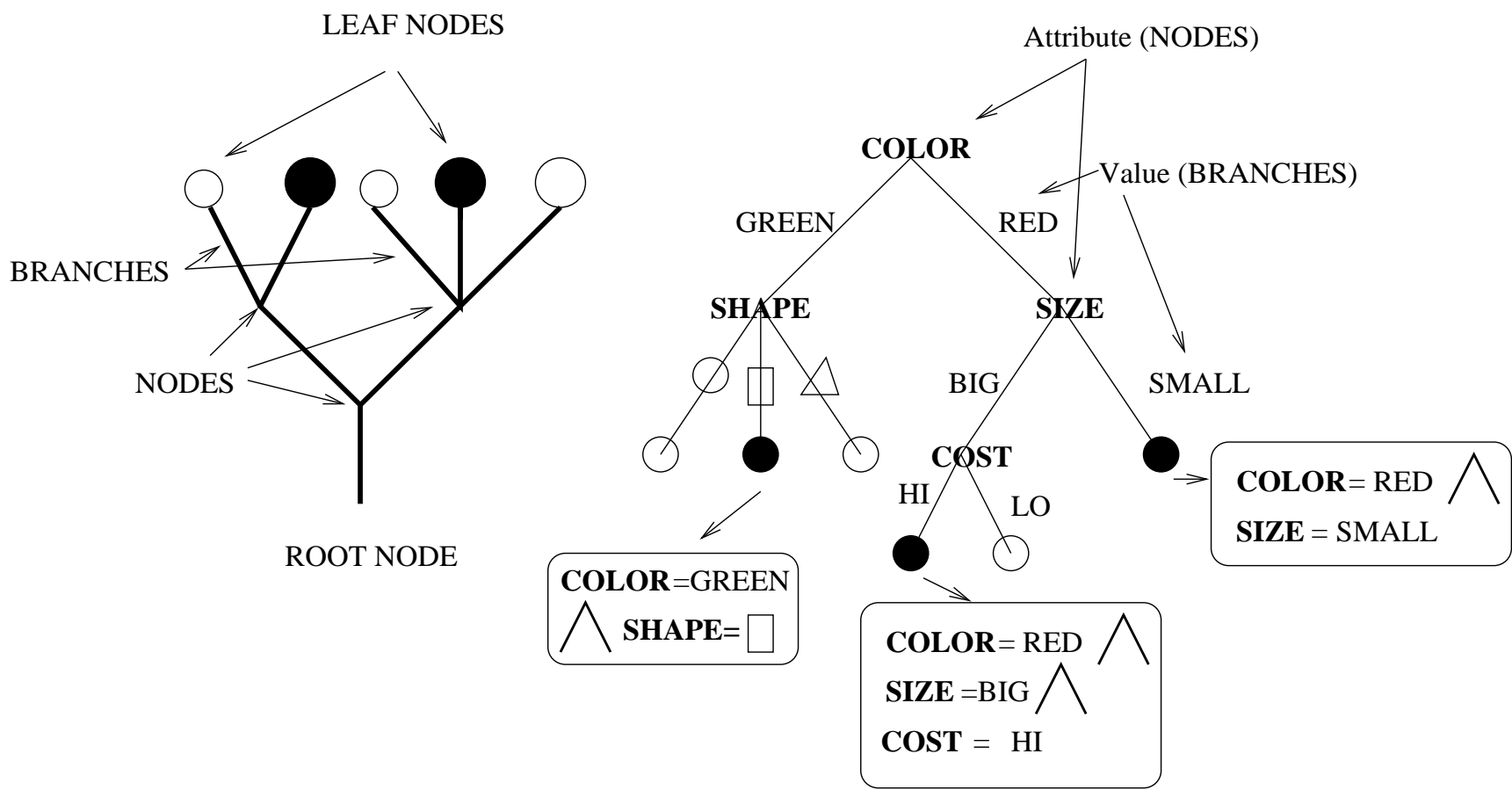
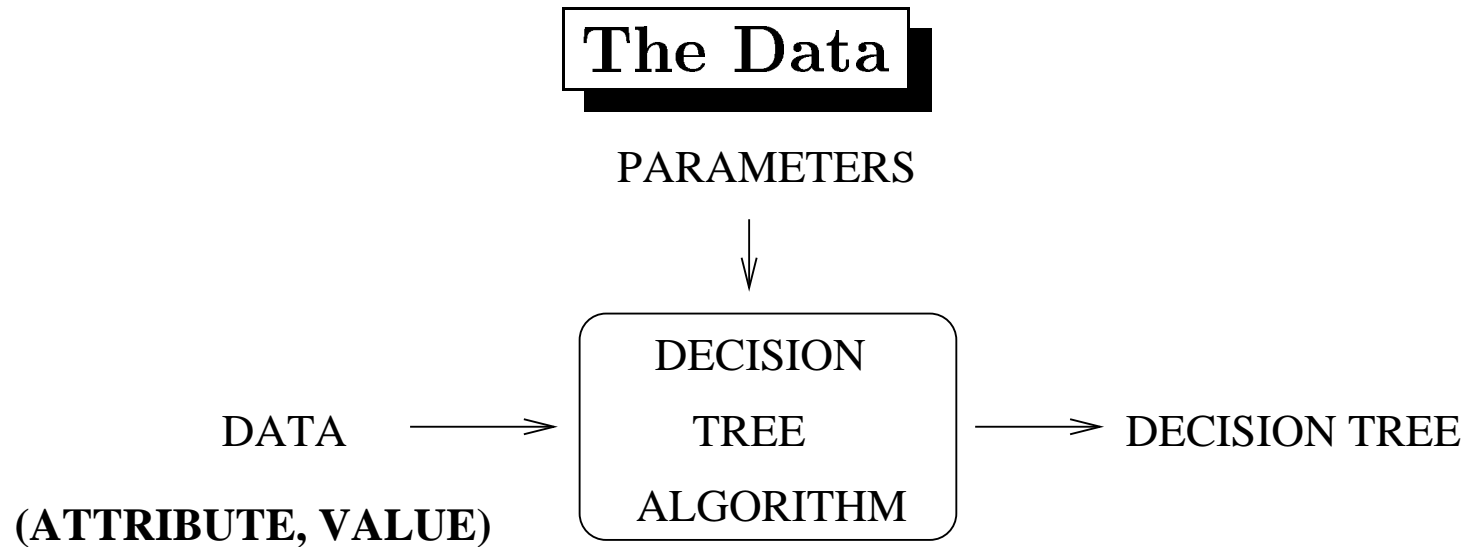


Decision Trees



Decision Tree: Representation of a CONCEPT

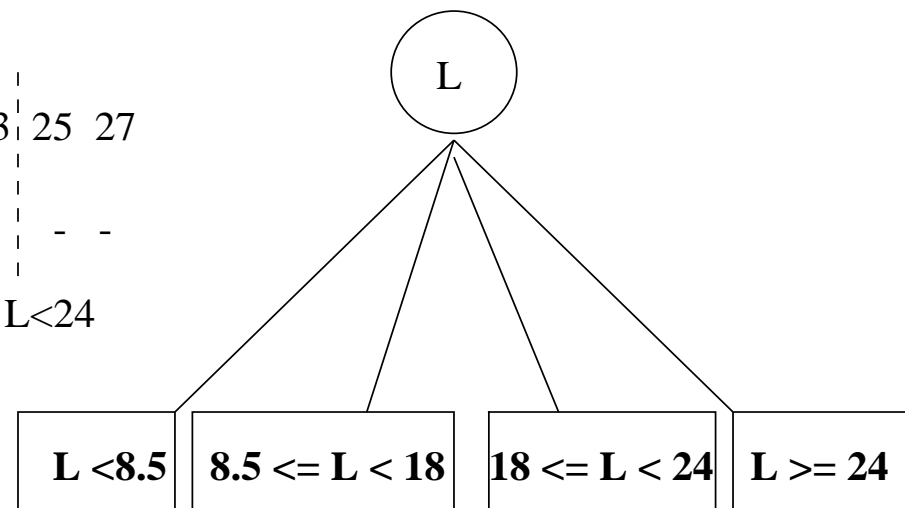


- ATTRIBUTES = (SIZE, SHAPE, COLOR, COST)
 - SIZE \in {big, small}, SHAPE \in {square, circle, triangle}
 - COLOR \in {red, green}, COST \in [0.99, 499.99]\$
- *Nominal (Discrete) Attributes:* SIZE, SHAPE, COLOR
- *Continuous (Real Valued) Attributes:* COST \rightarrow Discretized
- (COST < \$ 100) = LO and (COST \geq \$ 100) = HI

Discretizing Continuous Attributes

- Each node branches into a finite number of subtree
- Need a mechanism to discretize continuous attributes (COST)
- Sort the attributes and look at class label boundaries:

LENGTH	5	7	10	13	16	20	23	25	27
CLASS	+	+	-	-	-	+	+	-	-
			L < 8.5		L < 18		L < 24		



Example Dataset

Each Instance = (ATTR-VAL-SET) \rightarrow CLASS-LABEL

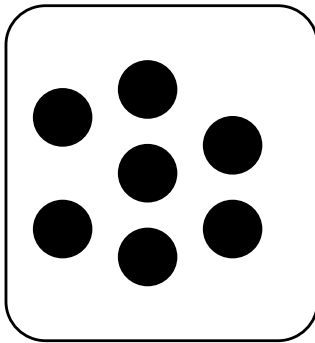
- (BIG, RED, SQUARE, LO) \rightarrow +
- (BIG, GREEN, TRIANGLE, LO) \rightarrow +
- (SMALL, RED, CIRCLE, HI) \rightarrow -
- (SMALL, GREEN, SQUARE, HI) \rightarrow -
- (SMALL, GREEN, TRIANGLE, LO) \rightarrow +
- (BIG, RED, CIRCLE, HI) \rightarrow -

Learning Decision Tree

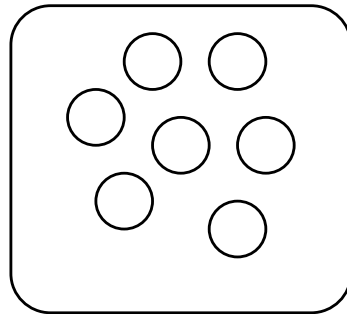
- Each Node n in a DT is associated with:
 - A SUBSET $\mathcal{X}(n)$ of the data set \mathcal{X}
 - A REGION (subspace) $\mathcal{I}(n)$ in the input space \mathcal{I}
- Root Node $n = 0$
 - $\mathcal{X}(0) = \mathcal{X}$ (all training samples)
 - $\mathcal{I}(0) = \mathcal{D}_1 \times \mathcal{D}_2 \times \dots \times \mathcal{D}_m$ (whole input space)
- Goal of DT Learning:
 - Partition \mathcal{I} (\mathcal{R}) into *PURE* regions (subsets)

What's Purity of node n

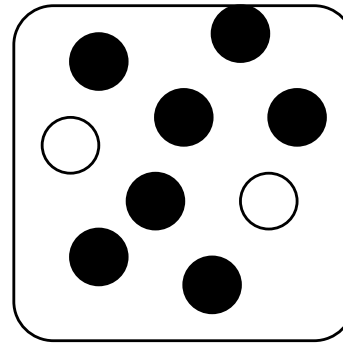
- Purity(n) depends on class labels of all examples $\in \mathcal{X}(n)$
- If all class labels are SAME (+ or -) then Purity(n) = 1
- If half of each class are present, Purity(n) = 0



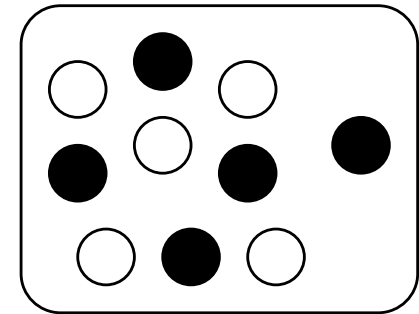
PURITY = 1



PURITY = 1



PURITY = 0.6



PURITY = 0

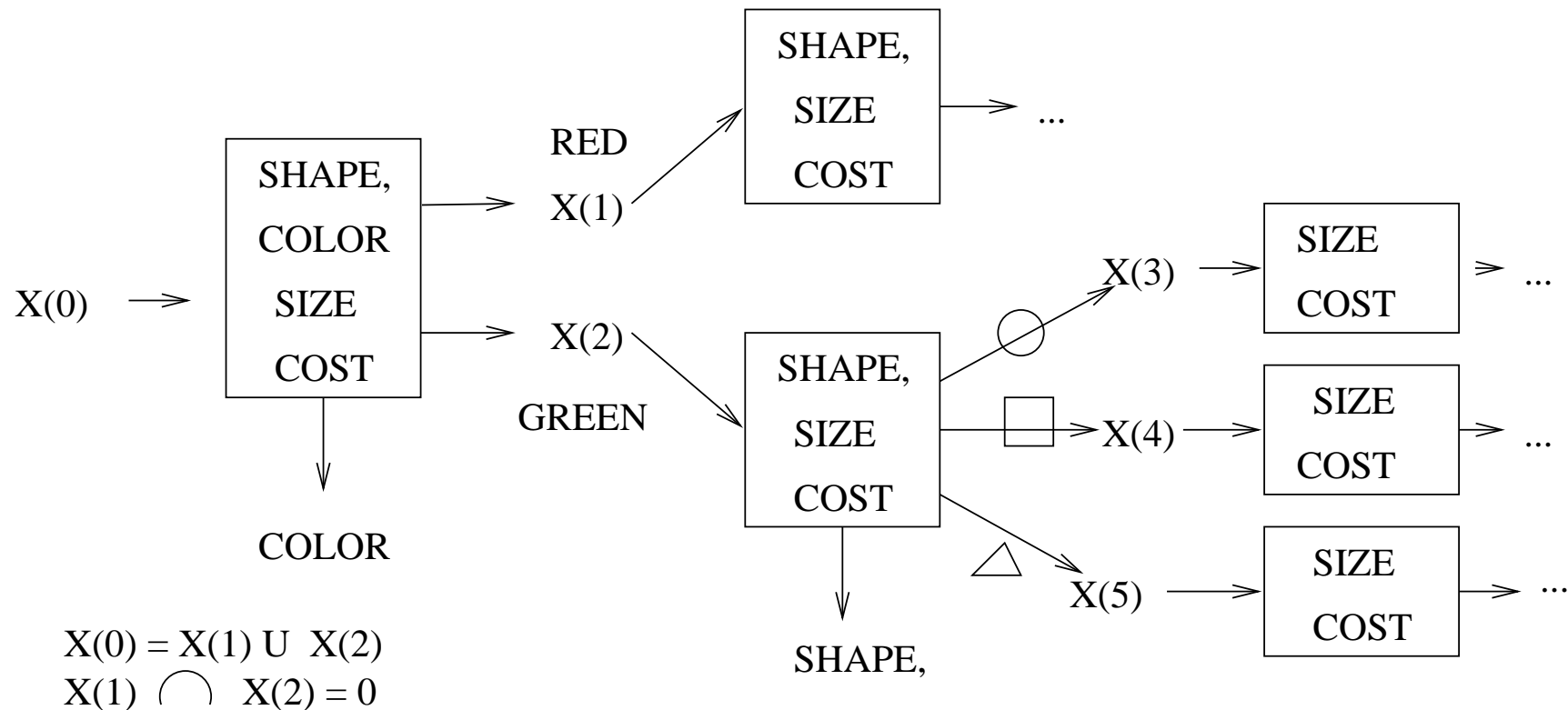
Entropy: A measure of IMpurity

- Let $\Omega = \{\omega_1, \dots, \omega_C\}$ = set of C classes
- $freq(\omega, n) = \#$ Examples of class ω at node n
- Then Entropy(n) is

$$\text{Entropy}(n) = - \sum_{i=1}^C \frac{freq(\omega_i, n)}{|X(n)|} \log_C \left(\frac{freq(\omega_i, n)}{|X(n)|} \right) \in [0, 1] \quad (1)$$

- and Purity(n) = $1 - \text{Entropy}(n)$
- i.e. More entropy \rightarrow Less purity

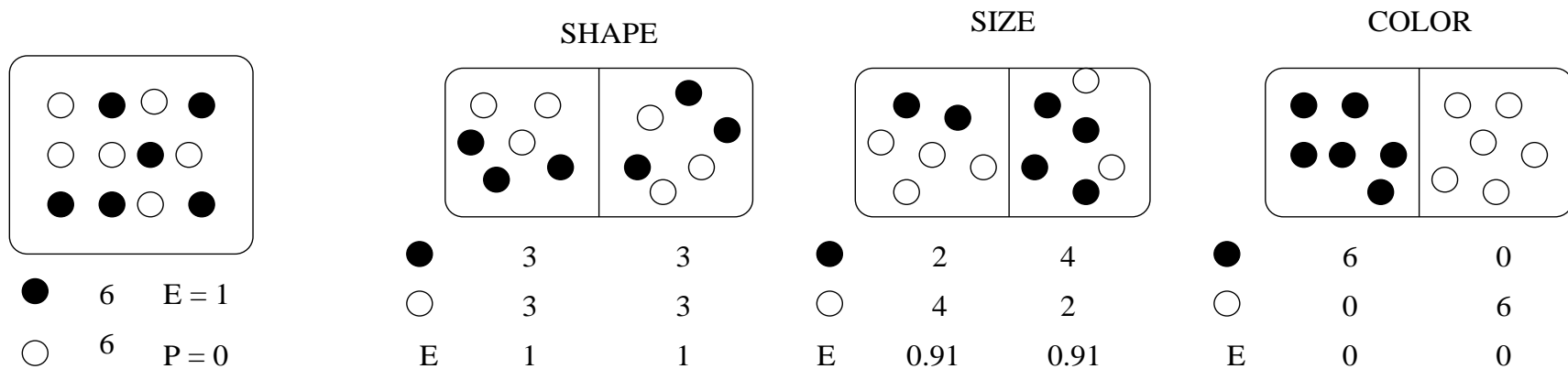
DT Learning Algorithm



- $\text{Purity}(X(n)) < \theta_P$ and $\frac{|X(n)|}{|X|} < \theta_S$ for $n = 0, 1, 2, 3, 4, 5, \dots$
- $\theta_P = \text{Purity parameter}$, $\theta_S = \text{size parameter}$

Choosing Best Attribute at a node

- Each node n has a set of attributes $\mathcal{A}(n)$ to choose from
- For each Attribute $A_j \in \mathcal{A}(n)$ compute $Gain-Ratio(A_j)$
- Pick the One with maximum $Gain-Ratio$



Gain Ratio

Gain-Ratio(X, A) depends on 2 factors: ($A \in \mathcal{A}(n)$ and $X = \mathcal{X}(n)$)

- *Gain*: Increase in expected purity of children nodes

$$\text{Gain}(X, A) = \text{Entropy}(X) - \sum_{j=1}^{m(A)} \frac{|X_j|}{|X|} \text{Entropy}(X_j) \quad (2)$$

- *Split-Info*: Size variability among children nodes

$$\text{SplitInfo}(X, A) = - \sum_{j=1}^{m(A)} \frac{|X_j|}{|X|} \log_2 \left(\frac{|X_j|}{|X|} \right) \quad (3)$$

- We need HIGH *Gain* and LOW *Split-Info* (Uneven split)

$$\text{GainRatio}(X, A) = \frac{\text{Gain}(X, A)}{\text{SplitInfo}(X, A)} \quad (4)$$

Pruning : DT Generalization

- Perfect Classification on Training Data does NOT imply good generalization (output for unseen examples). Reasons:
 - Noise in Training data
 - Over Complex Hypothesis (increased generalization error)
- **Pruning Methods**
 - **Prepruning:** Adjust θ_P and θ_S
 - **Postpruning:** Remove subtrees from a fully grown tree
- **Pruning evaluation** Cross-validation, MDL, Statistical tests

Conclusions

- Greedy Algorithms for a Powerful concept representation
- Very interpretable DNF rules can be extracted
- Computation Complexity = $O(|\mathcal{X}||\mathcal{A}|^2)$
- ID3, CART, C4.5 (latest implementation)
- Can handle continuous and missing variables
- Can be used for regression ALSO.