Decision Trees (D3)

Ahmet Sacan

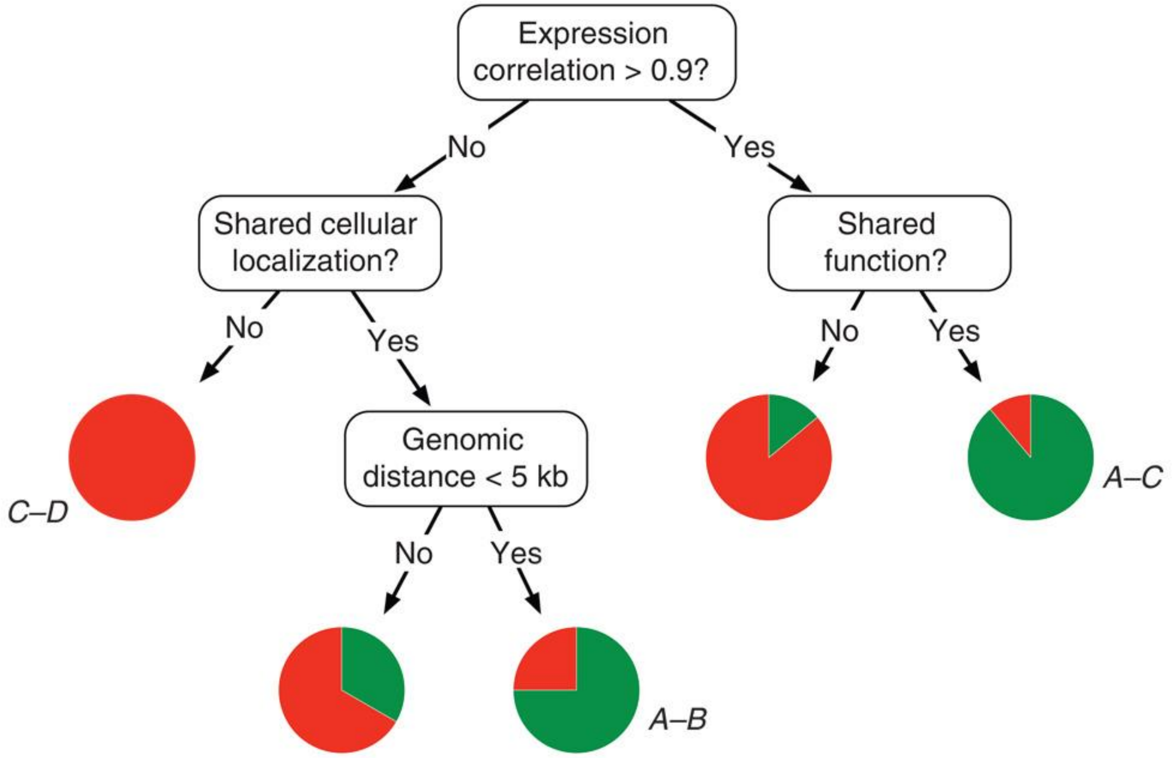
Some content/slides are from:Machine Learning by Tom Mitchell.

Decision Tree Example

a

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb
⋮					

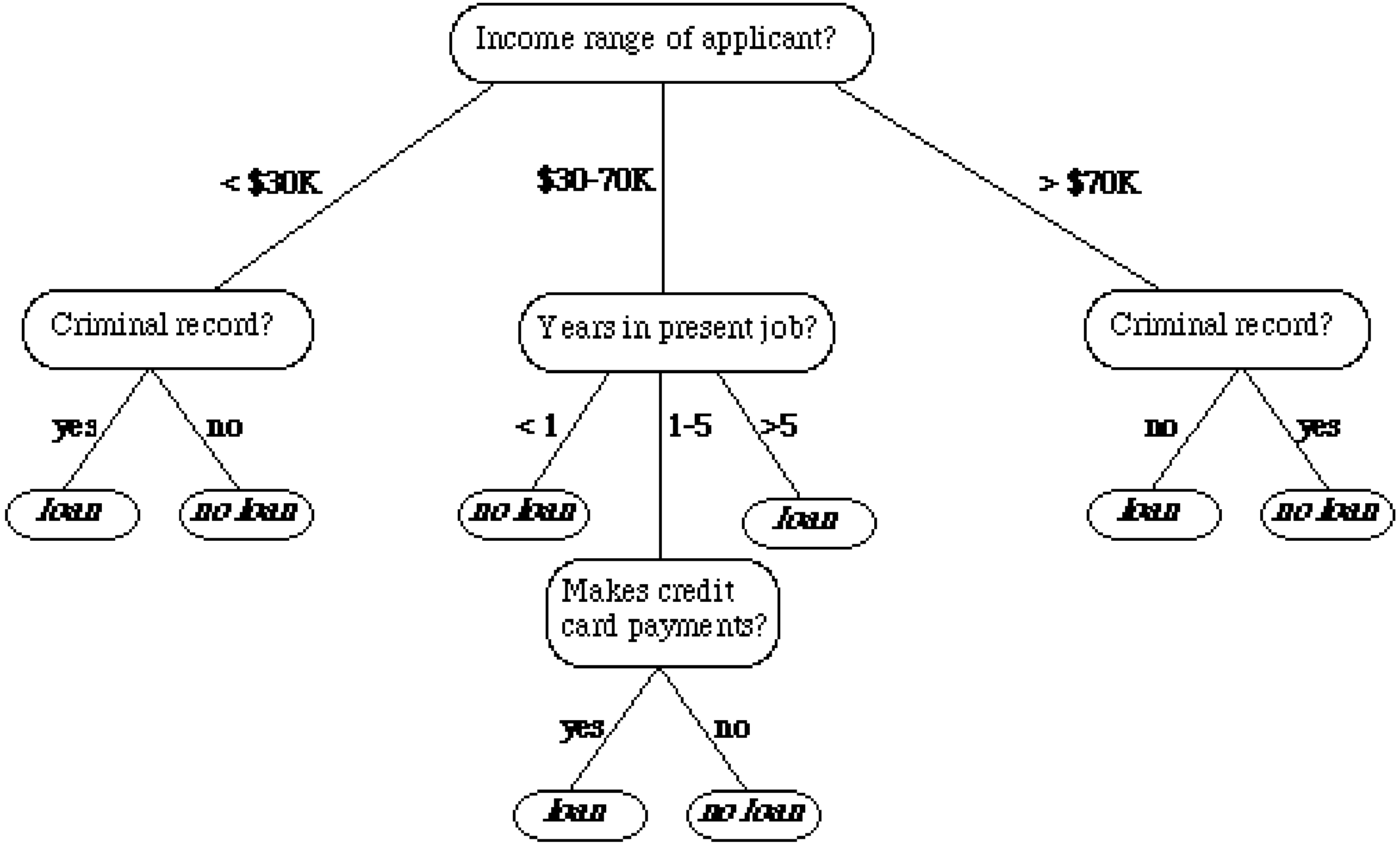
b



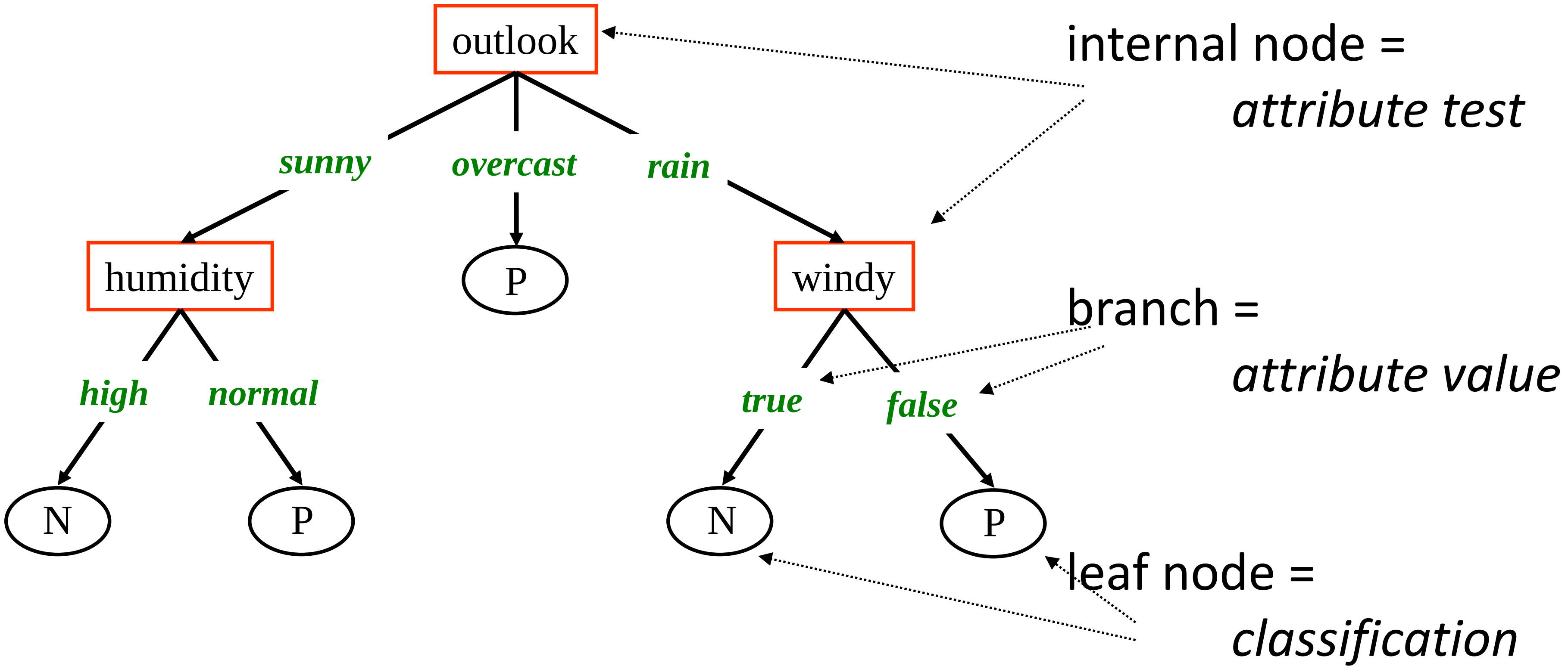
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701298/

Decision Tree Example

- Whether to issue a loan to a customer:



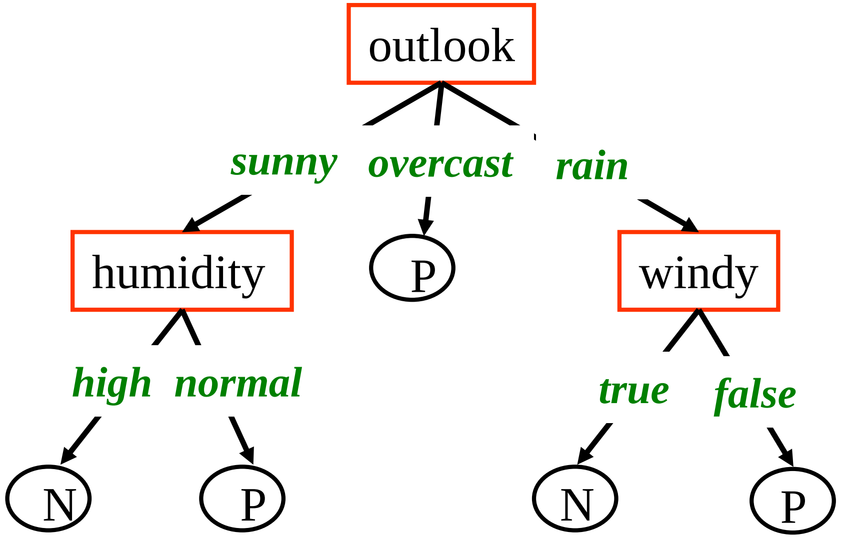
Structure of a Decision Tree



Logical Rules represented by D3

- Decision Trees represent a disjunction (OR) of conjunctions (AND) of constraints on the attribute values

$$
\begin{aligned}
& (Outlook = Sunny \ \wedge \ Humidity = Normal) \\
\vee \quad & (Outlook = Overcast) \\
\vee \quad & (Outlook = Rain \ \wedge \ Wind = Weak)
\end{aligned}
$$



Training Instances

- Is it a good day to play soccer?
- Attributes:

outlook:	sunny, overcast, rain
temperature:	cool, mild, hot
humidity:	high, normal
windy:	true, false

- Training instance:

<**overcast**, **hot**, **normal**, **false**>: **play**

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Random learner

- Arbitrarily pick an attribute to branch on, split the dataset by that attribute and repeat for each resulting node.
- Why is that a bad idea?

Occam's Razor

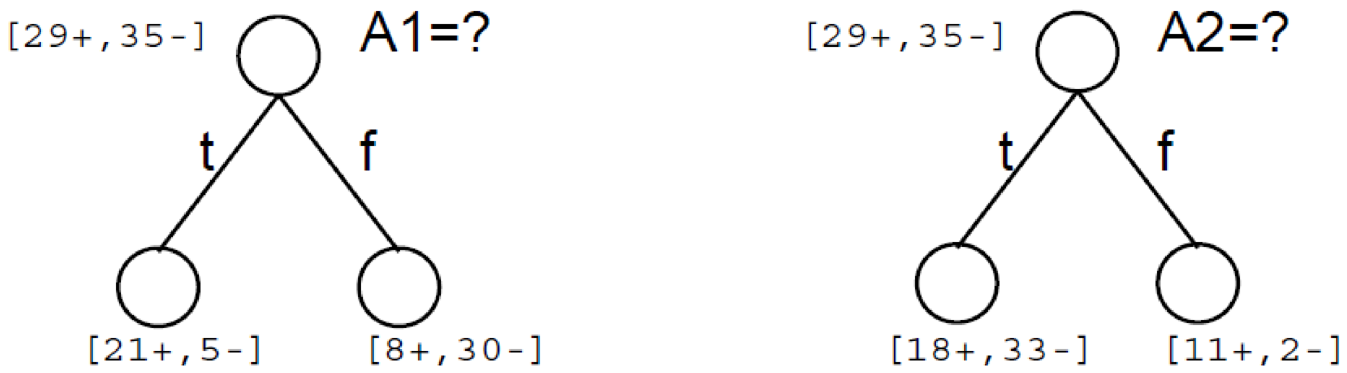
- Prefer simpler/shorter hypotheses/theories/explanations.
- Argument: There are fewer short hypotheses. Short hypotheses that fit data are unlikely to be coincidence

Top-Down Induction of D3s

- For data in each node:
  - Find best attribute to split by
  - Split data with that attribute
  - Repeat until all training examples are perfectly classified.

Split Criteria

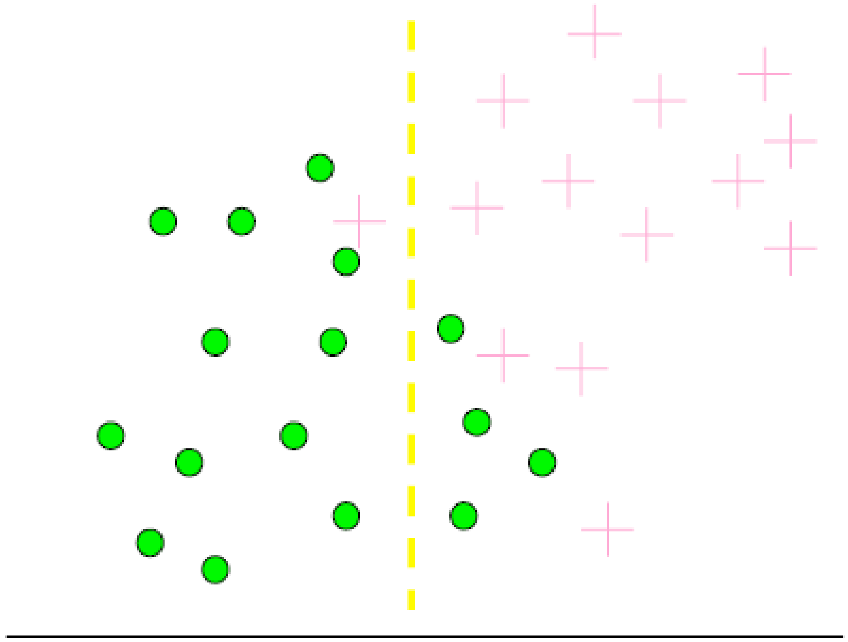
- Which attribute is best?



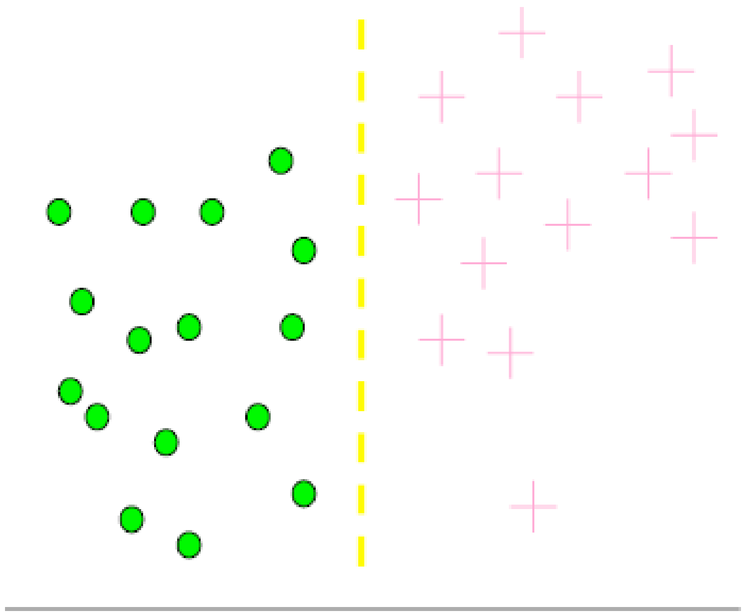
Split Criteria

Which test is more informative?

Split over whether Balance exceeds 50K



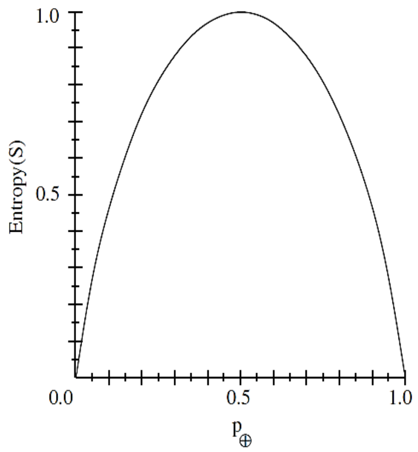
Split over whether applicant is employed

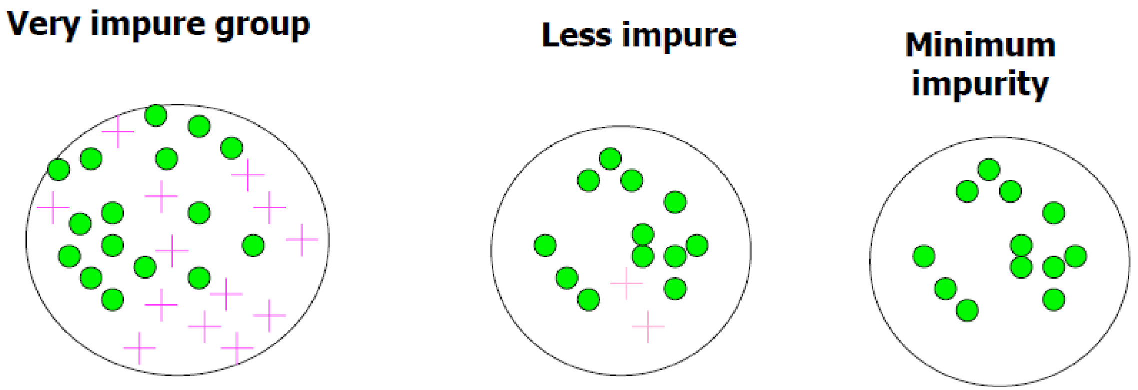


Based on slide by Pedro Domingos

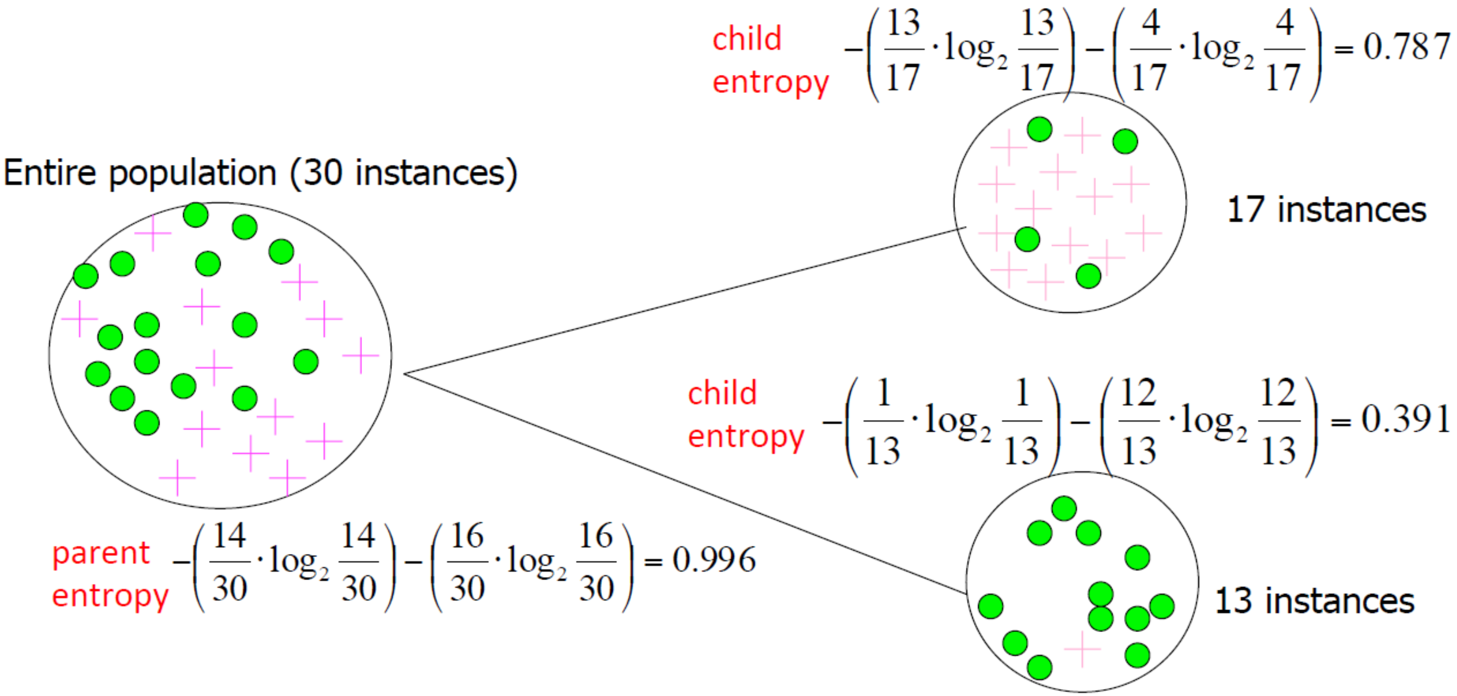
Entropy (disorder)

- Entropy is a measure of impurity/uncertainty in the data.
  - Multivalued attribute:
    - $entropy = -\sum_i p_i \log_2(p_i)$
  - Binary attribute: p+q=1.
    - $entropy = -(plog_2(p) + qlog_2(q))$

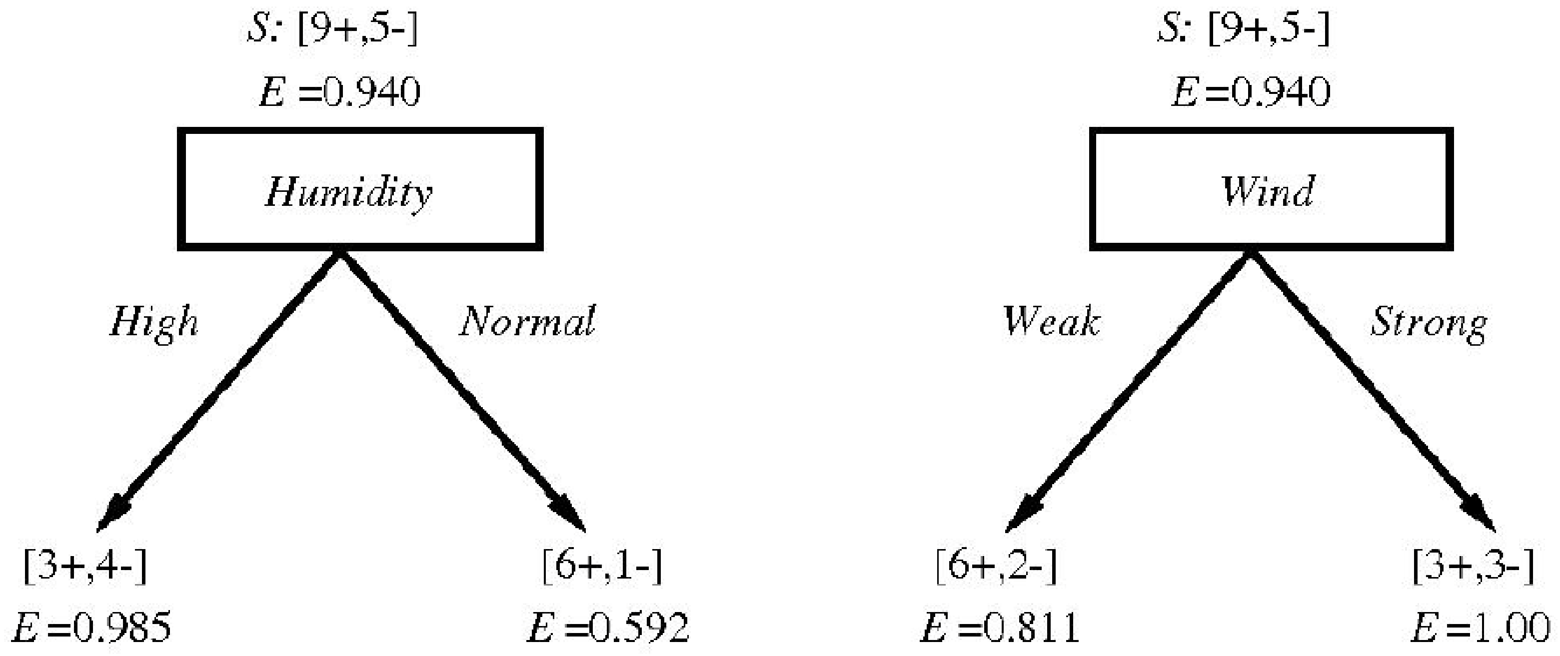




Entropy

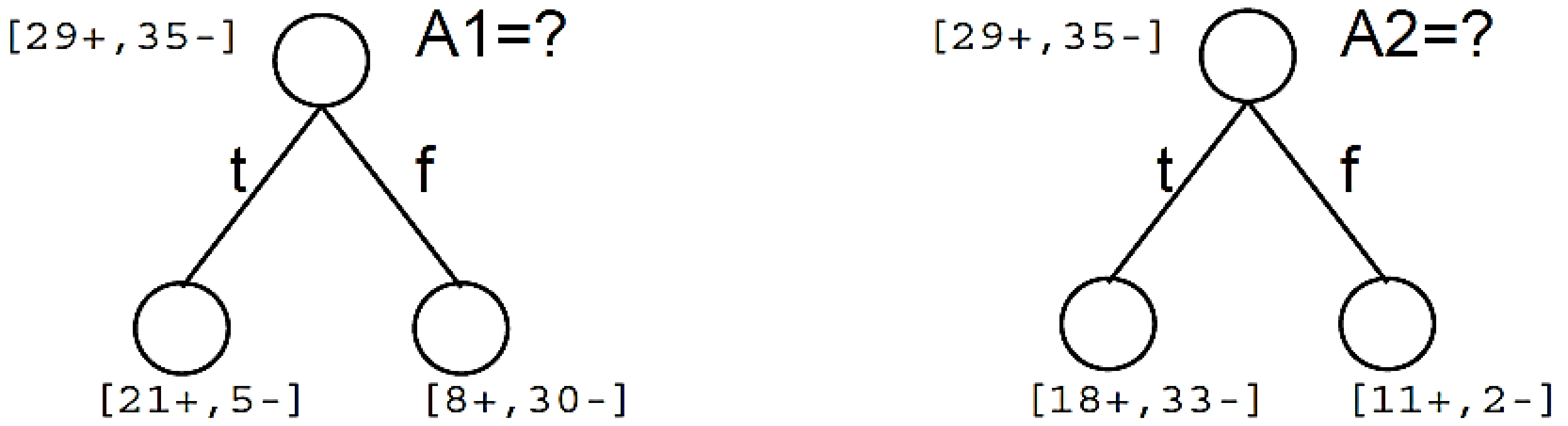

child entropy $-\left(\frac{13}{17}\cdot\log_2\frac{13}{17}\right)-\left(\frac{4}{17}\cdot\log_2\frac{4}{17}\right)=0.787$
Entire population (30 instances)
17 instances
child entropy $-\left(\frac{1}{13}\cdot\log_2\frac{1}{13}\right)-\left(\frac{12}{13}\cdot\log_2\frac{12}{13}\right)=0.391$
13 instances
parent entropy $-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$

Entropy


S: [9+,5-]
E =0.940
Humidity
High
Normal
[3+,4-]
E =0.985
[6+,1-]
E =0.592
S: [9+,5-]
E =0.940
Wind
Weak
Strong
[6+,2-]
E =0.811
[3+,3-]
E =1.00

Exercise

- Calculate the entropy of each of the nodes below.



Information Gain

- The goal of a split is to minimize total entropy.
- Information Gain: Expected reduction in entropy due to splitting on an attribute S:

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

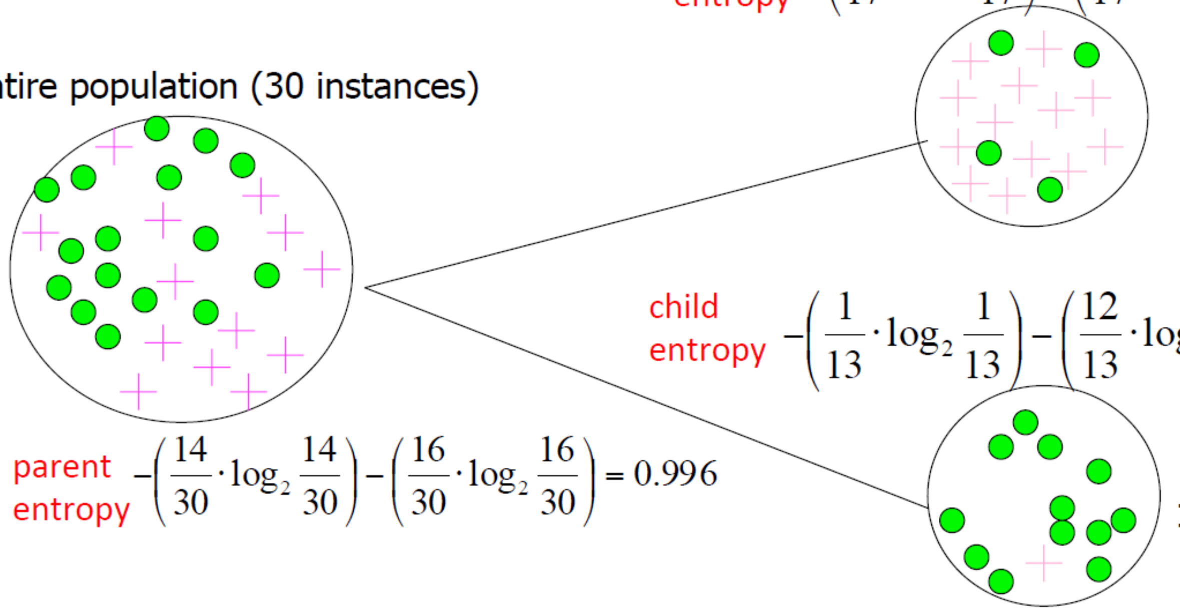
  - $Values(A)$: the set of all possible values for attribute A
  - S_v: subset of S for which attribute A has value v

Information Gain

Information Gain = entropy(parent) – [average entropy(children)]

child entropy $-\left(\frac{13}{17}\cdot\log_2\frac{13}{17}\right)-\left(\frac{4}{17}\cdot\log_2\frac{4}{17}\right)=0.787$

Entire population (30 instances)



17 instances

child entropy $-\left(\frac{1}{13}\cdot\log_2\frac{1}{13}\right)-\left(\frac{12}{13}\cdot\log_2\frac{12}{13}\right)=0.391$

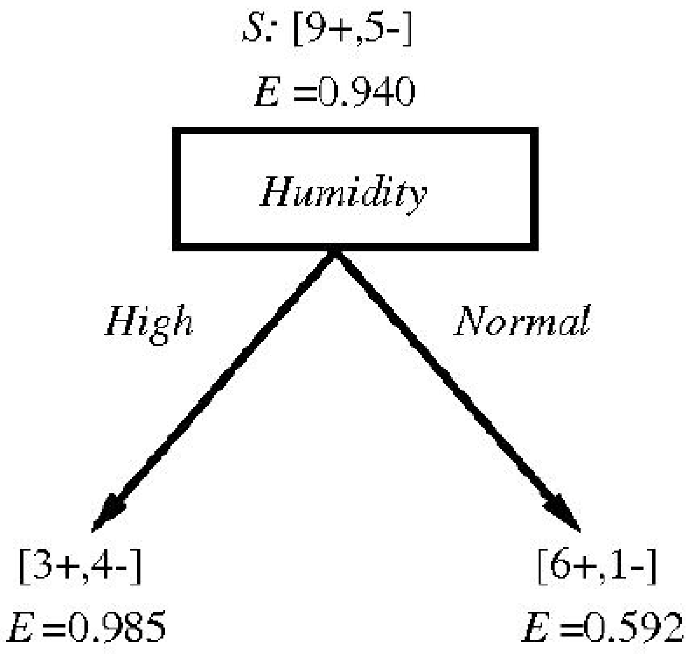
13 instances

parent entropy $-\left(\frac{14}{30}\cdot\log_2\frac{14}{30}\right)-\left(\frac{16}{30}\cdot\log_2\frac{16}{30}\right)=0.996$

(Weighted) Average Entropy of Children = $\left(\frac{17}{30}\cdot 0.787\right)+\left(\frac{13}{30}\cdot 0.391\right)=0.615$

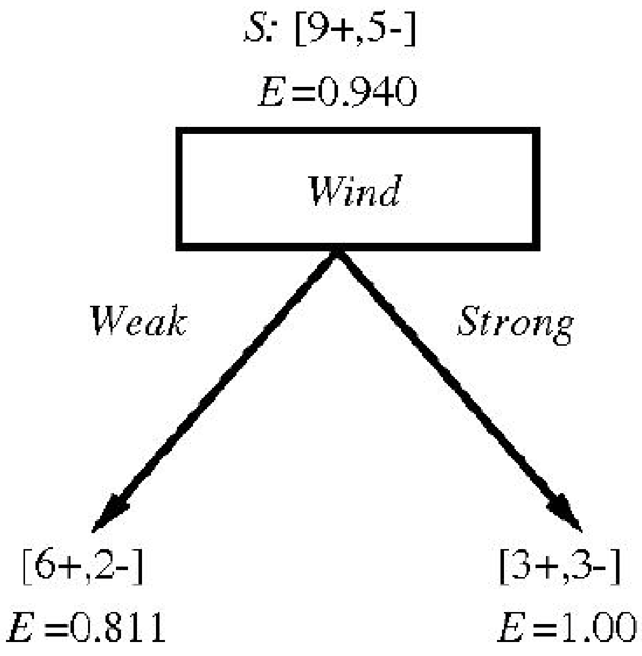
Information Gain= 0.996 - 0.615 = 0.38

Information Gain



Gain (S, Humidity)

= .940 - (7/14).985 - (7/14).592
= .151

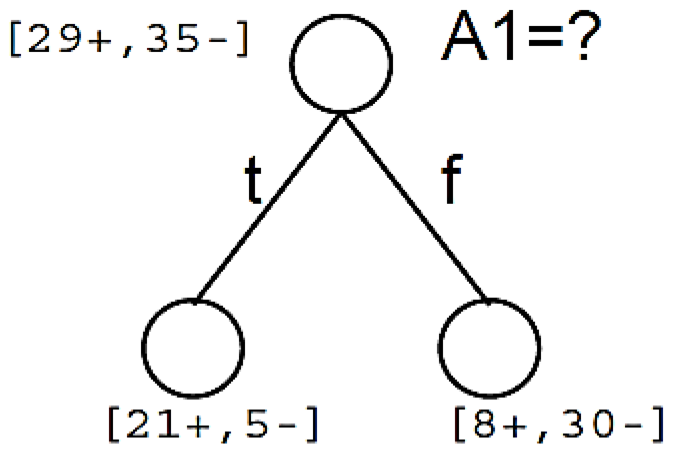


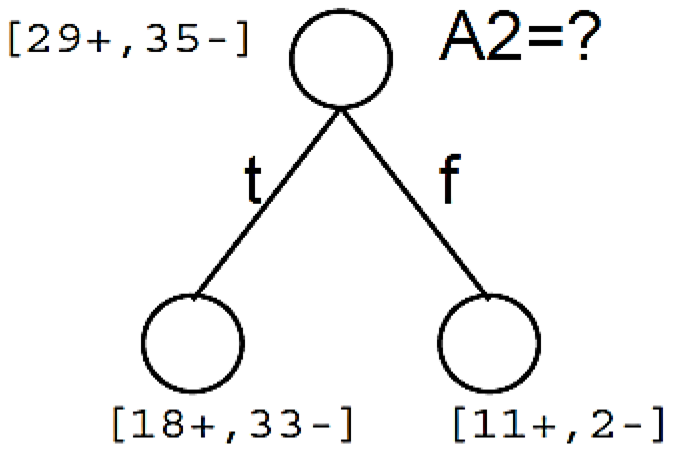
Gain (S, Wind)

= .940 - (8/14).811 - (6/14)1.0
= .048

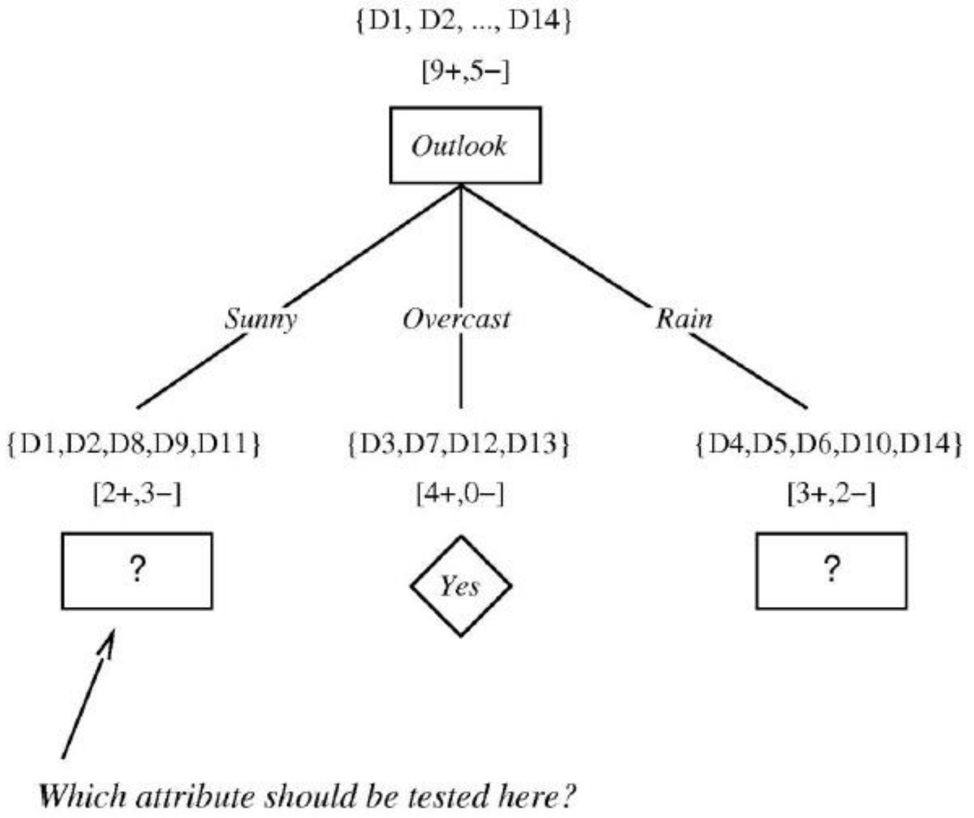
Exercise

- Information Gain A1:
- Information Gain A2:





ID3 (Iterative Dichotomizer) Algorithm



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

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

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

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

Exercise

- Build a decision tree to predict whether two genes interact, using the sample data below. Build the tree using information gain as the split criteria.

Gene pair	e: Expression correlation >=0.5?	s: Subcellular co-localization	f: Shared function	Interact?
A-B	0	0	0	NO
C-D	0	0	1	YES
E-F	0	1	0	YES
G-H	0	1	1	NO
I-J	1	0	0	YES
K-L	1	1	0	NO

Gene pair	e: Expression correlation >=0.5?	s: Subcellular co-localization	f: Shared function	Interact?
A-B	0	0	0	NO
C-D	0	0	1	YES
E-F	0	1	0	YES
G-H	0	1	1	NO
I-J	1	0	0	YES
K-L	1	1	0	NO

Gene pair	e: Expression correlation >=0.5?	s: Subcellular co-localization	f: Shared function	Interact?
A-B	0	0	0	NO
C-D	0	0	1	YES
E-F	0	1	0	YES
G-H	0	1	1	NO
I-J	1	0	0	YES
K-L	1	1	0	NO

Gene pair	e: Expression correlation >=0.5?	f: Shared function	Interact?
A-B	0	0	NO
C-D	0	1	YES
I-J	1	0	YES

Gene pair	e: Expression correlation >=0.5?	f: Shared function	Interact?
E-F	0	0	YES
G-H	0	1	NO
K-L	1	0	NO

Gene pair	e: Expression correlation >=0.5?	s: Subcellular co-localization	f: Shared function	Interact?
A-B	0	0	0	NO
C-D	0	0	1	YES

Gene pair	e: Expression correlation >=0.5?	s: Subcellular co-localization	f: Shared function	Interact?
I-J	1	0	0	YES

Gene pair	e: Expression correlation >=0.5?	f: Shared function	Interact?
A-B	0	0	NO
C-D	0	1	YES
I-J	1	0	YES

Problems with information gain

- It prefers attributes with MANY values
- Solution: "GainRatio" to penalize multiple-valued attributes.
  - Used in C4.5

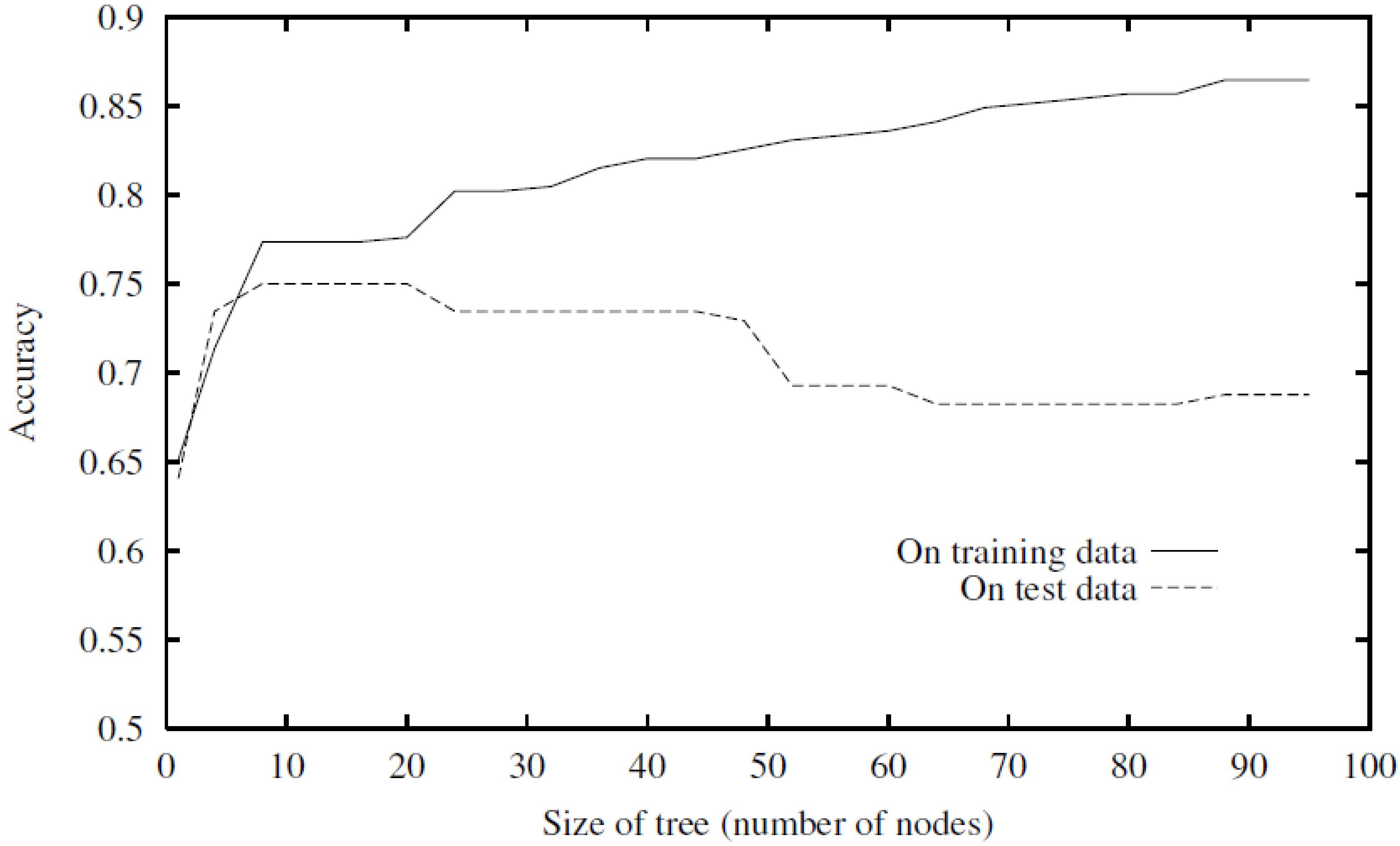
$$SplitInfo(S,A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

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

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

  - Attribute with the highest gain ratio is selected for the next split.

Overfitting



Avoiding Overfitting

- During tree construction:
  - Stop growing when classification is "good enough" rather than when it is perfect.
  - Grow full tree, then post-prune (works better).
- Selecting the "best" tree:
  - Use a validation set to evaluate performance of alternatives
  - Minimum Description Length (MDL)
    - SizeOfTree + NumberOfMisclassifications

Node post-pruning

- For each node:
  - Evaluate performance on validation set when this node is pruned out
- Remove the node whose removal gives the best performance on the validation set.
- Repeat until further pruning is harmful.

Rule post-pruning

- Convert the tree into equivalent set of rules
- Prune each rule independently of others.
  - Remove condition(s) whose removal does not worsen the accuracy.
- Gives a chance to remove a branch from a specific rule (whereas in node-pruning, removing a branch removes it from all descendants).
- Gives better classification accuracy then node-pruning. When you prune rules, they may no longer form a single decision tree.

Handling Continuous-Valued Attributes

- Find the partitioning of the continuous attribute that gives the best separation (e.g., using information gain criteria) of positive and negative samples.

Temperature:	40	48	60	72	80	90
PlayTennis:	No	No	Yes	Yes	Yes	No

$$\frac{48+60}{2} = 54 \qquad \frac{80+90}{2} = 85$$

Handling missing values

- Fill in the missing value by examining other samples sorted to a node.
  - Assign most common value for that attribute.
  - Or, assign the most common value for that attribute among the samples having the same target class.

Attributes with Costs

- Figuring out the value of an attribute may be costly. Consider:
  - cost of blood test: \$100
  - cost of fMRI scan: \$1000
- Can we optimize the tree so it prefers "cheaper" tests (without undermining the predictive quality) ?
- Use splitting criteria that integrate Gain and Cost:
  - Tan and Schlimmer (unweighted)

$$\frac{Gain^2(S,A)}{Cost(A)}$$

  - Nunez (weighted)

$$\frac{(2^{Gain(S,A)}-1)}{(Cost(A)+1)^w}$$

    - where $w \in [0,1]$ determines importance of cost.

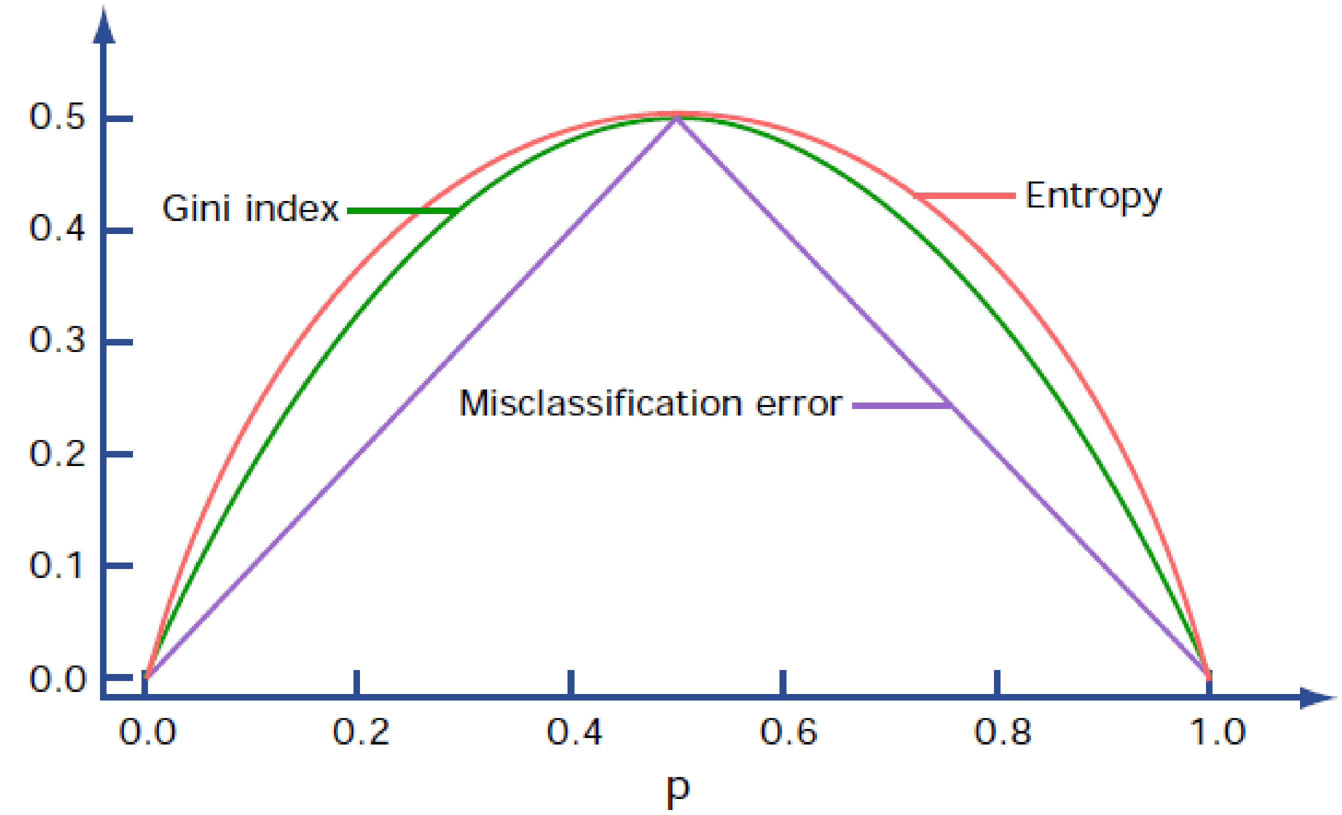
Commonly Used Implementations

- C4.5: Extension of ID3 to account for missing values, continuous attributes, tree pruning, and rule pruning.
- CART (Classification and Regression Trees) uses Gini Index
  - Gini measure "impurity" of the data.
  - $Gini(S) = 1 - \sum_i p_i^2$
  - Gini index of a binary split on attribute A

$$Gini(S, A) = \frac{S_1}{S} Gini(S_1) + \frac{S_2}{S} Gini(S_2)$$

  - Maximize the reduction in Gini index.

Entropy vs. Gini Index



Node impurity measures for two-class classification, as a function of the proportion p in class 2. Cross-entropy has been scaled to pass through (0.5, 0.5).

Image by MIT OpenCourseWare, adapted from Hastie et al., *The Elements of Statistical Learning*, Springer, 2009.

Attribute Selection Strategies

- Information Gain
  - Biased toward multi-valued attributes
- Gini Index
  - ??Biased toward multi-valued attributes
  - Problematic when number of classes is large
  - Tends to give balanced (equal-sized, equal-purity) partitions.
- Gain Ratio
  - Tends to give unbalanced partitions

Decision Trees vs. Others

- Biggest advantage is interpretability.
  - Easy to state and understand classification rules.
- Fast learning
- Scalability is an issue for large datasets.
  - Need to distribute samples to partitions at each split and recalculate the gain criteria for each partition.