"CO5241$_{Q}uestion_{1}$

Onuh Justus Izuchukwu

March 2025

CO5241 Machine Learning Assignment Solution to Question 1: Information Gain

1 Problem Statement

Calculate the information gain for splitting `CreditScore` at 650 in a decision tree classification task. Determine if this is a good root node split.

2 Solution

2.1 Step 1: Calculate Entropy Before Split

Entropy is defined as:

$$H(S) = -p_1 \log_2(p_1) - p_2 \log_2(p_2) \tag{1}$$

From the dataset:

- Total records: 8
- Low Risk: 4, High Risk: 4
- Probabilities: $P(Low) = 4/8 = 0.5$, $P(High) = 4/8 = 0.5$

$$H(S) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1.0 \tag{2}$$

2.2 Step 2: Calculate Entropy After Split at CreditScore = 650

Splitting results in two groups:

- Left split (≤ 650): 3 records (High Risk: 3, Low Risk: 0)
- Right split (> 650): 5 records (High Risk: 1, Low Risk: 4)

Entropy of the left group:

$$H(Left) = -(3/3 \log_2(3/3) + 0/3 \log_2(0/3)) = 0.0 \tag{3}$$

Entropy of the right group:

$$H(Right) = -(4/5 \log_2(4/5) + 1/5 \log_2(1/5)) \approx 0.72 \tag{4}$$

2.3 Step 3: Compute Weighted Entropy After Split

$$H(Split) = \frac{3}{8}(0.0) + \frac{5}{8}(0.72) = 0.45 \tag{5}$$

2.4 Step 4: Compute Information Gain

$$IG = H(S) - H(Split) = 1.0 - 0.45 = 0.55 \tag{6}$$

2.5 Step 5: Interpretation

- Since $IG = 0.55$, splitting at `CreditScore = 650` provides moderate information gain. - A better root split may exist, so other features should be checked.

3 Conclusion

Since the information gain is moderate (0.55), this split may be useful but should be compared with other features. The next step is to examine different features for a higher information gain.