Information Gain Calculation for Splitting on CreditScore at 650

Tran Quoc Thai
Student ID: 2370759

22/03/2025

1 Introduction

In decision tree classification, the selection of the optimal feature split is determined by maximizing the information gain (IG), which measures the reduction in entropy after partitioning the dataset based on a feature threshold. Entropy quantifies the impurity or unpredictability of a dataset. A higher entropy indicates a more uncertain distribution, while a lower entropy indicates a more homogeneous set.

This document details the calculation of information gain when splitting on `CreditScore` at 650.

2 Entropy Calculation

The entropy $H(S)$ of a dataset with binary classification labels $\{\text{Low}, \text{High}\}$ is computed using Shannon's entropy formula:

$$H(S) = -p_{\text{Low}} \log_2 p_{\text{Low}} - p_{\text{High}} \log_2 p_{\text{High}} \tag{1}$$

From the training dataset, we have:

- 4 instances classified as `Low` risk.
- 4 instances classified as `High` risk.

Thus, the probabilities are:

$$p_{\text{Low}} = \frac{4}{8} = 0.5,$$
$$p_{\text{High}} = \frac{4}{8} = 0.5.$$

Substituting these into the entropy formula:

$$\begin{aligned} H(S) &= -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) \\ &= -(0.5 \times -1 + 0.5 \times -1) \\ &= 1.000. \end{aligned}$$

3 Entropy After Splitting on `CreditScore` at 650

Splitting the dataset at `CreditScore` = 650 results in:

- **Left subset** (`CreditScore` $\leq$ 650): Contains instances {2, 4, 6, 8}, all classified as High risk.
- **Right subset** (`CreditScore` $>$ 650): Contains instances {1, 3, 5, 7}, all classified as Low risk.

3.1 Left Subset Entropy Calculation

Since all instances in the left subset belong to the same class (High risk), entropy is:

$$H(S_{\text{left}}) = 0.000.$$

3.2 Right Subset Entropy Calculation

Similarly, since all instances in the right subset belong to the same class (Low risk), entropy is:

$$H(S_{\text{right}}) = 0.000.$$

4 Information Gain Calculation

The information gain formula is given by:

$$IG = H(S) - \left(\frac{|S_{\text{left}}|}{|S|} H(S_{\text{left}}) + \frac{|S_{\text{right}}|}{|S|} H(S_{\text{right}}) \right). \tag{2}$$

Substituting values:

$$\begin{aligned} IG &= 1.000 - \left(\frac{4}{8} \times 0.000 + \frac{4}{8} \times 0.000 \right) \\ &= 1.000 - (0 + 0) \\ &= 1.000. \end{aligned}$$

5 Discussion on Root Node Selection

The high information gain of 1.000 suggests that splitting on `CreditScore` at 650 effectively reduces entropy, making it a strong candidate for the root node of the decision tree. Since this split results in completely homogeneous subsets, it perfectly separates the classes in this dataset. However, further evaluation on a larger dataset may be required to generalize this conclusion.