Decision Trees

Calculating Node Impurity

- Calculate the Entropy, Gini Index, and Classification Error for each of the following nodes

Node N_1	Count
Class = 0	0
Class = 1	6

$$\text{Gini} = 1 - \left[\left(\frac{0}{6}\right)^2 + \left(\frac{6}{6}\right)^2\right] = 0$$

$$\text{Entropy} = -\left[\left(\frac{0}{6}\right) \cdot \log_2\left(\frac{0}{6}\right) + \left(\frac{6}{6}\right) \cdot \log_2\left(\frac{6}{6}\right)\right] = 0$$

$$\text{Classification Error} = 1 - \max\left(\left(\frac{0}{6}\right), \left(\frac{6}{6}\right)\right) = 0$$

Node N_2	Count
Class = 0	1
Class = 1	5

$$\text{Gini} = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right] = 0.278$$

$$\text{Entropy} = -\left[\left(\frac{1}{6}\right) \cdot \log_2\left(\frac{1}{6}\right) + \left(\frac{5}{6}\right) \cdot \log_2\left(\frac{5}{6}\right)\right] = 0.650$$

$$\text{Classification Error} = 1 - \max\left(\left(\frac{1}{6}\right), \left(\frac{5}{6}\right)\right) = 0.167$$

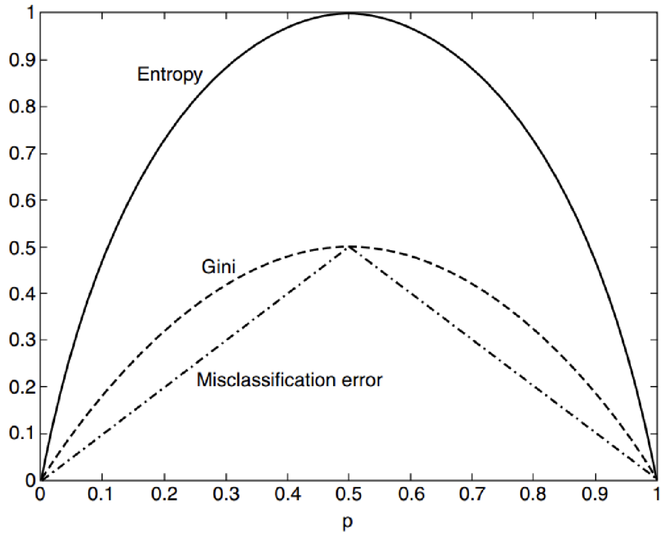
Node N_3	Count
Class = 0	3
Class = 1	3

$$\text{Gini} = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 0.5$$

$$\text{Entropy} = -\left[\left(\frac{3}{6}\right) \cdot \log_2\left(\frac{3}{6}\right) + \left(\frac{3}{6}\right) \cdot \log_2\left(\frac{3}{6}\right)\right] = 1$$

$$\text{Classification Error} = 1 - \max\left(\left(\frac{3}{6}\right), \left(\frac{3}{6}\right)\right) = 0.5$$

Node Impurity



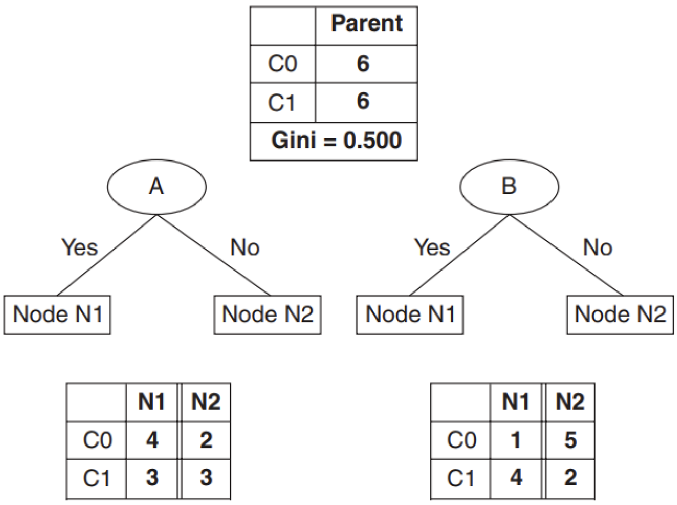
- Figure compares the values of the impurity measures for binary classification problems
- p refers to the fraction of records that belong to one of the two classes
- Observe that all three measures attain their maximum value when the class distribution is uniform (i.e., when $p = 0.5$)
- The minimum values for the measures are attained when all the records belong to the same class (i.e., when $p = 0$ or $p = 1$)
- To determine how well a test condition performs:
  - Compare the impurity of the parent node (before splitting) with the impurity of the child nodes (after splitting)
  - The larger their difference, the better the test condition
  - The gain, Δ, is a criterion that can be used to determine the goodness of a split:
  $$\Delta = I\left(\text{parent}\right) = \sum_{j=1}^{k} \frac{N\left(v_j\right)}{N} I\left(v_j\right)$$
  - Where $I\left(\cdot\right)$ is the impurity measure of a given node, N is the total number of records at the parent node, k is the number of children, and $N\left(v_j\right)$ is the number of records associated with the child node, v_j

Measuring Node Impurity

- Decision tree induction algorithms often choose a test condition that maximizes the gain Δ
- Since $I\left(\text{parent}\right)$ is the same for all test conditions, maximizing the gain is equivalent to minimizing the weighted average impurity measures of the child nodes
- When entropy is used as the impurity measure, the difference in entropy is known as the information gain, Δinfo

Splitting of Binary Attributes

- We will use Gini Index to select the best splitting of binary attributes
- A binary attribute splits into two partitions:
  - The parent node has 12 objects
  - There are two binary attributes, A and B, and their splitting results are shown in the figure below
- Which attribute (A or B) is better to split on?



Test	Node	Gini of Individual Children	Weighted Avg. Impurity of all Children
A	N1	$1-\left(\left(\frac{4}{7}\right)^2+\left(\frac{3}{7}\right)^2\right)=0.49$	$\left(\frac{7}{12}\cdot 49\right)+\left(\frac{5}{12}\cdot 48\right)=0.486$
	N2	$1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)=0.32$	
B	N1	$1-\left(\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2\right)=0.32$	$\left(\frac{5}{12}\cdot 0.32\right)+\left(\frac{7}{12}\cdot 0.41\right)=0.373$
	N2	$1-\left(\left(\frac{5}{7}\right)^2+\left(\frac{2}{7}\right)^2\right)=0.41$	

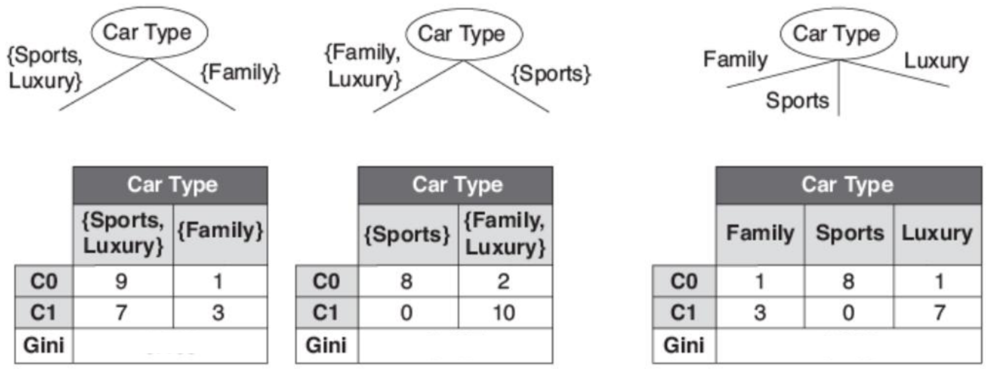
- Gain:

$$\Delta A = 0.5 - 0.486 = 0.014$$

$$\Delta B = 0.5 - 0.373 = 0.127$$

- We want to maximize the gain by minimizing the children's degree impurity

Splitting of Nominal Attributes



(a) Binary split (b) Multiway split

Test	Node	Gini of Individual Children	Weighted Avg. Impurity of all Children
$A.1$	Sport / Luxury	$1-\left(\left(\frac{9}{16}\right)^2+\left(\frac{7}{16}\right)^2\right)=0.492$	$\left(\frac{16}{20}\cdot 0.492\right)+\left(\frac{4}{20}\cdot 0.375\right)$
	Family	$1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)=0.375$	$=0.468$
$A.2$	Sport	$1-\left(\left(\frac{8}{8}\right)^2+\left(\frac{0}{8}\right)^2\right)=0$	$\left(\frac{8}{20}\cdot 0\right)+\left(\frac{12}{20}\cdot 0.278\right)$
	Family / Luxury	$1-\left(\left(\frac{2}{12}\right)^2+\left(\frac{10}{12}\right)^2\right)=0.278$	$=0.167$
B	Family	$1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)=0.375$	$\left(\frac{4}{20}\cdot 0.375\right)+\left(\frac{8}{20}\cdot 0\right)+$
	Sports	$1-\left(\left(\frac{8}{8}\right)^2+\left(\frac{0}{8}\right)^2\right)=0$	$\left(\frac{8}{20}\cdot 0.219\right)=0.163$
	Luxury	$1-\left(\left(\frac{1}{8}\right)^2+\left(\frac{7}{8}\right)^2\right)=0.219$	