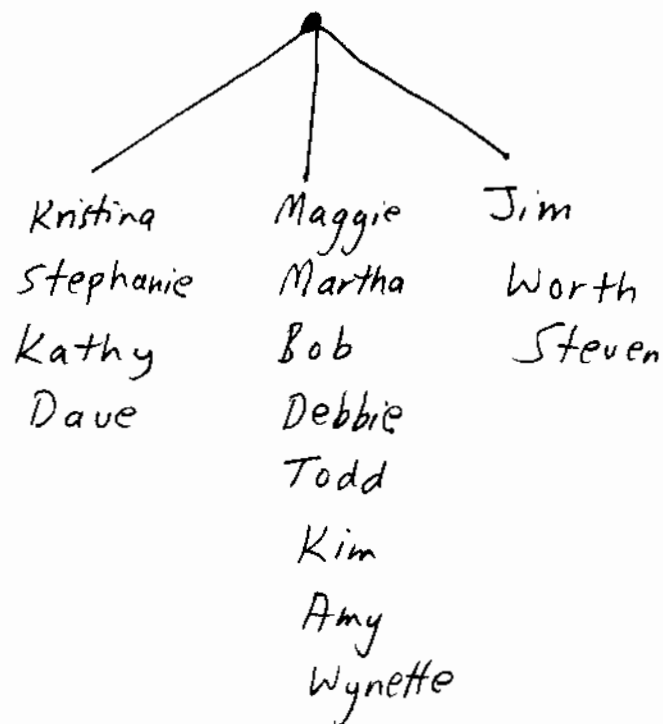


Gain vs. Gain Ratio

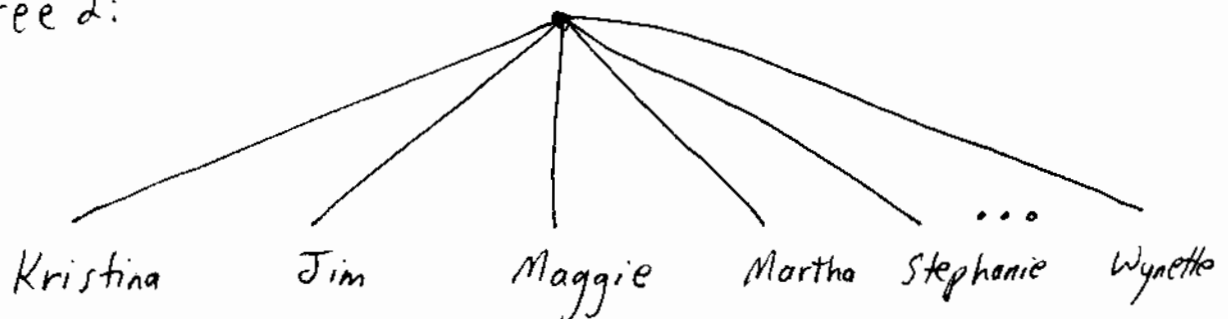
CSE 8331

⌘ given two different decision trees for data from Table 4.1 (output1)

Tree 1:



Tree 2:



Note that both behave perfectly on the training data.

Tree 1: $\text{Gain}(D, S) = H(D) - \sum_{i=1}^3 P(D_i) H(D_i)$

$$H(D) = \sum_{i=1}^3 p_i \ln \frac{1}{p_i}$$

$$= \frac{4}{15} \ln \frac{15}{4} + \frac{8}{15} \ln \frac{15}{8} + \frac{3}{15} \ln \frac{15}{3}$$

$$= 0.352 + 0.335 + 0.322$$

$$= 1.009$$

$$\text{Gain} = 1.009 - \left(\frac{4}{15} H(D_1) + \frac{8}{15} H(D_2) + \frac{3}{15} H(D_3) \right)$$

$$= 1.009 - 0 = 1.009$$

Note $H(D_1) = 1 \ln 1 + 0 + 0 = 0 = H(D_2) = H(D_3)$

$$\text{Gain Ratio} = \frac{\text{Gain}}{H\left(\frac{4}{15}, \frac{8}{15}, \frac{3}{15}\right)} = \frac{1.009}{1.009} = 1$$

Note: I'm using \ln throughout

Tree 2: $\text{Gain}(D, S) = H(D) - \sum_{i=1}^S P(D_i) H(D_i)$

Here $S = 15$

$$\text{Gain} = 1.009 - \left(\frac{1}{15} H(D_1) + \frac{1}{15} H(D_2) \dots + \frac{1}{15} H(D_{15}) \right)$$

Here $H(D_1) = H(D_2) = \dots = H(D_{15}) = 0$

$\therefore \text{Gain} = 1.009$

$$\text{Gain Ratio} = \frac{\text{Gain}}{H\left(\frac{1}{15}, \frac{1}{15}, \dots, \frac{1}{15}\right)}$$

$$= \frac{1.009}{15 \cdot \frac{1}{15} \cdot \ln 15} = \frac{1.009}{2.708} = 0.373$$

Notice that Gain Ratio is much smaller than Gain. This is because the perfect ordering is achieved primarily because each $|D_i| = 1$.