Best Splits (Cont.)

- Another measure of purity is the **entropy** measure, which is defined as

$$\text{entropy }(S) = -\sum_{i-1}^{k} p_i \log_2 p_i$$

- When a set S is split into multiple sets Si, I=1, 2, …, r, we can measure the purity of the resultant set of sets as:

$$\text{purity}(S_1, S_2, \ldots, S_r) = \sum_{i=1}^{r} \frac{|S_i|}{|S|} \text{ purity }(S_i)$$

- The information gain due to particular split of S into S_i, i = 1, 2, …., r

  Information-gain $(S, \{S_1, S_2, \ldots, S_r) = \text{purity}(S) - \text{purity}(S_1, S_2, \ldots S_r)$