information Gain (IG) on a feature F

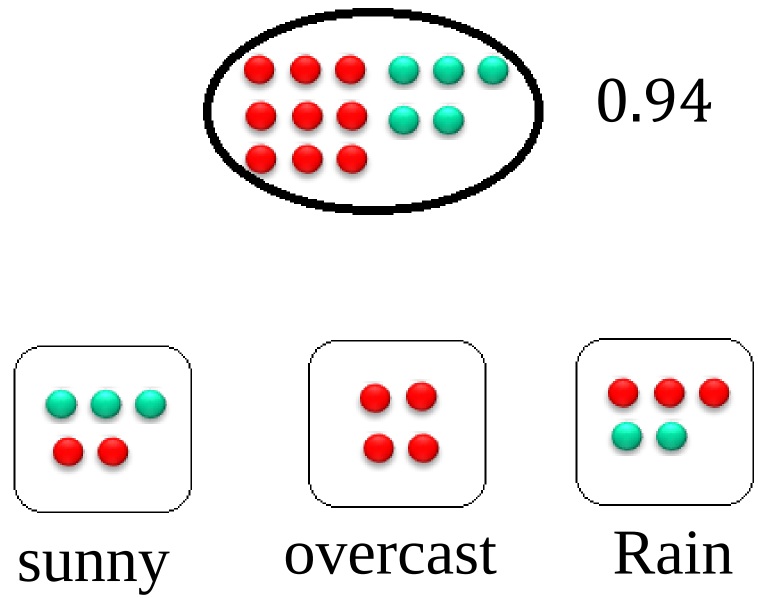
$$IG(S,F) = H(S) - \sum_{f \in F} \frac{|S_f|}{|S|} H(S_f)$$

- S_f number of elements of S with feature F having *value* f
- $IG(S,F)$ measures the increase in our certainty about S once we know the value of F

Computing Information Gain (outlook)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$E(sunny) = -\frac{2}{5}\log(_{2=yes/no})\frac{2}{5} - \frac{3}{5}\log(_{2=yes/no})\frac{3}{5} = 0.97$$

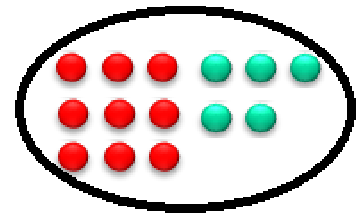
$$E(overcast) = 0$$

$$E(overcast) = 0.97$$

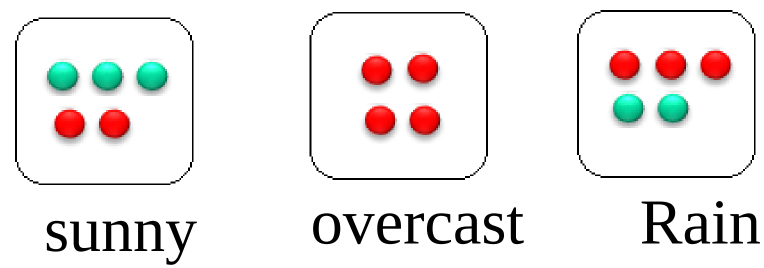
Computing Information Gain (outlook)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



0.94



$$IG(S,F) = H(S) - \sum_{f \in F} \frac{|S_f|}{|S|} H(S_f)$$

$$E(sunny) = 0.97$$

$$E(overcast) = 0$$

$$E(overcast) = 0.97$$

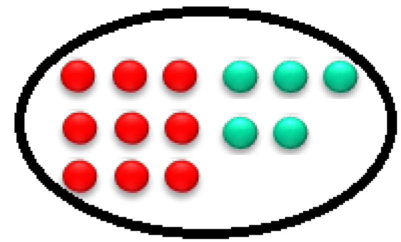
$$IG = 0.94 - \frac{5}{14} \times 0.97 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.97$$

$$IG = 0.246$$

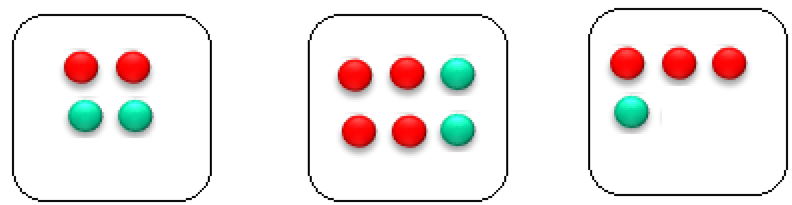
Computing Information Gain (Temp)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



0.94



Hot Mild cool

$$E(Hot) = -\frac{2}{4}\log_{(2=yes/no)}\frac{2}{4} - \frac{2}{4}\log_{(2=yes/no)}\frac{2}{4} = 1$$

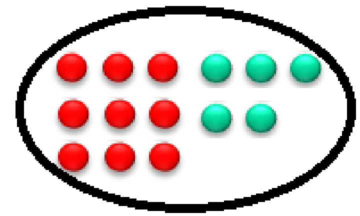
$$E(Mild) = -\frac{4}{6}\log_{(2=yes/no)}\frac{4}{6} - \frac{2}{6}\log_{(2=yes/no)}\frac{2}{6} = 0.92$$

$$E(cool) = -\frac{3}{4}\log_{(2=yes/no)}\frac{3}{4} - \frac{1}{4}\log_{(2=yes/no)}\frac{1}{4} = 0.811$$

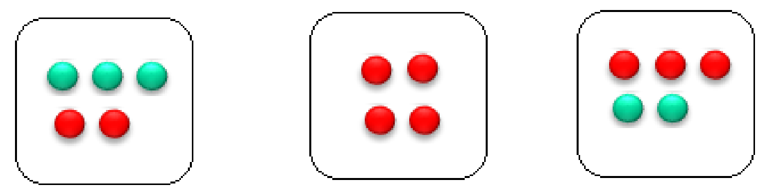
Computing Information Gain (Temp)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



0.94



sunny overcast Rain

$$IG(S,F) = H(S) - \sum_{f \in F} \frac{|S_f|}{|S|} H(S_f)$$

$E(Hot) = 1$

$E(Mild) = 0.92$

$E(cool) = 0.811$

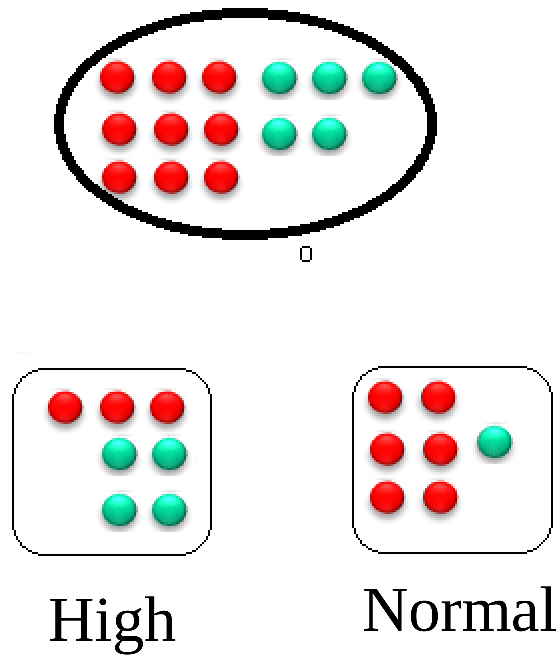
$$IG = 0.94 - \frac{5}{14} \times 1 - \frac{4}{14} \times 0.92 - \frac{5}{14} \times 0.811$$

$$IG = 0.029$$

Computing Information Gain (Humidity)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



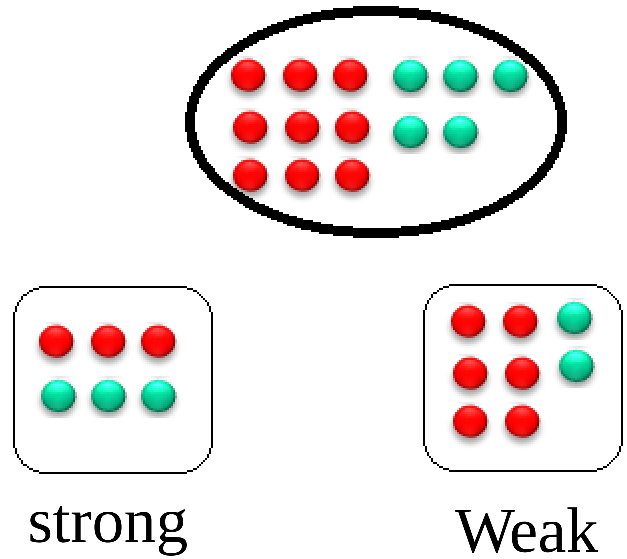
$$E(High) = -\frac{4}{7}\log(_{2=yes/no})\frac{4}{7} - \frac{3}{7}\log(_{2=yes/no})\frac{3}{7} = 0.985$$

$$E(Normal) = -\frac{6}{7}\log(_{2=yes/no})\frac{6}{7} - \frac{1}{7}\log(_{2=yes/no})\frac{1}{7} = 0.59$$

Computing Information Gain (Humidity)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$E(High) = 0.985$$

$$E(Normal = 0.59$$

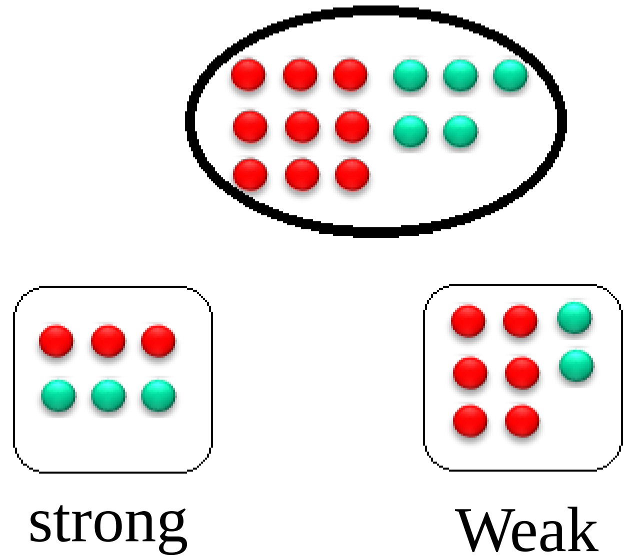
$$IG = 0.94 - \frac{6}{14} \times 0.985 - \frac{8}{14} \times 0.59$$

$$IG = 0.15$$

Computing Information Gain (wind)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$E(strong) = -\frac{3}{6}\log(_{2=yes/no})\frac{3}{6} - \frac{3}{6}\log(_{2=yes/no})\frac{3}{6} = 1$$

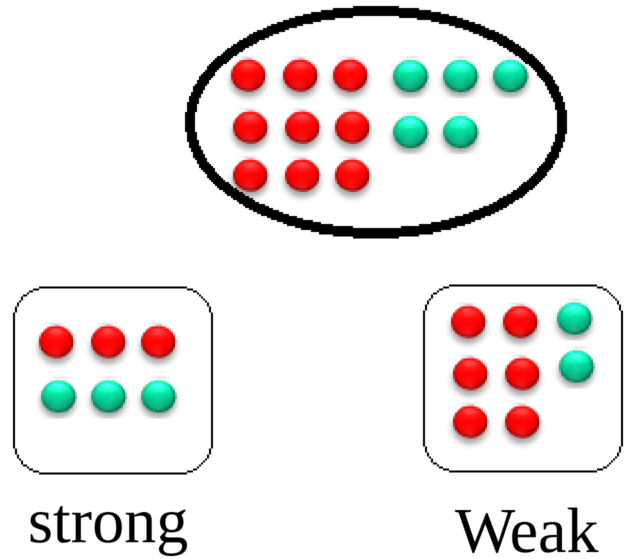
$$E(weak) = -\frac{6}{8}\log(_{2=yes/no})\frac{6}{8} - \frac{2}{8}\log(_{2=yes/no})\frac{2}{8} = 0.811$$

- $S_{weak} = [6+, 2-] \Longrightarrow H(S_{weak}) = 0.811$
- $S_{strong} = [3+, 3-] \Longrightarrow H(S_{strong}) = 1$

Computing Information Gain (wind)

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



$$E(strong) = 1$$

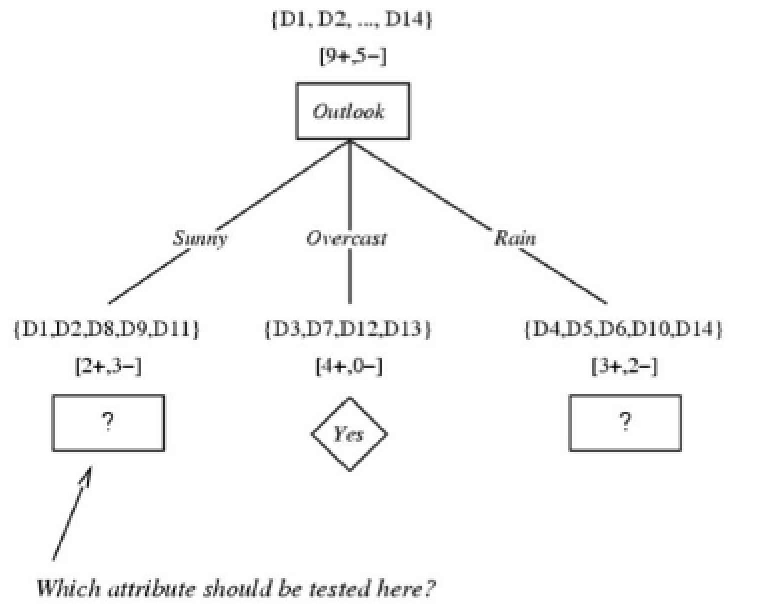
$$E(weak) = 0.811$$

$$IG = 0.94 - \frac{6}{14} \times 1 - \frac{8}{14} \times 0.811$$

$$IG = 0.48$$

Choosing the most informative feature

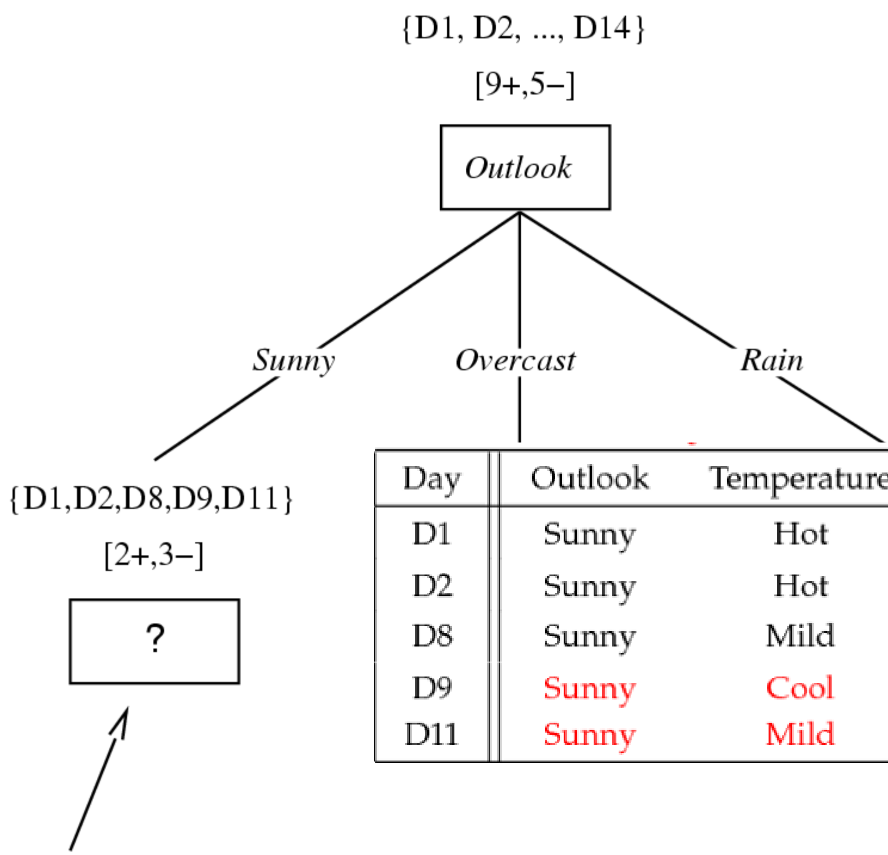
- At the root node, the information gains are:
  - $IG(S, \text{wind}) = 0.048$ (we already saw)
  - $IG(S, \text{outlook}) = 0.246$
  - $IG(S, \text{humidity}) = 0.151$
  - $IG(S, \text{temperature}) = 0.029$
- "outlook" has the maximum $IG \Longrightarrow$ chosen as the root node
- Growing the tree:
  - Iteratively select the feature with the highest information gain for each child of the previous node



Training Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

Which attribute should be tested here?

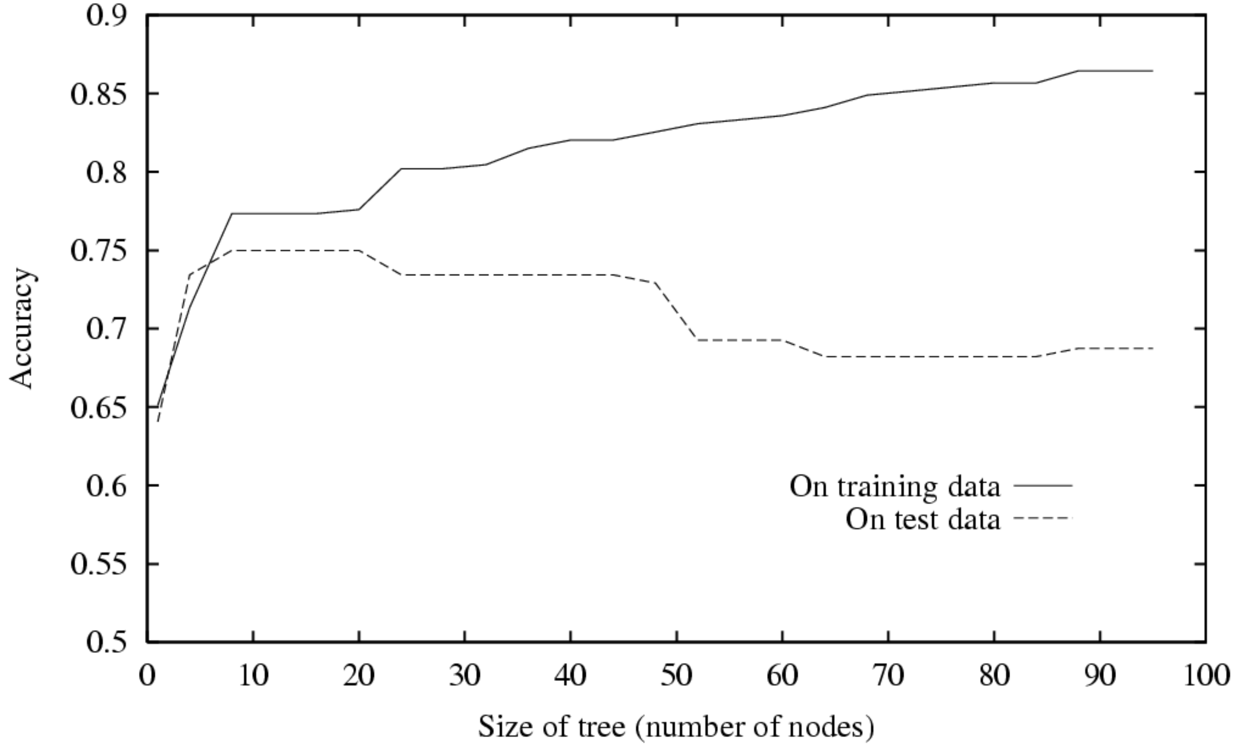
S_{sunny} = {D1,D2,D8,D9,D11}

$$E(sunny) = -\frac{2}{5}\log_{(2=yes/no)}\frac{2}{5} - \frac{3}{5}\log_{(2=yes/no)}\frac{3}{5} = 0.97$$

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

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

Overfitting in Decision Trees



Discrete vs. Continuous Attributes

- Continuous variables attributes - problems for decision trees
  - increase computational complexity of the task
  - promote prediction inaccuracy
  - lead to overfitting of data
- Convert continuous variables into discrete intervals
  - “greater than or equal to” and “less than”
  - optimal solution for conversion
  - difficult to determine discrete intervals ideal
    - size
    - number

Pruning a decision tree

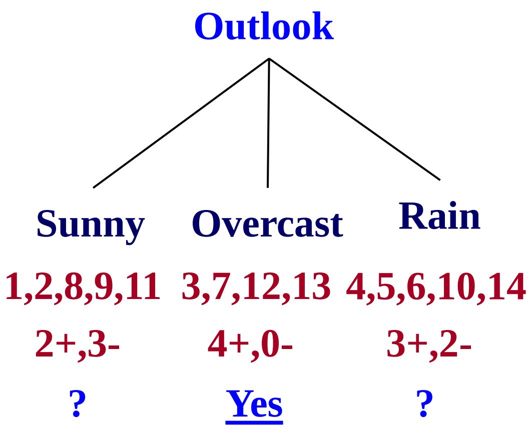
- Prune = remove leaves and assign majority label of the parent to all items
- Prune the children of node s if:
  - all children are leaves, and
  - the accuracy on the validation set does not decrease if we assign the most frequent class label to all items at s.

Two basic approaches

Pre-pruning: Stop growing the tree at some point during construction when it is determined that there is not enough data to make reliable choices.

Post-pruning: Grow the full tree and then remove nodes that seem not to have sufficient evidence.

Missing Values



$Gain(S_{sunny}, Temp) =$ **.97- 0-(2/5) 1 = .57**

$Gain(S_{sunny}, Humidity) =$

- Fill in: assign the most likely value of X_i to **s**:
  argmax $_k$ $P(X_i = k)$: Normal
  - **97-(3/5) Ent[+0,-3] -(2/5) Ent[+2,-0] = .97**
- Assign fractional counts $P(X_i = k)$ for each value of X_i to **s**
  - **.97-(2.5/5) Ent[+0,-2.5] - (2.5/5) Ent[+2,-.5] < .97**

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
8	Sunny	Mild	???	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes