Determine the Minimum Entropy Split

In this activity, you will calculate the entropy to determine how to build the most effective decision tree, given a specific scenario and data set.

Scenario Setup

Imagine you are building a decision tree to predict whether a personal loan given to a person would result in a **payoff** (i.e., the person pays off the loan) or **default** (the person fails to pay back the loan).

- Your entire dataset consists of 30 instances:
  - 16 belong to the "**default**" class
  - 14 belong to the "**payoff**" class
- The data points contain two features, “Balance” and "Residence".
  - "Balance" refers to the amount of money the person has in their savings and checking accounts at the time of the loan, which can take on two values: “< $50K” or “≧ $50K.”
  - “Residence” refers to whether or not the person owns their home or rents and can take on two values: “OWN” or “RENT”.

The entropy H over a leaf containing a set S of points is $H(S) = -\sum_{k=1}^{c} p_k \log(p_k)$.

The entropy H over an intermediate or a root node of a set S of points with two branches S_L and S_R is $H(S) = \frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$.

If we don't divide the set of 30 points and treat it as a leaf, the entropy will be:
$H(\text{undivided}) = -\frac{16}{30}\log_2\left(\frac{16}{30}\right) - \frac{14}{30}\log_2\left(\frac{14}{30}\right) \approx 0.99$.

Activity

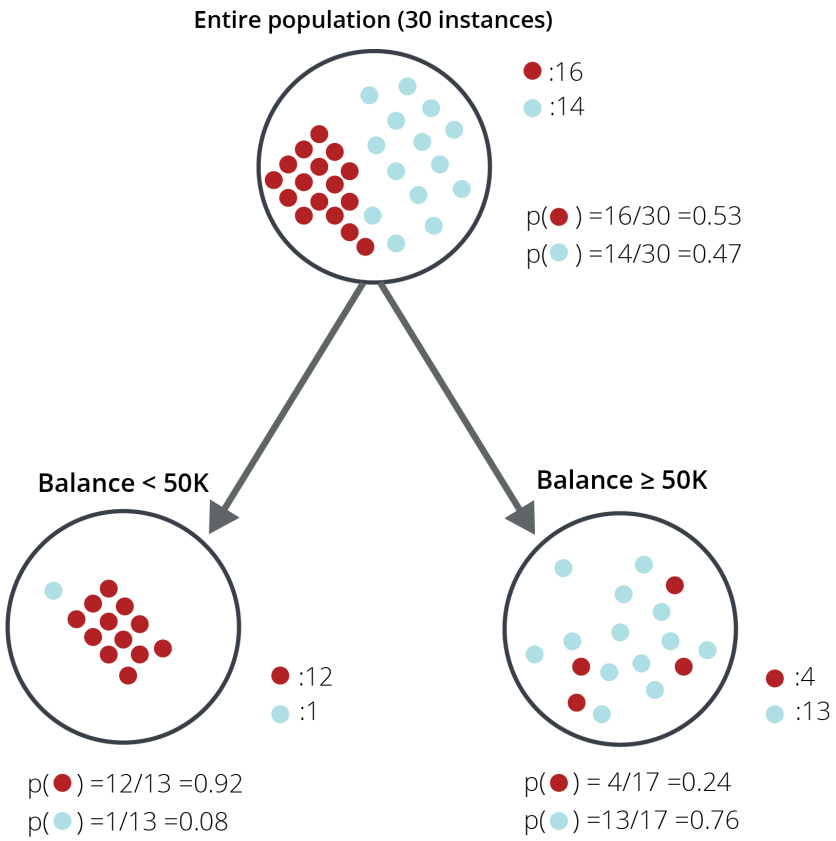
Your task is to determine which feature, "Balance" or "Residence", provides the lowest entropy split. There are two scenarios below to review. Once you've answered the question, click "Show Solution" to see if your answer is correct.

Scenario 1 | Scenario 2

Split on Balance

Calculate the entropy for the parent node and see how much uncertainty exists by splitting on "Balance".

The **blue circles** represent people from the "**payoff**" class and the **red circles** are people from the "**default**" class. Splitting the parent node on the "Balance" attribute gives us two child nodes: "< $50K” or "≧ $50K." Review the graphic below to see how the data is split.



Question: What is the entropy of the root node in this split?

Hide Solution

$$H(\text{Balance } < \$50K) = -\frac{12}{13}\log_2\left(\frac{12}{13}\right) - \frac{1}{13}\log_2\left(\frac{1}{13}\right) \approx 0.39$$

$$H(\text{Balance } \geq \$50K) = -\frac{4}{17}\log_2\left(\frac{4}{17}\right) - \frac{13}{17}\log_2\left(\frac{13}{17}\right) \approx 0.79$$

$$H(\text{divide on Balance}) = \frac{13}{30} \cdot 0.39 + \frac{17}{30} \cdot 0.79 = 0.62$$

Before splitting, the entropy of the tree is 0.99. After splitting, the entropy of the tree is 0.62. So, by splitting on "Balance" the entropy is reduced by (0.99 - 0.62) = **0.37**