6.034 Recitation 8: Identification Trees (Nov. 1)

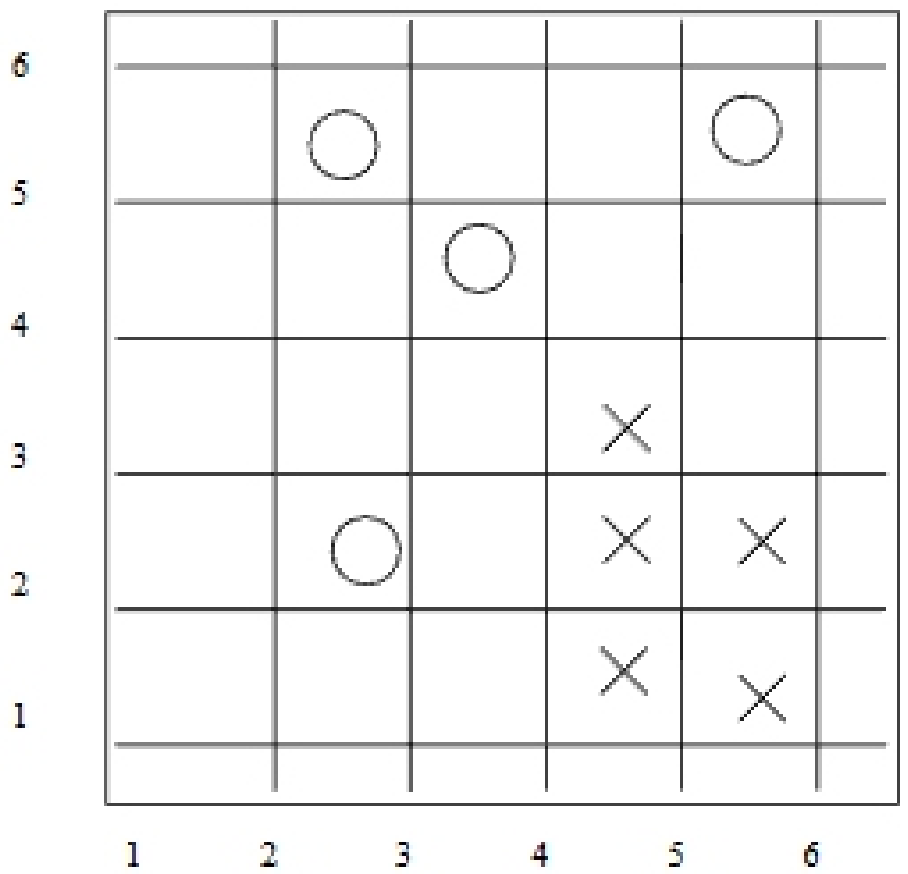
Kimberle Koile

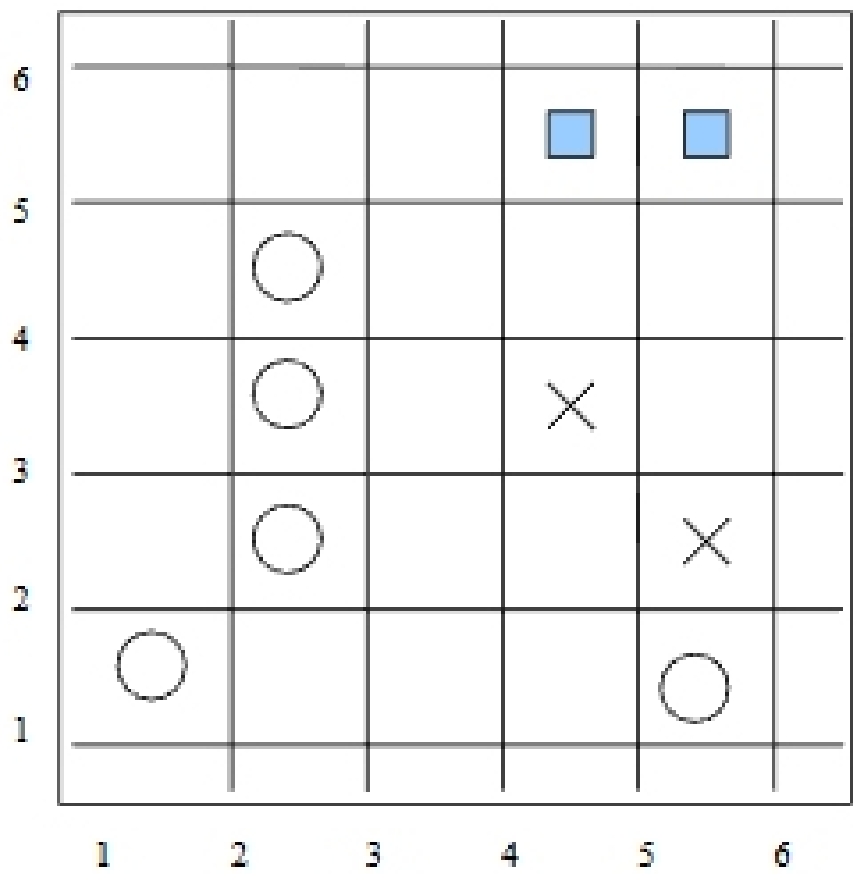
Use the entropy graphs on page 2 to determine the best splits for building an ID tree for the each of the following data sets, and draw the trees.

Recall that the best split for a set of data minimizes the average disorder:

$$\text{Average disorder} = \sum_b \left(\frac{n_b}{n_t}\right) \times \left(\sum_c -\frac{n_{bc}}{n_b} \log_2\left(\frac{n_{bc}}{n_b}\right)\right)$$

n_b is the total number of samples in a region b
n_t is the total number of samples in all regions
n_{bc} is the total of samples in region b of class c





Entropy: E = $-a \log_2 a - b \log_2 b - c \log_2 c \ldots$